Statistical Papers 42, 171–185 (2001)

Statistical Papers

On distributions of exceedances associated with order statistics and record values for arbitrary distributions

Ismihan G. Bairamov[1] and Samuel Kotz[2]

[1] Faculty of Science, Department of Statistics, Ankara University,
06100, Tandogan, Ankara, Turkey
[2] Department of Engineering Management and System Engineering,
The George Washington University, Washington, DC 20052, USA

Received: October 12, 1999; revised version: April 6, 2000

Exceedance statistics associated with the order statistics and record values for the generalized Bernoulli model are investigated for the case of an arbitrary underlying distribution. Distributions of these statistics are studied. Also the distribution of the second record value is obtained when the underlying distribution function contains atoms.

Key words: Order statistics, record value, distribution function, atom.

1. Introduction

This paper attempts to provide further insight into properties of records, exceedances and placement statistics for the case of arbitrary distributions. The importance of these concepts in construction of non-parametric tests of equality (identity) of distributions has been demonstrated by Katzenbeisser (1985), (1986), Matveychuk and Petunin (1990), Johnson and Kotz (1991), (1994). The concept of exceedances is closely connected with the negative (inverse) hypergeometric distribution; it is originally due to Condorcet in 1785 (see Todhunter (1865)) and was developed in the works of Gumbel and Schelling (1950) and Sarkadi (1957) .

Let $X = (X_1, X_2, ..., X_n)$ be a random sample of size n from an absolutely continuous distribution with a common cumulative distribution function (c.d.f) $F_X(.)$. Denote the order statistics of X by $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$. Let $X_{n+1}, X_{n+2}, ..., X_{n+m}$

be another sample of size m from the same distribution, independent from X. It is well known and can easily be derived that for $1 \leq i < j \leq n$,

$$P\{X_{n+1} \in (X_{i:n}, X_{j:n})\} = \frac{j-i}{n+1}$$

i.e. $(X_{i:n}, X_{j:n})$ is a distribution free confidence interval containing a future observation in the class of all absolutely continuous distribution functions $\Im_c$. It is also known that if f_1 and f_2 are two continuous and symmetric functions of n arguments satisfying

$$f_1(u_1, u_2, ..., u_n) \leq f_2(u_1, u_2, ..., u_n) \text{ for all } (u_1, u_2, ..., u_n) \in R^n,$$

then the probability $P\{X_{n+1} \in (f_1(X_1, X_2, ..., X_n), f_2(X_1, X_2, ..., X_n))\}$ is the same for all absolutely continuous distributions if and only if $f_1(X_1, X_2, ..., X_n) = X_{i:n}$ and $f_2(X_1, X_2, ..., X_n) = X_{j:n}$ for some $1 \leq i < j \leq n$ (see, e.g., Bairamov and Petunin (1990)). Matveychuk and Petunin (1991) and Johnson and Kotz (1991) studied a generalized Bernoulli model defined in terms of placement statistics from two random samples. Let additionally $Y_1, Y_2, ..., Y_m$ be a random sample of size m from a population with the c.d.f. $F_Y(.)$. Define the interval $J_{i,q} = (X_{i:n}, X_{i+q:n})$. Let T be the number of Y's falling into $J_{i,q}$. When $F_X(.) = F_Y(.)$, the distribution of T is

$$\begin{aligned} P\{T = t\} &= \binom{m}{t} \frac{q(q+1)...(q+t-1)(n+1-q)...(n+m-t-q)}{(n+1)(n+2)...(n+m)} \\ (t &= 0, 1, ..., m). \end{aligned}$$

Katzenbeisser (1985), (1986) obtained a formula for the distribution of T when the interval $J_{i,q}$ is one sided interval $(-\infty, X_{q:n})$ and proposed a test criterion for testing the null hypothesis $H_0 : F_X(x) = F_Y(x)$ versus the Lehmann alternatives $F_Y(x) = [F_X(x)]^\theta, \theta \neq 1$. Next, he extended these results to shift alternatives. Matveychuk and Petunin (1991) suggested a test criterion for testing the hypothesis $H_0 : F_X(x) = F_Y(x)$ using T. The critical region of the test is of the form

$$\frac{|T - E[T \mid H_0]|}{\sigma(T \mid H_0)} > t_\beta,$$

where t_β is chosen in such a manner that an approximate level of significance of the test equals 2β. Johnson and Kotz (1994) analyzed the test of homogeneity for two population introduced by Matveychuk and Petunin (1990) and provided a symmetrical test criterion.

The test of homogeneity described above is applicable when the underlying distributions are absolutely continuous. The results obtained in this paper extend the range of applicability of these tests to include data generated by not necessary continuous distributions. This data is of course known to be widespread in engineering and especial social sciences. Consider now a more general situation when the distribution functions $F_X(u)$ and $F_Y(u)$ possess atoms, i.e. there exist on the real line points of discontinuity $a_1, a_2, ..., a_l$, where $F(a_i - 0) < F(a_i)$. For these points probability of the event "the sample value X exactly equals a_i " is not zero, while for the points of continuity this probability equals zero. Of course if a sample is sufficiently large these atoms can be determined from the sample. In this case atoms are those sample values which occur in the sample more than ones. In Section 2 of this paper we shall investigate the distribution of T when the interval $J_{i,q}$ is one sided interval $(-\infty, X_{q:n})$ as it is the case in Katzenbeisser (1985) .

Let $X_1, X_2, ..., X_n, ...$ be a sequence of independent, identically distributed random variables with a common c.d.f. $F_X(.)$. The sequences of record times $U(n)$ and record values $X_{U(n)}$ are defined recursively as follows: $U(1) = 1, X_{U(1)} = X_1, U(n+1) = \min\left\{i : i > U(n), X_i > X_{U(n)}\right\}, n > 1$. A substantial interest in records is due perhaps to the fact that we often come across them in our everyday life and the record values are usually single out. Numerous results on record theory have been reviewed in Arnold, Balakrishnan and Nagaraja (1998). Distributions of record values, which is of substantial theoretical as well as practical interest have been studied in the literature only for either continuous or discrete distributions. But in many practical applications we encounter the situation when some observed values are being repeated. This attests that the underlying distribution function possesses point of discontinuity. Such an experiment can be modelled by the distribution having simultaneously both continuous and discrete components. In Section 3 we obtain the distribution of the second record value for distributions possessing atoms.

Let $X_{U(r)}$ be the r th record value, $X_{U(r)+1}, X_{U(r)+2}, ..., X_{U(r)+m}$ be the observations following $X_{U(r)}$. It is known that $X_{U(r)+1}, X_{U(r)+2}, ..., X_{U(r)+m}$ are mutually independent and identically distributed with the c.d.f $F_X(.)$ for any r. Furthermore, $X_{U(r)}$ and $X_{U(r)+k}$ are independent for $k = 1, 2, ...$ and

$$P\left\{X_{U(r)} < X_{U(r)+1} < X_{U(s)}\right\} = \frac{1}{2^r} - \frac{1}{2^s}, \quad r < s.$$

Denote by $S_m(r)$ the number of $X_{U(r)+1}, X_{U(r)+2}, ..., X_{U(r)+m}$ falling into the interval $(-\infty, X_{U(r)})$. For any $m, r = 1, 2, ...$

$$\begin{aligned} P\{S_m(r) = k\} &= = \frac{\binom{m}{k}}{(r-1)!} \int_0^\infty e^{-z(m-k+1)}(1 - e^{-z})^k z^{r-1} dz , \\ k &= 0, 1, ..., m, \end{aligned}$$

and for any r

$$\lim_{m\to\infty} \sup_{0\le x\le 1} \left| P\left\{\frac{S_m(r)}{m} \le x\right\} - \frac{1}{(r-1)!} \int_0^x \left[\ln(\frac{1}{1-u})\right]^{r-1} du \right| = 0$$

(see Bairamov (1997)). Numerous aspects of exceedances in record threshold models have been investigated in Wesolowsky and Ahsanullah (1998). In Section 4 of the present paper we shall also study the distribution of record exceedance statistics $S_m(r)$ when the underlying distribution function contains an atom at some point

$a \in R$. The simplicity of the models coupled with high level of universality of results presented in this paper provide a basis to expect their wide applicability in many different fields such as mixture models, approximations to distributions, characterizations as well as estimation and hypothesis testing.

2. Distributional properties of statistics based on order statistics for arbitrary distribution

Let X be a random variable defined on a probability space $\{\Omega, \Im, P\}$ with c.d.f. $F(x) = P\{X \le x\}$. Throughout this paper we will assume that F is an arbitrary distribution, i.e. F may contain a discrete, absolutely continuous and singular components simultaneously. Let $M = \{a_1, a_2, ..., a_l\}$, $(a_1 < a_2 < ... < a_l)$ be the set of atoms of the distribution. The following lemma will be useful for our investigations.

Lemma 2.1. Let $A \in \Im$ and $P\{A \mid X = x\}$ exist for all $x \in R$. Then

$$P\{A\} = \sum_{k=0}^{l} \int_{a_k}^{a_{k+1}-0} P\{A \mid X = x\} dF(x) + \sum_{k=1}^{l} P\{A \mid X = a_k\} P\{X = a_k\}, \quad (2.1)$$

where $a_0 = -\infty, a_{l+1} = \infty$.

The proof is a direct application of the total probability rule.

It is clear that if we have only one atom, say a then (2.1) becomes

$$P\{A\} = \int_{-\infty}^{a-0} P\{A \mid X = x\} dF(x) + \int_{a}^{\infty} P\{A \mid X = x\} dF(x)$$

$$+P\{A \mid X = a\} P\{X = a\}. \quad (2.2)$$

Now let $X_1, X_2, ..., X_n$ be a sample from distribution with the c.d.f. F, and X_{n+1}, $X_{n+2}, ..., X_{n+m}$ be the another sample from the same distribution independent of the first. Let $X_{1:n}, X_{2:n}, ..., X_{n:n}$ be the order statistics of $X_1, X_2, ..., X_n$. Consider the random variables $\xi_1(r), \xi_2(r), ..., \xi_m(r)$ defined as follows:

$$\xi_i(r) = \begin{cases} 1, & if \quad X_{n+i} < X_{r:n} \\ 0 & if \quad X_{n+i} \ge X_{r:n} \end{cases}, i = 1, 2, ..., m; \ 1 \le r \le n.$$

Finally define as above

$$S_m(r) = \sum_{i=1}^{m} \xi_i(r).$$

Evidently $S_m(r)$ is the number of observations $X_{n+1}, X_{n+2}, ..., X_{n+m}$ falling into the interval $(-\infty, X_{r:n})$. It is well known that the distribution function of $X_{r:n}$ is

$$F_r(x) = P\{X_{r:n} \le x\} = \sum_{i=r}^{n} \binom{n}{i} F^i(x)(1 - F(x))^{n-i} = I_{F(x)}(r, n - r + 1),$$

where $I_p(c,d) = \int_0^p t^{c-1}(1-t)^{d-1}dt/B(c,d)$ is the incomplete beta function ratio (see, e.g. David (1981)).

Theorem 2.1. Using the notation above:

$$P\{S_m(r)=k\} = \binom{m}{k}\left\{\frac{1}{B(r,n-r+1)}\sum_{j=0}^{l}\int_{F(a_j)}^{F(a_{j+1}-0)} t^{k+r-1}(1-t)^{m+n-k-r}dt\right.$$

$$\left.+\sum_{j=1}^{l} F^k(a_j-0)\,(1-F(a_j-0))^{m-k}\,(F_r(a_j)-F_r(a_j-0))\right\}, \qquad (2.3)$$

$$k = 0,1,\dots,m.$$

Proof. For simplicity we shall provide a proof in the case when the set M contains one element a only. It follows from the definition of $S_m(r)$ that

$$P\{S_m(r)=k\}$$

$$= \sum_{i_1,i_2,\dots,i_m} P\{A_{i_1}\cap A_{i_2}\cap\dots\cap A_{i_k}\cap \bar{A}_{i_{k+1}}\cap \bar{A}_{i_{k+2}}\cap\dots\cap\bar{A}_{i_m}\}, \qquad (2.4)$$

where A_{i_j} are the events $\left\{X_{n+i_j} < X_{r:n}\right\}$, $i_j \in \{1,2,\dots,m\}$; $i_j \neq i_l$ if $j \neq l$ and $\bar{A}_{i_j}$ denotes the complement of event A_{i_j}. Moreover

$$P\{A_{i_1}\cap A_{i_2}\cap\dots\cap A_{i_k}\cap \bar{A}_{i_{k+1}}\cap \bar{A}_{i_{k+2}}\cap\dots\cap\bar{A}_{i_m}\} =$$

$$= P\left\{X_{n+i_1} < X_{r:n},\dots,X_{n+i_k} < X_{r:n}, X_{n+i_{k+1}} \geq X_{r:n},\dots,X_{n+i_m} \geq X_{r:n}\right\}. \qquad (2.5)$$

Consider the probability appearing on the r.h.s. side of (2.5). By the symmetry

$$P\left\{X_{n+i_1} < X_{r:n},\dots,X_{n+i_k} < X_{r:n}, X_{n+i_{k+1}} \geq X_{r:n},\dots,X_{n+i_m} \geq X_{r:n}\right\}$$

$$= P\{X_{n+1} < X_{r:n},\dots,X_{n+k} < X_{r:n}, X_{n+k+1} \geq X_{r:n},\dots,X_{n+m} \geq X_{r:n}\}.$$

Denote $C = \{X_{n+1} < X_{r:n},\dots,X_{n+k} < X_{r:n}, X_{n+k+1} \geq X_{r:n},\dots,X_{n+m} \geq X_{r:n}\}$. Utilizing (2.2) one can write

$$P\{C\} =$$

$$= \int_{-\infty}^{a-0} P\{C \mid X_{r:n} = x\}\,dF_r(x) + \int_{a}^{\infty} P\{C \mid X_{r:n} = x\}\,dF_r(x)$$

$$+P\{C \mid X_{r:n} = a\}\, P\{X_{r:n} = a\}$$

$$= \int_{-\infty}^{a-0} P\{X_{n+1} < x, \dots, X_{n+k} < x, X_{n+k+1} \geq x, \dots, X_{n+m} \geq x\}\, dF_r(x)$$

$$+ \int_{a}^{\infty} P\{X_{n+1} < x, \dots, X_{n+k} < x, X_{n+k+1} \geq x, \dots, X_{n+m} \geq x\}\, dF_r(x)$$

$$+P\{X_{n+1} < a, \dots, X_{n+k} < a, X_{n+k+1} \geq a, \dots, X_{n+m} \geq a\}\, P\{X_{r:n} = a\}$$

$$= \frac{1}{B(r, n-r+1)} \int_{-\infty}^{a-0} F^k(x)(1-F(x))^{m-k} F^{r-1}(x)(1-F(x))^{n-r} dF(x)$$

$$+ \frac{1}{B(r, n-r+1)} \int_{a}^{\infty} F^k(x)(1-F(x))^{m-k} F^{r-1}(x)(1-F(x))^{n-r} dF(x)$$

$$+F^k(a-0)(1-F(a-0))^{m-k}\left[F_r(a) - F_r(a-0)\right]$$

Combining the terms we have

$$P\{C\} = \frac{1}{B(r, n-r+1)} \left\{ \int_{0}^{F(a-0)} t^{k+r-1}(1-t)^{m+n-k-r} dt \right.$$

$$\left. + \int_{F(a)}^{1} t^{k+r-1}(1-t)^{m+n-k-r} dt \right\}$$

$$+F^k(a-0)(1-F(a-0))^{m-k}\left[F_r(a) - F_r(a-0)\right]. \tag{2.6}$$

The number of summands in (2.4) is equal to $\binom{m}{k}$ each having the same probability (2.6). Hence

$$P\{S_m(r) = k\} = \binom{m}{k} \left\{ \frac{1}{B(r, n-r+1)} \times \right.$$

$$\times \left[\int_{0}^{F(a-0)} t^{k+r-1}(1-t)^{m+n-k-r} dt + \int_{F(a)}^{1} t^{k+r-1}(1-t)^{m+n-k-r} dt \right]$$

$$\left. + F^k(a-0)(1-F(a-0))^{m-k}\left[F_r(a) - F_r(a-0)\right] \right\}.$$

The theorem is thus proved.

3. Distribution of the second record value when underlying distribution function contains an atom

Let F has an atom at the point a and be continuous otherwise in $\mathbf{R}$. We shall define the record times in the usual manner (see, e.g. Arnold, Balakrishnan and Nagaraja (1998)): $U(1) = 1$ and $U(n) = min\left\{i : i > U(n-1), X_i > X_{U(n-1)}\right\}, n > 1$ and $X_{U(1)}, X_{U(2)}, \ldots$ be corresponding record values. $X_{U(1)} = X_1$ is a first (trivial) record value by definition. Denote $R(u) = -\ln(1 - F(u))$, $r(a) = R(a) - R(a-0)$, $f(a) = F(a) - F(a-0)$. We shall investigate the distribution of the second (first nontrivial) record value.

Theorem 3.1. Under assumptions stipulated above the distribution of second record value is

$$F_2(x) = P\left\{X_{U(2)} \le x\right\}$$

$$= \begin{cases} \int\limits_{-\infty}^{a-0} R(v)dF(v) + \int\limits_{a}^{\infty} R(v)dF(v) - r(a)\times & \\ \times\left[F(x) - F(a)\right] + \left[F(x) - F(a)\right] f(a)/\left[1 - F(a)\right] & \\ +R(a-0)f(a), & if \quad x \ge a \\ \int\limits_{-\infty}^{x} R(v)dF(v), & if \quad x < a. \end{cases}$$

Proof. By definition of the second record value

$$P\left\{X_{U(2)} \le x\right\} = \sum_{k=2}^{\infty} P\left\{X_k \le x, X_1 \ge X_2, \ldots, X_1 \ge X_{k-1}, X_1 < X_k\right\}. \tag{3.1}$$

Denote

$$A_k = \left\{X_k \le x, X_1 \ge X_2, \ldots, X_1 \ge X_{k-1}, X_1 < X_k\right\}. \tag{3.1a}$$

We shall use the following total probability formula which is a direct application of Lemma 1.1 :

$$P\left\{A_k\right\} = \int\limits_{-\infty}^{a-0}\int\limits_{-\infty}^{a-0} P\left\{A_k \mid X_1 = u, X_k = v\right\} dF(u)dF(v)$$

$$+\int\limits_{-\infty}^{a-0}\int\limits_{a}^{\infty} P\left\{A_k \mid X_1 = u, X_k = v\right\} dF(u)dF(v)$$

$$+\int\limits_{a}^{\infty}\int\limits_{-\infty}^{a-0} P\left\{A_k \mid X_1 = u, X_k = v\right\} dF(u)dF(v)$$

$$+\int\limits_{a}^{\infty}\int\limits_{a}^{\infty} P\left\{A_k \mid X_1 = u, X_k = v\right\} dF(u)dF(v)$$

$$+\int_{-\infty}^{a-0} P\{A_k \mid X_1 = a, X_k = v\} P\{X_1 = a\} dF(v)$$

$$+\int_{a}^{\infty} P\{A \mid X_1 = a, X_k = v\} P\{X_1 = a\} dF(v)$$

$$+\int_{-\infty}^{a-0} P\{A_k \mid X_1 = u, X_k = a\} P\{X_k = a\} dF(u)$$

$$+\int_{a}^{\infty} P\{A_k \mid X_1 = u, X_k = a\} P\{X_k = a\} dF(u)$$

$$+P\{A_k \mid X_1 = a, X_k = a\} P\{X_1 = a, X_k = a\} \tag{3.2}$$

However,

$$\int_{-\infty}^{a-0} \int_{a}^{\infty} P\{A_k \mid X_1 = u, X_k = v\} dF(u)dF(v) = 0, \tag{3.3}$$

$$\int_{-\infty}^{a-0} P\{A_k \mid X_1 = a, X_k = v\} P\{X_1 = a\} dF(v) = 0, \tag{3.4}$$

$$P\{A_k \mid X_1 = a, X_k = a\} P\{X_1 = a, X_k = a\} = 0 \tag{3.5}$$

and

$$\int_{a}^{\infty} P\{A_k \mid X_1 = u, X_k = a\} P\{X_k = a\} dF(u) = 0 \tag{3.6}$$

Consider the first integral on the r.h.s. of (3.2). The contribution to (3.1) from this probability is:

$$I \equiv \sum_{k=2}^{\infty} \int_{-\infty}^{a-0} \int_{-\infty}^{a-0} P\{A_k \mid X_1 = u, X_k = v\} dF(u)dF(v)$$

$$= \sum_{k=2}^{\infty} \int_{-\infty}^{\min(a-0,x)} \int_{-\infty}^{v} F^{k-2}(u)dF(u)dF(v) = \sum_{k=2}^{\infty} \int_{-\infty}^{\min(a-0,x)} \frac{F^{k-1}(v)}{k-1} dF(v) \tag{3.7}$$

Because of $F(a-0) < 1$ the series $\sum_{k=1}^{\infty} \frac{z^k}{k}$ converges uniformly on $E \equiv \{z : 0 \leq z \leq F(\min(x, a-0))\}$ due to the Weierstrass theorem (see, e.g. Rudin, Theorem 7.10, P. 119). That is the series in (3.7) can be integrated term by term. Hence from (3.7) it follows

$$I = \int_{-\infty}^{\min(a-0,x)} R(v)dF(v) = \begin{cases} \int_{-\infty}^{x} R(v)dF(v) & if \quad x < a \\ \int_{-\infty}^{a-0} R(v)dF(v) & if \quad x \geq a \end{cases} . \tag{3.8}$$

Recalling the definition of the event A_k in (3.1a) one observes that the second and the fifth probability contribute zero. From the third probability on the right hand side of (3.2) we have the contribution:

$$\sum_{k=2}^{\infty}\int_{a}^{\infty}\int_{-\infty}^{a-0} P\{A \mid X_1 = u, X_k = v\}\, dF(u)dF(v) = \sum_{k=2}^{\infty}\int_{a}^{x}\int_{-\infty}^{a-0} F^{k-2}(u)dF(u)dF(v)$$

$$= R(a-0)\int_{a}^{x} dF(v) = R(a-0)\,[F(x)-F(a)]\,,\ x \geq a. \tag{3.9}$$

Analogously the fourth one yields

$$\sum_{k=2}^{\infty}\int_{a}^{\infty}\int_{a}^{\infty} P\{A \mid X_1 = u, X_k = v\}\, dF(u)dF(v) = \sum_{k=2}^{\infty}\int_{a}^{x}\int_{a}^{v} F^{k-2}(u)dF(u)dF(v)$$

$$= \int_{a}^{x} [R(v)-R(a)]\, dF(v) = \int_{a}^{x} R(v)dF(v) - R(a)\,[F(x)-F(a)]\,, x \geq a. \tag{3.10}$$

In the same manner for the sixth probability in (3.2) we have:

$$\sum_{k=2}^{\infty}\int_{a}^{x} P\{A \mid X_1 = a, X_k = v\}\, P\{X_1 = a\}\, dF(v)$$

$$= \sum_{k=2}^{\infty}\int_{a}^{x} F^{k-2}(a)dF(v)\,[F(a)-F(a-0)]$$

$$= \frac{[F(x)-F(a)]\,[F(a)-F(a-0)]}{1-F(a)}, x \geq a. \tag{3.11}$$

For the seventh probability in (3.2) one obtains:

$$\sum_{k=2}^{\infty}\int_{-\infty}^{a-0} P\{A \mid X_1 = u, X_k = a\}\, P\{X_k = a\}\, dF(u)$$

$$= \begin{cases} 0, & if \quad x < a \\ \sum_{k=2}^{\infty} \int_{-\infty}^{a-0} F^{k-2}(u)dF(u)P\{X_k = a\} & \\ \quad = R(a-0)\,[F(a)-F(a-0)] & if \quad x \geq a. \end{cases} \tag{3.12}$$

Combining (3.3)-(3.12) in (3.1) we have

$$F_2(x) = P\left\{X_{U(2)} \leq x\right\} = \int_{-\infty}^{a-0} R(v)dF(v) + \int_{a}^{x} R(v)dF(v)$$

$$-[R(a)-R(a-0)]\,[F(x)-F(a)]+[F(x)-F(a)]\times$$
$$\times\,[F(a)-F(a-0)]\,/\,[1-F(a)]+R(a-0)\,[F(a)-F(a-0)],$$

for $x \geq a$ and

$$F_2(x)=\int_{-\infty}^{x} R(v)dF(v)$$

for $x < a$.

The theorem is thus proved.

Evidently $F_2(-\infty)=0$ and $F_2(\infty)=1$. In fact

$$F_2(\infty)=-R(a)\,(1-F(a))+R(a-0)\,[1-F(a-0)]$$

$$+\,[F(a)-F(a-0)]+\int_{-\infty}^{a-0} R(v)dF(v)+\int_{a}^{\infty} R(v)dF(v)$$

$$=\int_{-\infty}^{a-0} R(v)dF(v)+\int_{a}^{\infty} R(v)dF(v)+\int_{a-0}^{a} R(y)dF(y)=1.$$

4. Distribution of exceedances in a record model

Let F has an atom at the point a and is continuous otherwise in R. Let $X_{U(2)}$ be the second record value, $X_{U(2)+1}, X_{U(2)+2}, ..., X_{U(2)+m}$ be the observations following $X_{U(2)}$. Evidently $X_{U(2)+1}, X_{U(2)+2}, ..., X_{U(2)+m}$ are mutually independent and identically distributed with c.d.f $F(x)$. Furthermore $X_{U(2)}$ and $X_{U(2)+1}$ are also independent. Denote by $\tilde{S}_m$ the number of $X_{U(2)+1}, X_{U(2)+2}, ..., X_{U(2)+m}$ falling into $(-\infty, X_{U(2)})$. $\tilde{S}_m$ denotes the number of $X_{U(2)+i}$'s falling below the threshold $X_{U(2)}$. The following theorem gives a distribution of $\tilde{S}_m$ in an integral form.

Theorem 4.1. Under the assumptions stipulated above for any $m \geq 1$

$$P\left\{\tilde{S}_m=l\right\}=\binom{m}{l}\left\{\int_{0}^{F(a-0)} t^l(1-t)^{m-l}\ln\frac{1}{1-t}dt+\int_{F(a)}^{1} t^l(1-t)^{m-l}\ln\frac{1}{1-t}dt\right.$$

$$+\,[R(a-0)-R(a)]\int_{F(a)}^{1} t^l(1-t)^{m-l}dt$$

$$+\frac{F(a)-F(a-0)}{1-F(a)}\int_{F(a)}^{1} t^l(1-t)^{m-l}dt+F^l(a-0)\,(1-F(a-0))^{m-l}\times$$

$$\left.\times\,R(a-0)\,[F(a)-F(a-0)]\right\},\ l=0,1,2,...,m.$$

Proof. Analogously to the proof of Theorem 2.1

$$P\left\{\tilde{S}_m = l\right\} =$$

$$= \sum_{i_1, i_2, \ldots, i_m} P\left\{A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_l} \cap \bar{A}_{i_{l+1}} \cap \bar{A}_{i_{l+2}} \cap \ldots \cap \bar{A}_{i_m}\right\},$$

where $A_{i_j} = \left\{X_{n+i_j} < X_{U(2)}\right\}$, $i_j \in \{1, 2, \ldots, m\}$ and $\bar{A}_{i_j}$ denotes the complement of event A_{i_j}. By the symmetry

$$P\left\{A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_l} \cap \bar{A}_{i_{l+1}} \cap \bar{A}_{i_{l+2}} \cap \ldots \cap \bar{A}_{i_m}\right\}$$

$$= P\left\{X_{U(2)+1} < X_{U(2)}, \ldots, X_{U(2)+l} < X_{U(2)}, X_{U(2)+l+1} \geq X_{U(2)}, \ldots, X_{U(2)+m} \geq X_{U(2)}\right\}$$

Thus

$$P\left\{\tilde{S}_m = l\right\} = \binom{m}{l} \times$$

$$\times P\left\{X_{U(2)+1} < X_{U(2)}, \ldots, X_{U(2)+l} < X_{U(2)}, X_{U(2)+l+1} \geq X_{U(2)}, \ldots, X_{U(2)+m} \geq X_{U(2)}\right\} \tag{4.1}$$

Evidently by definition of $X_{U(2)}$,

$$P\left\{X_{U(2)+1} < X_{U(2)}, \ldots, X_{U(2)+l} < X_{U(2)}, X_{U(2)+l+1} \geq X_{U(2)}, \ldots, X_{U(2)+m} \geq X_{U(2)}\right\}$$

$$= \sum_{k=2}^{\infty} P\{X_{k+1} < X_k, \ldots, X_{k+l} < X_k, X_{k+l+1} \geq X_k, \ldots, X_{k+m} \geq X_k,$$

$$X_1 \geq X_2, \ldots, X_1 \geq X_{k-1}, X_1 < X_k\} \tag{4.2}$$

Denote

$$B_k = \{X_{k+1} < X_k, \ldots, X_{k+l} < X_k, X_{k+l+1} \geq X_k, \ldots, X_{k+m} \geq X_k,$$

$$X_1 \geq X_2, \ldots, X_1 \geq X_{k-1}, X_1 < X_k\}.$$

Analogously to the probability $P\{A_k\}$ in the proof of Theorem 3.1 we have

$$P\{B_k\} = \int_{-\infty}^{a-0} \int_{-\infty}^{a-0} P\{B_k \mid X_1 = u, X_k = v\}\, dF(u) dF(v)$$

$$+ \int_{-\infty}^{a-0} \int_{a}^{\infty} P\{B_k \mid X_1 = u, X_k = v\}\, dF(u) dF(v)$$

$$+ \int_{a}^{\infty} \int_{-\infty}^{a-0} P\{B_k \mid X_1 = u, X_k = v\}\, dF(u) dF(v)$$

$$+\int_a^{\infty}\int_a^{\infty} P\{B_k \mid X_1 = u, X_k = v\}\, dF(u)dF(v)$$

$$+\int_{-\infty}^{a-0} P\{B_k \mid X_1 = a, X_k = v\}\, P\{X_1 = a\}\, dF(v)$$

$$+\int_a^{\infty} P\{B \mid X_1 = a, X_k = v\}\, P\{X_1 = a\}\, dF(v)$$

$$+\int_{-\infty}^{a-0} P\{B_k \mid X_1 = u, X_k = a\}\, P\{X_k = a\}\, dF(u)$$

$$+\int_a^{\infty} P\{B_k \mid X_1 = u, X_k = a\}\, P\{X_k = a\}\, dF(u)$$

$$+P\{B_k \mid X_1 = a, X_k = a\}\, P\{X_1 = a, X_k = a\}. \tag{4.3}$$

We shall evaluate each summand in (4.3) separately.

1)

$$\sum_{k=2}^{\infty}\int_{-\infty}^{a-0}\int_{-\infty}^{a-0} P\{B_k \mid X_1 = u, X_k = v\}\, dF(u)dF(v)$$

$$= \sum_{k=2}^{\infty}\int_{-\infty}^{a-0}\int_{-\infty}^{v} F^l(v)\,(1 - F(v))^{m-l}\, F^{k-2}(u)dF(u)dF(v)$$

$$= \int_{-\infty}^{a-0} F^l(v)\,(1 - F(v))^{m-l}\, R(v)dF(v) = \int_0^{F(a-0)} t^l\,(1-t)^{m-l}\, R(t)dt. \tag{4.4}$$

2)

$$\sum_{k=2}^{\infty}\int_{-\infty}^{a-0}\int_a^{\infty} P\{B_k \mid X_1 = u, X_k = v\}\, dF(u)dF(v) = 0 \tag{4.5}$$

3)

$$\sum_{k=2}^{\infty}\int_a^{\infty}\int_{-\infty}^{a-0} P\{B_k \mid X_1 = u, X_k = v\}\, dF(u)dF(v)$$

$$= \sum_{k=2}^{\infty}\int_a^{\infty}\int_{-\infty}^{a-0} F^l(v)\,(1 - F(v))^{m-l}\, F^{k-2}(u)dF(u)dF(v)$$

$$= R(a-0)\int_a^{\infty} F^l(v)\,(1 - F(v))^{m-l}\, dF(v) = R(a-0)\int_{F(a)}^{1} t^l\,(1-t)^{m-l}\, dt \tag{4.6}$$

4)

$$\sum_{k=2}^{\infty}\int_a^{\infty}\int_a^{\infty} P\{B_k \mid X_1 = u, X_k = v\}\, dF(u)dF(v)$$

$$= \sum_{k=2}^{\infty}\int_a^{\infty}\int_a^{v} F^l(v)\,(1-F(v))^{m-l}\, F^{k-2}(u)dF(u)dF(v)$$

$$= \int_a^{\infty} F^l(v)\,(1-F(v))^{m-l}\,[R(v)-R(a)]\, dF(v) = \int_{F(a)}^{1} t^l\,(1-t)^{m-l}\, R(t)dt$$

$$-R(a)\int_{F(a)}^{1} t^l\,(1-t)^{m-l}\, dt \tag{4.7}$$

5)

$$\sum_{k=2}^{\infty}\int_{-\infty}^{a-0} P\{B_k \mid X_1 = a, X_k = v\}\, P\{X_1 = a\}\, dF(v) = 0 \tag{4.8}$$

6)

$$\sum_{k=2}^{\infty}\int_a^{\infty} P\{B_k \mid X_1 = a, X_k = v\}\, P\{X_1 = a\}\, dF(v)$$

$$= \sum_{k=2}^{\infty}\int_a^{\infty} F^l(v)\,(1-F(v))^{m-l}\, F^{k-2}(a)dF(v)P\{X_1 = a\}$$

$$= \frac{F(a)-F(a-0)}{1-F(a)}\int_{F(a)}^{1} t^l\,(1-t)^{m-l}\, dt \tag{4.9}$$

7)

$$\sum_{k=2}^{\infty}\int_{-\infty}^{a-0} P\{B_k \mid X_1 = u, X_k = a\}\, P\{X_k = a\}\, dF(u)$$

$$= \sum_{k=2}^{\infty}\int_{-\infty}^{a-0} F^l(a-0)\,(1-F(a-0))^{m-l}\, F^{k-2}(u)dF(u)\,[F(a)-F(a-0)]$$

$$= F^l(a-0)\,(1-F(a-0))^{m-l}\, R(a-0)\,[F(a)-F(a-0)] \tag{4.10}$$

8)

$$\int_a^{\infty} P\{B_k \mid X_1 = u, X_k = a\}\, P\{X_k = a\}\, dF(u) = 0 \tag{4.11}$$

9)
$$P\{B_k \mid X_1 = a, X_k = a\} P\{X_1 = a, X_k = a\} = 0 \tag{4.12}$$

Upon using (4.4)-(4.12) in (4.3) and (4.2) one obtains the assertion of the theorem.

Note. One can easily verify that $\sum_{k=0}^{m} P\left\{\tilde{S}_m = k\right\} = 1$. Indeed

$$\sum_{k=0}^{m} P\left\{\tilde{S}_m = k\right\} = \int_{-\infty}^{a-0} R(v)dF(v) + \int_{a}^{\infty} R(v)dF(v)$$
$$+[R(a-0) - R(a)]\,[1 - F(a)] + F(a) - F(a-0)$$
$$+R(a-0)\,[F(a) - F(a-0)]$$
$$= \int_{-\infty}^{a-0} R(v)dF(v) + \int_{a}^{\infty} R(v)dF(v) + \int_{a-0}^{a} R(v)dF(v) = 1.$$

Acknowledgment

The authors express their thanks to the anonymous referee and to the Coordinating Editor for their most valuable comments and suggestions.

References

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

Bairamov, I.G., Petunin, Yu.I. (1990) Structure of invariant confidence intervals containing the main distributed mass. *Theorya Veroyatnostey i ee Primeneniya.*Vol **35**, No 1, P.15-26.*(English translation in Theor. Probab. Appl.* 1991, Vol 35, No 1, P.15-26.)

Bairamov, I.G. (1997) Some distribution free properties of statistics based on record values and characterizations of the distributions through a record. *J. Appl. Statist. Sci.*. Vol. **5**, N. 1, P.17-25.

David, H. A. (1981) *Order Statistics.* Second edition, John Wiley and Sons, New York.

Gumbel, E.J. and Schelling, H, von (1950) The distribution of the number of exceedances. *Ann. Math.. Statist.* **21**, 247-262.

Johnson, N. and Kotz, S. (1991) Some generalizations of Bernoulli and Polya-Eggenberger contagion models. *Statist. Papers* **32**, P.1-17.

Johnson, N. and Kotz, S. (1994) Further comments on Matveychuk and Petunin's generalized Bernoulli model, and nonparametric tests of homogeneity. *J. Statist. Plann. Infer.* **41** , P. 61-72.

Katzenbeisser, W. (1985) The distribution of two sample location exceedance test statistics under Lehmann alternatives. *Statist. Hefte* (now *Papers*) **26**,131-138.

Katzenbeisser, W. (1986) The exact power of two-sample location tests based on exceedances against shift alternatives. *Math. Op. Forsch. (Statistics)* **20**, 47-54.

Matveychuk, S.A., Petunin Yu.I.(1990/1991) A generalization of the Bernoulli model arising in order statistics. I/II *Ukr. Math. J.* Vol. **42**, P. 518-528/ **43**, P. 779-785.

Rudin W. (1959) *Principles of Mathematical Analysis.* Mc Graw-Hill Book Company, INC. New York.

Sarkadi, K. (1957) On the distribution of the number of exceedances. *Ann. Math. Stat.,* **28**, 1021-1022.

Todhunter, I. (1865) *A History of the Mathematical Theory of Probability.* Cambridge: Macmillan (Reprinted 1965, New York: Chelsea).

Wesolowsky, J. and Ahsanullah, M. (1998) Distributional properties of exceedance statistics. *Ann. Inst. Statist. Math.* Vol. **50**, N_3, 543-565.