AUSTRIAN JOURNAL OF STATISTICS
Volume 40 (2011), Number 1 & 2, 157–166

On the Joint Distribution of Precedences and Exceedances for the Two-Sample Problem

Eugenia Stoimenova[1] and Narayanaswamy Balakrishnan[2]
[1]Bulgarian Academy of Sciences, Bulgaria
[2] McMaster University, Canada

Abstract: This paper studies the joint behaviour of precedences of one sample with respect to a threshold from the other sample and exceedances of second sample with respect to a threshold from the first sample. Exact joint distribution of the number of precedences and exceedances is derived for the case of equal distributions of the two samples.

Keywords: Exceedance statistics, Precedence statistics, Random threshold, Distribution-free properties, Stochastic ordering.

1 Introduction

Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be random samples from continuous distribution functions F and G, respectively. Let X_T and Y_T be two random variables serving as random thresholds. Suppose $X_T = f_1(X_1, X_2, \dots)$ is specified by the first sample and $Y_T = f_1(Y_1, Y_2, \dots)$ is specified by the second sample. We define the exceedance statistics A_T as the number of Y's larger than X_T, and B_T as the number of X's smaller than Y_T.

These statistics are potentially useful for testing whether the two random samples are from the same population. For example, the precedence test (van der Laan and Chakraborti, 2001; Balakrishnan and Ng, 2006) is based on the number of observations in the X-sample that are smaller than the rth order statistic $Y_{(r)}$ from the second sample. Large values of this statistic lead to rejection of the null hypothesis about equality of the two distributions against stochastic ordering. Precedence tests are useful in life-testing experiments where the data become available naturally an ordered way. The experiment is terminated after a certain number of failures.

In the general situation of testing equality of two distributions, the precedences and exceedances are both necessary with respect to thresholds from both samples. The number of exceedances in the Y-sample with respect to a threshold from the X-sample could be used along with the number of precedences in the X-sample with respect to a threshold from the Y-sample. Such a test has been defined by Stoimenova and Balakrishnan (2010) using order statistics as thresholds.

In this paper, we derive the joint distribution of the number of precedences B_r with respect to the Y-sample and the number of exceedances A_r with respect to the X-sample. We derive these distributions in the case when the two population distributions are the same. We consider examples for the joint distribution of B_r and A_r when there is shift in location of the two sample distributions.

2 Exact Distributions of Precedences and Exceedances

2.1 Exceedances from one Sample

Many special type of exceedance properties have been discussed in the literature. Recently, Bairamov and Eryilmaz (2008) derived the joint distribution of precedence and exceedance statistics from one sample with respect to lower and upper random thresholds from a second sample. The sample and the thresholds were assumed to be independent. More precisely, let Z_L and Z_R be two random variables termed the lower and upper random thresholds. Let $Y_1, Y_2, \ldots$ be a sequence of independent random variables from a common distribution G, and independent from Z_L and Z_R. As Bairamov and Eryilmaz (2008) discussed this problem, these two thresholds are generally dependent random variables, and may be viewed as a function of a random sample $X_1, X_2, \ldots$, and hence, $Z_L = f_1(X_1, X_2, \ldots)$ and $Z_R = f_2(X_1, X_2, \ldots)$.

Define the exceedance statistics

$$S_n(Z_L) = \#\{i \leq n : Y_i < Z_L\} \quad \text{and} \quad S_n(Z_R) = \#\{i \leq n : Y_i > Z_R\}. \tag{1}$$

It is clear that $S_n(Z_L)$ and $S_n(Z_R)$ define the number of observations in the sample $Y_1, Y_2, \ldots, Y_n$ which precede (exceed) the level Z_L (Z_R) (see Figure 1, left).

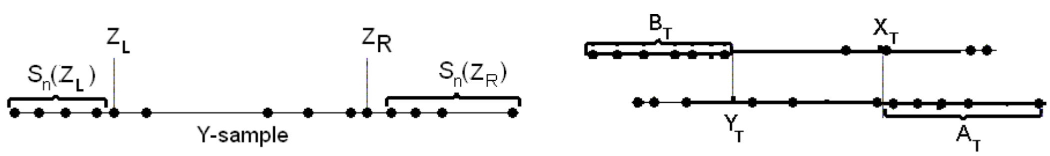

Figure 1: Thresholds and exceedance statistics for the case of one sample (left) and the case of two samples (right).

The following result (Bairamov and Eryilmaz, 2008) provides the joint probability mass function of $S_n(Z_L)$ and $S_n(Z_R)$. For $k + i \leq n$, we have

$$\begin{aligned} &P(S_n(Z_L) = k, S_n(Z_R) = i) \\ &= \frac{n!}{k!i!(n-k-i)!}\mathbb{E}\{F^k(Z_L)[1-F(Z_R)]^i[F(Z_R)-F(Z_L)]^{n-k-i}\}, \end{aligned}$$

where the expectation is taken with respect to the joint distribution of Z_L and Z_R.

The special case of $Z_R = \infty$ corresponds to the popular distribution of a single exceedance statistic with respect to a random threshold. The exact distribution of this statistic has been given in an integral form by Wesołowski and Ahsanullah (1998) as follows:

$$P(S_n(Z_L) = k) = \binom{n}{k}\mathbb{E}\{G(x)^k(1-G(x))^{n-k}\},$$

where the expectation is taken with respect to the distribution of Z_L.

2.2 Exceedances from two Samples

Consider two independent random samples of equal size from continuous distributions F and G with probability densities f and g, respectively. Let $X_{(1)} < \cdots < X_{(n)}$ be the ordered first sample and $Y_{(1)} < \cdots < Y_{(n)}$ be the ordered second sample.

For $0 \leq r < n$, specify the threshold variables based on both samples to be the $(1+r)$-th order statistic from the Y-sample and $(n-r)$-th order statistic from the X-sample (see Figure 1, right). Then, the exceedance statistics of interest from the two samples are

$$\begin{aligned} A_r &= \text{the number of } Y\text{'s} \ \text{ larger than } \ X_{(n-r)} \\ B_r &= \text{the number of } X\text{'s smaller than } Y_{(1+r)}. \end{aligned} \tag{2}$$

For unbalanced sample size case we can modify these definitions suitably. Instead of taking r, we need to take a proportion, say p, and take $r_1 = [(m+1)(1-p)]$ and $r_2 = [(n+1)p]$, where $[\cdot]$ denotes the integer part. Clearly, these r_1 and r_2 will correspond to $(1-p)$-th and p-th sample quantiles from the two samples. In this section we have results for the balanced sample size case. For the unbalanced case, we derive an asymptotic result in the Section 4.

First, we will derive the joint distribution of A_r and B_r for arbitrary continuous distributions F and G. For a fixed r, $0 \leq r < n$, two cases arise for $P(A_r = k, B_r = i)$ according to the ordering of the observations. Note that $i < n-r$ iff $k < n-r$.

Consider k exceedances in the Y-sample with respect to $X_{(n-k)}$ and i precedences in the X-sample with respect to $Y_{(1+r)}$. Event $\{A_r = k\}$ means that the $(n-r)$-th ordered observations from the X-sample is between the $(n-k)$-th and $(n-k+1)$-th ordered observations from the Y-sample, while event $\{B_r = i\}$ means that the $(1+r)$-th ordered observation from the Y-sample is between the i-th and $(i+1)$-th ordered observations from the X-sample.

Theorem 1. *1. The joint distribution of A_r and B_r for $1 \leq k \leq n-r-2$ and $0 \leq i \leq n-r-1$ is given by*

$$\begin{aligned} P(A_r = k, B_r = i) &= \sum_{q=0}^{r} \sum_{t=0}^{n-i-r-1} \frac{n!}{i!t!(n-i-q-t)!q!} \\ &\times \mathbb{E}\left\{[F(Y_1)]^i [F(Y_2) - F(Y_1)]^t [F(Y_3) - F(Y_2)]^{n-i-q-t} [1 - F(Y_3)]^q\right\}, \end{aligned} \tag{3}$$

where Y_1, Y_2 and Y_3 represent random variables from distribution G and the expectation is taken with respect to the joint density of the order statistics $Y_{(1+r)}$, $Y_{(n-k)}$ and $Y_{(n-k+1)}$ from distribution G;

2. For $0 \leq i \leq n-r-1$, we have

$$\begin{aligned} &P(A_r = n-r-1, B_r = i) \\ &= \sum_{q=0}^{r} \frac{n!}{i!q!(n-i-q)!} \mathbb{E}\left\{[F(Y_1)]^i [F(Y_2) - F(Y_1)]^{n-i-q} [1 - F(Y_2)]^q\right\}, \end{aligned} \tag{4}$$

where the expectation is taken with respect to the joint density of $Y_{(1+r)}$ and $Y_{(2+r)}$;

3. *For* $0 \leq i \leq n-r-1$, *we have*

$$P(A_r = 0, B_r = i) = \sum_{t=0}^{n-r-i-1} \frac{n!}{i!t!(n-i-t)!} \mathbb{E}\left\{[F(Y_1)]^i [F(Y_2) - F(Y_1)]^t [1 - F(Y_2)]^{n-i-t}\right\}, \quad (5)$$

where the expectation is taken with respect to the joint density of $Y_{(1+r)}$ *and* $Y_{(n)}$.

Proof: The proof of (3) uses order statistics approach. Conditional on

$$Y_{(1+r)} = y_1, \; Y_{(n-k)} = y_2, \; Y_{(n-k+1)} = y_3, \quad (6)$$

we consider the following events:

$$W_{q,t} := \begin{cases} i \; X\text{-observations preceding } y_1 \\ q \; X\text{-observations exceeding } y_3 \\ t \; X\text{-observations between } y_1 \text{ and } y_2 \\ n-i-q-t \; X\text{-observations between } y_2 \text{ and } y_3 \end{cases}$$

for any $0 \leq q \leq r$ and $0 \leq t \leq n-i-1$.

The probability of $W_{q,t}$ is evidently given by the multinomial probability

$$P(W_{q,t}) = \frac{n!}{i!t!(n-i-q-t)!q!} [F(y_1)]^i [F(y_2) - F(y_1)]^t \times [F(y_3) - F(y_2)]^{n-i-q-t} [1 - F(y_3)]^q, \qquad \text{for } y_1 < y_2 < y_3. \quad (7)$$

The conditional probability of $\{A_r = k, B_r = i\}$, given (6), is obtained by summing (7) over all $q = 0, \ldots, r$ and $t = 0, \ldots, n-i-1$. Hence, the unconditional probability of $\{A_r = k, B_r = i\}$, with respect to the joint density of $Y_{(1+r)}$, $Y_{(n-k)}$, and $Y_{(n-k+1)}$, is

$$P(A_r = k, B_r = i) = \sum_{q=0}^{r} \sum_{t=0}^{n-i-r-1} \frac{n!}{i!t!(n-i-q-t)!q!} \int_{-\infty}^{\infty} \int_{y_1}^{\infty} \int_{y_2}^{\infty} [F(y_1)]^i \times [F(y_2) - F(y_1)]^t [F(y_3) - F(y_2)]^{n-i-q-t} [1 - F(y_3)]^q \times g_{1+r,n-k,n-k+1:n}(y_1, y_2, y_3) \, dy_3 \, dy_2 \, dy_1, \quad (8)$$

where $g_{1+r,n-k,n-k+1:n}$ is the joint density of the three order statistics $Y_{(1+r)}$, $Y_{(n-k)}$, and $Y_{(n-k+1)}$ from the Y-sample given by (see David and Nagaraja, 2003; Arnold, Balakrishnan, and Nagaraja, 2008)

$$g_{1+r,n-k,n-k+1:n}(y_1, y_2, y_3) = \frac{n!}{r!(n-k-r-2)!(k-1)!} \times [G(y_1)]^r [G(y_2) - G(y_1)]^{n-k-r-2} [1 - G(y_3)]^{k-1} g(y_1) g(y_2) g(y_3), \quad y_1 < y_2 < y_3, \quad (9)$$

and g is the density corresponding to G. We, therefore, obtain (3).

Similarly, the joint distribution can be represented for $0 \leq i \leq n-r-1$ and $k = n-r-1$ or $k = 0$ using the joint density of two order statistics from distribution G, thus obtaining the expressions in (4) and (5) □

The theorem is true for any two continuous distributions F and G, and can be applied to specific distributions.

In the case when $n-r \le i,\, k \le n$, the ordering of the observations can be viewed as a symmetric image of the ordering for the first case with the following switches: $F \leftrightarrow G$; $(r+1) \leftrightarrow (n-r)$; $i \leftrightarrow (n-k)$; $k \leftrightarrow (n-i)$. Analogous expressions to (3), (4) and (5) can, therefore, be easily derived in this case.

3 Null Distribution

Suppose that the distribution functions F and G of the samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are equal which is often the null hypothesis of interest. Then, the joint distribution of the number of precedences and exceedances is as follows.

Theorem 2. *The joint distribution of A_r and B_r, under $H_0 : F(x) = G(x)$, is given by*

$$P(A_r = k, B_r = i | H_0) = \frac{\binom{r+k}{r}\binom{r+i}{r}}{\binom{2n}{n}} \binom{2n-2r-i-k-2}{n-r-i-1}, \qquad (10)$$

for $0 \le i,\, k \le n-r$, and

$$P(A_r = k, B_r = i | H_0) = \frac{\binom{2n-r-k-1}{n-r-1}\binom{2n-r-i-1}{n-r-1}}{\binom{2n}{n}} \binom{k+i-2n+2r}{i-n+r}, \qquad (11)$$

for $n-r \le i,\, k \le n$.

Proof: The proof of (10) follows by substituting $G(x) = F(x)$ in the representation (3) of the joint distribution of A_r and B_r. The proof for $k = n-r-1$ and $k = 0$ is similar by using representations (4) and (5) of the joint distribution.

Let $1 \le i \le n-r-1$ and $1 \le k \le n-r-1$ be fixed. Substituting $G(x) = F(x)$ in the representation (3) of the joint distribution of A_r and B_r, we obtain

$$\begin{aligned} P(A_r = k, B_r = i) &= \frac{n!n!}{r!i!(n-k-r-2)!(k-1)!} \sum_{q=0}^{r} \sum_{t=0}^{n-i-r-1} \frac{1}{t!(n-i-q-t)!q!} \\ &\quad \times \int_{-\infty}^{\infty}\int_{y_1}^{\infty}\int_{y_2}^{\infty} [F(y_1)]^{i+r}[F(y_2)-F(y_1)]^{t+n-k-r-2}[F(y_3)-F(y_2)]^{n-i-q-t} \\ &\quad \times [1-F(y_3)]^{q+k-1} f(y_1)f(y_2)f(y_3)\, dy_3\, dy_2\, dy_1\,. \qquad (12) \end{aligned}$$

Upon substituting $u_i = 1 - F(y_i)$, $i = 1,2,3$, and $du_i = -f(y_i)\,dy_i$, the integral in (12) is simplified as

$$Z = \int_0^1\int_0^{u_1}\int_0^{u_2} (1-u_1)^{i+r}(u_1-u_2)^{t+n-k-r-2}(u_2-u_3)^{n-i-q-t} u_3^{q+k-1}\, du_3\, du_2\, du_1\,.$$

Next, by substituting $w_1 = u_3/u_2$ with $du_3 = u_2\,dw_1$, and further $w_2 = u_2/u_1$ with $du_2 = u_1\,dw_2$, we obtain

$$
\begin{aligned}
Z &= \int_0^1 \int_0^{u_1} \int_0^1 (1-u_1)^{i+r}(u_1-u_2)^{t+n-k-r-2} u_2^{n-i-q-t}(1-w_1)^{n-i-q-t} u_2^{q+k-1} \\
&\qquad\qquad\qquad\qquad \times w_1^{q+k-1} u_2\,dw_1\,du_2\,du_1 \\
&= B(q+k, n-i-q-t+1) \int_0^1 \int_0^{u_1} (1-u_1)^{i+r}(u_1-u_2)^{t+n-k-r-2} u_2^{n-i-t+k}\,du_2\,du_1 \\
&= B(q+k, n-i-q-t+1) \int_0^1 \int_0^1 (1-u_1)^{i+r} u_1^{t+n-k-r-2}(1-w_2)^{t+n-k-r-2} \\
&\qquad\qquad\qquad\qquad \times u_1^{n-i-t+k} w_2^{n-i-t+k} u_1\,dw_2\,du_1 \\
&= B(q+k, n-i-q-t+1) B(n-i-t+k+1, t+n-k-r-1) \\
&\qquad\qquad\qquad\qquad \times B(2n-r-i, i+r+1)\,,
\end{aligned}
$$

where $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$ denotes the complete beta function.

Now substituting the above expression for Z in (12), and then expressing the beta functions in terms of gamma functions and performing some simple algebra, we get

$$
P(A_r = k, B_r = i | H_0) = \frac{\binom{r+k}{r}\binom{r+i}{r}}{\binom{2n}{n}} \binom{2n-2r-i-k-2}{n-r-i-1}\,. \tag{13}
$$

Further, as we have noted, the distribution of A_r and B_r for the case $i \geq n-r$ and $k \geq n-r$ is symmetrical to the case $i, k \leq n-r-1$ with appropriate replacements. So, if we denote $\psi_0(n, r, k, i)$ to be the right side of (13), then for $n-r \leq i, k \leq n$, we get

$$
\begin{aligned}
P(A_r = k, B_r = i | H_0) &= \psi_0(n, n-r-1, n-i, n-k) \\
&= \frac{\binom{2n-r-k-1}{n-r-1}\binom{2n-r-i-1}{n-r-1}}{\binom{2n}{n}} \binom{k+i-2n+2r}{i-n+r}\,.
\end{aligned} \tag{14}
$$

Now, by combining (13) and (14), we obtain the required result. □

4 Notes on the Limiting Distributions

In this section, we derive asymptotic properties of the distributions of properly normalized exceedance statistics connected with infinitely increasing sample sizes.

4.1 Exceedances from one Sample

We start with the limiting behaviour of the exceedance statistic $S_n(Z_L)$ which is equal to the number of observations in the sample $Y_1, Y_2, \ldots, Y_n$ preceding the level Z_L (see

Equation (1)). The result is due to Wesołowski and Ahsanullah (1998), and is as follows. For $n \to \infty$, we have

$$\frac{1}{n} S_n(Z_L) \xrightarrow{d} G(Z_L) \,,$$

where G is the distribution function of the sequence Y.

The asymptotic joint distribution of $(S_n(Z_L)/n, S_n(Z_R)/n)$ is due to Bairamov and Eryilmaz (2008). For $n \to \infty$, they showed that

$$\left(\frac{S_n(Z_L)}{n}, \frac{S_n(Z_R)}{n} \right) \xrightarrow{d} (G(Z_L), G(Z_R)) \,.$$

4.2 Exceedances from two Samples

The limiting behaviour of the joint distribution of exceedance statistics with respect to two samples depends on the tails of the parent distributions. Note that the order statistic $Y_{(1+r)}$ is a point estimate of the p-th quantile of the Y-distribution, where $1 + r = [np]$. Similarly, $X_{(n-r)}$ is a point estimate of the $(1-p)$-th quantile of the X-distribution. Thus, B_r counts the number of observations from the X-sample before $G^{-1}(p)$ and A_r counts the number of observations from the Y-sample after $F^{-1}(1-p)$. Hence, as $n, m \to \infty$ and $r/m \to p$, we have

$$\frac{B_r}{m} \xrightarrow{d} F(G^{-1}(p)) \qquad \text{and} \qquad \frac{A_r}{n} \xrightarrow{d} G(F^{-1}(1-p)) \,.$$

5 Monte Carlo Simulation for Location-Shift Alternative

In order to evaluate the distributional properties of different exceedance statistics specified by r, we consider the location-shift alternative of the form $F(x) = G(x + \theta)$ for some $\theta > 0$, where θ is the shift in locatiion between the two distributions. We generate samples of size $n = 25$ from the following three distributions.

1. Uniform distribution in [0, 1],
2. Standard normal distribution,
3. Standard exponential distribution with cdf $F(x; \theta) = 1 - e^{-x/\theta}$, $x > 0$, $\theta > 0$.

The values of (A_r, B_r), with $r = 0, 4$ and 8, were determined through Monte Carlo simulations when the shift is $\theta = 0, 0.5$, and 0.75.

In Figure 2 we have presented the joint distribution of A_r and B_r for the Uniform distribution [0, 1] when $n = 25$. In order to obtain the estimated exceedance statistics A_r and B_r, we generated 100000 sets of data from $U[0, 1]$ and $U[0 + \theta, 1 + \theta]$ for the first and second samples, respectively. The upper row plots show the joint distribution in case of equal distributions, i.e., no shift in the second sample. From left to right, we have the distributions of (A_1, B_1), (A_4, B_4) and (A_8, B_8). The middle row plots show the same joint distribution for a shift of 0.5 in the second distribution while the Lower row corresponds to a shift of 0.75 in the second distribution. In Figures 3 and 4, we have presented similar plots for the cases of Normal and Exponential distributions, respectively.

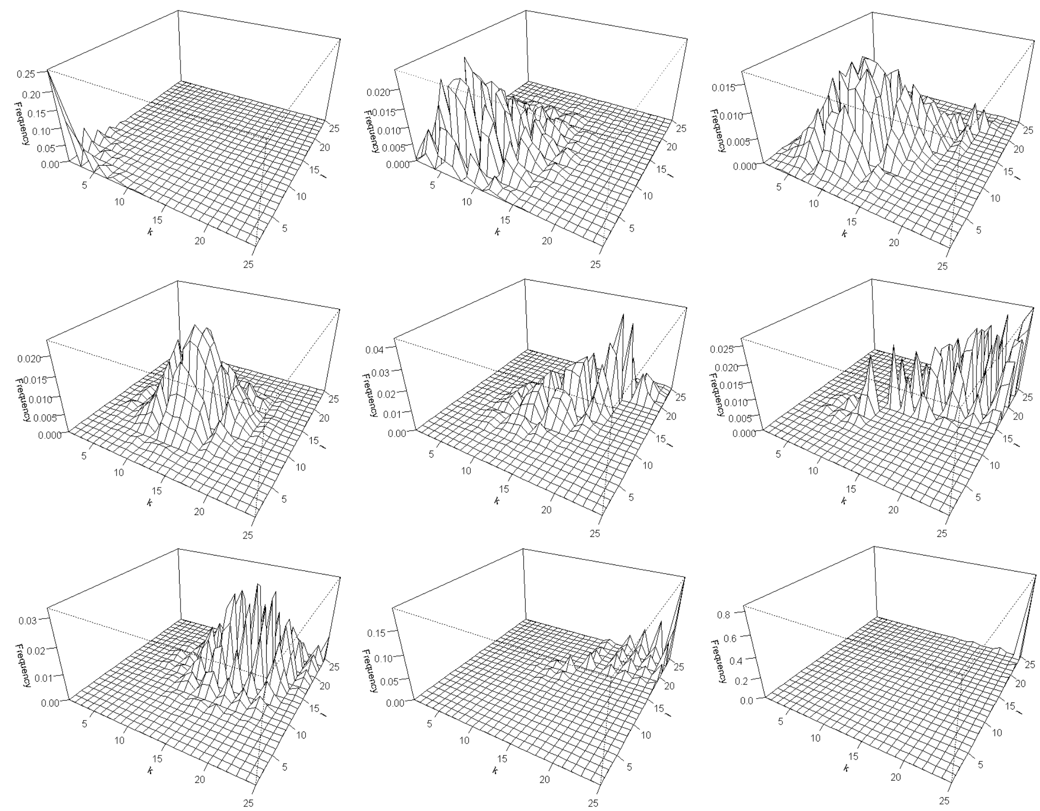

Figure 2: Joint distribution for the Uniform distribution [0,1] when $n = 25$. Upper row: no shift in the second distribution. From left to right, the r parameter is 0, 4 and 8. Middle row: shift of 0.5 in the second distribution. Lower row: shift of 0.75 in the second distribution.

From Figures 2 – 4 we see that the joint distribution of (A_r, B_r) is quite sensitive to the size of the shift for any of the parent distributions. For $n = 25$, the distribution for different r is located at different centers. Clearly, for larger r the amount of information relative to the sample size is greater. However, for small samples, such as 8 to 15, the exceedance statistics for $r = 0$ gives enough information for the shift in either symmetric or right-skewed distributions. In other words, this means that when $n = 8$, the number of observations from the X-sample before the first observation in the Y-sample and the number of observations from the Y-sample after the last observation in the X-sample, provide enough information to make a good decision about the difference between the two distributions, (in terms of location, of course).

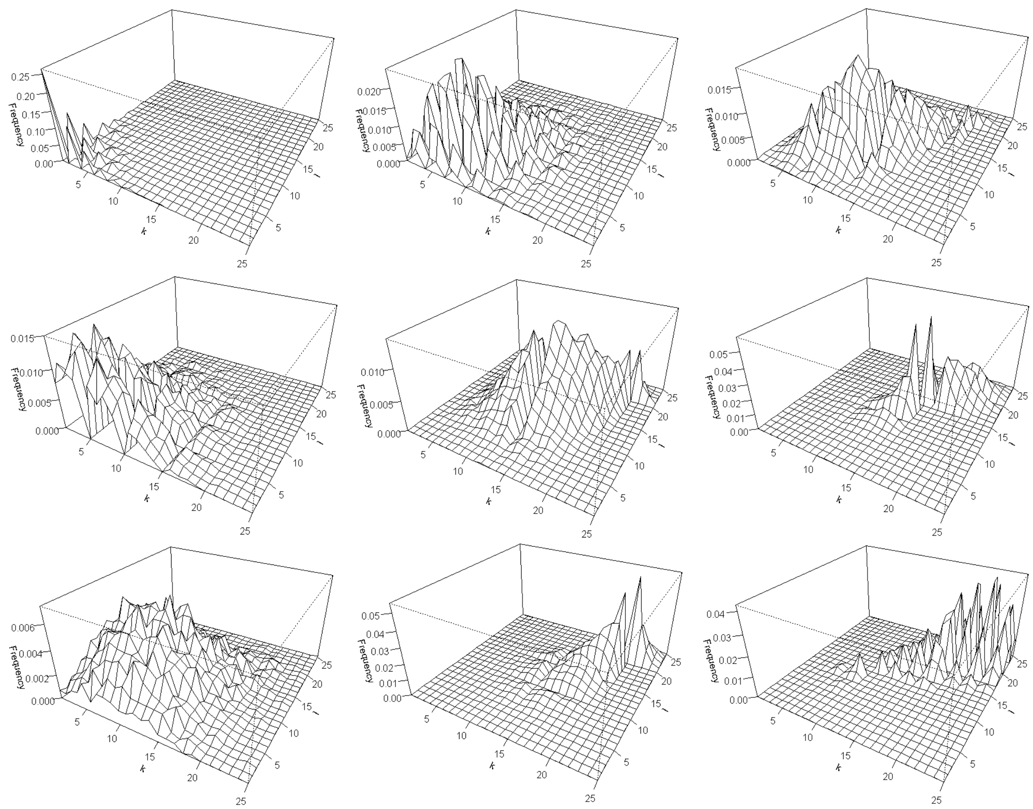

Figure 3: Joint distribution for the Normal distribution [0,1] when $n = 25$. Upper row: no shift in the second distribution. From left to right, the r parameter is 0, 4 and 8. Middle row: shift of 1 in the second distribution. Lower row: shift of 1.5 in the second distribution.

References

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (2008). *A First Course in Order Statistics. Reprint of the 1992 original edition*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Bairamov, I., and Eryilmaz, S. (2008). Joint behaviour of precedences and exceedances in random threshold models. *Australian and New Zealand Journal of Statistics*, *50*, 209-216.

Balakrishnan, N., and Ng, H. K. T. (2006). *Precedence-Type T-tests and Applications*. Hoboken, NJ: John Wiley & Sons.

David, H. A., and Nagaraja, H. N. (2003). *Order Statistics* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Stoimenova, E., and Balakrishnan, N. (2010). *A class of exceedance-type statistics for the two-sample problem.* ((submitted for publication))

van der Laan, P., and Chakraborti, S. (2001). Precedence tests and Lehmann alternatives. *Statistical Papers*, *42*, 301-312.

Wesołowski, J., and Ahsanullah, M. (1998). Distributional properties of exceedance statistics. *Annals of the Institute of Mathematical Statistics*, *50*, 543-565.

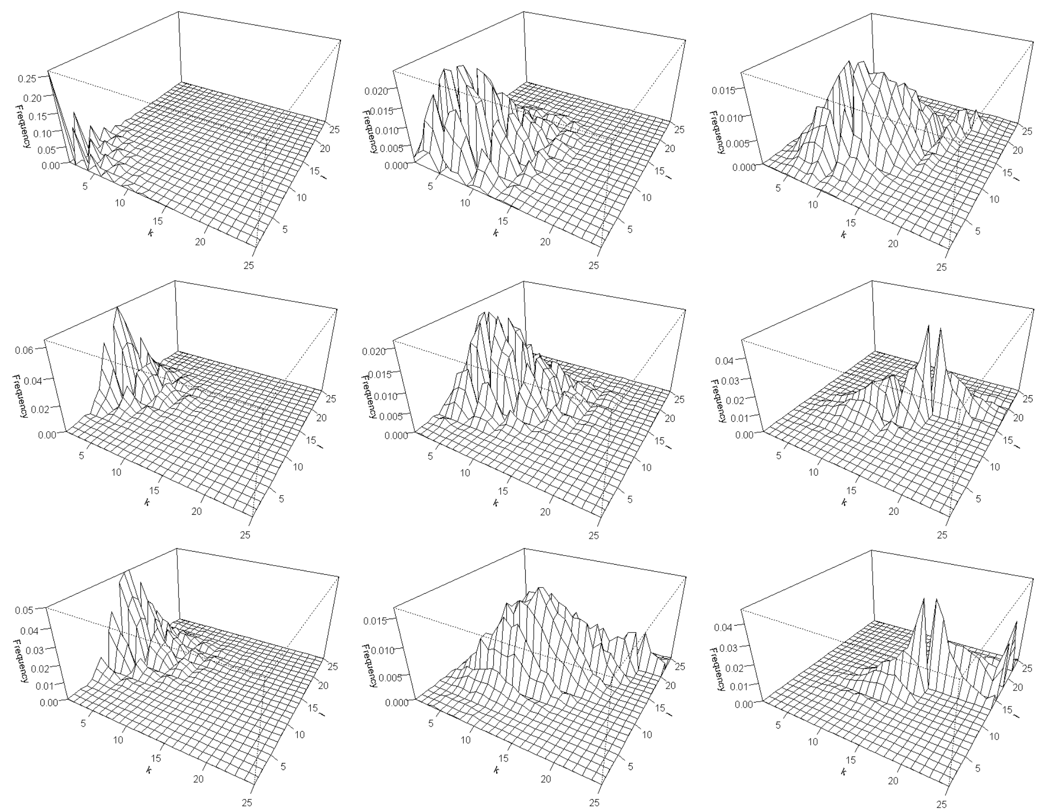

Figure 4: Joint distribution for the Exponential distribution E(1) when $n = 25$. Upper row: no shift in the second distribution. From left to right, the r parameter is 0, 4 and 8. Middle row: shift of 0.5 in the second distribution. Lower row: shift of 0.75 in the second distribution.

Authors' addresses:

Eugenia Stoimenova
Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bontchev str., block 8
1113 Sofia
Bulgaria
E-email: jeni@math.bas.bg

Narayanaswamy Balakrishnan
Department of Mathematics and Statistics
McMaster University
Hamilton, Ontario
Canada L8S 4K1
E-email: bala@mcmaster.ca