Aplikace matematiky

Igor Očka

Simple random walk and rank order statistics

Aplikace matematiky, Vol. 22 (1977), No. 4, 272–290

Persistent URL: http://dml.cz/dmlcz/103703

SIMPLE RANDOM WALK AND RANK ORDER STATISTICS

IGOR OČKA

(Received May 5, 1976)

INTRODUCTION

The present article is closely related to Dwass's article [1]. The method used in [1] is based on the analogy of rank order statistics and functions on a simple random walk, and it is applied to the case of equal sample sizes in the two-sample problem. Here the method will be extended to the case of arbitrary sample sizes. This approach simplifies much the calculation of the distributions of rank order statistics compared to the combinatorial approach used e.g. in Reimann-Vincze [4]. Another extension of Dwass's appeared in Mohanty-Handa [3], namely for two samples where one sample size is a multiple of the other.

I. THE METHOD

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be two samples from the same distribution with a continuous distribution function. Let the combined sample of $n + m$ values arranged in the increasing order be denoted by $Z_1, Z_2, \ldots, Z_{n+m}$. Let us replace the X's by 1's and Y's by -1's in this sequence. We call such a sequence of l's and -1's a *sequence of rank order indicators*, and we shall denote it by $V_1, V_2, \ldots, V_{n+m}$. In addition, we define $V_0 = 0$.

There are $\binom{n+m}{n}$ different sequences of rank order indicators, and they have the same probability $\binom{n+m}{n}^{-1}$. We can define different random variables on the sequence $V_1, V_2, \ldots, V_{n+m}$, e.g.

$$\max_{k=0,1\ldots,n+m} \sum_{i=0}^{k} V_i ; \quad \min_{k=0,1,\ldots,n+m} \sum_{i=0}^{k} V_i .$$

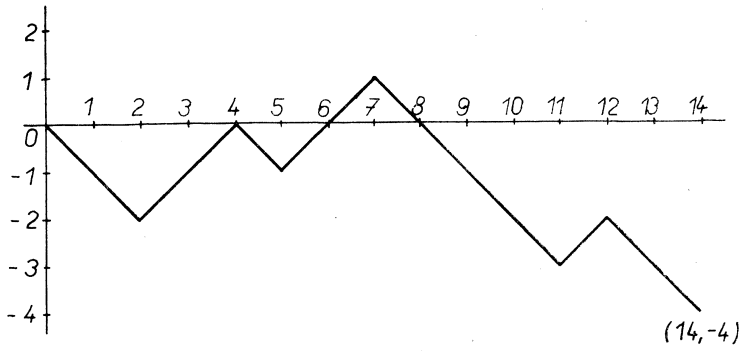
In this article we shall find the distribution functions of many such random variables. The method used here is a slight generalization of Dwass's method from [1], based on the relation between a simple random walk and a sequence of rank order indicators. Dwass [1] deals with the special case $n = m$, while in the present article the distributions of all statistics mentioned by Dwass [1] will be derived for the case of arbitrary n and m. Moreover, some other statistics will be studied.

We can suppose $m \geqq n$, and we put $d = m - n$. Then $n + m = 2n + d$, and $m = n + d$, where $d \geqq 0$. In the sequel, we consider d to be an arbitrary but fixed constant, while n will change; this is the basic step for the generalization of Dwass's theory.

Definition. *A random variable U_n which is a function of $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_{n+d}$ only through the rank order indicators is called a rank order statistic.*

Further, let $F_n(x) = (\text{number of } X_i's \leqq x)/n$ be the empirical distribution function of $X_1, X_2, \ldots, X_n$, similarly $G_m(x)$ the empirical distribution function of $Y_1, Y_2, \ldots, Y_m$, and write $H_{n,m}(x) = n F_n'x) - m G_m(x)$. Finally, $\chi_{\mathbf{A}}$ will denote the characteristic function of the set $\mathbf{A}$.

We can consider each sequence of rank order indicators to be equivalent to a simple random walk which starts at $(0, 0)$ and ends at the point $(2n + d, -d)$. For example: for $n = 5$, $d = 4$, and the sequence of rank order indicators $-1, -1, 1, 1, -1, 1, 1, -1, -1, -1, -1, 1, -1, -1$ we have the following random walk.



Let us derive the relation between rank order statistics and analogous functions on a simple random walk.

Let $W_1, W_2, \ldots$ be independent random variables with the same distribution given by

$$W_i = 1, \quad \text{with probability} \quad p;$$

$$= -1, \quad \text{with probability} \quad 1 - p = q,$$

and let $W_0 = 0$.

We define a simple random walk by the random variables $S_n = \sum_{i=0}^{n} W_i$. Since the sequence of rank order indicators always ends at the point $(2n + d, -d)$, we shall need a conditional random walk passing through this point.

Lemma 1. *For any $p \in (0, 1)$, the conditional distribution of $W_1, W_2, \ldots, W_{2n+d}$ given that $S_{2n+d} = -d$ assigns equal probabilities to each of the $\binom{2n+d}{n}$ possible sequences $W_1, W_2, \ldots, W_{2n+d}$.*

Proof is easy by direct calculation of the conditional probability.

If $p < \frac{1}{2}$ the simple random walk S_n is transient and passes through the point $(2n + d, -d)$ for some n with probability 1. However, also with probility 1 it passes through points $(2n + d, -d)$, n arbitrary, only finitely many times. Thus we can define a random variable T by

$$T = \max_{n=0,1,\ldots} (2n + d; \sum_{i=0}^{2n+d} W_i = -d), \text{ if this maximum exists,}$$
$$= 0, \text{ otherwise.}$$

Definition. *Let U be a function defined on the random walk S_n. We say that it satisfies Assumption A if the value of U is completely determined by $W_1, W_2, \ldots, W_T$ and does not depend on $W_{T+1}, W_{T+1}, \ldots$, whenever $T > 0$.*

The following Lemma 2 gives the relation between the distribution of U and that of a rank order statistic.

Lemma 2. *a) The conditional distribution of $W_1, W_2, \ldots, W_T$ given that $T = 2n + d$ assigns equal probabilities to each of the $\binom{2n+d}{n}$ possible sequences of n* numbers 1 *and $n + d$ numbers -1. If a function U satisfies Assumption A, the conditional distribution of U given that $T = 2n + d$ is exactly that of a rank order statistic.*

b) Conversely, suppose U_n is a rank order statistic defined for every $n = 1, 2, \ldots$. Then there is a function U satisfying Assumption A, such that the conditional distribution of U given that $T = 2n + d$ is exactly the distribution of U_n.

Proof. Part a) follows from Lemma 1. To prove part b), we define U arbitrarily for $T = 0$, and

$U(W_1, W_2, \ldots, W_T, \ldots) = U_n(W_1, W_2, \ldots, W_{2n+d})$ for $T > 0$, $T = 2n + d$. This U satisfies Assumption A and it is defined for all n, which proves the lemma.

The following theorem enables us to determine the distributions of rank order statistics.

Theorem. *Suppose U_n is a rank order statistic for every n and U is the related function satisfying Assumption A. Define*

$$E(U) = h(p), \quad 0 \leqq p < \tfrac{1}{2}.$$

Then the following equality is valid for $p \in [0, \frac{1}{2})$, where the right-hand side is a power series in powers of $p(1-p) = pq$:

$$\frac{h(p)}{(1-2p)\,q^d} = \sum_{n=0}^{\infty} E(U_n)\binom{2n+d}{n}(pq)^n. \tag{1}$$

Proof. Clearly,

$$P(T = 2n+d) = P\Big(\sum_{i=1}^{2n+d} W_i = -d\Big)\, P\Big(\forall_{k>0} \sum_{i=1}^{2n+d+k} W_i \neq -d \,\Big|\, \sum_{i=1}^{2n+d} W_i = -d\Big) =$$

$$= P\Big(\sum_{i=1}^{2n+d} W_i = -d\Big)\, P\Big(\forall_{k>0} \sum_{i=2n+d+1}^{2n+d+k} W_i \neq 0\Big) = p^n q^{n+d}\binom{2n+d}{n}(1-2p),$$

since the sum $\sum_{i=1}^{2n+d} W_i = -d$ must be composed of $n+d$ addends -1 and n addends 1 and the number of such sequences is $\binom{2n+d}{n}$, and since the probability that the simple random walk never returns to the origin is $1-2p$ (see [2], Chap. XI). Hence, using Lemma 2, we obtain

$$h(p) = E(U) = \sum_{n=0}^{\infty} E(U \mid T = 2n+d)\, P(T = 2n+d) = \sum_{n=0}^{\infty} E(U_n)\binom{2n+d}{n} \cdot (pq)^n q^d (1-2p),$$

which proves the theorem.

Since the right-hand side of (1) is a power series in powers of pq, it is sufficient to express the fraction $h(p)/[(1-2p)\,q^d]$ as a power series in powers of pq with coefficients a_n; on comparing the coefficients we obtain

$$E(U_n) = \binom{2n+d}{n}^{-1}.$$

In deriving individual distributions we shall work with functions $\varphi(U)$ where φ will be usually a function on the set of values of U_n. It is clear that $\varphi(U_n)$ will be also rank order statistic. For φ being the characteristic function of a set $\boldsymbol{A}$ we have

$$E(\varphi(U_n)) = P(U_n \in \boldsymbol{A})$$

and equation (1) becomes

$$\frac{P(U \in \boldsymbol{A})}{(1-2p)\,q^d} = \sum_{n=0}^{\infty} P(U_n \in \boldsymbol{A})\binom{2n+d}{n}(pq)^n. \tag{2}$$

In Feller [2] and Dwass [1] the probabilities $P(U \in \mathbf{A})$ are derived for different U and $\mathbf{A}$. It then depends on our ability to find the power series (2) in individual cases. Some distributions of rank order statistics are derived in the following part.

II. DISTRIBUTIONS OF RANK ORDER STATISTICS

In each derivation of the distribution of a rank order statistic the proper choice of the function φ and the definition of the statistic U is mentioned.

II.1. **Distribution of $N_{n,m}$**

First we define $Z_0 = -\infty$ in addition to the sequence $Z_1, Z_2, \ldots, Z_{2n+d}$. We put

$$N_{n,m} = \textit{number of indices } 0 \leqq i \leqq 2n + d \quad \text{for which} \quad H_{n,m}(Z_i) = 0 .$$

We choose $U = \varphi(U')$, $U_n = \varphi(U_n')$, where $U_n' = N_{n,m}$, U' = number of indices $i \geqq 0$ for which $S_i = 0$, and $\varphi = \chi_{\{x>k\}}$, $k \geqq 0$ integer. According to the preceding part we have

$$E(U) = P(U' > k) = h(p) , \quad E(U_n) = P(N_{n,m} > k) .$$

In [1], appendix (3), one can find that $P(U' > k) = (2p)^k$. We derive the power series for the expression

$$\frac{(2p)^k}{(1 - 2p)\, d^d} \tag{3}$$

in powers of pq. In part III.1 we shall prove that

$$(2p)^k = 2^k \sum_{n=k}^{\infty} \binom{2n + d - k}{n - k} (pq)^n \, q^d (1 - 2p) ,$$

and therefore

$$\frac{(2p)^k}{(1 - 2p)\, q^d} = \sum_{n=k}^{\infty} 2^k \binom{2n + d - k}{n - k} (pq)^n , \tag{4}$$

and

$$P(N_{n,m} > k) = 2^k \binom{2n + d - k}{n - k} \binom{2n + d}{n}^{-1} .$$

If we put $d = 0$ we obtain the same result as Dwass [1].

II.2. **Distribution of $N_{n,m}^+$ and $N_{n,m}^-$ — number of positive and negative sojourns**

Let $0 = i_0 < i_1 < \ldots < i_{N_{n,m}}$ be all the indices for which $H_{n,m}(Z_i) = 0$. The part of the sequence of the rank order indicators between i_{j-1}, i_j will be called a *sojourn*.

If $H_{n,m}(Z_i) > 0$ for $i_{j-1} < i < i_j$ we say that the j-th sojourn is pozitive, and if $H_{n,m}(Z_i) < 0$ for $i_{j-1} < i < i_j$ we say that the j-th sojourn is negative. We define

$$N^+_{n,m} = \textit{number of positive sojourns}\,,$$

$$N^-_{n,m} = \textit{number of negative sojourns}\,.$$

To find the distribution of $N^+_{n,m}$ we take U' = number of positive sojourns in the simple random walk, $\varphi = \chi_{\{x \geqq k\}}$, $U = \varphi(U')$, $U_n = \varphi(N^+_{n,m})$. If we use the equality $P(U' \geqq k) = (p/q)^k$, $p < q$, from [1], and the result from III.2, formula (2) becomes

$$\frac{p^k}{(1-2p)\,q^k q^d} = \sum_{n=k}^{\infty} \binom{2n+d}{n-k} (pq)^n\,,$$

so that

$$P(N^+_{n,m} \geqq k) = \binom{2n+d}{n-k}\binom{2n+d}{n}^{-1} \quad \text{for} \quad k = 1, 2, \ldots, n\,.$$

For $d = 0$ this is the same result as in [1].

Let the sequence of rank order indicators $V_0, V_1, \ldots, V_{2n+d}$ have the sum $\sum_{i=1}^{2n+d} V_i = -d$. Let i be such index that $V_0 + \ldots + V_i = 0$ and $V_0 + \ldots + V_j \neq 0$ for all $j > i$. Then the sequence of rank order indicators $-V_0, -V_1, \ldots, -V_i, V_{i+1}, \ldots, V_{2n+d}$ has the same number of negative sojourns as is the number of positive sojourns in the sequence $V_0, V_1, \ldots, V_{2n+d}$. Both sequences have the same probability so that the distribution of $N^-_{n,m}$ must be equal to the distribution of $N^+_{n,m}$.

II.3. **Distribution of $N_{n,m}(r)$ – number of visits to height r**

We denote

$$N_{n,m}(r) = \textit{number of indices } i\,,\ 0 \leqq i \leqq 2n + d, \textit{ for which}$$

$$H_{n,m}(Z_i) = r\,, \quad r \geqq 0\,,$$

the rank order statistic called the number of visits to height r. Let U' = number of indices $i \geqq 0$, for which $S_i = r$, $r \geqq 0$, $\varphi = \chi_{\{x>k\}}$, $U_n = \varphi(N_{n,m}(r))$, $U = \varphi(U')$. From [1], appendices (6) and (3), it follows that

$$P(U' > k) = (p/q)^r\,(2p)^k \quad \text{for} \quad r \geqq 0\,.$$

By III.1 we get

$$\frac{h(p)}{(1-2p)\,q^d} = \frac{2^k}{(1-2)\,q^d}\,\frac{p^{k+2r}}{(pq)^r} =$$

$$= 2^k \sum_{n=k+2r}^{\infty} \binom{2n+d-k-2r}{n-k-2r} (pq)^{n-r}\,.$$

Hence the distribution we are looking for has the form

$$P(N_{n,m}(r) > k) = 2^k \binom{2n + d - k}{n - k - r} \binom{2n + d}{n}^{-1}, \tag{5}$$

$$\text{for} \quad r = 0, 1, \ldots, n, \quad k = 0, 1, \ldots, n - r.$$

In [1] one can find the same result for $d = 0$.

II.4. Distribution of $N^+_{n,m}(r)$ – upcrossings of r

We define the rank order statistic called the upcrossings of r as

$$N^+_{n,m}(r) = \textit{number of indices } i,\ 0 < i \leqq 2n + d, \textit{ for which}$$
$$H_{n,m}(Z_i) = r + 1 \quad \text{and} \quad H_{n,m}(Z_{i-1}) = r, \quad r \geqq 0.$$

We choose U' = number of indices $i > 0$ in the simple random walk for which $S_{i-1} = r$, $S_i = r + 1$, $r \geqq 0$, $\varphi = \chi_{\{x \geqq k\}}$, $U_n = \varphi(N^+_{n,m}(r))$, $U = \varphi(U')$. In [1], p. 1052, it is proved that

$$E[\varphi(U')] = h(p) = (p/q)^{k+r}, \quad k > 0.$$

This is analogous to the preceding case and we obtain

$$P(N^+_{n,m}(r) \geqq k) = \binom{2n + d}{n - k - r} \binom{2n + d}{n}^{-1}, \quad \text{for} \quad r \geqq 0, \quad k = 1, 2, \ldots n - r.$$

II.5. Distribution of $N^*_{n,m}(r)$ – number of crossings of r

We shall find the distribution of the statistic

$$N^*_{n,m}(r) = \text{number of indices } i, \quad 0 < i < 2n + d, \text{ for which}$$
$$H_{n,m}(Z_{i-1}) < r \quad \text{and} \quad H_{n,m}(Z_{i+1}) > r, \quad \text{or}$$
$$H_{n,m}(Z_{i-1}) > r \quad \text{and} \quad H_{n,m}(Z_{i+1}) < r, \quad \text{for} \quad r \geqq 0,$$

called the number of crossings of r. The functions U, U_n in this case are U' = number of indices $i > 0$ for which $S_{i-1} < r$ and $S_{i+1} > r$; or $S_{i-1} > r$ and $S_{i+1} < r$, $\varphi = \chi_{\{x \geqq k\}}$, $U = \varphi(U')$, $U_n = \varphi(N^*_{n,m}(r))$. It is proved in [1], p. 1048, that

$$P(U' \geqq 2k) = (p/q)^{r+2k-1}, \quad \text{for} \quad r > 0, \quad k > 0,$$
$$P(U' \geqq 2k) = 2p^{k+1}/q^k, \quad \text{for} \quad r = 0, \quad k > 0.$$

First, we shall study the case $r > 0$. By the same procedure as in II.3. it can be calculated that

$$(p/q)^{r+2k-1} \frac{1}{(1 - 2p)\, d^d} = \sum_{n=r+2k-1}^{\infty} \binom{2n + d}{n - r - 2k + 1} (pq)^n.$$

The distribution of the rank order statistic is

$$P(N^*_{n,m}(r) \geqq 2k) = \binom{2n + d}{n - r - 2k + 1}\binom{2n + d}{n}^{-1}, \quad \text{for} \quad r > 0, \quad k > 0.$$

Second, in the case $r = 0$ we obtain by (4)

$$\frac{2p^{k+1}}{q^k}\frac{1}{(1 - 2p)\,q^d} = \frac{2p^{2k+1}}{(pq)^k(1 - 2p)\,q^d} = 2\sum_{n=2k+1}^{\infty}\binom{2n + d - 2k - 1}{n - 2k - 1}(pq)^{n-k} =$$

$$= 2\sum_{n=k+1}^{\infty}\binom{2n + d - 1}{n - k - 1}(pq)^k$$

and the distribution is

$$P(N^*_{n,m}\,0 \geqq 2k) = 2\binom{2n + d - 1}{n - k - 1}\binom{2n + d}{n}^{-1}, \quad \text{for} \quad k > 0.$$

If we put $d = 0$ in both cases we obtain the same results as Dwass [1].

II.6. **Distribution of $D^+_{n,m}$ upper side maximum deviation**

Let us find the distribution of the upper side maximum deviation

$$D^+_{n,m} = \max_x (nF_n(x) - mG_m(x)) = \max_x H_{n,m}(x).$$

We take functions $U = \varphi(U')$, $U_n = \varphi(D^+_{n,m})$ where $\varphi = \chi_{\{x \geqq k\}}$, $U' = \max\{0, S_1, S_2, \ldots\}$. We can find in [1], p. 1051, that $E(U) = (p/q)^k$. It is easy to prove, similarly as in II.2, that

$$(6) \qquad P(D^+_{n,m} \geqq k) = \binom{2n + d}{n - k}\binom{2n + d}{n}^{-1}, \quad \text{for} \quad k = 0, 1, \ldots, n.$$

This result agrees with Dwass's formula for $n = m$. Reimann-Vincze [4] prove that

$$P(D^+_{n,m} = k) = \frac{m - n + 2k + 1}{m + k + 1}\binom{m + n}{n - k}\binom{m + n}{n}^{-1},$$

$$\text{for} \quad k = 0, 1, \ldots, n,$$

which is not difficult to verify again by (6).

II.7. **Distribution $Q_{n,m}$ – number of times $D^+_{n,m}$ is achieved**

The statistic $Q_{n,m}$ is defined as

$$Q_{n,m} = \textit{number of indices } i \textit{ for which } H_{n,m}(Z_i) = D^+_{n,m}.$$

We derive the distribution of the rank order statistic $Q_{n,m}$ by means of the two dimensional statistic $(D^+_{n,m}, Q_{n,m})$. In order to do it we put

$$U' = (D^+_{n,m}, Q_{n,m})$$

$$D^+ = \max\{0, S_1, S_2, \ldots\}$$

$$Q = \textit{number of indices } i \geqq 0 \textit{ for which } S_i = D^+ ,$$

$$U' = (D^+, Q),$$

$$\varphi(x, y) = \chi_{\{(x,y);\, x \geqq k, y \geqq r\}}(x, y)$$

$$U_n = \varphi(U'_n), \quad U = \varphi(U')$$

In [1], p. 1052, the following formula is proved:

$$P(D^+ \geqq k,\ Q \geqq r) = (p/q)^k\, p^{r-1} .$$

The desired power series equals

$$\frac{p^{2k+r-1}}{(1-2p)\, q^d (pq)^k} = \sum_{n=2k+r-1}^{\infty} \binom{2n+d-2k-r+1}{n-2k-r+1} (pq)^{n-k} =$$

$$= \sum_{n=k+r-1}^{\infty} \binom{2n+d-r+1}{n-k-r+1} (pq)^n$$

so that the distribution of the two-dimensional statistic is given by

$$P(D^+_{n,m} \geqq k,\ Q_{n,m} \geqq r) = \binom{2n+d-r+1}{n-k-r+1} \binom{2n+d}{n}^{-1},$$

$$\text{for} \quad r = 1, 2, \ldots, n+1, \quad k = 0, 1, \ldots, n-r+1,$$

and consequently

$$P(Q_{n,m} \geqq r) = P(D^+_{n,m} \geqq 0,\ Q_{n,m} \geqq r) = \binom{2n+d-r+1}{n-r+1} \binom{2n+d}{n}^{-1},$$

$$\text{for} \quad r = 1, 2, \ldots, n+1 .$$

II.8a. **Distribution of $Q_{n,m,k}$ — position of the k-th zero**

If the statistic $N_{n,m}$ defined in II.2 is larger than or equal to k the definition

$$Q_{n,m,k} = \textit{the index } i,\ 0 \leqq i \leqq 2n+d, \textit{ for which } H_{n,m}(Z_i) = 0 \textit{ for the } k\textit{-th time}$$

has a sense. The distribution of $Q_{n,m,k}$ will be obtained from the distribution of the two-dimensional statistic $(Q_{n,m,k}, N_{n,m})$. We define

$$U'_n = (Q_{n,m,k}, N_{n,m}), \quad U' = (Q_k, N)$$

where

$$Q_k = \textit{the index } i \geqq 0 \textit{ for which } S_i = 0 \textit{ for the } k\textit{-th time},$$

$$N = \textit{number of indices } j \geqq 0 \textit{ for which } S_j = 0 .$$

The formula

$$\sum_i P(Q_k = i,\ N = k + r)\, t^i = \left[1 - (1 - 4pqt^2)^{1/2}\right]^k (2p)^r (1 - 2p)$$

given in [1] suggests to define the suitable φ as

$$\varphi(x, y, t) = \sum_i \chi_{\{(x,y);\, x=i, y=k+r\}}(x, y) \,.\, t^i , \quad t \in (0, 1)$$

and to put $U = \varphi(U', t)$, $U_n = \varphi(U'_n, t)$. The function $h(p)$ equals the power series $\sum_i P(Q_k = i, N = k + r) \,.\, t^i$. We obtain by III.3

$$\frac{\left[1 - (1 - 4pqt^2)^{1/2}\right]^k (2p)^r}{q^d} =$$

$$= \frac{2^{k+r}}{(pq)^d} \,.\, 2^{-k}\left[1 - (1 - 4pqt^2)^{1/2}\right]^k \,.\, 2^{-(r+d)}\left[1 - (1 - 4pq)^{1/2}\right]^{r+d} =$$

$$= 2^{k+r} \sum_{i=r+d}^{\infty} \sum_{h=k}^{\infty} k(r + d)(2h - k)^{-1} \binom{2h - k}{h - k} (2i - r - d)^{-1} \binom{2i - r - d}{i - r - d} \,.$$

$$t^{2h}(pq)^{i+h-d} = 2^{k+r} \sum_{s=k}^{\infty} \sum_{n=r+s}^{\infty} k(r + d)(2s - k)^{-1} \binom{2s - k}{s - k} (2n + d - 2s - r)^{-1} \,.$$

$$\binom{2n + d - 2s - r}{n - s - r} (pq)^n t^{2s} .$$

The last equality follows by the substitution $i + h - d = n$, $h = s$. The statistic $Q_{n,m,k}$ cannot be equal to an odd number. The probability distribution of the two-dimensional statistic is

$$P(Q_{n,m,k} = 2i,\ N_{n,m} = k + r) =$$

$$= 2^{k+r} k(r + d)(2i - k)^{-1} (2n + d - 2i - r)^{-1} \,.$$

$$\binom{2i - k}{i - k} \binom{2n + d - 2i - r}{n - i - r} \binom{2n + d}{n}^{-1} ,$$

$$\text{for} \quad r = 1, 2, \ldots, n - k, \quad i = k, k + 1, \ldots, n .$$

II.8b. **Distribution of $Q_{m,n,k}$ – position of the k-th zero**

Let U'_n, U' be defined as in the preceding paragraph, but the function φ will be now

$$\varphi(x, y, t) = \sum_i \chi_{\{(x,y);\, x=i, y \geqq k\}}(x, y) \,.\, t^i , \quad t \in (0, 1) ,$$

and we put $U_n = \varphi(U_n', t)$, $U = \varphi(U', t)$. A consequence if the fact that the function $h(p)$ equals $[1 - (1 - 4pqt^2)^{1/2}]^k$ is

$$2^k(\tfrac{1}{2})^k [1 - (1 - 4pqt^2)^{1/2}]^k \cdot \frac{p^0}{(1 - 2p)\, q^d} =$$

$$= 2^k k \sum_{i=k}^{\infty} \sum_{j=0}^{\infty} (2i - k)^{-1} \binom{2i - k}{i - k} \binom{2j + d}{j} t^{2i} (pq)^{j+i} =$$

$$= 2^k k \sum_{i=k}^{\infty} \sum_{n=i}^{\infty} (2i - k)^{-1} \binom{2i - k}{i - k} \binom{2n - 2i + d}{n - i} (pq)^n t^{2i},$$

by III.1 and III.3 and the substitution $n = i + j$. Following the general theory we get the distribution

$$P(Q_{n,m,k} = 2i, N_{n,m} \geqq k) = 2^k k (2i - k)^{-1} \binom{2n - 2i + d}{n - i} \binom{2i - k}{i - k} \binom{2n + d}{n}^{-1},$$

$$\text{for} \quad k = 1, 2, \ldots, n, \qquad i = k, k + 1, \ldots, n.$$

If we put $d = 0$ both formulas from II.8a and b agree with Dwass's results.

II.9. Distribution of $R_{n,m}^+$ – index where $D_{n,m}^+$ is first achieved

We define the rank order statistic

$$R_{n,m}^+ = \min \{i \mid H_{n,m}(Z_i) = D_{n,m}^+, \quad i = 1, \ldots, 2n + d\}, \quad \text{if} \quad D_{n,m}^+ > 0,$$
$$= 0, \quad \text{for} \quad D_{n,m}^+ = 0.$$

We shall find the distribution of the two dimensional statistic

$$U_n' = (R_{n,m}^+, D_{n,m}^+).$$

We denote

$$D^+ = \max \{0, S_1, S_2, \ldots\},$$
$$R^+ = \min \{i \mid S_i = D^+, \; i = 0, 1, \ldots\},$$
$$U' = (R^+, D^+),$$
$$\varphi(x, y, t) = \sum_i \chi_{\{(x,y):\, x = i, y = k\}}(x, y) \cdot t^i, \quad t \in (0, 1),$$
$$U_n = \varphi(U_n', t), \quad U = \varphi(U', t).$$

In [1], p. 1051, we can find that

$$h(p) = (2qt)^{-k} [1 - 4(1 - pqt^2)^{1/2}]^k (1 - p/q).$$

We are able to calculate the distribution of the statistic $(R^+_{n,m}, D^+_{n,m})$ by means of III.3

$$\frac{h(p)}{(1-2p)\,q^d} = (\tfrac{1}{2})^k\left[1-(1-4pqt^2)^{1/2}\right]^k \frac{p^{d+k+1}}{(pq)^{d+k+1}\,t^k} =$$

$$= k(d+k+1)\sum_{j=k}^{\infty}\sum_{i=d+k+1}^{\infty}(2j-k)^{-1}\binom{2j-k}{j-k}(2i-d-k-1)^{-1}\cdot$$

$$\cdot\binom{2i-d-k-1}{i-d-k-1}t^{2j-k}(pq)^{j+i-d-k-1} = k(d+k+1)\sum_{j=k}^{\infty}\sum_{n=j}^{\infty}(2j-k)^{-1}\cdot$$

$$\cdot\binom{2j-k}{j-k}(2n-2j+d+k+1)^{-1}\binom{2n-2j+d+k+1}{n-j}t^{2j-k}(pq)^n$$

(by substitution $j+i-k-1=n$). Since the sum $r+k$ is always even $(2j-k=r)$ we can put $j=(r+k)/2$. Consequently, the distribution of the rank order statistics is

$$P(R^+_{n,m}=r,\ D^+_{n,m}=k) = k(d+k+1)\,r^{-1}(2n-r+d+1)^{-1}\cdot$$

$$\cdot\binom{r}{\frac{1}{2}(r-k)}\binom{2n-r+d+1}{n-\frac{1}{2}(r+k)}\binom{2n+d}{n}^{-1},$$

$$\text{for}\quad r+k\quad \text{even},\quad k=1,2,\ldots,n,\quad r=k,k+1,\ldots,2n-k-d.$$

This formula specialized for $d=0$ is not equal to the formula given in [1]. It seems there is a mistake in [1] because the distribution given there is not equal to the coefficient a_n in the power series by $\binom{2n}{n}$.

II.10. Distribution of $L_{n,m}$ – length of positive sojourns

Let us investigate the distribution of the rank order statistic

$$L_{n,m} = \sum_{j\in A}(i_{j+1}-i_j),$$

where

$$A = \{j : H_{n,m}(Z_{i_j}) = H_{n,m}(Z_{i_{j+1}}) = 0 \quad\text{and}\quad H_{n,m}(Z_i) > 0,\quad \text{for}\quad i_j < i < i_{j+1}\},$$

called the length of positive sojourns. Denote

$$U' = \sum_{j\in B}(i_{j+1}-i_j),$$

where

$$B = \{j : S_{i_j} = S_{i_{j+1}} = 0 \quad\text{and}\quad S_i > 0 \quad\text{for}\quad i_j < i < i_{j+1}\},$$

$$\varphi(x,t) = \sum_k \chi_{\{x=k\}}(x)\,.\,t^k,\quad t\in(0,1),$$

$$U = \varphi(U',t),\quad U_n = \varphi(L_{n,m},t).$$

The function $h(p)$ equals

$$h(p) = E(U) = \sum P(U' = k)\, t^k . \tag{7}$$

Dwass [1] proved the formula

$$h(p) = (1 - 2p)\,[q - \tfrac{1}{2}[1 - (1 - 4pqt^2)^{1/2}]]^{-1} . \tag{8}$$

The random variable $L_{n,m}$ can take on the even numbers $0, 2, 4, \ldots, 2n$ only. Since the expression (8) does not depend on k the probabilities for $L_{n,m}$ do not depend on k as well the distribution is

$$P(L_{n,m} = 2k) = \frac{1}{n+1}, \quad \text{for} \quad k = 0, 1, \ldots, n .$$

II.11. **Distribution of $R^+_{n,m} + D^+_{n,m}$**

Let us define random variables $M_1, M_2, \ldots, M_{D^+}$ on the simple random walk by

$$\begin{aligned} M_1 &= k , \quad \text{if} \quad S_k = 1 \quad \text{and} \quad S_i < 1 \quad \text{for} \quad i < k , \\ &= 0 , \quad \text{if} \quad S_i < 1 \quad \text{for all} \quad i \geqq 0 , \\ M_j &= k , \quad \text{if} \quad M_{j-1} = r \quad \text{and} \quad S_{k+r} - S_r = 1 \quad \text{and} \quad S_{k+r} - S_s < 1 \\ &\quad \text{for} \quad s = r+1 , \quad r+2, \ldots, k+r-1 . \end{aligned}$$

(The statistic D^+ has been defined in II.9.) The sum

$$(M_1 + 1) + (M_2 + 1) + \ldots + (M_{D^+} + 1)$$

is equal to $R^+ + D^+$ (see II.9). The generating function of this sum is

$$\sum_{k=0}^{\infty} [[1 - (1 - 4pqt^2)^{1/2}]\,(2q)^{-1}]^k\,(1 - pq^{-1}) ,$$

as shown in Feller [2]. By elementary algebra we can prove that the preceding formula is equal to (8) so that the distribution is

$$P(R^+_{n,m} + D^+_{n,m} = 2k) = \frac{1}{n+1}, \quad \text{for} \quad k = 0, 1, \ldots, n .$$

II.12. **Distribution of $D_{n,m}$ – two side maximum deviation**

The statistic of the two side maximum deviation is defined as

$$D_{n,m} = \max_{-\infty < x < \infty} |H_{n,m}(x)| = \max_{0 \leqq i \leqq n+m} |H_{n,m}(Z_i)| .$$

The distribution of $D_{n,m}$ will be obtained from the distribution of the two-dimensional statistic

$$U'_n = \left(D^+_{n,m}, -D^-_{n,m}\right),$$

where $D^+_{n,m}$ was introduced in II.6 and

$$D^-_{n,m} = \max_{-\infty < x < \infty} \{-H_{n,m}(x)\} = \max_{0 \leqq i \leqq n+m} \{-H_{n,m}(Z_i)\}.$$

We have to define suitable functions U and U_n; put

$$U' = \left(\max_{1 \leqq j \leqq T} S_j, \min_{1 \leqq j \leqq T} S_j\right), \quad \text{for} \quad T > 0$$

$$= (0, 0), \quad \text{for} \quad T = 0$$

$$\varphi(x, y) = \chi_{\{(x,y): x < k, y > -(s+d)\}}(x, y), \quad \text{for} \quad k > 0, \quad s > 0,$$

$$U_n = \varphi(U'_n), \quad U = \varphi(U').$$

The function $h(p)$ equals

$$h(p) = P\left(\max_{1 \leqq j \leqq T} S_j < k, \quad \min_{1 \leqq j \leqq T} S_j > -(s + d)\right), \quad T > 0,$$

and consequently

$$h(p) = P\{(-(s + d) < S_j < k; \quad j = 1, \ldots, T) \cup (T = 0)\} = P\{A\},$$

where A is the event described in the brackets. Then the complementary event $\bar{A} = B \cup C$, where the events B, C are defined as

$$B = \{\textit{the simple random walk reaches the point } k \textit{ without reaching the point } -(s + d) \textit{ before, and then it returns to the point } -d\},$$

$$C = \{\textit{the simple random walk reaches the point } -(s + d) \textit{ without recahing the point } k \textit{ before, and then it returns to the point } -d\}.$$

The probability $P(B)$ is equal to

$$P(B) = \left(\frac{p}{q}\right)^k \frac{1 - (p/q)^{s+d}}{1 - (p/q)^{s+d+k}}, \quad p < q,$$

(see [1], p. 1051). The probability $P(C)$ is equal to the product of $1 - P(B)$ = the probability of reaching $-(s + d)$ without reaching the point k before, and of $(p/q)^s$ = the probability of reaching $-d$ from the point $-(s + d)$. Hence

$$P(C) = \left(\frac{p}{q}\right)^s \left[1 - \left(\frac{p}{q}\right)^k \frac{1 - (p/q)^{s+d}}{1 - (p/q)^{s+d+k}}\right], \quad p < q,$$

and finally

$$P(A) = 1 - P(B) - P(C) = [1 - (p/q)^{k+s+d}]^{-1} \cdot$$

$$\cdot [1 - (p/q)^s - (p/q)^k + (p/q)^{k+s}] = 1 - \sum_{i=1}^{\infty} [(p/q)^{is+(i-1)(k+d)} +$$

$$+ (p/q)^{ik+(i-1)(s+d)} - (p/q)^{i(k+s)+(i-1)d} - (p/q)^{i(k+s+d)}] .$$

Since $h(p) = P(A)$, from the preceding equality and in view of III.1 and III.2 we get

$$\frac{h(p)}{(1-2p)\,q^d} = \sum_{n=0}^{\infty} \binom{2n+d}{n} (pq)^n - \sum_{i=1}^{\infty} \left[\sum_{n=is+(i-1)(k+d)}^{\infty} \binom{2n+d}{n-is-(i-1)(k+d)} \cdot \right.$$

$$\cdot (pq)^n + \sum_{n=ik+(i-1)(s+d)}^{\infty} \binom{2n+d}{n-ik-(i-1)(s+d)} (pq)^n -$$

$$- \sum_{n=ik+is+(i-1)d}^{\infty} \binom{2n+d}{n-i(k+s)-(i-1)d} (pq)^n -$$

$$\left. - \sum_{n=i(k+s+d)}^{\infty} \binom{2n+d}{n-i(k+s+d)} (pq)^n \right] =$$

$$= \sum_{n=0}^{\infty} \left[\binom{2n+d}{n} - \sum_{i=1}^{\infty} \left[\left(\binom{2n+d}{n-is-(i-1)(k+d)} + \binom{2n+d}{n-ik-(i-1(s+d)} - \right. \right. \right.$$

$$\left. \left. \left. - \binom{2n+d}{n-i(k+s)-(i-1)d} - \binom{2n+d}{n-i(k+s+d)} \right) \right] \right] (pq)^n ,$$

$$\text{for} \quad s = 1, 2, \ldots, n , \quad k = 1, 2, \ldots, n .$$

The distribution of the two dimensional statistic U_n' is then given by

$$P(D_{n,m}^{+} < k,\ D_{n,m}^{-} < s + d) =$$

$$= 1 - \sum_{i=1}^{\infty} \left[\left(\binom{2n+d}{n-is-(i-1)(k+d)} + \binom{2n+d}{n-ik-(i-1)(s+d)} - \right. \right.$$

$$\left. \left. - \binom{2n+d}{n-i(k+s)-(i-1)d} - \binom{2n+d}{n-i(k+s+d)} \right) \right] \binom{2n+d}{n}^{-1} ,$$

$$\text{for} \quad s = 1, 2, \ldots, n , \quad k = 1, 2, \ldots, n .$$

Consequently the distribution of the rank order statistic $D_{n,m}$ is clearly

$$P(D_{n,m} < k) = P(D_{n,m}^{+} < k,\ D_{n,m}^{-} < k) , \quad \text{for} \quad k > d ,$$

$$= 0 , \quad \text{for} \quad k \leqq d .$$

If we put $d = 0$ in both formulas we have the same as in [1].

II.13. **Distribution of $D^-_{n,m}$ – lover side maximum deviation**

The lower side maximum deviation was defined in II.12. Its distribution wil be derived by means of the statistic U'_n defined in II.12. Clearly,

$$P(D^+_{n,m} = n,\ D^-_{n,m} < s + d) = \binom{2n + d}{n}^{-1} = P(D^+_{n,m} < k,\ D^-_{n,m} = n + d),$$

$$\text{for} \quad k = 1, 2, \ldots, n + 1, \quad s = 1, 2, \ldots, n + 1,$$

because in both cases there is only one sequence of rank order indicators. Hence

$$P(D^-_{n,m} < s + d) = P(D^+_{n,m} < n + 1,\ D^-_{n,m} < s + d) =$$
$$= P(D^+_{n,m} < n,\ D^-_{n,m} < s + d) + P(D^+_{n,m} = n,\ D^-_{n,m} < s + d) =$$
$$= P(D^+_{n,m} < s,\ D^-_{n,m} < n + d) + P(D^+_{n,m} < s,\ D^-_{n,m} = n + d) =$$
$$= P(D^+_{n,m} < s,\ D^-_{n,m} < n + 1 + d) = P(D^+_{n,m} < s), \quad \text{for} \quad s = 1, \ldots, n.$$

Taking into account the result for $D^+_{n,m}$ from II.6, we see that

$$P(D^-_{n,m} \geqq k + d) = \binom{2n + d}{n - k}\binom{2n + d}{n}^{-1}, \quad \text{for} \quad k = 0, 1, \ldots, n.$$

II.14. **Distribution of $B_{n,m}$**

The statistic $B_{n,m}$ has been defined in Reimann-Vincze [4], p. 294, as

$$B_{n,m} = \max_{-\infty < x < \infty} \left|H_{n,m}(x) + \tfrac{1}{2}(m - n)\right| - \tfrac{1}{2}(m - n) =$$
$$= \max_{0 \leqq i \leqq 2n+d} \left|H_{n,m}(Z_i) + \tfrac{1}{2}(m - n)\right| - \tfrac{1}{2}(n - m),$$

and its distribution given in [4], p. 296. We can find the distribution of $B_{n,m}$ by means of the distribution of the statistic U'_n from II.12. It is easy to see that

$$P(B_{n,m} < k) = P(\max_i \left|H_{n,m}(Z_i) + \tfrac{1}{2}d\right| - \tfrac{1}{2}d < k) =$$
$$= P(\max_i (H_{n,m}(Z_i) + \tfrac{1}{2}d) < k + \tfrac{1}{2}d;\ -\min_i (H_{n,m}(Z_i) + \tfrac{1}{2}d) < k + \tfrac{1}{2}d) =$$
$$= P(D^+_{n,m} < k;\ D^-_{n,m} < k + d).$$

Therefore the distribution of $B_{n,m}$ is

$$P(B_{n,m} < k) = 1 - \sum_{i=1}^{\infty}\left[2\binom{2n + d}{n - ik - (i - 1)(k + d)} - \binom{2n + d}{n - 2ik - (i - 1)d} - \right.$$
$$\left. - \binom{2n + d}{n - i(2k + d)}\right]\binom{2n + d}{n}^{-1}, \quad \text{for} \quad k = 1, 2, \ldots, n.$$

The distribution in [4] is given in the following different form

$$(9) \qquad P(B_{n,m} = k) = \binom{m+n}{n}^{-1} \sum_{\gamma=-\infty}^{\infty} \left[\binom{m+n}{m+\gamma s} - \binom{m+n}{m+k+\gamma s} \right],$$

where $s = 2k + m - n$.

We have not been able to show the equivalence of our formula and formula (9). However, if we put $k = n$ in (9) we obtain

$$P(B_{n,m} = n) = 1 - 2\binom{m+n}{n}^{-1},$$

and this is not true, since evidently $P(B_{n,m} = n) = 2\binom{m+n}{n}^{-1}$; thus it seems there is some error in (9).

III. APPENDIX

III.1. Let $W_1, W_2, \ldots, W_k$ be the random variables defined in part I. Then

$$p^k = P(W_1 = W_2 = \ldots = W_k = 1) = \sum_{n=k}^{\infty} P(W_1 = W_2 = \ldots = W_k =$$

$$= 1 \mid T = 2n + d)\, P(T = 2n + d) = \sum_{n=k}^{\infty} P(W_1 = W_2 = \ldots = W_k =$$

$$= 1 \mid T = 2n + d) \binom{2n+d}{n} p^n q^{n+d} (1 - 2p)$$

by the proof of the Theorem. Since

$$P(W_1 = W_2 = \ldots = W_k = 1 \mid T = 2n + d) = \binom{2n+d-k}{n-k} \binom{2n+d}{n}^{-1}$$

the following formula holds

$$p^k = \sum_{n=k}^{\infty} \binom{2n+d-k}{n-k} (pq)^n q^d (1-2p), \quad \text{for} \quad p \in (0, 1/2),$$

$$k = 0, 1, \ldots, n.$$

III.2. Let us verify the formula

$$\frac{p^k}{q^k(1-2p)\, q^d} = \sum_{n=k}^{\infty} \binom{2n+d}{n-k} (pq)^n.$$

To do it, we use the result from III.1 for p^{2k}, and we get

$$\frac{p^k}{q^k(1-2p)\, q^d} = \sum_{n=2k}^{\infty} \binom{2n+d-2k}{n-2k} (pq)^{n-k} = \sum_{n=k}^{\infty} \binom{2n+d}{n-k} (pq)^n.$$

III.3. We are going to prove the formula

$$p^k = \left[\tfrac{1}{2}\left[1 - (1-4pq)^{1/2}\right]\right]^k = k\sum_{n=k}^{\infty}(2n-k)^{-1}\binom{2n-k}{n-k}(pq)^n$$

which has been proved in [1], too. Since $1 = p^2 + 2pq + q^2$, and $1 - 4pq = p^2 - 2pq + q^2 = q - p^2$, the following formula is valid

$$p = \tfrac{1}{2}(1 - q + p) = \tfrac{1}{2}\left[1 - (1 - 4pq)^{1/2}\right], \quad \text{for} \quad q > p\,.$$

The $1-1$ mapping $pq = t$ of the interval $\langle 0, 1/2)$ onto the interval $\langle 0, 1/4)$ has the inverse

$$p = \tfrac{1}{2}\left[1 - (1-4t)^{1/2}\right].$$

Using the result III.1 for p^{k-1}, $d = 0$, we have

$$\frac{p^{k-1}}{1-2p} = \frac{\left(\frac{1}{2}\right)^{k-1}\left[1 - (1-4t)^{1/2}\right]^{k-1}}{(1-4t)^{1/2}} = \sum_{n=k-1}^{\infty}\binom{2n-k+1}{n-k+1}t^n\,.$$

By integrating both sides with respect to t,

$$\left(\tfrac{1}{2}\right)^k\left[1 - (1-4t)^{1/2}\right]^k = k\sum_{n=k-1}^{\infty}(n+1)^{-1}\binom{2n-k+1}{n-k+1}t^{n+1} =$$

$$= \sum_{n=k}^{\infty}(2n-k)^{-1}\binom{2n-k}{n-k}t^n\,.$$

The integration constant is determined from $t = 0$.

Acknowledgement. The author's thanks are due to RNDr. Zbyněk Šidák, DrSc., for his guidance and advice.

References

[1] *M. Dwass:* Simple random walk and rank order statistics. Ann. Math. Statist. 38 (1967), 1042—1053.

[2] *W. Feller:* An introduction to probability theory and its applications. 2nd edition. J. Wiley, New York 1967.

[3] *S. G. Mohanty, B. R. Handa:* Rank order statistics related to a generalized random walk. Studia Sci. Math. Hung. 5 (1970), 267—276.

[4] *J. Reimann, I. Vincze:* On the comparison of two samples with slightly different sizes. A Magyar Tud. Akad. matem. Kutató Intezetének Közleményei 5 (1960), 293—300.

Souhrn

JEDNODUCHÁ NÁHODNÁ PROCHÁZKA A POŘADOVÉ STATISTIKY

IGOR OČKA

Článek obsahuje zobecnění Dwassovy metody [1] výpočtu rozdělení dvouvýběrových pořadových statistik, která je založena na analogii funkcí na jednoduché náhodné procházce a dvouvýběrových pořadových statistik.

První část práce obsahuje odvození metody. Rozšíření Dwassova postupu na dva výběry o m a n prvcích, m, n libovolné, je založeno na tom, že rozdíl $m - n = d$ je považován za libovolnou, ale pevně danou konstantu, zatímco m není fixováno. První část je zakončena větou, která udává vztah mezi rozdělením funkcí na jednoduché náhodné procházce a odpovídajícími pořadovými statistikami. V druhé části jsou spočtena rozdělení vybraných pořadových statistik.

Author's address: RNDr. Ing. *Igor Očka*, Ústav ekonomiky a organizace stavebnictví, Zbraslavská 5, 152 57 Praha 5.