Fast computation by block permanents of cumulative distribution functions of order statistics from several populations*

D. H. Glueck†§ A. Karimpour-Fard† J. Mandel†
L. Hunter† K. E. Muller‡

February 1, 2008

*Deborah H. Glueck is Assistant Professor, Department of Preventive Medicine and Biometrics, University of Colorado at Denver and Health Sciences Center, Campus Box B119, 4200 East Ninth Avenue, Denver, Colorado 80262 (e-mail: Deborah.Glueck@uchsc.edu). Anis Karimpour-Fard is a graduate student in Bioinformatics, Department of Preventive Medicine and Biometrics, University of Colorado at Denver and Health Sciences Center, Campus Box B119, 4200 East Ninth Avenue, Denver, Colorado 80262 (e-mail: Anis Karimpour-Fard@uchsc.edu). Jan Mandel is Professor, Department of Mathematics, Adjunct Professor, Department of Computer Science, and Director of the Center for Computational Mathematics, University of Colorado at Denver and Health Sciences Center, Campus Box 170, Denver, Colorado 80217-3364 (e-mail:Jan.Mandel@cudenver.edu). Larry Hunter is Associate Professor of Biology, Computer Science, Pharmacology, and Preventive Medicine and Biometrics, and Director of the Center for Computational Pharmacology, University of Colorado at Denver and Health Sciences Center, PO Box 6511, MS 8303, Aurora, CO 80045-0511 (e-mail: Larry.Hunter@uchsc.edu). Keith E. Muller is Professor and Director of the Division of Biostatistics, Department of Epidemiology and Health Policy Research, University of Florida, 1329 SW 16th Street Room 5125, PO Box 100177 Gainesville, FL 32610-0177 (e-mail:Keith.Muller@biostat.ufl.edu)

Glueck was supported by NCI K07CA88811. Mandel was supported by nsf-cms 0325314. Muller was supported by NCI P01 CA47 982-04, NCI R01 CA095749-01A1 and NIAID 9P30 AI 50410. Hunter was supported by NIAAA 1U01 AA13524-02 and NCI 5 P30 CA46934-15.

The authors thank Professor Gary Grunwald for his helpful comments.

†University of Colorado at Denver and Health Sciences Center

‡University of Florida

Abstract

The joint cumulative distribution function for order statistics arising from several different populations is given in terms of the distribution function of the populations. The computational cost of the formula in the case of two populations is still exponential in the worst case, but it is a dramatic improvement compared to the general formula by Bapat and Beg. In the case when only the joint distribution function of a subset of the order statistics of fixed size is needed, the complexity is polynomial, for the case of two populations.

Keywords: block matrix, computational complexity, multiple comparison.

1 INTRODUCTION

The Benjamini and Hochberg (1995) procedure represents one of what has become a rather large class of techniques in which we would like to be able to calculate order statistics arising from several populations. The complexity of the calculations implied by such approaches has remained a barrier to accurate probability statements. We provide tools which greatly extend the range of computable cases.

Order statistics obtained by sampling from two different populations occur, e.g., when p-values arise from null or alternative hypotheses, from men or women, or from two different types of cancer.

The distribution of order statistics for independent, identically distributed random variables is well known, and appears in every basic statistics book; for example, Hogg and Craig (1978, Chapter 4, Section 6). David and Nagaraja (2003) and Balakrishnan and Rao (1998) provide a thorough review of order statistics. For identically distributed random variables, the cumulative distribution function is concise and fast to compute.

For independent, but not identically distributed random variables, a formula for computing the joint cumulative distribution function of the order statistics was given by Bapat and Beg (1989). However, this formula is computationally intractable, because it involves an exponential number of permanents of the size of the number of random variables. In addition, the complexity of the computation of the permanent by the best algorithms grows exponentially (Knuth, 1998, p. 499). Approximate algorithms for computing the permanent (Valiant, 1979; Forbert and Marx, 2003; Jerrum et al., 2004) with lower asymptotic complexity are still not practical.

We show that the computational cost of the formula in the case of two populations is still exponential, but is a dramatic improvement compared to the general formula by Bapat and Beg. In the case when only the joint distribution function of a subset of the order statistics of fixed size is needed, we show that the complexity is polynomial, in the case of two populations.

2 NOTATION AND PRELIMINARIES

For an $m \times m$ matrix $\boldsymbol{A}$, with entries a_{ij}, the permanent is given by Aitken (1939, p. 30)

$$\operatorname{per}[\boldsymbol{A}] = \sum_{\pi} \prod_{i=1}^{m} a_{i,\pi(i)} \ . \tag{1}$$

where π ranges over all permutations of $\{1, 2, \ldots, m\}$. Hence, the permanent is defined much like the determinant, but with all signs positive. The permanent can be expanded by row or columns exactly like the determinant. The computational cost of evaluating the permanent by expansion is $O(m!)$ operations. The computational cost using the best algorithms is exponential Knuth (1998, p. 499).

The following notation will be used in all theorems and proofs in this paper without further explicit reference. X_i, $i = 1, \ldots, m$ are independent real valued random variables with cumulative distribution functions $F_i(x)$. The order statistics $Y_1, Y_2, \ldots, Y_m$ are random variables defined by sorting the values of X_i. In particular, $Y_1 \leq Y_2 \leq \ldots \leq Y_m$. The arguments of the joint cumulative distribution function of order statistics are customarily written omitting redundant arguments; thus let n , $1 \leq n_1 < n_2 < \cdots < n_k \leq m$, denote the indices of the remaining arguments and $y_1 \leq y_2 \leq \cdots \leq y_k$ their values. Finally, define the index vector $\mathbf{i} = (i_0, i_1, \ldots i_{k+1})$ and the summation index set

$$\mathcal{I} = \left\{ \mathbf{i} : \begin{array}{c} 0 = i_0 \leq i_1 \leq \cdots \leq i_k \leq i_{k+1} = m, \\ \text{and } i_j \geq n_j \text{ for all } 1 \leq j \leq k \end{array} \right\}. \tag{2}$$

Writing summation over the set $\mathcal{I}$ in terms of loops is straightforward. Using the set $\mathcal{I}$ instead of the loop in this paper allows an insight into the structure of the method and its complexity, and it does not tie the mathematical formulation to any particular implementation.

The joint cumulative distribution function of the set $\{Y_{n_1}, Y_{n_2}, \ldots, Y_{n_k}\}$, which is a subset of the complete set of order statistics, is defined as

$$F_{Y_{n_1},\ldots Y_{n_k}}(y_1, \ldots, y_k) = \Pr\left\{(Y_{n_1} \leq y_1) \wedge (Y_{n_2} \leq y_2) \wedge \cdots \wedge (Y_{n_k} \leq y_k)\right\}. \tag{3}$$

For two sequences a_m and b_m, let $a_m \sim b_m$ denote $\lim_{m\to\infty} a_m/b_m = 1$. Let const be a generic positive constant independent of m; that is, const can have a different value every time it is used. Now $a_m = O(b_m)$ can be written as $|a_m| \leq \text{const}\, b_m$.

3 JOINT CUMULATIVE DISTRIBUTION FUNCTION OF ORDER STATISTICS

First consider the distribution of the order statistics of a random sample where each sample member is taken from a possibly different population with its own distribution.

Theorem 1 (Bapat and Beg (1989), Theorem 4.2) *The cumulative distribution function of the order statistics satisfies*

$$F_{Y_{n_1},\ldots Y_{n_k}}(y_1,\ldots,y_k)=\sum_{\mathbf{i}\in\mathcal{I}}\frac{P_{i_1,\ldots,i_k}(y_1,\ldots,y_k)}{(i_1-i_0)!\,(i_2-i_1)!\cdots(i_{k+1}-i_k)!},\tag{4}$$

where

$$\begin{aligned}&P_{i_1,\ldots,i_k}(y_1,\ldots,y_k)\\&\quad=\operatorname{per}\left[\left[F_i(y_j)-F_i(y_{j-1})\right]_{(i_j-i_{j-1})\times 1}\right]_{j=1,i=1}^{j=k,i=m}\end{aligned}\tag{5}$$

is the permanent of the block matrix with the block row index j and block column index i. The blocks have (i_j-i_{j-1}) rows, and 1 column each, which is denoted by the subscript $(i_j-i_{j-1})\times 1$. Each block has only one distinct entry, which is $[F_i(y_j)-F_i(y_{j-1})]$. We take $F_i(y_0)=0,\quad F_i(y_{k+1})=1$.

In expanded form, the permanent (5) can be written as

$$
\operatorname{per}\begin{bmatrix}
F_1(y_1) & F_2(y_1) & \cdots & F_m(y_1) \\
\vdots & \vdots & & \vdots \\
F_1(y_1) & F_2(y_1) & \cdots & F_m(y_1) \\
-\,-\,-\,-\,- & -\,-\,-\,- & - & -\,-\,- \\
F_1(y_2)-F_1(y_1) & F_2(y_2)-F_2(y_1) & \cdots & F_m(y_2)-F_m(y_1) \\
\vdots & \vdots & & \vdots \\
F_1(y_2)-F_1(y_1) & F_2(y_2)-F_2(y_1) & \cdots & F_m(y_2)-F_m(y_1) \\
-\,-\,-\,- & -\,-\,-\,- & - & -\,-\,-\,- \\
\vdots & \vdots & & \vdots \\
-\,-\,-\,- & -\,-\,-\,- & - & -\,-\,-\,- \\
F_1(y_k)-F_1(y_{k-1}) & F_2(y_k)-F_2(y_{k-1}) & \cdots & F_m(y_k)-F_m(y_{k-1}) \\
\vdots & \vdots & & \vdots \\
F_1(y_k)-F_1(y_{k-1}) & F_2(y_k)-F_2(y_{k-1}) & & F_m(y_k)-F_m(y_{k-1}) \\
-\,-\,-\,- & -\,-\,-\,- & - & -\,-\,-\,- \\
[1-F_1(y_k)] & [1-F_2(y_k)] & \cdots & [1-F_m(y_k)] \\
\vdots & \vdots & & \vdots \\
[1-F_1(y_k)] & [1-F_2(y_k)] & \cdots & [1-F_m(y_k)]
\end{bmatrix}, \tag{6}
$$

where the j-th group, $j = 1, \ldots, k+1$, contains $i_j - i_{j-1}$ repetitions of the same row.

Proof. The theorem is stated, but not proved in Bapat and Beg (1989). We provide a proof for the sake of completeness, and to prepare the ground for our result.

Define $y_0 = -\infty$, and $y_{k+1} = \infty$. Note that for $i \in \{1, 2, \ldots, m\}$, $F_i(y_0) = 0$, and $F_i(y_{k+1}) = 1$, since the F_i are cumulative distribution functions. Denote $A = F_{Y_{n_1},\ldots Y_{n_k}}(y_1, \ldots, y_k)$. Then we have

$$
A = \Pr\left(\bigcap_{j=1}^{k}\left\{Y_{n_j} \le y_j\right\}\right) = \Pr\left(\bigcap_{j=1}^{k}\{\text{at least } n_j \text{ of } X_i \le y_j\}\right). \tag{7}
$$

Denote by I_j the random variable equal to the number of X_i such that $X_i \le y_j$. Then $I_1 \le I_2 \le \cdots \le I_k$, and the condition that at least n_j of $X_i \le y_j$ is equivalent to $I_j \geq n_j$. Thus,

$$A = \Pr\left(\bigcap_{j=1}^{k} \{I_j \geq n_j\}\right) = \Pr\left(\bigcup_{\mathbf{i}\in\mathcal{I}}^{i_2} \bigcap_{j=1}^{k} \{I_j = i_j\}\right), \tag{8}$$

and, since the events $\bigcap_{j=1}^{k} \{I_j = i_j\}$ for different $\mathbf{i}$ are disjoint,

$$A = \sum_{\mathbf{i}\in\mathcal{I}} \Pr\left(\bigcap_{j=1}^{k} \{I_j = i_j\}\right) \tag{9}$$

$$= \sum_{\mathbf{i}\in\mathcal{I}} \Pr\left(\bigcap_{j=1}^{k+1} \{\text{exactly } i_j - i_{j-1} \text{ of } X_i \in (y_{j-1}, y_j]\}\right). \tag{10}$$

Now fix $\mathbf{i}$ and write an arbitrary permutation of $\{1, 2, \ldots, m\}$ as

$$\pi = (\pi_1, \pi_2, \ldots, \pi_k, \pi_{k+1}), \tag{11}$$

where each subsequence π_j has exactly $i_j - i_{j-1}$ terms. We will use $\{\pi_j\}$ to denote the set of the terms. Then,

$$\exists\pi\forall j \in \{1, 2, \ldots, k+1\} : \text{exactly } i_j - i_{j-1} \text{ of } X_i \in (y_{j-1}, y_j] \tag{12}$$

$$\iff \exists\pi\forall j \in \{1, 2, \ldots, k+1\} : \forall i \in \{\pi_j\} : X_i \in (y_{j-1}, y_j]. \tag{13}$$

Hence,

$$\Pr\left(\bigcap_{j=1}^{k+1} \{\text{exactly } i_j - i_{j-1} \text{ of } X_i \in (y_{j-1}, y_j]\}\right) \tag{14}$$

$$= \frac{\sum_\pi \Pr\left(\bigcap_{j=1}^{k+1} \bigcap_{i\in\{\pi_j\}} \{X_i \in (y_{j-1}, y_j]\}\right)}{(i_1 - i_0)! \cdots (i_{k+1} - i_k)!} \tag{15}$$

$$= \frac{\sum_\pi \prod_{j=1}^{k+1} \prod_{i\in\{\pi_j\}} [F_i(y_j) - F_i(y_{j-1})]}{(i_1 - i_0)! \cdots (i_{k+1} - i_k)!}, \tag{16}$$

because the events in the intersection are independent: there is one event for each X_i, which are independent random variables. Substituting into (9) and comparing with the definition of the permanent (1) concludes the proof. ■

As noted in the introduction, using a general algorithm for permanents is prohibitively expensive. Given simplifying assumptions, however, the problem becomes easier. In the case when the variables X_1, $X_2, \ldots, X_m$ are independent and identically distributed (that is, the classical case of sampling from a single population), Theorem 1 reduces to the following well-known result (David and Nagaraja, 2003, p. 11).

Theorem 2 *Suppose that $F_i = F$ for all i. Then the joint cumulative distribution function of the order statistics satisfies*

$$F_{Y_{n_1},\ldots Y_{n_k}}(y_1, \ldots, y_k) = \sum_{\mathbf{i}\in\mathcal{I}} m! \prod_{j=1}^{k+1} \frac{[F(y_j) - F(y_{j-1})]^{i_j - i_{j-1}}}{(i_j - i_{j-1})!}. \tag{17}$$

Now consider drawing a random sample from two populations, each with a different cumulative distribution function, say $F(x)$, and $G(x)$. Sample the first n random variables from the first population with the distribution function F, and then $m-n$ from the second population with the distribution function G. Then the permanents from Equation 4 (Bapat and Beg (1989)) simplify to the block form with constant blocks,

$$\begin{aligned} &P_{i_1,\ldots,i_k}(y_1, \ldots, y_k) \\ &= \operatorname{per}\begin{bmatrix} [F(y_1) - F(y_0)]_{(i_1-i_0)\times n} & [G(y_1) - G(y_0)]_{(i_1-i_0)\times(m-n)} \\ [F(y_2) - F(y_1)]_{(i_2-i_1)\times n} & [G(y_2) - G(y_1)]_{(i_2-i_1)\times(m-n)} \\ \vdots & \vdots \\ [F(y_{k+1}) - F(y_k)]_{(i_{k+1}-i_k)\times n} & [G(y_{k+1}) - G(y_k)]_{(i_{k+1}-i_k)\times(m-n)} \end{bmatrix}, \end{aligned} \tag{18}$$

where the subscripts indicate the dimensions of blocks created by the repetition of the term in the brackets, and we take

$$F(y_0) = G(y_0) = 0, \quad F(y_{k+1}) = G(y_{k+1}) = 1. \tag{19}$$

In expanded form, the permanent (18) can be written as

$$
\operatorname{per}\left[\begin{array}{cccccc}
F(y_1) & \cdots & F(y_1) & G(y_1) & \cdots & G(y_1) \\
\vdots & & \vdots & \vdots & & \vdots \\
F(y_1) & \cdots & F(y_1) & G(y_1) & \cdots & G(y_1) \\
----- & - & ---- & --- & - & --- \\
F(y_2)-F(y_1) & \cdots & F(y_2)-F(y_1) & G(y_2)-G(y_1) & \cdots & G(y_2)-G(y_1) \\
\vdots & & \vdots & \vdots & & \\
F(y_2)-F(y_1) & \cdots & F(y_2)-F(y_1) & G(y_2)-G(y_1) & \cdots & G(y_2)-G(y_1) \\
---- & - & ---- & ---- & - & --- \\
\vdots & & \vdots & \vdots & & \vdots \\
---- & - & ---- & ---- & - & --- \\
F(y_k)-F(y_{k-1}) & \cdots & F(y_k)-F(y_{k-1}) & G(y_k)-G(y_{k-1}) & \cdots & G(y_k)-G(y_{k-1}) \\
\vdots & & \vdots & \vdots & & \vdots \\
F(y_k)-F(y_{k-1}) & \cdots & F(y_k)-F(y_{k-1}) & G(y_k)-G(y_{k-1}) & \cdots & G(y_k)-G(y_{k-1}) \\
---- & - & ---- & ---- & - & --- \\
1-F\left(y_k\right) & \cdots & 1-F\left(y_k\right) & 1-G\left(y_k\right) & \cdots & 1-G\left(y_k\right) \\
\vdots & & \vdots & \vdots & & \vdots \\
1-F\left(y_k\right) & \cdots & 1-F\left(y_k\right) & 1-G\left(y_k\right) & \cdots & 1-G\left(y_k\right)
\end{array}\right]. \tag{20}
$$

This special form of the permanent allows us to evaluate the joint distribution of the order statistic more efficiently.

Theorem 3 *Suppose that $F_i\left(x\right) = F\left(x\right)$, for all $1 \le i \le n$, and $F_i\left(x\right) = G\left(x\right)$, for all $n+1 \le i \le m$. Then*

$$
\begin{aligned}
&F_{Y_{n_1},\ldots Y_{n_k}}\left(y_1,\ldots,y_k\right) = \\
&\sum_{\mathbf{i}\in\mathcal{I}}\sum_{\boldsymbol{\lambda}}\prod_{j=1}^{k+1}\frac{n!\,(m-n)!}{\lambda_j!\,(i_j-i_{j-1}-\lambda_j)!} \\
&\qquad\cdot\left[F\left(y_j\right)-F\left(y_{j-1}\right)\right]^{\lambda_j}\left[G\left(y_j\right)-G\left(y_{j-1}\right)\right]^{i_j-i_{j-1}-\lambda_j}, \qquad (21)
\end{aligned}
$$

where $\boldsymbol{\lambda} = (\lambda_1,\lambda_2,\ldots,\lambda_{k+1})$ ranges over all integer vectors such that

$$
\lambda_1+\lambda_2+\cdots+\lambda_{k+1} = n, \quad 0 \le \lambda_j \le i_j - i_{j-1}. \tag{22}
$$

Proof. We evaluate the permanents $P_{i_1,\ldots,i_k}\left(y_1,\ldots,y_k\right)$ from (18). Let $S_1 = \{1,2,\ldots,n\}$ and $S_2 = \{n+1,n+2,\ldots,m\}$. Write a permutation

Interval	$(-\infty, y_1]$	$(y_1, y_2]$	$\cdots$	(y_k, ∞)	Total
$\# \in S_1$	λ_1	λ_2	$\cdots$	λ_{k+1}	n
$\# \in S_2$	$i_1 - \lambda_1$	$i_2 - i_1 - \lambda_2$	$\cdots$	$m - i_k - \lambda_{k+1}$	$m - n$
Total	i_1	$i_2 - i_1$	$\cdots$	$m - i_k$	m

Table 1: Total number of order statistics in each interval, and number from population 1 and 2 in each interval.

of $\{1, 2, \ldots, m\}$ as $\pi = (\pi_1, \pi_2, \ldots, \pi_k, \pi_{k+1})$, where each subsequence π_j has exactly $i_j - i_{j-1}$ terms. The subsequence π_j is a list of the subscripts of the random variables that fall in the interval (y_{j-1}, y_j). Then the term in the definition of the permanent (1) associated with π is

$$\prod_{i=1}^{m} a_{i,\pi(i)} = \prod_{j=1}^{k+1} \left[F\left(y_j\right) - F\left(y_{j-1}\right)\right]^{\lambda_j} \left[G\left(y_j\right) - G\left(y_{j-1}\right)\right]^{i_j - i_{j-1} - \lambda_j}, \qquad (23)$$

where λ_j is the number of random variables with subscripts listed in $\{\pi_j\}$ that are in S_1. For illustration, the intervals and the number of order statistics of each type in them are shown in Table 1.

The number of permutations π such that λ_j is the number of the elements from $\{\pi_j\}$ that are in S_1 is found as the product ABC, where

$$A = \frac{n!}{\prod_{j=1}^{k+1} \lambda_j!} \qquad (24)$$

is the number of ways to distribute the n elements of S_1 so that set j has λ_j elements (the multinomial coefficient),

$$B = \frac{(m-n)!}{\prod_{j=1}^{k+1} (i_j - i_{j-1} - \lambda_j)!} \qquad (25)$$

is the number of ways to distribute the $m - n$ elements of S_1 so that set j has $i_j - i_{j-1} - \lambda_j$ elements, and

$$C = \prod_{j=1}^{k+1} (i_j - i_{j-1})! \qquad (26)$$

is the number of permutations that do not change the distribution of the elements S_1 and S_2 into those sets. Thus,

$$P_{i_1,\ldots,i_k}(y_1,\ldots,y_k) = \sum_{\pi}\prod_{i=1}^{m} a_{i,\pi(i)}$$

$$= \sum_{\boldsymbol{\lambda}}\prod_{j=1}^{k+1}\frac{(i_j - i_{j-1})!}{\lambda_j!\,(i_j - i_{j-1} - \lambda_j)!}$$

$$\cdot\left[F(y_j) - F(y_{j-1})\right]^{\lambda_j}\left[G(y_j) - G(y_{j-1})\right]^{i_j - i_{j-1} - \lambda_j}, \tag{27}$$

with the sum over all $\boldsymbol{\lambda}$ that satisfy (22). The result now follows from Theorem 1. ■

The proof of Theorem 3 easily carries over to the general case of order statistics of a sample selected from an arbitrary number of populations. The proof of the next theorem can therefore be omitted.

Theorem 4 *Suppose that $F_i = G_1$ for the first m_1 indices i, $F_i = G_2$ for the next m_2 indices i, etc., and $F_i = G_N$ for the last m_N indices i, with*

$$m_1 + \cdots + m_N = m, \quad m_s > 0 \text{ for all } s. \tag{28}$$

Then

$$F_{Y_{n_1},\ldots Y_{n_k}}(y_1,\ldots,y_k) = \tag{29}$$

$$= \sum_{\mathbf{i}\in\mathcal{I}}\sum_{[\lambda_{js}]}\prod_{j=1}^{k+1}\prod_{s=1}^{N}\frac{m_s!}{\lambda_{js}!}\left[G_s(y_j) - G_s(y_{j-1})\right]^{\lambda_{js}} \tag{30}$$

where the summation is over all integer matrices $[\lambda_{js}]$ size $k+1$ by N such that

$$\lambda_{js} \geq 0 \quad \text{for all } j \text{ and all } s, \tag{31}$$

$$\sum_{j=1}^{k+1}\lambda_{js} = m \quad \text{for all } s, \tag{32}$$

$$\sum_{s=1}^{N}\lambda_{js} = i_j - i_{j-1} \quad \text{for all } j, \tag{33}$$

and we take $G_s(y_0) = 0$, $G_s(y_{k+1}) = 1$.

Theorem 4 covers all of the theorems above. In the particular case when all $m_i = 1$, i.e., every distribution is different because it comes from a different population, it gives exactly the same result as Theorem 1. With two populations, the complexity of Theorem 4 reduces to the complexity of Theorem 3. The complexity of Theorem 3 is less than that of the Theorem 1 from Bapat and Beg (1989), as discussed in the next section.

4 COMPLEXITY

We will now compare the relative complexity of Theorem 1, from Bapat and Beg (1989), and our formula, Theorem 3. We assume that the evaluation of the cumulative distribution function of each of the statistics takes a constant number of operations.

For $1 \le n_1 < n_2 < \cdots < n_k \le m$, denote the number of elements of the index set $\mathcal{I}$ by

$$\nu\left(n_1, n_2, \cdots, n_k; m\right) = |\mathcal{I}| = \sum_{i_k=n_k}^{m} \sum_{i_{k-1}=n_{k-1}}^{i_k} \cdots \sum_{i_1=n_1}^{i_2} 1. \tag{34}$$

Theorem 5 *The number* $\nu\left(n_1, n_2, \cdots, n_k; m\right)$ *of the Bapat-Beg permanents in Theorem 1 is bounded by*

$$\nu\left(n_1, n_2, \ldots, n_k; m\right) \le \nu\left(1, 2, \ldots, k; m\right) \le \nu\left(1, 2, \ldots, m; m\right) = C_m, \tag{35}$$

where

$$\nu\left(1, 2, \ldots, k; m\right) = \binom{m+k}{k}\left(1 - \frac{k}{m+1}\right), \tag{36}$$

and

$$C_m = \frac{1}{m+1}\binom{2m}{m} = \frac{(2m)!}{(m+1)!\,m!}. \tag{37}$$

Proof. The inequalities in (35) are obtained by taking the smallest numbers for $n_1, n_2, \ldots, n_k$ and the largest possible value for k, which both give the largest number of terms. We now prove that

$$\nu\left(1, 2, \ldots, k; m\right) = \binom{m+k}{k} - \binom{m+k}{k-1} \tag{38}$$

by induction over k. For $k = 1$, (38) follows from

$$\nu(1;m) = \sum_{i_1=1}^{m} 1 = m \tag{39}$$

and

$$\binom{m+1}{1} - \binom{m+1}{1-1} = (m+1) - 1 = m. \tag{40}$$

Now assume that (38) holds for some k and we will show that

$$\nu(1,2,\ldots,k+1,m) = \binom{m+k+1}{k+1} - \binom{m+k+1}{k}. \tag{41}$$

From the definition (34) and the induction assumption (38), it follows that

$$\nu(1,2,\ldots,k+1;m) = \sum_{i_{k+1}=k+1}^{m} \nu(1,2,\ldots,k;i_{k+1}) \tag{42}$$

$$= \sum_{i=k+1}^{m} \binom{i+k}{k} - \binom{i+k}{k-1} \tag{43}$$

$$= \sum_{i=k+1}^{m} \left[\binom{i+k+1}{k+1} - \binom{i+k}{k+1}\right] \tag{44}$$

$$- \sum_{i=k+1}^{m} \left[\binom{i+k}{k} - \binom{i+k+1}{k}\right], \tag{45}$$

where we have used the identity

$$\binom{n}{r} - \binom{n-1}{r} = \binom{n-1}{r-1} \tag{46}$$

twice. Both sums telescope, and we get

$$\nu(1,2,\ldots,k+1;m) = \left[\binom{m+k+1}{k+1} - \binom{2k+1}{k+1}\right] \tag{47}$$

$$- \left[\binom{m+k+1}{k} + \binom{2k+1}{k}\right], \tag{48}$$

which, noting that

$$\binom{2k+1}{k+1} = \frac{(2k+1)!}{(k+1)!k!} = \binom{2k+1}{k}, \tag{49}$$

gives (41). Equations (36) and (37) follow from (38) by a direct computation:

$$\binom{m+k}{k} - \binom{m+k}{k-1} = \frac{m+k}{1}\frac{m+k-1}{2}\cdots\frac{m+2}{k-1}\frac{m+1}{k} \tag{50}$$

$$- \frac{m+k}{1}\frac{m+k-1}{2}\cdots\frac{m+2}{k-1} \tag{51}$$

$$= \binom{m+k}{k}\left(1-\frac{k}{m+1}\right), \tag{52}$$

and

$$\binom{m+m}{m} - \binom{m+m}{m-1} = \binom{2m}{m}\left(1-\frac{m}{m+1}\right) = \frac{1}{m+1}\binom{2m}{m}, \tag{53}$$

which concludes the proof. ■

The numbers C_m defined by (37) are known as the Catalan numbers (Stanley, 1999), and the numbers $a_{k,m} = \nu(1, 2, \ldots, k; m)$ are called the Catalan triangle (Shapiro, 1976). From the Stirling approximation $m! \sim \sqrt{2\pi m}\, m^m/e^m$, the growth of Catalan numbers is exponential,

$$C_m \sim \text{const}\, m^{-3/2}4^m > \text{const}\, \alpha^m, \tag{54}$$

for any $1 < \alpha < 4$ (with a different const for each α).

Theorem 6 *The worst case complexity of computing the distribution function of the order statistics from Theorem 1 is*

$$\text{const}\, C_m m K^m \sim \text{const}\, m^{-1/2}4^m P(m), \tag{55}$$

where $P(m)$ is the number of operations for computing permanent of order m.

Proof. The denominator in (4) requires at most $O(m)$ operations, and there are at most C_m terms in the sum by Theorem 5. ■

It is known that the complexity of computing the permanent is bounded by

$$P(m) = O(m^a 2^m)$$

for some a, e.g., from the Ryser's formula (Knuth, 1998). So, the complexity of the computation of the distribution function from Theorem 1 is exponential in m. Therefore, the computation is practical only for small m.

Fortunately, a drastic reduction of complexity is possible in the case when the order statistics come from two populations. In fact, the complexity reduces still farther when we need only a small number k of order statistics.

Theorem 7 *Let $C(k, m, n)$ be the number of operations in Theorem 3 to evaluate the joint distribution function of k order statistics from m random variables from two populations, with $n \leq m$ of the variables from the first population. Then*

$$C(k, m, n) \leq \text{const}\, k \binom{m+k}{k} \binom{n+k}{k} \left(1 - \frac{k}{m+1}\right). \tag{56}$$

In the worst case over all k and n, the complexity is bounded by

$$C(k, m, n) \leq \text{const}\, m \frac{(2m)^2}{(m!)^4} \sim \text{const}\, 16^m, \tag{57}$$

For any fixed k, the complexity is bounded by

$$C(k, m, n) = O\left(m^k n^k\right). \tag{58}$$

i.e., the complexity is polynomial in m.

Proof. The complexity is bounded by const CLM, where $C = \binom{m+k}{k}\left(1 - \frac{k}{m}\right)$ is the number of terms in the sum over $\mathbf{i}$, L is the number of possible index vectors $\boldsymbol{\lambda}$ satisfying (22), and M is the complexity of evaluating the products in one term of the sum, which is $M = O(k)$. To bound L, drop the upper bounds in (22). Thus L is bounded above by the number of all integer vectors $\boldsymbol{\lambda}$ such that

$$\lambda_1 + \lambda_2 + \cdots + \lambda_{k+1} = n, \quad \lambda_j \geq 0 \text{ for all } j, \tag{59}$$

which is the same as the number of ways to distribute n indistinguishable objects to $k+1$ distinguishable bins, which equals to $\binom{n+k}{k}$. This gives (56).

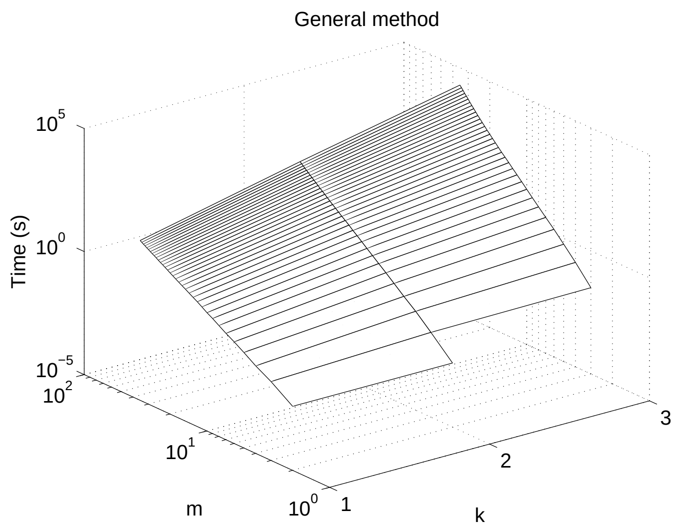

Figure 1: Times for evaluating the joint cumulative distribution function of the first k order statistics of m random variables from two distributions, using the general Bapat-Beg formula (Theorem 1).

The bound (57) follows by taking a pessimistic value of k in each term (56) - twice $k = m$, then $k = 0$, and pessimistic value $n = m$. The second part of (57) follows from the Stirling formula.

The polynomial bound (58) follows from (56) and the inequality

$$\binom{p+k}{k} = \frac{(p+k)\,(p+k-1)\cdots(p+1)}{1\cdot 2\cdots k} \leq \mathrm{const}(k)p^k$$

applied with $p = m$ and $p = n$. ■

Although the complexity of evaluating the cumulative distribution function of order statistics from Theorem 1 is exponential in the general case, we have shown in Theorem 7 that the complexity is bounded by a polynomial of a small degree when there are only two populations, and the number of order statistics considered, k, is fixed and small. The complexity also depends on n, the number of random variables from the first population, S_1. In general, n is fixed by the state of nature.

To confirm and illustrate the result, we have conducted a timing experiment. We calculated the joint distribution function in the case of two popula-

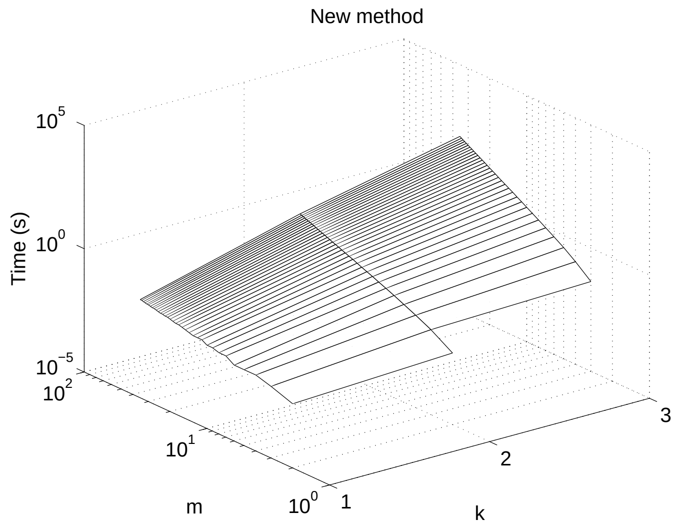

Figure 2: Times for evaluating the joint cumulative distribution function of the first k order statistics of m random variables from two distributions, using the new formula from Theorem 3.

Bapat-Beg formula Theorem 1, Fig. 1	New formula Theorem 3, Fig. 2	Improvement Fig. 3
$10^{-2.9-0.36k} m^{2.0+1.1k}$	$10^{-2.6-0.01k} m^{0.06+1.02k}$	$10^{-0.30-0.34k} m^{1.93+0.09k}$

Table 2: Fit of timing in Mathematica of the evaluation of the joint distribution of the first k statistics of m variables from two populations (n=1 from one population, $m - n$ from the other). For fixed k, regression was used to fit the logarithm of the time with a linear function of $\log m$, and regression was then used again to fit the coefficients by linear functions of k.

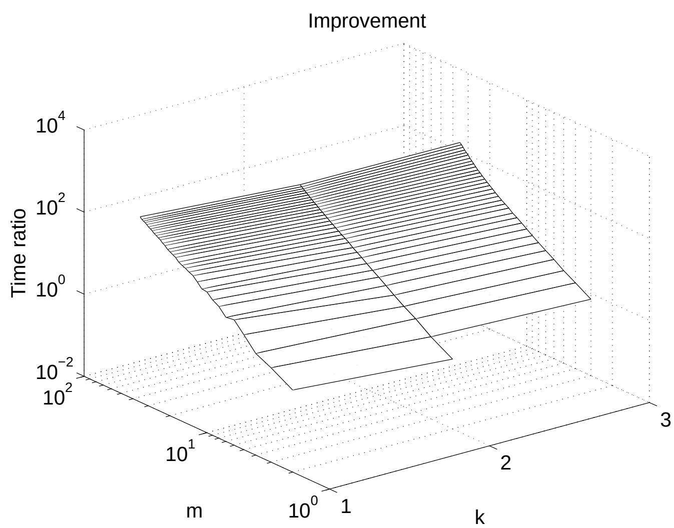

Figure 3: Ratio of times for evaluating the joint cumulative distribution function of the first k order statistics of m random variables from two distributions, using the Bapat-Beg formula (Theorem 1) and the new formula from Theorem 3.

tions. We considered $k = 1$, $k = 2$, and $k = 3$, and fixed $n = 1$. We measured the amount of time it took to compute the joint distribution function using the general Bapat Beg formula with permanents (Fig. 1) and the new special formula (Fig. 2). Both theorems were implemented in Mathematica . The permanents were computed in Mathematica using the code

```
Permanent[A_List] :=With[v = Array[x, Length[A]],
                    Coefficient[Times@@(A.v), Times@@v]
```

from Weisstein (2006). This function computes the permanent of matrix A by Vardi's formula as the coefficient of $x_1 \cdots x_m$ in

$$\prod_{i=1}^{m} (a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{im}x_m) \, ,$$

using symbolic manipulation with automatic caching of partial results by the Mathematica kernel. Amazingly, calculating the permanent from (18) in Mathematica results in times that grow polynomially with m, the number of rows in the permanent. Consequently, for two populations, while the theoretical complexity of Bapat Beg is exponential, the actual time observed while calculating the formulas in Mathematica was polynomial (Fig. 1). Graphing the time versus the log of m produces almost straight lines in a log-log plot. We attribute this speedup to the reuse of partial results by the Mathematica kernel.

Mathematica calculates the Bapat Beg formula more rapidly than predicted. In the timing experiment, the observed times for the new formula (Theorem 3) are much faster than the Bapat Beg formula. The observed improvement was quite dramatic (Fig. 3). The observed improvement is of the order m^2 (Table 2). The observed complexity of the new formula for two populations was of the order m^k, which confirms the result of Theorem 7 for constant $n = 1$.

All calculations were done using a custom New Tech Solutions workstation with 4 AMD Opteron 848 processors running Mathematica 5.2, under the SuSE Linux Enterprise Server 10 operating system.

Mathematica code to calculate the cumulative distribution function for arbitrary collections of order statistics of independent random variables which may have different distributions is available free from the authors. Examples demonstrating the use of the software are also available from the authors.

References

Aitken, A. C. (1939), *Determinants and Matrices*. New York: Oliver and Boyd.

Balakrishnan, N. and C. R. Rao (1998), "Order statistics: An introduction," *Order statistics: Theory & Methods* (Vol. 16:*Handbook of Statist.*) Amsterdam: North-Holland, 3–24.

Bapat, R. B. and M. I. Beg (1989), "Order statistics for non identically distributed variables and permanents," *Sankhyā Ser. A*, **51**, 79–93.

Benjamini, Y. and Y. Hochberg (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. Roy. Statist. Soc. Ser. B*, **57**, 289–300.

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

Forbert, H. and D. Marx (2003), "Calculation of the permanent of a sparse positive matrix," *Computer Physics Communications*, **150**, 267–273.

Hogg, R. V. and A. T. Craig (1978), *Introduction to Mathematical Statistics* (4th ed.), New York: Macmillan Publishing Co., Inc.

Jerrum, M., A. Sinclair, and E. Vigoda (2004), "A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries," *Journal of the ACM*, **51**, 671–697.

Knuth, D. E. (1998), *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms* (3rd ed), New York: Addison-Wesley.

Shapiro, L. W.(1976), "A Catalan triangle," *Discrete Math.*, **14**, 83–90.

Stanley, R. P. (1999), *Enumerative combinatorics. Vol. 2*, volume 62 of *Cambridge Studies in Advanced Mathematics*, Cambridge: Cambridge University Press.

Valiant, L. G. (1979), "The complexity of computing the permanent," *Theoretical Computer Science*, **8**, 189–201.

Weisstein, E. W. (2006), “Permanent.” From MathWorld – A Wolfram Web Resource. http://mathworld.wolfram.com/Permanent.html.