Combinatorics, computer algebra and Wilcoxon-Mann-Whitney test

Memorandum COSOR 96-24, 1996, Eindhoven University of Technology

Combinatorics, computer algebra and Wilcoxon-Mann-Whitney test

A. Di Bucchianico
Department of Mathematics and Computing Science
Eindhoven University of Technology
P. O. Box 513
5600 MB Eindhoven, The Netherlands
sandro@win.tue.nl
URL: http://www.win.tue.nl/win/math/bs/statistics/bucchianico

Abstract

We show the combinatorics behind the Wilcoxon-Mann-Whitney two-sample test. This yields new combinatorial proofs of recurrences for its null distribution given recently by Brus and Chang, as well as new recurrences. It is shown how to convert these recurrences into generating functions. These generating functions are used to obtain closed expressions for the null distribution when one of the sample sizes is fixed and to compute moments. We also show how to perform these calculations with the aid of the computer algebra system Mathematica.

Keywords Wilcoxon-Mann-Whitney two-sample test, partitions, Gaussian binomial coefficients, computer algebra, generating functions, recurrences.

AMS classification 05A15, 05A17, 62-04, 62E15, 62E30, 62G10, 65U05

1 Introduction

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples with continuous distribution functions F and G, respectively. In order to test whether X_1 is stochastically larger than Y_1, Wilcoxon introduced in [14] the statistic

$$W_{m,n} = \text{sum of the ranks of the } X_i\text{'s in the combined sample}$$

Mann and Whitney introduced in [9] the equivalent statistic

$$M_{m,n} = \sum_{i=1}^{m} \#\{j : Y_j < X_i\}$$

The equivalence of these statistics can be seen as follows. Let $X_{(i)}$ denote the ith order statistic of $X_1, \ldots, X_m$. Then for $i = 2, \ldots, m$, the ranks of $X_{(1)}, \ldots, X_{(i-1)}$ are included in $W_{m,n}$, but not in $M_{m,n}$. Hence, $W_{m,n} = M_{m,n} + \sum_{i=1}^{m} i = M_{m,n} + \frac{1}{2}m(m+1)$.
In order to compute critical values and moments of their statistic, Mann and Whitney gave a recurrence relation. A somewhat different recurrence relation was used by Fix and Hodges

in [7]. Recently, Brus ([3]) and Chang ([5]) gave new recurrence relations for these statistics. Unfortunately, most of the proofs in [3] and [5] are calculations that do not give insight in the structure of these new recurrence relations. The aim of this paper is to give a combinatorial explanation of these recurrence relations. By doing so, we also find new recurrences and use them to obtain generating functions. From these generating functions we derive moments and solve the open problems on closed formulas for small sample sizes posed by Chang ([5]). We show how these calculations can be performed with the computer algebra system Mathematica[1]. It transpires that the use of computer algebra opens new horizons for nonparametric statistics. Instead of time-consuming calculations with recurrences, exact distributions can be found very fast from generating functions with the aid of a computer algebra system.

2 Partitions

In this section we link the distribution of the Mann-Whitney statistic to partitions of integers. We use this combinatorial interpretation of the Mann-Whitney statistic in order to explain its properties. In particular, we explain the recurrence relations given by Brus and Chang in [3] and [5], respectively.

Under $H_0 : F = G$, all rank orders in the combined sample are equiprobable. Thus,

$$\mathbf{P}(M_{m,n} = k) = \frac{f(m,n,k)}{\binom{m+n}{n}}, \tag{1}$$

where $f(m,n,k)$ denotes the number ways we can choose a subset of $\{0,1,\ldots,n\}$ with m elements such that the elements of this subset add up to k. In combinatorial terminology, $f(m,n,k)$ is nothing but the number of partitions of k with at most m non-zero blocks of maximal size n (see [1] or [6])[2]. This connection was already noted by Wilcoxon himself, but is hardly used in the statistical literature.
The favourite tool of in combinatorics for studying partitions is the Ferrers diagram (see [1] and [6]). This is a graphical way to represent a partition (see example below). Its statistical counterpart is known in nonparametric statistics as the Gnedenko path (other names are pair chart or PP-plot, see e.g. [11]). The Gnedenko path of the samples $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ is defined as follows. The Gnedenko path is a path from $(0,0)$ to (m,n) with unit steps to the east direction or north direction. If the ith value of the ordered combined sample comes from $X_1,\ldots,X_m$, then our path goes one unit step east, and one unit step north otherwise. Since we assume that F and G are continuous, the probability of a tie (i.e. the event $X_i = Y_j$) equals zero. Hence, the Gnedenko path is well-defined almost surely. In terms of the Gnedenko path, the value of the statistic $M_{m,n}$ is nothing but the area below the Gnedenko path. This interpretation is the basic idea of our approach.

Example Let $m = 4$ and $n = 3$ and let the ranks of the first sample be $1, 3, 4$, and 6. Then we have the following Gnedenko path:

[1] Mathematica is a registered trademark of Wolfram Research, Inc.

[2] This interpretation is equivalent to the interpretation of $M_{m,n}$ in terms of inversions (cf. [2] or [10]). For yet another combinatorial interpretation in terms of Young lattices, see [12].

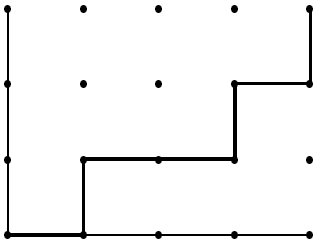

Figure 1: Gnedenko path

The corresponding partition in this case is $0+1+1+2=4$ with the following Ferrers diagram.

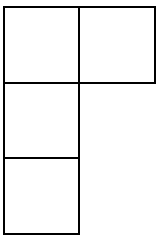

Figure 2: Ferrers diagram

We now use the Gnedenko path to give simple proofs for properties of the statistic $M_{m,n}$. All proofs could also be given in terms of partitions. Since Gnedenko paths have a clear statistical interpretation, we state our proofs in terms of Gnedenko paths and briefly mention the partition interpretation after each proof.

The first result is the well-known symmetry property of $M_{m,n}$. In spite of its simplicity, it turns out to be useful in Section 4.

Proposition 2.1 *The distribution of $M_{m,n}$ is symmetric under $H_0 : F = G$, i.e.* $\mathbf{P}(M_{m,n} = k) = \mathbf{P}(M_{m,n} = mn - k)$ *for* $k = 0, \ldots, mn$.

Proof: Each Gnedenko path has a unique representation as an m-tuple $\langle v_1, \ldots, v_m \rangle$, where v_i is the maximal vertical distance of the path to the point $(i, 0)$. Now associate to each Gnedenko path $w = \langle v_1, \ldots, v_m \rangle$ a new path $w^* = \langle n - v_m, \ldots, n - v_1 \rangle$. If the area below w equals k, then the area below w^* equals $mn - k$. Since the map $w \to w^*$ is a bijection on the set of Gnedenko paths from $(0, 0)$ to (m, n), the result follows. □

The partition analogue of this proof is to consider the map that sends a partition $(\lambda_1, \ldots, \lambda_j)$ to the partition $(n - \lambda_j, \ldots, n - \lambda_1)$.

A similar argument yields another symmetry property that is useful for making statistical tables.

Proposition 2.2 *If $F = G$, then* $\mathbf{P}(M_{m,n} = k) = \mathbf{P}(M_{n,m} = k)$ *for* $k = 0, \ldots, mn$.

Proof: Each Gnedenko path has a unique representation as an n-tuple $\langle h_1, \ldots, h_n \rangle$, where h_i is the maximal horizontal distance of the path to the point $(m, n-i)$. Now associate to each Gnedenko path $w = \langle h_1, \ldots, h_n \rangle$ from $(0,0)$ to (m,n) the unique path from $(0,0)$ to (n,m) such that h_i is the maximal vertical distance of the path to the point $(i,0)$. Since this map is an area preserving bijection from the set of Gnedenko paths from $(0,0)$ to (m,n) to the set of Gnedenko paths from $(0,0)$ to (n,m), the result follows. □

The partition analogue of this proof is to consider the conjugate partition (cf. [1, Theorem 1.5]).
As an immediate corollary we obtain the mean of the statistic $M_{m,n}$.

Corollary 2.3 *Under $H_0 : F = G$, we have $E\, M_{m,n} = \dfrac{mn}{2}$.*

Proof: The result follows directly from Proposition 2.1. □

The following lemma will be used often in the sequel (cf. [3, Lemma 2]).

Lemma 2.4 *Under $H_0 : F = G$, we have $\mathbf{P}(M_{1,n} = k) = \dfrac{1}{n+1}$ for $k = 0, \ldots, n$ and $\mathbf{P}(M_{m,1} = k) = \dfrac{1}{m+1}$ for $k = 0, \ldots, m$.*

Proof: There are $n+1$ paths from $(0,0)$ to $(1,n)$. Only the path that goes through both the points $(0,k)$ and $(1,k)$ has area k. This proves the first property. The second property follows from the first property by Proposition 2.2. □

In terms of partitions, Lemma 2.4 is the trivial assertion that there is only one partition of k with one block.

As we will see in Section 4, there is no closed expression for the distribution of the statistic $M_{m,n}$ under $H_0 : F = G$. Therefore, recursion formulas were used for computations. However, computations with recursions are time-consuming. We will use recursions to obtain closed expressions for generating functions, which lend themselves to fast computations with computer algebra systems. It is more convenient to give recursion formulas for $f(m,n,k)$ (see Formula (1)) than for the distribution $M_{m,n}$ itself. Most books on nonparametric statistics only give the following recursion formula which goes back to [9]:

Theorem 2.5 *With the initial and boundary conditions*

$$\begin{aligned} f(m,n,k) &= 0 && \text{if } k < 0 \text{ or } m < 0 \text{ or } n < 0, \text{ or } k > mn \\ f(m,n,0) &= 1 && \text{if } m \geq 0 \text{ and } n \geq 0, \end{aligned}$$

we have

$$f(m,n,k) = f(m-1,n,k-n) + f(m,n-1,k) \tag{2}$$

Proof: Note that $f(m,n,k)$ equals the number of paths from $(0,0)$ to (m,n) with area k. These paths must pass through either $(m-1,n)$ or through $(m,n-1)$. In the former case the path from $(0,0)$ to $(m-1,n)$ has area $k-n$, in the latter case the path from $(0,0)$ to $(m,n-1)$ has area k. □

The partition analogue of this proof is to look whether the largest block of a partition has size n. Of course, this way of conditioning is crude. Using more refined ways of conditioning, we rediscover the new recurrence formulas of Brus and Chang in [3] and [5].

The first refinement of the conditioning that led to Formula (2) is to condition on the point of the line $x = m - 1$ where the path goes east. This yields Formula 7 of [3].

Theorem 2.6 (Brus) *With the initial and boundary conditions*

$$\begin{array}{ll} f(m,n,k) = 0 & \text{if } k < 0 \text{ or } m < 0 \text{ or } n < 0, \text{ or } k > mn \\ f(m,n,0) = 1 & \text{if } m \geq 0 \text{ and } n \geq 0, \end{array}$$

we have

$$f(m,n,k) = \sum_{i=0}^{n} f(m-1,i,k-i) \tag{3}$$

Proof: If $(m-1,i)$ is the point where a path with area k goes east, then the part of this path up to $(m-1,i)$ must have area $k-i$ and blocks of size not exceeding i. □

The partition analogue of this proof is to look at the size of the largest block. Instead of looking at the largest block, we may also look at the size of the j largest blocks $(1 \leq j \leq n-1)$. Formula (3) is more useful than (2), since it only involves terms with $m-1$.

Theorem 2.7 (Brus) *With the initial and boundary conditions*

$$\begin{array}{ll} f(m,n,k) = 0 & \text{if } k < 0 \text{ or } m < 0 \text{ or } n < 0, \text{ or } k > mn \\ f(m,n,0) = 1 & \text{if } m \geq 0 \text{ and } n \geq 0, \end{array}$$

we have

$$f(m,n,k) = \sum_{i_1=0}^{n} \sum_{i_2=0}^{i_1} \ldots \sum_{i_j=0}^{i_{j-1}} f(m-j, i_j, k-i_1-\ldots-i_j) \tag{4}$$

Proof: Fix an arbitrary Gnedenko path. Let $(m-\ell, i_\ell)$ be the point on the line $x = m-\ell$ where the path goes east. Then the part of the path up to $(m-j, i_j)$ has area $k-i_1-\ldots-i_j$. This yields the result. □

Instead of looking at the end of the Gnedenko path (or the largest block of the associated partition), we may also look at the beginning (or the smallest block). This yields the following recursion formulas:

Theorem 2.8 *With the initial and boundary conditions*

$$\begin{array}{ll} f(m,n,k) = 0 & \text{if } k < 0 \text{ or } m < 0 \text{ or } n < 0, \text{ or } k > mn \\ f(m,n,0) = 1 & \text{if } m \geq 0 \text{ and } n \geq 0, \end{array}$$

we have

$$f(m,n,k) = f(m-1,n,k) + f(m,n-1,k-m) \tag{5}$$

Proof: The number of paths that pass through $(1,0)$ and have area k is equal to $f(m-1,n,k)$. The other paths must go through $(0,1)$. The number of these paths equals the number of paths from $(0,1)$ to (m,n) with area $k-m$ between the path and the line $y=1$. Hence, their number equals $f(m,n-1,k-m)$. □

The partition analogue of this proof is to look whether a partition has precisely m blocks[3].

Refining this way of conditioning, we obtain Formula 6 of [3] and a new recurrence, which is an analogue of (4).

Theorem 2.9 (Brus) *With the initial and boundary conditions*

$$\begin{array}{ll} f(m,n,k)=0 & \text{if } k<0 \text{ or } m<0 \text{ or } n<0, \text{ or } k>mn \\ f(m,n,0)=1 & \text{if } m\geq 0 \text{ and } n\geq 0, \end{array}$$

we have

$$f(m,n,k)=\sum_{i=0}^{n} f(m-1,n-i,k-im) \tag{6}$$

Proof: Fix an arbitrary Gnedenko path. If $(0,i)$ is the point where a path with area k goes east, then the remaining path from $(1,i)$ to (m,n) must have area $k-im$ between the path and the line $y=i$. □

Theorem 2.10 *With the initial and boundary conditions*

$$\begin{array}{ll} f(m,n,k)=0 & \text{if } k<0 \text{ or } m<0 \text{ or } n<0, \text{ or } k>mn \\ f(m,n,0)=1 & \text{if } m\geq 0 \text{ and } n\geq 0, \end{array}$$

we have

$$f(m,n,k)=\sum_{i_j=0}^{n}\sum_{i_{j-1}=0}^{i_j}\ldots\sum_{i_1=0}^{i_2} f(m-j,n-i_j,k-i_1-\ldots-i_{j-1}-i_j(m-j+1)) \tag{7}$$

Proof: Fix an arbitrary Gnedenko path. Let $(\ell-1,i_\ell)$ be the point on the line $x=\ell-1$ where the path goes east. Then the part of the path up to (j,i_j) has area $k-i_1-\ldots-i_j$. The part of the path from (j,i_j) to (m,n) has area $k-(m-j)i_j$ between the line $y=i_j$ and the path. Combining this yields the result. □

It is possible to obtain more recurrences by other ways of conditioning (*e.g.*, on the size of the largest square below the Gnedenko path) or by applying Proposition 2.1. However, such recurrences do not seem to be useful for statistical purposes.

So far we only considered recurrences for the probability mass function of the statistic $M_{m,n}$. Summing these recurrences, we see that the same recurrences (but with different initial and boundary conditions) also hold for the cumulative distribution function of $M_{m,n}$. We summarize these recurrences in the next theorem.

[3] In combinatorics, blocks of size 0 are usually not allowed in the definition of partition.

Theorem 2.11 *Let $A(m,n,k)$ be the number of partitions of all integers not exceeding k with at most m nonzero blocks, each of size at most n. Under the initial and boundary conditions*

$$\begin{aligned}
A(m,n,k) &= 0 && \text{if } k<0 \text{ or } m<0 \text{ or } n<0\\
A(m,n,0) &= 1 && \text{if } m\geq 0 \text{ and } n\geq 0\\
A(0,n,k) &= 1 && \text{if } k\geq 0 \text{ and } n\geq 0\\
A(m,0,k) &= 1 && \text{if } k\geq 0 \text{ and } m\geq 0\\
A(m,n,k) &= \binom{m+n}{n} && \text{if } k>mn \text{ and } m>0 \text{ and } n>0
\end{aligned}$$

we have

$$A(m,n,k) = A(m-1,n,k-n)+A(m,n-1,k) \tag{8}$$

$$A(m,n,k) = A(m-1,n,k)+A(m,n-1,k-m) \tag{9}$$

$$A(m,n,k) = \sum_{i=0}^{n} A(m-1,i,k-i) \tag{10}$$

$$A(m,n,k) = \sum_{i=0}^{n} A(m-1,n-i,k-im) \tag{11}$$

$$A(m,n,k) = \sum_{i_1=0}^{n}\sum_{i_2=0}^{i_1}\ldots\sum_{i_j=0}^{i_{j-1}} A(m-j,i_j,k-i_1-\ldots-i_j) \tag{12}$$

$$A(m,n,k) = \sum_{i_j=0}^{n}\sum_{i_{j-1}=0}^{i_j}\ldots\sum_{i_1=0}^{i_2} A(m-j,n-i_j,k-i_1-\ldots-i_{j-1}-i_j(m-j+1)) \tag{13}$$

$$\tag{14}$$

Proof: Sum the recurrences (2), (3), (4), (5), (6), and (7) with respect to k. □.

For their calculation of significance probabilities of the Mann-Whitney statistic, Fix and Hodges ([7]) used another approach to obtain recurrences for the cumulative distribution function of $M_{m,n}$. Their idea, which goes back to [14], is to express the function $A(m,n,k)$, which counts restricted partitions, in terms of unrestricted partitions. Fix and Hodges only used the first of the following recurrences; the second and third recurrences were given in [3].

Theorem 2.12 *Let $A_0(r,m)$ be the number of partitions of integers not exceeding r with at most m blocks. With the initial and boundary conditions*

$$\begin{aligned}
A_0(r,m) &= 0 && \text{if } r<0 \text{ or } m\leq 0\\
A_0(0,m) &= 1 && \text{if } m>0\\
A_0(r,1) &= r+1 && \text{if } r\geq 0
\end{aligned}$$

we have

$$A_0(r,m) = A_0(r,m-1) + A_0(r-m,m) \tag{15}$$

$$A_0(r,m) = \sum_i A_0(r-mi, m-1) \tag{16}$$

$$A_0(r,m) = \sum_{k_1} \ldots \sum_{k_j} A_0(r-k_1-\ldots-k_j-(m-j), m-j) \tag{17}$$

$$A_0(r,m) = \sum_i A(r-i, m-1, i) \tag{18}$$

$$A_0(r,m) = \sum_{k_1=1}^{r} \sum_{k_2=1}^{k_1} \ldots \sum_{k_j=1}^{k_{j-1}} A(r-k_1-\ldots-k_j, m-j, k_j) \tag{19}$$

Proof: The first three recurrences come from conditioning on the size of the smallest blocks as follows:

- Check whether the partition has exactly m non-zero blocks, i.e. the first block must be non-zero.
- Condition on the size of the smallest block.
- Condition on the size of the j smallest blocks.

The remaining two recurrences come from conditioning on the size of the largest blocks[4]. Note that removing the largest blocks put restrictions on the sizes of the remaining blocks.

- Condition on the size of the largest block.
- Condition on the size of the j largest blocks.

□

3 Moments of the Mann-Whitney statistic

Generating functions are important in both statistics and combinatorics. Their importance is growing due to availability of computer algebra systems. Let us look at the generating function of $f(m,n,k)$ with respect to k. This generating function goes back to Gauss (see [1, p. 51]) and was rediscovered in the context of lattice path counting by Pólya (see [10]).

Theorem 3.1 *The probability generating function of the Mann-Whitney statistic $M_{m,n}$ is given by*

$$\sum_{k=0}^{mn} \mathbf{P}(M_{m,n}=k)\, q^k = \frac{1}{\binom{m+n}{n}} \frac{(q;q)_{n+m}}{(q;q)_n (q;q)_m} = \frac{\begin{bmatrix} n \\ m \end{bmatrix}_q}{\binom{m+n}{n}} \tag{20}$$

where $(a;b)_n = (1-a)(1-ab)\ldots(1-ab^{n-1})$. In particular, $(q;q)_n = (1-q)(1-q^2)\ldots(1-q^n)$.

[4]Note that there is no analogue of Theorem 2.8, because there is only one partition of r such that the size of the largest block equals r.

Proof: See [1, Chapter 3] for a proof based on recurrences or [2, Chapter 11, pp. 203-204] for a proof based on inversions. □

The number $\begin{bmatrix} n \\ m \end{bmatrix}_q$ is called Gaussian binomial coefficient. It is a generalization of the ordinary binomial coefficient, since if q tends to 1, then the limit is the ordinary binomial coefficient (see *e.g.*, [1, Theorem 3.2]). Note that the Gaussian binomial coefficient is a polynomial in q of degree mn, since $k \leq mn$. A simple proof for the asymptotic normality of the statistic $M_{m,n}$ based on Theorem 3.1 is given in [13]; the original proof of this result can be found in [9]. For another use of Gaussian binomial coefficients in statistics, see [8].

We will now show how to use Theorem 3.1 for computing moments of $M_{m,n}$. Since the calculations are very laborious, we will use the computer algebra system Mathematica. For the sake of illustration, we recalculate the mean (cf. Corollary 2.3). The following calculation is a slight improvement on a calculation shown to me by René Swarttouw (personal communication).

Recall that the mean is the derivative of the right-hand side of (20). We first define

$$G_{m,n}(q) := \frac{(1-q^{n+1})\ldots(1-q^{n+m})}{(1-q)\ldots(1-q^m)}.$$

Thus,

$$\log G_{n,m}(q) = \sum_{k=1}^{m} \log\,(1-q^{n+k}) - \sum_{k=1}^{m} \log\,(1-q^k).$$

It now follows that

$$\frac{\frac{d}{dq}G_{n,m}(q)}{G_{n,m}(q)} = \frac{d}{dq}\log G_{n,m}(q) = \sum_{k=1}^{m} \frac{k\,q^{k-1}}{1-q^k} - \sum_{k=1}^{m} \frac{(n+k)\,q^{n+k-1}}{1-q^{n+k}}.$$

This yields the following expression for the derivative of $G_{m,n}$:

$$\frac{d}{dq}G_{n,m}(q) = G_{n,m}(q) \sum_{k=1}^{m} \frac{k\,q^{k-1}(1-q^{n+k}) - (n+k)\,q^{n+k-1}\,(1-q^k)}{(1-q^k)\,(1-q^{n+k})}.$$

Since $G_{n,m}$ is a polynomial in q, we may take the limit $q \to 1$ in order to find $G'_{m,n}(1)$. The factor $G_{n,m}(q)$ tends to $\binom{n+m}{n}$ as $q \to 1$. Hence, it remains to calculate

$$\lim_{q\to 1} \frac{k\,q^{k-1}\,(1-q^{n+k}) - (n+k)\,q^{n+k-1}(1-q^k)}{(1-q^k)\,(1-q^{n+k})}.$$

This limit could be evaluated by applying L' Hôpital's rule twice, as done by René Swarttouw. However, this involves a tedious computation of second derivatives, which can be avoided as follows. First simplify the numerator by pulling out a factor q^{k-1} and expanding the remaining terms. In this way we may rewrite the numerator as $k(1-q^n) - nq^n(1-q^k)$. Then simplify the denominator by using that $(1-q^\ell) = (1-q)(1+q+\ldots+q^{\ell-1})$. The limit then reduces to

$$\frac{1}{k(n+k)} \lim_{q\to 1} \frac{k\,(1+\ldots+q^{n-1}) - n\left(q^n+\ldots+q^{n+k-1}\right)}{1-q},$$

which by L' Hôpital's rule evaluates to $\frac{-1}{k(n+k)}\left(\frac{1}{2}kn(n-1)-n(kn+\frac{1}{2}k(k-1)\right)=n/2$. The case $k=1$, which must be treated separately, can be treated in a similar way.

Similar (but unwieldy) calculations yield higher moments. The appendix contains a Mathematica program for performing these calculations. With this program, central moments are easily computed. These moments are needed for Edgeworth expansions (see [7]). The central moments up to order four were already computed in [9]; the sixth central moment was calculated in [7]. To illustrate our program, we now give the 8th central moment.

8th central moment of $M_{m,n}=\dfrac{mn(1+m+n)}{34560}\,P(m,n)$ with $P(m,n)=$

$$\begin{aligned}
&-96\,n+96\,n^2+240\,n^3-240\,n^4-432\,n^5-144\,n^6 && +\\
&\left(-96+192\,n+224\,n^2-540\,n^3-100\,n^4+780\,n^5+404\,n^6\right) && m\ +\\
&\left(96+224\,n-600\,n^2-200\,n^3+900\,n^4-48\,n^5-420\,n^6\right) && m^2\ +\\
&\left(240-540\,n-200\,n^2+1095\,n^3-395\,n^4-735\,n^5+175\,n^6\right) && m^3\ +\\
&\left(-240-100\,n+900\,n^2-395\,n^3-630\,n^4+525\,n^5\right) && m^4\ +\\
&\left(-432+780\,n-48\,n^2-735\,n^3+525\,n^4\right) && m^5\ +\\
&\left(-144+404\,n-420\,n^2+175\,n^3\right) && m^6
\end{aligned}$$

Higher moments can be calculated in reasonable time, *e.g.* it takes two minutes on a Sun SPARCstation5 to compute the 12th moment.

4 Closed expressions for small sample sizes

In this section, we study closed expressions for $A_0(r,m)$ and $f(m,n,k)$. A closed expression for $A(2,n,k)$ was given in [4]. In [3], a closed formula for $f(2,n,k)$ was derived from (6). Closed formulas for $A_0(r,m)$ $(m\le 4)$ and $f(3,n,k)$ were given in [5]. The purpose of this section is to extend these results, which solves the open problems posed in [5].

The closed formulas for $A_0(r,m)$ in [5] were partly obtained by experimentation. We now give a generating function, from which one can compute closed formulas for $A_0(r,m)$ for any fixed m.

Theorem 4.1 *The generating function of* $A_0(r,m)$ *w.r.t.* r *is given by*

$$\sum_{r=0}^{\infty}A_0(r,m)\,z^r=\begin{cases}\dfrac{1}{(1-z)^2} & \text{for } m=1,\\[2ex] \dfrac{1}{(1-z)^2}\,\dfrac{1}{1-z^2}\cdots\dfrac{1}{1-z^m} & \text{for } m=2,3,\ldots\ .\end{cases}\tag{21}$$

Proof: By Theorem 2.12, we have $A_0(r,1) = r+1$ for $r \geq 0$. Hence, $\sum_{r=0}^{\infty} A_0(r,1)\, z^r = (1-z)^{-2}$.
Now define $\mathcal{A}(m) := \sum_{r=0}^{\infty} A_0(r,m)\, z^r$. Using the first recurrence of Theorem 2.12, we obtain $\mathcal{A}(m) = \mathcal{A}(m-1) + z^m\, \mathcal{A}(m)$ for $m = 1, 2, \ldots$. Thus $\mathcal{A}(m) = \frac{1}{1-z^m}\, \mathcal{A}(m-1)$, from which our result follows directly. □

Thus, finding closed formulas for $A_0(r,m)$ for fixed m is just a matter of expanding (21) into partial fractions. With the help of a computer algebra system this is not too hard. For example, let us compute a closed formula for $A_0(r,5)$. Expanding (21) into partial fractions[5] and then finding the coefficient of z^5 of the terms[6], we obtain after considerable simplification

$$A_0(r,5) = \frac{8861}{10800} + \frac{(-1)^r}{16} + \frac{317\, r}{384} + \frac{(-1)^r\, r}{128} + \frac{5\, r^2}{18} + \frac{89\, r^3}{2160} + \frac{r^4}{360} + \frac{r^5}{14400} +$$

$$\left(\frac{z}{16} \sum_{r=0}^{\infty} (-1)^r\, z^{2r} + \frac{1}{27}(1-z^2) \sum_{r=0}^{\infty} (-1)^r\, z^{3r} + \frac{1}{25}(2+z-z^3-2z^4) \sum_{r=0}^{\infty} (-1)^r\, z^{5r} \right) [[z^r]],$$

where $f(z)\,[[z^r]]$ denotes the coefficient of z^r in $f(z)$.

As can be seen from Theorem 3.1, a general closed expression for the distribution of $M_{m,n}$ does not exist. Lemma 2.4 gives the well-known elementary closed formula for $m = 1$. A closed formula for $m = 2$ was given is [3]. This results also follows directly from the closed formula for $A(2,n,k)$ given earlier in [4] for the Wilcoxon rank sum statistic. We now give a new simple proof of these expressions based on the generating function (21). This proof

Theorem 4.2 *For $m = 2$, the distribution under $H_0 : F = G$ of $M_{2,n}$ is given by*

$$\mathbf{P}\,(M_{2,n} = k) \;=\; \begin{cases} \dfrac{k + \frac{3}{2} + \frac{(-1)^k}{2}}{(n+1)\,(n+2)} & \text{if } 0 \leq k \leq n, \\ 1 - \mathbf{P}\,(M_{2,n} = 2n-k) & \text{if } n+1 \leq k \leq 2n. \end{cases}$$

Proof: By Theorem 3.1, the probability generating function of $M_{2,n}$ equals

$$\frac{2}{\binom{n+2}{2}}\, \frac{(1-q^{n+1})\,(1-q^{n+2})}{(1-q)\,(1-q^2)} \tag{22}$$

The partial fraction decomposition equals

$$\frac{2}{(n+1)\,(n+2)} \left(q^{2\,n} + \frac{1-q^{2\,n}}{4\,(1+q)} + \frac{1-2\,q^n+q^{2\,n}}{2\,(1-q)^2} + \frac{1+4\,q^n-5\,q^{2\,n}}{4\,(1-q)} \right) \tag{23}$$

It follows from the form of the numerators that for $0 \leq k < n$, the coefficient of q^k equals the coefficient of q^k in

$$\frac{2}{(n+1)\,(n+2)} \left(\frac{1}{4\,(1+q)} + \frac{1}{2\,(1-q)^2} + \frac{1}{4\,(1-q)} \right).$$

[5] In Mathematica, this can be done with the command `Apart`.
[6] In Mathematica, this can be done using the command `DiscreteMath‘RSolve‘SeriesTerm`.

A simple calculation yields that this coefficient equals $\dfrac{k+\frac{3}{2}+\frac{(-1)^k}{2}}{(n+1)(n+2)}$. Thus we have proven that $(n+1)(n+2)\,\mathbf{P}(M_{2,n}=k) = k+\frac{3}{2}+\frac{(-1)^k}{2}$ for $0 \le k < n$. Another look at (23) reveals that the coefficient of q^n in (23) equals the coefficient of q^n in

$$\frac{1}{(n+1)(n+2)}\left(\frac{1}{4(1+q)}+\frac{1}{2(1-q)^2}+\frac{1}{4(1-q)}-\frac{q^n}{(1-q)^2}+\frac{q^n}{1-q}\right), \tag{24}$$

which equals $\dfrac{n+\frac{3}{2}+\frac{(-1)^n}{2}-1+1}{(n+1)(n+2)} = \dfrac{n+\frac{3}{2}+\frac{(-1)^n}{2}}{(n+1)(n+2)}$. The proof is now completed by applying Proposition 2.1. □

A closer look at (22) reveals that $M_{2,n}$ can be decomposed as a sum of two independent uniform random variables, viz. $M_{2,n} = U[0,1,2,\ldots,n] + U[0,2,4,\ldots,n/2]$ if n is even and $M_{2,n} = U[0,1,2,\ldots,n+1] + U[0,2,4,\ldots,(n-1)/2]$ if n is odd. In a similar way we may try to prove that $M_{3,n}$ is the sum of three independent uniform random variables. This is easily seen to be true, except for the case $n = 6k+4$. In this case there does not exist such a factorization. For example, if we take $n = 4$, then the probability generating function factors as follows:

$$\frac{1}{35}(1-q+q^2)(1+q+q^2+q^3+q^4)(1+q+q^2+q^3+q^4+q^5+q^6)$$

Since polynomials with rational or real coefficients factor uniquely, this means that there exist no decomposition of $M_{3,4}$ as sum of three independent discrete random variables[7]. We now derive a closed formula for $M_{4,n}$ (a closed formula for $M_{3,n}$ can be found in [5]).

Recall that $f(m,n,k) = \binom{m+n}{n}\mathbf{P}(M_{m,n}=k)$. From formula (20) we deduce that

$$\sum_{k=0}^{\infty} f(4,n,k)\,q^k = \frac{(1-q^n)(1-q^{n+1})(1-q^{n+2})(1-q^{n+3})(1-q^{n+4})}{(1-q)(1-q^2)(1-q^3)(1-q^4)} \tag{25}$$

Using Mathematica, we can easily decompose this expression into partial fractions (for the result, see Appendix B). A similar treatment of this decomposition[8] as in Theorem 4.2 reveals that closed formulas can only be given when we restrict the values of k to intervals of length n. For example, if $0 \le k < n$, then we find that

$$f(4,n,k) = \begin{cases} \dfrac{144+72\,k+15\,k^2+k^3}{144} & \text{if } k = 0 \mod 12 \\ \dfrac{65+63\,k+15\,k^2+k^3}{144} & \text{if } k = 1 \mod 12 \\ \dfrac{76+72\,k+15\,k^2+k^3}{144} & \text{if } k = 2 \mod 12 \\ \text{etc.} \end{cases}$$

[7] For an approximation of $M_{m,n}$ by a sum of independent *continuous* uniform random variables, see [4]. This approximation is shown to perform better than the usual normal approximation.

[8] Unfortunately, Mathematica cannot factor $1+q+q^2$. Maple can, but cannot compute the coefficient of q^k for abstract k. So we have to extract the term with denominator $1+q+q^2$ and rewrite it as $(1-q)/(1-q^3)$.

We conclude this section by giving another generating function, viz. the generating function of $f(m,n,k)$ with respect to k and n.

Theorem 4.3 *The generating function of $f(m,n,k)$ w.r.t. to n and k is given by*

$$\sum_{n=0}^{\infty}\sum_{k=0}^{\infty} f(m,n,k)\,x^n\,y^k = \frac{1}{(1-x)\,(1-x\,y)\,\ldots\,(1-x\,y^m)} = \frac{1}{(x;y)_{m+1}} \tag{26}$$

Proof: Define $\mathcal{F}(m) := \sum_{n=0}^{\infty}\ \sum_{k=0}^{\infty}\ f(m,n,k)\,x^n\,y^k$. Note that the inner sum actually runs from 0 to mn. It follows from Lemma 2.4 that

$$\mathcal{F}(1) = \sum_{n=0}^{\infty}\sum_{k=0}^{n} x^n\,y^k = \sum_{n=0}^{\infty} x^n\,\frac{y^{n+1}-1}{y-1} = \frac{1}{(1-x)\,(1-xy)}$$

By Theorem 2.8, we have $\mathcal{F}(m) = \mathcal{F}(m-1) + xy^m\mathcal{F}(m)$, i.e. $\mathcal{F}(m) = \frac{1}{1-xy^m}\,\mathcal{F}(m-1)$. The result now follows by induction. □

5 Computational efficiency

In this section we investigate the computing efficiency of the various recurrence formulas for the statistic $M_{m,n}$. We implemented the following procedures for computing $\mathbf{P}(M_{m,n}) = k$ In Mathematica:

1. use of the recurrence formula (2)
2. use of the recurrence formula (3)
3. use of the recurrence formula (5)
4. use of the recurrence formula (6)
5. direct computation of the coefficient of q^k in the generating function (20).

m	n	k	Method 1	Method 2	Method 3	Method 4	Method 5
5	5	6	0.2	0.2	0.2	0.4	0.3
5	5	12	0.5	0.5	0.5	0.6	0.3
5	5	18	0.3	0.3	0.4	0.4	0.3
5	10	10	0.8	0.9	0.7	1.8	0.3
5	10	25	4.0	4.0	4.2	5.1	0.3
5	10	40	1.1	1.0	1.8	1.0	0.4
10	5	10	0.9	0.8	0.9	1.4	0.3
10	5	25	4.2	4.0	4.7	5.5	0.3
10	5	40	1.0	1.1	1.6	1.3	0.4
10	10	25	28	27	29	46	0.7
10	10	50	140	140	160	190	0.9
10	10	75	31	31	44	38	1.2

Table 1: Computation time in Mathematica on a SunSPARCstation 5 in seconds

From this table we may conclude that none of the methods 1 through 4 is superior to the others (cf. the remarks in [3, 5]). Method 4 is inferior to all other methods. Method 5 shows that symbolic computation is very fast. However, since Mathematica is very slow in computing recurrences (especially in checking boundary conditions), implementation of the recurrences in some other language may yield a fast implementation too.

Acknowledgements I would like to thank Daniel Loeb for pointing out errors in a preliminary version of this paper, Fred Simons for showing me how to compute limits in Mathematica efficiently and René Swarttouw for showing me how to differentiate (3.1)

A Mathematica procedures

This appendix contains some Mathematica procedures that can be used to calculate the moments of the Mann-Whitney statistic.

The limits are calculated by setting up a Taylor series around $z = 1$ instead of using the Mathematica command Limit. Fred Simons pointed out to me that Mathematica sometimes handles Taylor series much more efficiently than limits.

```
Needs["Algebra`SymbolicSum`"] (* for sums with abstract upper limit       *)

Protect[k,m,n,z,G]            (* protects the values of k,m,n,z, and G    *)

LogG[k_,n_:n,z_:z] := Log[1 - z^(n+k)] - Log[1 - z^k] (* n,z are defaults *)

DerivativeOfLogG[r_] := DerivativeOfLogG[r] = Module[{j,der},Sum[
Simplify[r! Coefficient[Normal[Series[LogG[k],{z,1,r+1}]],z-1,r]],{k,1,m}]]
(* expand LogG in Taylor series of order r in powers of z - 1 *)
(* Normal gets rid of the O(z-1)^r+1 symbol of the Taylor series *)

FactorialMoments[r_] := Module[{j,equations},
equations := Table[
ReplaceAll[Simplify[Together[D[Log[G[z]],{z,j}]]], G[z] ->1]
== DerivativeOfLogG[j],
                  {j,1,r}];
Flatten[
ReplaceAll[Table[D[G[z],{z,i}],{i,1,r}],
Solve[equations,Table[D[G[z],{z,i}],{i,1,r}]]]
       ]
                             ]

Moment[0] := 1                (* zeroth moment equals one by definition *)
Moment[r_] := Module[{j},
Simplify[Sum[StirlingS2[r,j] FactorialMoments[r][[j]],{j,1,r}]]]
(* calculate moments by expressing them in terms of factorial moments  *)
```

```
CentralMoment[r_?OddQ] := 0  (* odd central moments are always zero    *)

CentralMoment[r_?EvenQ] := Module[{x,j,mean=Moment[1],coeffs},
coeffs = CoefficientList[Expand[(x-mean)^r],x];
Factor[Simplify[Dot[coeffs,Table[Moment[j],{j,0,r}]]]]
                              ]
```

B The Mann-Whitney statistic for $m = 4$

With the Mathematica function `Apart` we can perform a partial fraction expa nsion on (25) with the following result (after rewriting the denominator $1 + q + q^2$ as $(1 - q)/(1 - q^3)$) :

$$\sum_{k=0}^{\infty} f(4,n,k)\,q^k = q^{4\,n} + \frac{2 + q^n + q^{2\,n} - q^{3\,n} - 3\,q^{4\,n}}{16\,(1+q)} + \frac{34 + 9\,q^n + 9\,q^{2\,n} + 119\,q^{3\,n} - 171\,q^{4\,n}}{144\,(1-q)} + \frac{1 - q^{4\,n}}{8\,(1+q^2)} + \frac{1 - 2\,q^{2\,n} + q^{4\,n}}{32\,(1+q)^2} + \frac{59 + 4\,q^n + 54\,q^{2\,n} - 356\,q^{3\,n} + 239\,q^{4\,n}}{288\,(q-1)^2} + \frac{1 - q^2 + q^n - q^{3\,n} - q^{4\,n} - q^{1+n} + q^{2+3\,n} + q^{1+4\,n}}{9\,(1-q^3)} + \frac{-3 + 2\,q^n + 12\,q^{2\,n} - 18\,q^{3\,n} + 7\,q^{4\,n}}{24\,(q-1)^3} + \frac{1 - 4\,q^n + 6\,q^{2\,n} - 4\,q^{3\,n} + q^{4\,n}}{24\,(q-1)^4}.$$

Removing all terms of order x^n and higher in the denominators and using the Mathematica function `SeriesTerm`, we obtain that for $0 \le k < n$

$$f(4,n,k) = \begin{cases} \dfrac{130 + 126\,k + 30\,k^2 + 2\,k^3}{288} + \dfrac{1 - q^2}{9\,(1-q^3)}\left[\left[q^k\right]\right] & \text{if } k \text{ is odd} \\ \dfrac{220 + 144\,k + 30\,k^2 + 2\,k^3 + 36\,(-1)^{k/2}}{288} + \dfrac{1 - q^2}{9\,(1-q^3)}\left[\left[q^k\right]\right] & \text{if } k \text{ is even.} \end{cases}$$

References

[1] G.E. Andrews, The theory of partitions, Enclycopedia of Mathematics and its Applications, Addison-Wesley, 1976.

[2] F.N. David and D.E. Barton, Combinatorial chance, Charles Griffin & Co., London, 1962.

[3] T. Brus, A recurrence formula for the distribution of the Wilcoxon rank sum statistic, Stat. Probab. Lett. 7 (1989), 161-165.

[4] N. Buckle, C.H. Kraft and C. van Eeden, An approximation to the Wilcoxon-Mann-Whitney distribution, J. Am. Statist. Assoc. 64 (1969), 591-599.

[5] D.K. Chang, A note on the distribution of the Wilcoxon rank sum statistic, Stat. Probab. Lett. 13 (1992), 343-349.

[6] L. Comtet, Advanced Combinatorics, Reidel, Dordrecht, 1974.

[7] E. Fix and J.L. Hodges, Jr., Significance probabilities of the Wilcoxon test, Ann. Math. Statist. 26 (1955), 301-312.

[8] B.R. Handa and S.G. Mohanty, On q-binomial coefficients and some statistical applications, SIAM J. Math. Anal. 11 (1980), 1027-1035.

[9] H.B. Mann and D.R. Whitney, On a test whether one of two random variables is stochastically larger than the other, Ann. Math. Statist. 18 (1947), 50-60.

[10] G. Pólya, Gaussian binomial coefficients and the enumeration of inversions, in: Proceedings of the Second Chapel Hill Conference on Combinatorial Mathematics and its Applications, University of North Carolina, Chapel Hill, 1970, 381-384.

[11] D. Quade, The pair chart, Statistica Neerlandica 27 (1973), 29-45.

[12] I.R. Savage, Contributions to the theory of rank order statistics: the two-sample case, Ann. Math. Statist. (27) 1956, 590-615.

[13] L. Takács, Some asymptotic formulas for lattice paths, J. Stat. Plan. Inf. 14 (1986), 123-142.

[14] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics 1 (1945), 80-83.