Statistica Sinica **7**(1997), 805-813

THE MANN WHITNEY WILCOXON DISTRIBUTION USING LINKED LISTS

Ying Kuen Cheung and Jerome H. Klotz

University of Wisconsin at Madison

Abstract: We give an improved algorithm for calculating the exact null distribution of the two sample Mann Whitney Wilcoxon rank sum statistic. The algorithm modifies the update method of Smid using a minimal linked list which directs calculation of only those intermediate probabilities required for the final value. Using an efficient shortened representation of the list of required intermediate values, exact probabilities for sample sizes of the order of 100 for each of the two samples can be computed on a personal computer for cases covering the range from many ties with few different values to few ties with many different values.

Key words and phrases: Distribution two sample rank sum test, exact P-value, linked list, ties, update algorithm.

1. Introduction and Notation

Let $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ be two independent samples with cumulative distribution functions F and G respectively. To test the hypothesis $H:\ F=G$ against shift alternatives, the Mann-Whitney (1947) form of the two sample Wilcoxon (1945) rank sum statistic adjusted for possible ties, rejects for extreme values of

$$U=\sum_{i=1}^{m}\sum_{j=1}^{n}(I[Y_j<X_i]+\frac{1}{2}I[Y_j=X_i]), \tag{1}$$

where the indicator $I[A]=1$ if the event A is true and $I[A]=0$ otherwise.

For the case of no tied values, the null distribution

$$P[U\le u|m,n]=A(u|N,m)/\left(\begin{array}{c}N\\ m\end{array}\right), \tag{2}$$

where $N=m+n$ and $A(u|N,m)$ is the number of possible arrangements of m Xs and n Ys that give a value of U that does not exceed u.

A variety of tables exist, such as Milton (1964); and Fix and Hodges (1955) use partition theory to extend the computation of exact probabilities. Despite this distribution coverage, problems occur when there are ties and sample sizes can be increased. It therefore seems worthwhile to provide an algorithm to compute exact values (2) for a wider range of sample sizes and also cover the more complicated distribution when ties are present.

2. Update Formulae

For the case of no ties, the update formula (e.g. Mann and Whitney (1947) or Lehmann (1975), page 51)

$$A(u|N,m) = A(u|N-1,m) + A(u-n|N-1,m-1) \tag{3}$$

has often been used to table the distribution. Formula (3) is derived using the largest value in the pooled sample is either a Y or an X observation respectively. Boundary conditions are

$$A(u|N,m) = 0 \text{ for } u < 0, \quad A(u|N,m) = \begin{pmatrix} N \\ m \end{pmatrix} \text{ for } u \geq mn,$$

$$A(u|N,m) = (u+1) \text{ for } m = 1 \text{ or } n = 1 \text{ and } 0 \leq u \leq mn.$$

By distribution symmetry, we can restrict values so that $u \leq mn/2$ and $m \leq n$ using

$$P[U \leq u|m,n] = 1 - P[U \leq mn-u-1|m,n] \quad \text{for } u > mn/2$$

and

$$P[U \leq u|m,n] = P[U \leq u|n,m] \quad \text{for } m > n.$$

When the combined sample has ties, let K be the number of distinct observation values $z_1 < z_2 < \cdots < z_K$, where $1 < K < N$. Define for $k = 1, \ldots, K$, the counts

$$R_k = \sum_{i=1}^{m} I[X_i = z_k], \quad S_k = \sum_{j=1}^{n} I[Y_j = z_k], \quad T_k = R_k + S_k.$$

Then the average rank version of the Mann-Whitney (1947) statistic (1) can be expressed in terms of the count vectors $\boldsymbol{R} = (R_1, \ldots, R_K)^T$, $\boldsymbol{S} = (S_1, \ldots, S_K)^T$ and the $K \times K$ matrix

$$Q = \begin{pmatrix} 1/2 & 0 & \cdots & 0 \\ 1 & 1/2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1/2 \end{pmatrix}$$

as the product

$$U = \boldsymbol{R}^T Q \boldsymbol{S}.$$

Because U can possibly take on half integer values in the case of ties, we instead multiply by 2 and use the integer valued statistic $W = 2U = 2\boldsymbol{R}^T Q\boldsymbol{S}$. If we partition Q at the last row and column

$$Q = \begin{pmatrix} Q_{11} & \mathbf{0} \\ \mathbf{1}^T & 1/2 \end{pmatrix},$$

where $\mathbf{1}^T = (1, 1, \ldots, 1)$ and $\mathbf{0} = (0, 0, \ldots, 0)^T$, then we can write

$$W = 2\boldsymbol{R}_{11}{}^T Q_{11}\boldsymbol{S}_{11} + R_K(a_K - 2m + R_K), \tag{4}$$

where $\boldsymbol{R}_{11}{}^T = (R_1, \ldots, R_{K-1})$, $\boldsymbol{S}_{11} = (S_1, \ldots, S_{K-1})^T$. The coefficients

$$a_k = 2(t_1 + t_2 + \cdots + t_{k-1}) + t_k \tag{5}$$

for $k = 2, 3, \ldots, K$ with $a_1 = t_1$. Denote conditionally given $\boldsymbol{T} = (T_1, \ldots, T_K)^T = \boldsymbol{t} = (t_1, \ldots, t_K)^T$

$$P[W \le w|\boldsymbol{T} = \boldsymbol{t}] = A(w|K, m, \boldsymbol{t}) / \begin{pmatrix} N \\ m \end{pmatrix}.$$

Then equation (4) gives the update formula of Smid (1956) rewritten for W as

$$A(w|K, m, \boldsymbol{t}) = \sum_{r_K = L_{(K)}}^{L^{(K)}} A(w - r_K(a_K - 2m + r_K)|K-1, m - r_K, \boldsymbol{t}_{11}) \begin{pmatrix} t_K \\ r_K \end{pmatrix}. \tag{6}$$

Here $\boldsymbol{t}_{11} = (t_1, \ldots, t_{K-1})^T$ and the lower and upper limits are respectively

$$L_{(K)} = \max(0, m - t_1 - t_2 - \cdots - t_{K-1}) \quad \text{and} \quad L^{(K)} = \min(m, t_K).$$

Boundary conditions on w are

$$A(w|K, m, \boldsymbol{t}) = 0 \ \text{ for } w < w_{\min}, \quad A(w|K, m, \boldsymbol{t}) = \begin{pmatrix} N \\ m \end{pmatrix} \ \text{ for } w \ge w_{\max},$$

where $w_{\min}$ and $w_{\max}$ are the smallest and largest possible values of W obtained from the extreme arrays. These values satisfy, $w_{\min} \ge 0$ and $w_{\max} \le 2mn$ with equality for the case of no ties. Boundary conditions on m are

$$A(w|K, 1, \boldsymbol{t}) = \begin{cases} 0, & \text{for } w < a_1 - 1, \\ \sum_{j=1}^{k} t_j, & \text{for } a_k - 1 \le w < a_{k+1} - 1, \\ N, & \text{for } a_K - 1 \le w, \end{cases}$$

and

$$A(w|K, N-1, \boldsymbol{t}) = \begin{cases} 0, & \text{for } w < 2N-1-a_1, \\ \sum_{j=1}^{k} t_j, & \text{for } 2N-1-a_k \le w < 2N-1-a_{k+1}, \\ N, & \text{for } 2N-1-a_K \le w, \end{cases}$$

where a_k are defined in equation (5).

For the case $K = 2$ the statistic simplifies to $W = R_2(t_1+t_2) + m(t_1-m)$. We thus have the starting condition

$$A(w|2, m, (t_1, t_2)) = \sum_{r_2} \binom{t_1}{m-r_2}\binom{t_2}{r_2}, \tag{7}$$

where these hypergeometric terms are summed over integer values of r_2 in the range

$$L_{(2)} = \max(0, m-t_1) \le r_2, \quad r_2(t_1+t_2) + m(t_1-m) \le w.$$

3. Computing the C.D.F.

Klotz (1966) discussed computing the exact null distribution of U given $\{T_k = t_k \text{ for } k = 1, \ldots, K\}$ and recommended the enumeration of all $2 \times K$ arrays with fixed margins as follows:

$$\begin{array}{|cccc|c|} \hline R_1 & R_2 & \cdots & R_K & m \\ S_1 & S_2 & \cdots & S_K & n \\ \hline t_1 & t_2 & \cdots & t_K & N \\ \hline \end{array} \;.$$

The exact upper tail P-value is obtained by calculating U for each possible table and accumulating the null probability

$$\begin{aligned} &P[R_1 = r_1, \ldots, R_K = r_K | T_1 = t_1, \ldots, T_K = t_K] \\ &= \binom{t_1}{r_1}\binom{t_2}{r_2}\cdots\binom{t_K}{r_K}\bigg/\binom{N}{m} \end{aligned}$$

for all tables that give a value of U greater than or equal the observed value. More recently, this method has been improved in the statistical package StatXact using a modification by Mehta, Patel, and Tsiatis (1984), pages 821-823 called the network algorithm. Their algorithm generates a set of tables by enumerating paths through a set of nodes determined by m, n, and $\boldsymbol{t}$. At intermediate nodes, upper bounds and lower bounds are calculated for the statistic and tables are skipped for cases in which the upper bound is less or the lower bound is at least as big as the observed statistic.

This approach has the virtue of simplicity and can be extended to the several sample test of Kruskal and Wallis (1952) (see Klotz and Teng (1977)). However, the algorithm still computes U values for many cases which do not exceed the bounds and which do not contribute to the final result. In contrast, the update algorithm can avoid computing extra cases using appropriate logic.

Current computer languages such as Pascal and C or C++ facilitate manipulating linked lists. We propose using the update algorithm of Smid (1956) along with a minimal linked list to specify only those intermediate calculations needed for the final probability value.

Consider the example with $(K, m, w) = (5, 10, 80)$ and $\boldsymbol{t} = (4, 5, 3, 4, 6)^T$. Using equation (6), the boundary conditions and equation (7), Figure 1 gives the minimal list of necessary intermediate values (K^*, m^*, w^*) in the order of their generation.

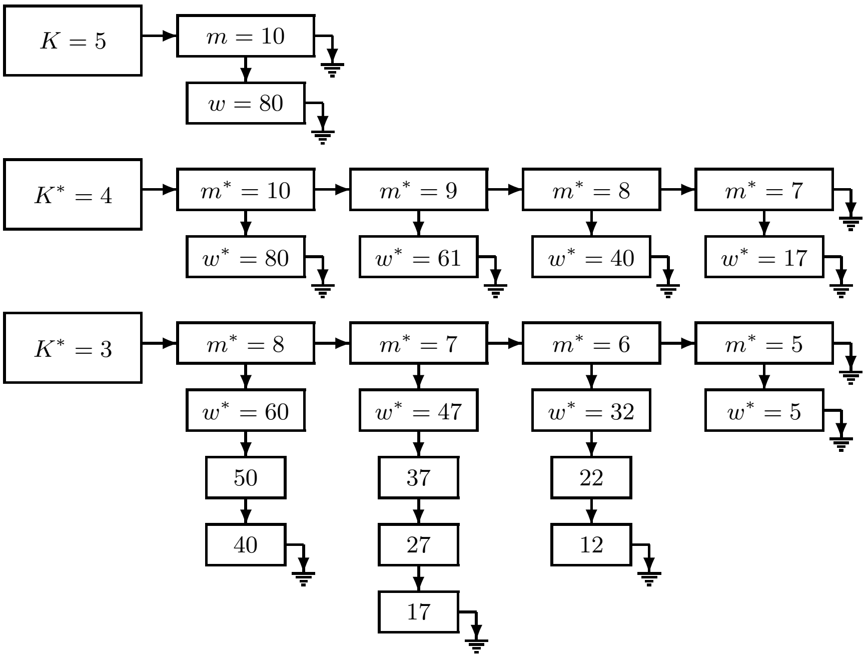

Figure 1. Linked list for $(K, m, w) = (5, 10, 80)$ and $\boldsymbol{t} = (4, 5, 3, 4, 6)$.

Table 1 gives intermediate $A(w^*|K^*, m^*, \boldsymbol{t})$ values corresponding to the list in Figure 1 (in the order of computation).

For example, using equation (6), the boundary conditions and equation (7), we have

$$A(47|3, 7, (4, 5, 3)) = \sum_{r_3=0}^{3} A(47 - r_3(7 + r_3)|2, 7 - r_3, (4, 5))\binom{3}{r_3}$$

$$= \binom{9}{7}\binom{3}{0} + \binom{9}{6}\binom{3}{1} + 105\binom{3}{2} + 21\binom{3}{3} = 624.$$

Table 1. Intermediate values for computing $A(80|5, 10, (4, 5, 3, 4, 6))$.

K^*	m^*	w^*	A	K^*	m^*	w^*	A
3	8	60	494	4	10	80	6864
		50	474		9	61	4865
		40	402		8	40	1266
	7	47	624		7	17	60
		37	412	5	10	80	56244
		27	201				
		17	60				
	6	32	458				
		22	68				
		12	10				
	5	5	5				

For the third term in the above sum we use equation (7)

$$A(29|2, 5, (4, 5)) = \sum_{r_2=1}^{3} \binom{4}{5 - r_2}\binom{5}{r_2} = 105.$$

Finally, using equation (6), previously computed values, and the boundary condition on the lower range of w^* values,

$$A(80|5, 10, (4, 5, 3, 4, 6)) = \sum_{r_5=0}^{6} A(80 - r_5(18 + r_5)|4, 10 - r_5, (4, 5, 3, 4)) \binom{6}{r_5}$$
$$= 6864 \times 1 + 4865 \times 6 + 1266 \times 15 + 60 \times 20 + 0 + 0 + 0 = 56244.$$

This gives

$$P[U \leq 40|m = 10,\ \boldsymbol{t} = (4, 5, 3, 4, 6)] = 56244/\binom{22}{10} \doteq 0.08697804\ .$$

4. Simplifying the Linked List

For large samples with few ties, the minimal list of intermediate integer (N^*, m^*, w^*) values that determines the calculation of $A(w|N, m, \boldsymbol{t})$ can be quite large. To save on required storage, we consider a shortened list representation that takes advantage of regularities.

To illustrate the regularities, consider the stage $K^* = 3$ and $m^* = 7$ in Figure 1. In this part of the linked list the intermediate values w^* range from 47 down to 17 in decrements of $\Delta = 10$. Thus the set of values $\{47, 37, 27, 17\}$ could be replaced by $[47, 17]$ and the decrement $\Delta = 10$.

To see why this regularity occurs, use equation (4) repeatedly to write the value

$$w = \sum_{i=1}^{K} r_i(a_i - 2m_i + r_i),$$

where $m_i = \sum_{j=1}^{i} r_j$. For intermediate (K^*, m^*) an intermediate value w^* is of the form

$$w^* = w - \sum_{i=K^*+1}^{K} r_i(a_i - 2m_i + r_i). \tag{8}$$

For fixed values of $(r_{K^*+3}, r_{K^*+4}, \ldots, r_K)$ and $m_{K^*+2} - m^*$, consider the values w^* given by (8) for all possible (r_{K^*+1}, r_{K^*+2}) such that $r_{K^*+1} + r_{K^*+2} = m_{K^*+2} - m^*$. Then after some algebra we have

$$w^* = C + r_{K^*+1}(t_{K^*+1} + t_{K^*+2}),$$

where the integer

$$C = w - \sum_{i=K^*+3}^{K} r_i(a_i - 2m_i + r_i) - (r_{K^*+1} + r_{K^*+2})(a_{K^*+2} - 2m_{K^*+2}) - (r_{K^*+1} + r_{K^*+2})^2$$

is a constant. Thus consecutive values differ by $\Delta = t_{K^*+1} + t_{K^*+2}$ and we can save considerable storage if we represent such sets by their upper and lower endpoints as in the example above.

To illustrate the savings, consider a case with $U = 13.5$, $m = 10$, $n = 9$ and tie vector $\boldsymbol{t} = (1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 3, 1, 1,)$ (from Daniel (1990), page 131, table 3.46). Table 2 describes the intermediate values of (K^*, m^*, w^*) and uses the notation for the set of integer values from b down to a in decrements of Δ given by $[b, a] = \{b, b - \Delta, b - 2\Delta, \ldots, a\}$. This is a case with 124 w^* values in the minimal list and $P[U \le 13.5|\boldsymbol{t}] = 370/92378 \doteq 0.00400528$.

For large samples this list representation saves considerable memory. Memory is also conserved by writing over earlier values of $A(w^*|K^*, m^*, \boldsymbol{t})$ keeping double precision values only for the most recently updated values with even and odd K^*.

To illustrate the capability of the algorithm, the data of Mehta, Patel and Tsiatis (1984), page 823 with $U = 5462$, $m = 107$, $n = 112$ and $\boldsymbol{t} = (35, 88, 43, 40, 13)$ was run using a C program on a personal computer and quickly obtained a P-value of $\hat{\alpha} \doteq 0.11927038$.

Table 2. Intermediate values for $U = 13.5$, $m = 10$, $n = 9$, with ties $\boldsymbol{t} = (1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 3, 1, 1)$.

K^*	Δ	m^*	w^*	K^*	Δ	m^*	w^*
15		10	27	9	2	9	[23,19], 15, [11,9]
14		10	27			8	[15,1]
		9	9			7	3
13	2	10	27	8	3	8	[15,3], 13, [11,5], 7, 1
		9	[11,9]			7	15, [13,1], [11,8], [6,0], 2
12	4	10	27			6	3, 1
		9	[15,11], 9	7	3	7	[13,1], [11,2], [9,0]
		8	1			6	11, [9,0], [7,1], 2
11	4	10	27	6	2	6	[11,1], [8,0]
		9	[19,11], 9			5	[7,3], 2
		8	[5,1]	5	2	5	[9,1], [6,2]
10	2	10	27			4	[3,1]
		9	[21,19], 15, [11,9]	4	3	3	[5,2], [3,0],1
		8	11, [7,1]	3	3	2	[3,0], 1

A zip file of an executable program for a Windows IBM-PC compatible system can be downloaded from the World Wide Web homepage

http://www.stat.wisc.edu/~klotz/klotz.html

by clicking on **Wilcox.zip** under the **Software:** heading.

Acknowledgement

Thanks go to the Wisconsin/Hilldale Undergraduate/Faculty Research Award for support.

References

Daniel, Wayne W. (1990). *Applied Nonparametric Statistics* (*2nd Edition*). PWS-Kent Publishing, Boston.

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

Klotz, J. H. (1966). The Wilcoxon, ties, and the computer. *J. Amer. Statist. Assoc.* **61**, 772-787.

Klotz, J. and Teng, J. (1977). One-way layout for counts and the exact enumeration of the Kruskal-Wallis H distribution with ties. *J. Amer. Statist. Assoc.* **72**, 165-169.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one criterion variance analysis. *J. Amer. Statist. Assoc.* **47**, 583-621.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based On Ranks.* Holden-Day, San Francisco.

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

Mehta, C. R., Patel, N. R. and Tsiatis, A. A. (1984). Exact significance testing to establish treatment equivalence with ordered categorical data. *Biometrics* **40**, 819-825.

Milton, Roy C. (1964). An extended table of critical values for the Mann-Whitney (Wilcoxon) two sample statistic. *J. Amer. Statist. Assoc.* **59**, 925-934.

Smid, L. J. (1956). On the distribution of the test statistics of Kendall and Wilcoxon's two sample test when ties are present. *Statist. Neerlandica* **10**, 205-214.

Wilcoxon, Frank (1945). Individual comparisons by ranking methods. *Biometrics* **1**, 80-83.

Department of Statistics, University of Wisconsin-Madison, 1210 W. Dayton St., Madison, WI 53706-1685, U.S.A.

E-mail: ken@stat.wisc.edu

E-mail: klotz@stat.wisc.edu

(Received August 1995; accepted February 1996)