NOTES

NOTE ON WILCOXON'S TWO-SAMPLE TEST WHEN TIES ARE PRESENT

By J. Hemelrijk

Mathematical Centre, Amsterdam

Wilcoxon's parameterfree two-sample test (cf. Wilcoxon [1]; H. B. Mann and D. R. Whitny [2]) depends on a statistic U with the following definition: If $x_1, \cdots, x_n$ and $y_1, \cdots, y_m$ are the two samples, U is the number of pairs (i, j) with $x_i > y_j$. The probability distribution of U, under the hypothesis that the samples have been drawn independently from the same *continuous* population, has been derived by Mann and Whitney. The influence of ties on this probability distribution has not been investigated as yet.

It is noteworthy that Wilcoxon's U is closely connected with the quantity S, which Kendall (cf. e.g. Kendall [3]) introduced in the theory of rank correlation. When r pairs of numbers (u_k, v_k) are given, S is computed by scoring:

$$\begin{aligned} -1, &\text{ if } (u_h - u_k)(v_h - v_k) < 0, \\ 0, &\text{ if } (u_h - u_k)(v_h - v_k) = 0, \\ +1, &\text{ if } (u_h - u_k)(v_h - v_k) > 0, \end{aligned}$$

and adding the scores for all pairs (h, k) with $h < k$. If, in this definition, we take $r = n + m$ and substitute the values $x_1, \cdots, x_n, y_1, \cdots, y_m$ in this order for $u_1, \cdots, u_n, u_{n+1}, \cdots, u_r$, and 0 or 1 respectively for v_k if $u_k = x_i$ for some i or $u_k = y_j$ for some j respectively, then the following relation holds:

$$2U + S = nm. \tag{1}$$

The simplest way to see this is by considering the total score of $2U + S$ for every pair (h, k). This score is equal to $+1$ if $v_h = 0$ and $v_k = 1$, and 0 otherwise. The sum of the scores is therefore nm.

Relation (1) holds if no ties are present among the two samples $x_1, \cdots, x_n$ and $y_1, \cdots, y_m$. It is natural to define U in general by extending (1) to the case when there are ties. Since for a pair (x_i, y_j) with $x_i = y_j$ the score of S is equal to zero, the score for U must be taken as $\frac{1}{2}$ for such a pair.

Now Kendall has derived the mean and the standard deviation of S under the hypothesis that for a given order of the quantities $v_1, \cdots, v_r$ all the $r!$ possible permutations of $u_1, \cdots, u_r$ are equally probable. This condition is fulfilled in our case if the samples $x_1, \cdots, x_n$ and $y_1, \cdots, y_m$ have been drawn at random from the same population (which need not be continuous anymore). Therefore, the mean and standard deviation of U under the null hypothesis may be derived from Kendall's formulas.