Covariances of Two Sample Rank Sum Statistics

Peter V. Tryon

Institute for Basic Standards, National Bureau of Standards, Boulder, Colorado 80302

(November 26, 1971)

This note presents an elementary derivation of the covariances of the \( c(c - 1)/2 \) two-sample rank sum statistics computed among all pairs of samples from \( c \) populations.

Key words: c Sample problem; covariances, Mann-Whitney-Wilcoxon statistics; rank sum statistics; statistics.

Mann-Whitney or Wilcoxon rank sum statistics, computed for some or all of the \( c(c - 1)/2 \) pairs of samples from \( c \) populations, have been used in testing the null hypothesis of homogeneity of distribution against a variety of alternatives [1, 3, 4, 5]. This note presents an elementary derivation of the covariances of such statistics under the null hypothesis.

The usual approach to such an analysis is the rank sum viewpoint of the Wilcoxon form of the statistic. Using this approach, Steel [3] presents a lengthy derivation of the covariances. In this note it is shown that thinking in terms of the Mann-Whitney form of the statistic leads to an elementary derivation. For comparison and completeness the rank sum derivation of Kruskal and Wallis [2] is repeated in obtaining the means and variances.

Let \( x_i^r, r = 1, 2, \ldots, n_i, i = 1, 2, \ldots, c, \) be the \( r \)th item in the sample of size \( n_i \) from the \( i \)th of \( c \) populations. Let \( M_{ij} \) be the Mann-Whitney statistic between the \( i \)th and \( j \)th samples defined by

\[
M_{ij} = \sum_{r=1}^{n_i} \sum_{s=1}^{n_j} z_{ij}^{rs} \tag{1}
\]

where

\[
 z_{ij}^{rs} = \begin{cases} 
 1, & x_j^r > x_i^s \\
 0, & x_j^r \leq x_i^s
\end{cases}
\]

Thus \( M_{ij} \) is the number of times items in the \( j \)th sample exceed items in the \( i \)th sample. Let \( W_{ij} \) be the Wilcoxon rank sum statistic defined by

\[
W_{ij} = \sum_{s=1}^{n_j} R_{ij}(x_j^s), \tag{2}
\]

where \( R_{ij}(x_j^s) \) is the rank of \( x_j^s \) in the combined \( i \)th and \( j \)th samples. Then \( M_{ij} \) and \( W_{ij} \) are related by

\[
M_{ij} = W_{ij} - n_f(n_j + 1)/2. \tag{3}
\]
Obviously \( E(M_{ij}) = E(W_{ij}) - n_j(n_j + 1)/2 \), the variances of \( M_{ij} \) and \( W_{ij} \) are equal and the covariances among the \( M_{ij} \) are equal to the covariances among the \( W_{ij} \).

**THEOREM:** Under the null hypothesis, the Mann-Whitney form of the statistic has the following moments:

\[
\begin{align*}
E(M_{ij}) &= n_j n_j/2 \\
V(M_{ij}) &= n_i n_j(n_i + n_j + 1)/12 \\
C(M_{ij}, M_{ik}) &= C(M_{ji}, M_{kj}) = n_i n_j n_k/12 \\
C(M_{ij}, M_{jk}) &= - n_i n_j n_k/12 \\
C(M_{ij}, M_{ik}) &= 0 \quad \text{ijk all different} \\
C(M_{ij}, M_{kl}) &= - n_i n_j n_k/12 \quad \text{ijkl all different.}
\end{align*}
\]

**PROOF:** Consider the population consisting of one of each of the integers 1, 2, \ldots, \( N_{ij} = n_i + n_j \). The population mean and variance are \((N_{ij} + 1)/2\) and \((N_{ij}^2 - 1)/12\), respectively. Now let \( R_{ij} \) be the mean of a random sample of size \( n_j \) drawn without replacement from the population. Then

\[
E(R_{ij}) = (N_{ij} + 1)/2,
\]

and

\[
V(R_{ij}) = \frac{(N_{ij}^2 - 1)}{12n_j} \left( \frac{N_{ij} - n_j}{N_{ij} - 1} \right) = \frac{(N_{ij} + 1)(N_{ij} - n_j)}{12n_j}
\]

where the quantity in brackets is sometimes called the finite sample correction factor. Under the null hypothesis, the distribution of \( W_{ij} \) is identical to that of \( n_j R_{ij} \). The mean and variance of \( M_{ij} \) follow immediately.

To obtain the covariances, let \( M_{(ij+k)} \) be the Mann-Whitney statistic for the combined \( j \)th and \( k \)th samples compared to the \( i \)th sample. Recalling that \( M_{ab} \) is the number of times items in the \( b \)th sample exceed items in the \( a \)th sample, it is easily seen that \( M_{(ij+k)} = M_{ij} + M_{ik} \). Now, since \( n_{j+k} = n_j + n_k \),

\[
V(M_{(ij+k)}) = V(M_{ij} + M_{ik}) = n_i(n_j + n_k)(n_i + n_j + n_k + 1)/12.
\]

But

\[
V(M_{ij} + M_{ik}) = V(M_{ij}) + V(M_{ik}) + 2C(M_{ij}, M_{ik}).
\]

It follows that

\[
C(M_{ij}, M_{ik}) = n_i n_j n_k/12.
\]

The remaining covariances follow from the identity \( M_{ij} = n_i n_j - M_{ji} \).

**References**


(Paper 76B1&2 — 36)