JOURNAL OF RESEARCH of the National Bureau of Standards — B. Mathematical Sciences
Volume 76B, Nos. 1 and 2, January-June 1972

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].[1] 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 rth item in the sample of size n_i from the ith of c populations. Let M_{ij} be the Mann-Whitney statistic between the ith and jth samples defined by

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

where

$$z_{ij}^{rs}=\left\{\begin{matrix} 1, x_j^s > x_i^r \\ 0, x_j^s \leqslant x_i^r \end{matrix}\right\}$$

Thus M_{ij} is the number of times items in the jth sample exceed items in the ith 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 ith and jth samples. Then M_{ij} and W_{ij} are related by

$$M_{ij}=W_{ij}-n_j(n_j+1)/2. \tag{3}$$

AMS Subject Classification: 6231.

[1] Figures in brackets indicate the literature references at the end of this paper.

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:*

$$E(M_{ij})=n_in_j/2$$

$$V(M_{ij})=n_in_j(n_i+n_j+1)/12$$

$$\left.\begin{array}{c} C(M_{ij}, M_{ik})=C(M_{ji}, M_{ki})=n_in_jn_k/12 \\ C(M_{ij}, M_{ki})=-\ n_in_jn_k/12 \end{array}\right\} \text{ijk } \textit{all different}$$

$$C(M_{ij}, M_{kl})=0 \qquad \text{ijkl } \textit{all different.}$$

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

$$E(R_{ij})=(N_{ij}+1)/2, \tag{4}$$

and

$$V(R_{ij})=\frac{(N_{ij}^2-1)}{12n_j}\left[\frac{N_{ij}-n_j}{N_{ij}-1}\right]=\frac{(N_{ij}+1)(N_{ij}-n_j)}{12n_j} \tag{5}$$

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_jR_{ij}. The mean and variance of M_{ij} follow immediately.

To obtain the covariances, let $M_{i(j+k)}$ 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 $M_{i(j+k)}=M_{ij}+M_{ik}$. Now, since $n_{j+k}=n_j+n_k$,

$$V(M_{i(j+k)})=V(M_{ij}+M_{ik})=n_i(n_j+n_k)(n_i+n_j+n_k+1)/12. \tag{6}$$

But

$$V(M_{ij}+M_{ik})=V(M_{ij})+V(M_{ik})+2C(M_{ij}, M_{ik}). \tag{7}$$

It follows that

$$C(M_{ij}, M_{ik})=n_in_jn_k/12. \tag{8}$$

The remaining covariances follow from the identity $M_{ij}=n_in_j-M_{ji}$.

References

[1] Jonckheere, A. R., A distribution free *K*-sample test against ordered alternatives, Biometrika **41**, 133-145 (1954).

[2] Kruskal, W. H., and Wallis, W. A., Use of ranks in one-criterion variance analysis, Journal of the American Statistical Association **47**, 583-621 (1952).

[3] Steel, R. D. G., A rank sum test for comparing all pairs of treatments, Technometrics **2**, 197-207 (1960).

[4] Terpstra, T. J., The asymptotic normality and consistency of Kendall's test against trend when ties are present in one rankir Proceedings Koninkljke Nederlandse Akademie v Wetenschoppen **55**, 327-333 (1952).

[5] Tryon, P. V., and Hettmansperger, T. P., A class of non-parame ric tests for homogeneity against ordered alternatives, u published paper submitted to a technical journal, (1970).

(Paper 76B1&2—36