SP 14 8260 TD

Reprinted from THE ANNALS OF MATHEMATICAL STATISTICS
Vol. 23, No. 1, March, 1952
Printed in U.S.A.

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:

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

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.

According to Kendall ([4], pp. 56 and 60), we have

$$E(S) = 0 \tag{2}$$

and

$$\begin{aligned}\operatorname{var}(S) = \tfrac{1}{18}\{r(r-1)(2r+5) - \sum_t t(t-1)(2t+5) \\ - \sum_s s(s-1)(2s+5)\} + \frac{1}{9r(r-1)(r-2)}\{\sum_t t(t-1)(t-2)\} \\ \cdot \{\sum_s s(s-1)(s-2)\} + \frac{1}{2r(r-1)}\{\sum_t t(t-1)\}\{\sum_s s(s-1)\},\end{aligned} \tag{3}$$

where summation $\sum_t$ takes place over the various ties among $u_1, \cdots, u_r$, and $\sum_s$ over the ties among $v_1, \cdots, v_r$; t and s respectively indicating the number of elements in every group of equal numbers among $u_1, \cdots, u_r$ and $v_1, \cdots, v_r$ respectively. From (1) we have

$$E(U) = \tfrac{1}{2}nm - E(S) = \tfrac{1}{2}nm \tag{4}$$

and

$$\operatorname{var}(U) = \tfrac{1}{4}\operatorname{var}(S). \tag{5}$$

The group $v_1, \cdots, v_r$ consists of n numbers 0 and m numbers 1; thus s in (3) takes the values n and m and we have

$$\sum_s s(s-1)(2s+5) = n(n-1)(2n+5) + m(m-1)(2m+5),$$

$$\sum_s s(s-1)(s-2) = n(n-1)(n-2) + m(m-1)(m-2),$$

$$\sum_s s(s-1) = n(n-1) + m(m-1).$$

Substituting in (3) and (5), we obtain after some reduction

$$\begin{aligned}\operatorname{var}(U) = \tfrac{1}{12}nm(n+m+1) - \tfrac{1}{72}\sum_t t(t-1)(2t+5) \\ + \frac{n(n-1)(n-2) + m(m-1)(m-2)}{36(n+m)(n+m-1)(n+m-2)}\sum_t t(t-1)(t-2) \\ + \frac{n(n-1)+m(m-1)}{8(m+n)(m+n-1)}\sum_t t(t-1),\end{aligned} \tag{6}$$

where $\sum_t$ takes place over the ties among the values $x_1, \cdots, x_n, y_1, \cdots, y_m$, taken together.

When no ties are present this reduces to results of Mann and Whitney [2]:

$$E(U) = \tfrac{1}{2}nm; \operatorname{var}(U) = \tfrac{1}{12}nm(n+m+1). \tag{7}$$

From (6) and (7) it is easy to prove (e.g., by induction) that var (U) is decreased by the presence of ties among the observations. These results constitute a first

step towards the possibility of using Wilcoxon's test for samples from *any* population.

REFERENCES

[1] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, Vol. 1 (1945), pp. 80–83.

[2] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 50–60.

[3] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, Vol. 30 (1938), pp. 81–93.

[4] M. G. Kendall, *Rank Correlation Methods*, Ch. Griffin and Co., London, 1948.

Reprinted from THE ANNALS OF MATHEMATICAL STATISTICS
Vol. 23, No. 1, March, 1952
Printed in U.S.A.

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:

$$-1, \text{ if } (u_h - u_k)(v_h - v_k) < 0,$$

$$0, \text{ if } (u_h - u_k)(v_h - v_k) = 0,$$

$$+1, \text{ if } (u_h - u_k)(v_h - v_k) > 0,$$

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.

According to Kendall ([4], pp. 56 and 60), we have

$$E(S) = 0 \tag{2}$$

and

$$\begin{aligned} \operatorname{var}(S) = \tfrac{1}{18}\{r(r-1)(2r+5) - \sum_t t(t-1)(2t+5) \\ - \sum_s s(s-1)(2s+5)\} + \frac{1}{9r(r-1)(r-2)}\{\sum_t t(t-1)(t-2)\} \\ \cdot \{\sum_s s(s-1)(s-2)\} + \frac{1}{2r(r-1)}\{\sum_t t(t-1)\}\{\sum_s s(s-1)\}, \end{aligned} \tag{3}$$

where summation $\sum_t$ takes place over the various ties among $u_1, \cdots, u_r$, and $\sum_s$ over the ties among $v_1, \cdots, v_r$; t and s respectively indicating the number of elements in every group of equal numbers among $u_1, \cdots, u_r$ and $v_1, \cdots, v_r$ respectively. From (1) we have

$$E(U) = \tfrac{1}{2}nm - E(S) = \tfrac{1}{2}nm \tag{4}$$

and

$$\operatorname{var}(U) = \tfrac{1}{4}\operatorname{var}(S). \tag{5}$$

The group $v_1, \cdots, v_r$ consists of n numbers 0 and m numbers 1; thus s in (3) takes the values n and m and we have

$$\sum_s s(s-1)(2s+5) = n(n-1)(2n+5) + m(m-1)(2m+5),$$

$$\sum_s s(s-1)(s-2) = n(n-1)(n-2) + m(m-1)(m-2),$$

$$\sum_s s(s-1) = n(n-1) + m(m-1).$$

Substituting in (3) and (5), we obtain after some reduction

$$\begin{aligned} \operatorname{var}(U) = \tfrac{1}{12}nm(n+m+1) - \tfrac{1}{72}\sum_t t(t-1)(2t+5) \\ + \frac{n(n-1)(n-2) + m(m-1)(m-2)}{36(n+m)(n+m-1)(n+m-2)}\sum_t t(t-1)(t-2) \\ + \frac{n(n-1) + m(m-1)}{8(m+n)(m+n-1)}\sum_t t(t-1), \end{aligned} \tag{6}$$

where $\sum_t$ takes place over the ties among the values $x_1, \cdots, x_n, y_1, \cdots, y_m$, taken together.

When no ties are present this reduces to results of Mann and Whitney [2]:

$$E(U) = \tfrac{1}{2}nm; \operatorname{var}(U) = \tfrac{1}{12}nm(n+m+1). \tag{7}$$

From (6) and (7) it is easy to prove (e.g., by induction) that var (U) is decreased by the presence of ties among the observations. These results constitute a first

step towards the possibility of using Wilcoxon's test for samples from *any* population.

REFERENCES

[1] F. WILCOXON, "Individual comparisons by ranking methods," *Biometrics Bull.*, Vol. 1 (1945), pp. 80–83.

[2] H. B. MANN AND D. R. WHITNEY, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 50–60.

[3] M. G. KENDALL, "A new measure of rank correlation," *Biometrika*, Vol. 30 (1938), pp. 81–93.

[4] M. G. KENDALL, *Rank Correlation Methods*, Ch. Griffin and Co., London, 1948.

Reprinted from THE ANNALS OF MATHEMATICAL STATISTICS
Vol. 23, No. 1, March, 1952
Printed in U.S.A.

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:

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

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.

According to Kendall ([4], pp. 56 and 60), we have

$$E(S) = 0 \tag{2}$$

and

$$\begin{aligned} \operatorname{var}(S) = \tfrac{1}{18}\{r(r-1)(2r+5) - \sum_t t(t-1)(2t+5) \\ - \sum_s s(s-1)(2s+5)\} + \frac{1}{9r(r-1)(r-2)}\{\sum_t t(t-1)(t-2)\} \\ \cdot \{\sum_s s(s-1)(s-2)\} + \frac{1}{2r(r-1)}\{\sum_t t(t-1)\}\{\sum_s s(s-1)\}, \end{aligned} \tag{3}$$

where summation $\sum_t$ takes place over the various ties among $u_1, \cdots, u_r$, and $\sum_s$ over the ties among $v_1, \cdots, v_r$; t and s respectively indicating the number of elements in every group of equal numbers among $u_1, \cdots, u_r$ and $v_1, \cdots, v_r$ respectively. From (1) we have

$$E(U) = \tfrac{1}{2} nm - E(S) = \tfrac{1}{2} nm \tag{4}$$

and

$$\operatorname{var}(U) = \tfrac{1}{4} \operatorname{var}(S). \tag{5}$$

The group $v_1, \cdots, v_r$ consists of n numbers 0 and m numbers 1; thus s in (3) takes the values n and m and we have

$$\sum_s s(s-1)(2s+5) = n(n-1)(2n+5) + m(m-1)(2m+5),$$

$$\sum_s s(s-1)(s-2) = n(n-1)(n-2) + m(m-1)(m-2),$$

$$\sum_s s(s-1) = n(n-1) + m(m-1).$$

Substituting in (3) and (5), we obtain after some reduction

$$\begin{aligned} \operatorname{var}(U) = \tfrac{1}{12}nm(n+m+1) - \tfrac{1}{72}\sum_t t(t-1)(2t+5) \\ + \frac{n(n-1)(n-2) + m(m-1)(m-2)}{36(n+m)(n+m-1)(n+m-2)}\sum_t t(t-1)(t-2) \\ + \frac{n(n-1) + m(m-1)}{8(m+n)(m+n-1)}\sum_t t(t-1), \end{aligned} \tag{6}$$

where $\sum_t$ takes place over the ties among the values $x_1, \cdots, x_n, y_1, \cdots, y_m$, taken together.

When no ties are present this reduces to results of Mann and Whitney [2]:

$$E(U) = \tfrac{1}{2} nm; \operatorname{var}(U) = \tfrac{1}{12} nm(n+m+1). \tag{7}$$

From (6) and (7) it is easy to prove (e.g., by induction) that var (U) is decreased by the presence of ties among the observations. These results constitute a first

step towards the possibility of using Wilcoxon's test for samples from *any* population.

REFERENCES

[1] F. WILCOXON, "Individual comparisons by ranking methods," *Biometrics Bull.*, Vol. 1 (1945), pp. 80–83.

[2] H. B. MANN AND D. R. WHITNEY, "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 50–60.

[3] M. G. KENDALL, "A new measure of rank correlation," *Biometrika*, Vol. 30 (1938), pp. 81–93.

[4] M. G. KENDALL, *Rank Correlation Methods*, Ch. Griffin and Co., London, 1948.

8260