KONINKL. NEDERL. AKADEMIE VAN WETENSCHAPPEN – AMSTERDAM
Reprinted from Proceedings, Series A, 55, No. 3 and Indag. Math., 14, No. 3, 1952

MATHEMATICS

THE ASYMPTOTIC NORMALITY AND CONSISTENCY OF KENDALL'S TEST AGAINST TREND, WHEN TIES ARE PRESENT IN ONE RANKING

BY

T. J. TERPSTRA

(Communicated by Prof. D. VAN DANTZIG at the meeting of March 29, 1952)

1. *Introduction*

If $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l$ [1]) are l random variables under observation, H. B. MANN [7] defines an upward trend by means of the inequality

$$\sum_{i<j}\sum \varepsilon_{j,i} > 0, \tag{1.1}$$

where

$$\varepsilon_{j,i} = 2\,\mathrm{P}[\boldsymbol{x}_i < \boldsymbol{x}_j] - 1, \quad -1 \leqq \varepsilon_{j,i} \leqq +1. \text{ [2])}$$

If only one observation x_i of $\boldsymbol{x}_i$ is given for each $i = 1, \ldots, l$, all observations being different from each other, the test of significance of rank correlation, given by M. G. KENDALL [4] and based on his ranking coefficient $\boldsymbol{S}$ can be considered as a test against trend. MANN has shown the consistency and unbiasedness of the test, using a statistic $\boldsymbol{T}$, which is defined as the number of pairs (i, j) with $i < j$ and $\boldsymbol{x}_i < \boldsymbol{x}_j$ and which, but for a constant, is identical with KENDALL'S $\boldsymbol{S}$.

In this paper, the asymptotic normality and consistency of KENDALL'S test is shown for the case, that any number of observations of each random variable is available, all observations being different from each other. Instead of KENDALL'S $\boldsymbol{S}$, another statistic is defined, which again, but for a constant term, proves to be identical with $\boldsymbol{S}$, this time for the case where ties are present in one ranking. In the definition of this statistic, use is made of the statistic $\boldsymbol{U}$ of WILCOXON'S two-sample test [9], as defined by MANN and WHITNEY [8], for compairing all samples, taken from the variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l$.

2. *The problem*

Let $\boldsymbol{x}_{i,h}$, $i = 1, \ldots, l$, $h = 1, \ldots, n_i$, $\sum_i n_i = N$, be N independent, continuously distributed random variables, all $\boldsymbol{x}_{i,h}$, $h \leqq n_i$, having the

[1]) Random variables will be distinguished from numbers (e.g. from the values, they take in an experiment), by printing them in bold type.

[2]) Unless explicitly stated, i, j, take the values $1, \ldots, l$.

same distribution function as a random variable $\boldsymbol{x}_i$. Let $x_{i,h}$ denote an observation of $\boldsymbol{x}_{i,h}$. We shall call the values $x_{i,h}$, $h \leqq n_i$, the sample taken from $\boldsymbol{x}_i$.

We want a test, based on the observations $x_{i,h}$ for the hypothesis H_0, stating that the random variables $\boldsymbol{x}_i$ all have the same continuous distribution function, against the alternative hypothesis H_1, stating that the variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l$ possess an upward trend, as defined by (1. 1).

Because of the continuity, all observations may be assumed to be different from each other, this being true with probability 1.

3. *The statistic used and its connection with* KENDALL's $\boldsymbol{S}$

Let $\boldsymbol{U}_{i,j}$ be the number of pairs (h, k), $h \leqq n_i$, $k \leqq n_j$, with $i < j$ and $\boldsymbol{x}_{i,h} < \boldsymbol{x}_{j,k}$ [3]). Then we define

$$\boldsymbol{T} = \sum_{i<j} \sum \boldsymbol{U}_{i,j}. \tag{3. 1}$$

For the case, that equal observations occur, this definition may be extended by increasing $\boldsymbol{U}_{i,j}$ with one half for each pair $(\boldsymbol{x}_{i,h}, \boldsymbol{x}_{j,k})$ of equal observations.

J. HEMELRIJK [3] has remarked that the statistic $\boldsymbol{U}$ of WILCOXON's two-sample test is connected with KENDALL's $\boldsymbol{S}$ for two rankings. In the same way it can be shown that $\boldsymbol{T}$ also is connected with $\boldsymbol{S}$. For this purpose we arrange all observations in order of increasing magnitude. A second ranking with l ties of the sizes n_i [4]) is obtained, by attributing to the n_i observations $x_{i,h}$ the same rank

$$\frac{1}{n_i} \sum_{h=1}^{n_i} \Big(\sum_{k=1}^{i-1} n_k + h \Big) = \sum_{k=1}^{i-1} n_k + \tfrac{1}{2}(n_i + 1).$$

If we compute KENDALL's $\boldsymbol{S}$ for these two rankings, a pair $(\boldsymbol{x}_{i,h}, \boldsymbol{x}_{j,k})$ $(i \leqq j)$ contributes to $\boldsymbol{S}$ a score $sgn\,(j - i)\,(\boldsymbol{x}_{j,k} - \boldsymbol{x}_{i,h})$, whereas only pairs $(\boldsymbol{x}_{i,h}, \boldsymbol{x}_{j,k})$ $(i < j)$ are considered for computation of $\boldsymbol{T}$, contributing to $\boldsymbol{T}$ a score 1, $\frac{1}{2}$ or 0, according to $(\boldsymbol{x}_{j,k} - \boldsymbol{x}_{i,h})$ being > 0, $= 0$, or < 0.

Consequently

$$2\,\boldsymbol{T} - \boldsymbol{S} = \sum_{i<j} \sum n_i\, n_j. \tag{3. 2}$$

To test the hypothesis H_0, against the alternative hypothesis H_1, a critical region $\{T \geqq T_a\}$ is defined, where T_a is the smallest integer with

$$\mathrm{P}\,[\boldsymbol{T} \geqq T_a \,|\, H_0] \leqq a.$$

For $T \geqq T_a$, H_0 is rejected, the probability of a wrong conclusion being at most a.

---

[3]) MANN and WHITNEY define $\boldsymbol{U}_{i,j}$ as the number of pairs $(\boldsymbol{x}_{i,h}, \boldsymbol{x}_{j,k})$ with $i < j$ and $\boldsymbol{x}_{i,h} > \boldsymbol{x}_{j,k}$.

[4]) The case $n_i = 1$ is included.

4. *The probability distribution of* $\mathbf{T}$ *under* H_0

Lemma 1: Taking for some j all $\mathbf{x}_h$ with $i < j$ and $h \leqq n_i$ together, we define $\mathbf{Z}_1 = 0$ and for $j \geqq 2$: $\mathbf{Z}_j = \mathbf{U}_{(1,\ldots,j-1),j}$ as the number of pairs (h, k) with $h \leqq n_i$, $i < j$, $k \leqq n_j$ and $\mathbf{x}_{i,h} < \mathbf{x}_{j,k}$. Then $\mathbf{T} = \sum_j \mathbf{Z}_j$.

Proof: From the definition of $\mathbf{U}_{i,j}$ and $\mathbf{U}_{(1,\ldots,j-1),j}$ follows:

$$(4.1) \qquad \begin{cases} \mathbf{U}_{(1,\ldots,j-1),j} = \mathbf{U}_{1,j} + \ldots + \mathbf{U}_{j-1,j}, \\ \mathbf{T} = \sum_{i<j} \sum \mathbf{U}_{i,j} = \sum_{j=2} \mathbf{U}_{(1,\ldots,j-1),j} = \sum_j \mathbf{Z}_j. \end{cases}$$

Lemma 2: If H_0 is true, the random variables $\mathbf{Z}_j$ are (completely) independent.

This lemma follows immediately from Theorem I (below) and the definition of $\mathbf{Z}_j$.

Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be n (completely) independent random variables with the same continuous probability distribution. Let R_n be the n-dimensional fundamental probability set and let for every point $E = (y_1, \ldots, y_n) \in R_n$ the ranks $r_1, \ldots, r_n$ be defined by $r_i = \frac{1}{2} \sum_j sgn\,(y_i - y_j) + \frac{1}{2}(n+1)$. Thus $\mathbf{r}_1, \ldots, \mathbf{r}_n$ have a simultaneous probability distribution on R_n.

Theorem I: *If the set of random variables* $\mathbf{y}_1, \ldots, \mathbf{y}_n$ *is split into two sub-sets* $\mathbf{y}_1, \ldots, \mathbf{y}_m$ *and* $\mathbf{y}_{m+1}, \ldots, \mathbf{y}_n$ *and if* $\{\mathbf{U}\}$ *denotes a set of statistics depending on the permutation of the ranks* $\mathbf{r}_1, \ldots, \mathbf{r}_m$ *only, when arranged according to increasing magnitude, and* $\{\mathbf{V}\}$ *a set of statistics depending on the ranks* $\mathbf{r}_{m+1}, \ldots, \mathbf{r}_n$ *only, then* $\{\mathbf{U}\}$ *is stochastically independent of* $\{\mathbf{V}\}$.

Proof: Let $K = K(r_{m+1}, \ldots, r_n)$ be the subset of R_n, consisting of all points $(y_1, \ldots, y_n)$ for which $y_{m+1}, \ldots, y_n$ have the ranks $r_{m+1}, \ldots, r_n$, then $\mathrm{P}[\{\mathbf{U}\} = \{U\} | K] = \mathrm{P}[\{\mathbf{U}\} = \{U\}]$. On K, the statistics $\{\mathbf{V}\}$ have constant values $\{V\}$. If $\Lambda = \Lambda\{V\}$ is the subset of R_n, consisting of all disjoint subsets K, corresponding with the set $\{V\}$, we have

$$\mathrm{P}[\{\mathbf{U}\} = \{U\} | \Lambda] = \frac{\sum_{(K)} \mathrm{P}[\{\mathbf{U}\} = \{U\} \wedge \mathbf{E} \in K]}{\sum_{(K)} \mathrm{P}[\mathbf{E} \in K]} =$$

$$= \frac{\sum_{(K)} \mathrm{P}[\{\mathbf{U}\} = \{U\} | K] \cdot \mathrm{P}[\mathbf{E} \in K]}{\sum_{(K)} \mathrm{P}[\mathbf{E} \in K]} = \mathrm{P}[\{\mathbf{U}\} = \{U\}].$$

Theorem II: *If* H_0 *is true, the mean and the variance of* $\mathbf{T}$ *are*

$$(4.2) \qquad \begin{cases} \mu_0 = \mathfrak{E}(\mathbf{T} | H_0) = \frac{1}{4}(N^2 - \sum_i n_i^2), \\ \sigma_0^2 = \mathrm{var}(\mathbf{T} | H_0) = \frac{1}{72}\{N(N+1)(2N+1) - \sum n_i(n_i+1)(2n_i+1)\} \end{cases}$$

where

$$N = \sum_i n_i.$$

Proof: The proof follows immediately from (3.2) and KENDALL's result (cf. M. G. KENDALL, 1938 and Rank Correlation Methods, p. 43 (4.4)) by remarking that $(n_i - 1)(2n_i + 5)$ and $(n_i + 1)(2n_i + 1)$ have a

constant difference, cancelling after summation over i against the corresponding part of the first term.

Theorem III: *If H_0 is true, $\boldsymbol{T}$ is asymptotically normally distributed with mean and variance given by* (4. 2).

Proof: Because of (4. 1) and Lemma 2, the characteristic function $K(t)$ of the reduced variable $\tilde{\boldsymbol{T}} = \boldsymbol{T} - \mathcal{E}(\boldsymbol{T} \mid H_0)$ is the product of the characteristic functions $K_j(t)$ of the variables $\tilde{\boldsymbol{Z}}_j = \boldsymbol{Z}_j - \mathcal{E}(\boldsymbol{Z}_j \mid H_0)$, thus (cf [2]), putting

$$s_n(t) = \prod_{k=1}^{n} (\sin kt)/kt,$$

$$K(t) = s_N(\tfrac{1}{2}t) / \prod_i s_{n_i}(\tfrac{1}{2}t).$$

The characteristic function of the variable $\boldsymbol{T}' = \boldsymbol{T}/\sigma_0$ evidently is $K(t/\sigma_0)$. [5])

Now we assume that n and (or) l depend on a parameter ν and tend to infinity as $\nu \to \infty$, whereas constants A and B exist with

(4. 3) $\quad 0 < A \leqq n^{-1} n_i \leqq B < \infty$ for all $i = 1, \ldots, l$ and all ν.

Then we have for large ν

$$\log K\left(\frac{t}{\sigma_0}\right) = -\tfrac{1}{6} \sum_{k=1}^{N} \left(k \frac{t}{2\sigma_0}\right)^2 + \tfrac{1}{6} \sum_i \sum_{k=1}^{n_i} \left(k \frac{t}{2\sigma_0}\right)^2 + O(l^{-1} n^{-1}),$$

consequently

$$\lim \log K\left(\frac{t}{\sigma_0}\right) = \lim \tfrac{1}{24} \frac{t^2}{\sigma_0^2} \left(-\sum_{k=1}^{N} k^2 + \sum_i \sum_{k=1}^{n_i} k^2\right) = -\tfrac{1}{2} t^2,$$

which is the logarithm of the characteristic function of a normally distributed variable with mean 0 and variance 1.

Because of (3. 2) we have the following

Corollary: KENDALL's *rank-correlation coefficient $\boldsymbol{S}$ for two rankings, one ranking consisting of l ties of the sizes n_i ($n_i = 1$ included), whereas in the other ranking all ranks are different from each other, is asymptotically normally distributed with mean* 0 *and variance* $4\sigma_0^2$.

5. *The consistency of the test*

To derive the class of alternative hypotheses H, against which the test is consistent, we define, analogous to D. VAN DANTZIG [1]

$$\tau_{j,k;\,i,h} = \iota(\boldsymbol{x}_{j,k} - \boldsymbol{x}_{i,h}),$$

where

$$\iota(z) = \begin{cases} 1, & \text{if } z \geqq 0, \\ 0, & \text{if } z < 0. \end{cases}$$

[5]) I owe this result to Prof. Dr D. VAN DANTZIG.

The statistic T can then be rewritten as

$$T = \sum_{i<j} \sum_{h=1}^{n_i} \sum_{k=1}^{n_j} \tau_{j,k;\,i,h}. \tag{5.1}$$

We first derive the mean of T and an upper bound for the variance of T under the hypothesis

$$H: \quad \mathrm{P}[x_{i,h} < x_{j,k}] = \tfrac{1}{2}(1 + \varepsilon_{j,i}), \quad -1 \leqq \varepsilon_{j,i} \leqq +1, \tag{5.2}$$

all $x_{i,h}$ being (completely) independent and continuously distributed. Because of the continuity

$$\mathrm{P}[x_{i,h} < x_{j,k}] = \mathrm{P}[x_{i,h} \leqq x_{j,k}] = \tfrac{1}{2}(1 + \varepsilon_{j,i}),$$

hence $\mathfrak{E}\tau_{j,k;\,i,h} = \frac{1}{2}(1 + \varepsilon_{j,i})$ and

$$\mu = \mathfrak{E}(T|H) = \tfrac{1}{2} \sum_{i<j} \sum n_i n_j (1 + \varepsilon_{j,i}). \tag{5.3}$$

As $\iota(z)^2 = \iota(z)$, $\mathfrak{E}(\tau_{j,k;\,i,h})^2 = \mathfrak{E}\tau_{j,k;\,i,h} = \frac{1}{2}(1 + \varepsilon_{j,i})$ and

$$\operatorname{var} \tau_{j,k;\,i,h} = \tfrac{1}{4}(1 - \varepsilon_{j,i}^2).$$

Further we remark that $\tau_{j,k;\,i,h}$ and $\tau_{j',k';\,i',h'}$ are only correlated, if the pairs $(x_{i,h}, x_{j,k})$ and $(x_{i',h'}, x_{j',k'})$ have a variable in common. Then:

$$\operatorname{cov}(\tau_{j',k';\,i,h}, \tau_{j,k;\,i,h}) = \mathrm{P}[x_{i,h} < x_{j,k} \wedge x_{i,h} < x_{j',k'}] - \tfrac{1}{4}(1 + \varepsilon_{j,i})(1 + \varepsilon_{j',i})$$
$$\leqq \min\{\tfrac{1}{2}(1 + \varepsilon_{j,i}), \tfrac{1}{2}(1 + \varepsilon_{j',i})\} - \tfrac{1}{4}(1 + \varepsilon_{j,i})(1 + \varepsilon_{j',i}) \leqq \tfrac{1}{4}(1 - \varepsilon_{j,i} \cdot \varepsilon_{j',i}),$$

and in the same way:

$$\operatorname{cov}(\tau_{j,k;\,i,h}, \tau_{j,k;\,i',h'}) \leqq \tfrac{1}{4}(1 - \varepsilon_{j,i} \cdot \varepsilon_{j,i'}),$$
$$\operatorname{cov}(\tau_{j',k';\,j,k}, \tau_{j,k;\,i,h}) = \mathrm{P}[x_{i,h} < x_{j,k} < x_{j',k'}] - \mathrm{P}[x_{i,h} < x_{j,k}] \cdot \mathrm{P}[x_{j,k} < x_{j',k'}] \leqq 0.$$ [6])

Because of (5.1) and the inequalities, derived above, we obtain

$$\sigma^2 = \operatorname{var}(T|H) \leqq \tfrac{1}{4} \sum_{i<j} \sum n_i n_j (n_i + n_j - 1) + 2 \sum_{i<j<k} \sum \sum n_i n_j n_k. \tag{5.4}$$

From (4.2) and (5.4) follows:

$$\sigma^2 \leqq 12\, \sigma_0^2. \tag{5.5}$$

We now proceed to state the following

Theorem IV: *Rejection of the hypothesis H_0, if and only if $T \geqq T_a$, is a consistent test of H_0 against the class of alternative hypotheses* (5.2), *satisfying the conditions*

a) $\lambda = \dfrac{1}{n^2 l(l-1)} \sum_{i<j} \sum n_i n_j \varepsilon_{j,i} > 0,$

(so that $\frac{1}{2}\lambda n^2 l(l-1) = \mu - \mu_0$*), and*

b) $\lambda^{-1} = \mathrm{o}((ln)^{1/3}),$

[6]) Cf. H. B. Mann [1], p. 251.

and for sufficiently small α against no other alternatives, belonging to the class of alternative hypotheses (5. 2).

Proof: The probability that H_0 is not rejected under the hypothesis H is $P[\boldsymbol{T} < T_\alpha \mid H]$, where T_α is the smallest integer with

$$\beta = P[\boldsymbol{T} \geqq T_\alpha \mid H_0] \leqq \alpha.$$

We can write $T = T_\alpha$ in the form $T = \mu_0 + c\sigma_0$, where $\mu_0 = \mathfrak{E}(\boldsymbol{T} \mid H_0)$ and $\sigma_0^2 = \text{var } (\boldsymbol{T} \mid H_0)$.

Because of the asymptotic normality of $\boldsymbol{T}$, $\lim \beta = \alpha$ and $\lim c = \xi_\alpha$ [7]), where

$$\frac{1}{\sqrt{2\pi}} \int_{\xi_\alpha}^{\infty} e^{-\frac{1}{2}x^2}\, dx = \alpha.$$

We then obtain

$$P[\boldsymbol{T} < T_\alpha | H] = P[\boldsymbol{T} - \mu < -\tfrac{1}{2}\lambda l(l-1)\, n^2 + c\sigma_0 | H]$$
$$\leqq \frac{\sigma^2}{\sigma_0^2}\{-\tfrac{1}{2}\lambda l(l-1)\, n^2\, \sigma_0^{-1} + c\}^{-2} \leqq 12\,\{\tfrac{1}{2}\lambda l\,(l-1)\, n^2\, \sigma_0^{-1} - c\}^{-2},$$

because of (5. 5) and BIENAYMÉ's inequality, $\mu_0 - \mu + c\sigma_0$ being negative for sufficiently large ν, as c is bounded, $\lambda > 0$. The expression between curved brackets tends to infinity with ν as $\lambda^{-1} = 0((ln)^{\frac{1}{2}})$, consequently $\lim P[\boldsymbol{T} < T_\alpha \mid H] = 0$ and the probability, that H_0 will be rejected under the hypothesis H tends to 1.

We now prove, as stated in Theorem IV, that the class of alternative hypotheses cannot be extended without loss of the property of consistency.

If condition (a) is not satisfied, then

$$P[\boldsymbol{T} \geqq T_\alpha | H] \leqq 12\,\{-\tfrac{1}{2}\lambda l(l-1)\, n^2\, \sigma_0^{-1} + c\}^{-2} \leqq 12\, c^{-2},$$

thus for sufficiently small α (e.g. sufficiently large c) there remains a positive probability $> 1 - 12c^{-2}$ that H_0 will not be rejected [8]).

If condition (b) is not satisfied, there is a constant c_1 and a sub-sequence of values ν, with $|\lambda(ln)^{\frac{1}{2}}| < c_1$. For this sub-sequence, we have for sufficiently small α

$$P[\boldsymbol{T} \geqq T_\alpha | H] \leqq 12\,\{-\tfrac{1}{2}\lambda l(l-1)\, n^2\, \sigma_0^{-1} + c\}^{-2} \leqq 12\, c_2^{-2},$$

where c_2 is a small constant. For sufficiently small α there again remains a positive probability $> 1 - 12\, c_2^{-2}$ that H_0 will not be rejected.

6. *The mean and the variance of $\boldsymbol{T}$, when equal observations are present*

In section 3, $\boldsymbol{T}$ was also defined for the case, that equal observations occur and it was shown that (3. 2) also holds.

---

[7]) All limits are taken under condition (4. 3).

[8]) If condition (b) is satisfied, this probability tends to 1, for each $\alpha < 1$.

The mean and the variance of T can thus be derived from the mean and the variance of KENDALL's S for two rankings, one ranking consisting of l ties of the sizes n_i and the other ranking consisting of k ties of the sizes m_j, representing the groups of equal observations.

We obtain (cf. M. G. KENDALL [5], p. 43 (4. 3))

$$\mathcal{E}(T|H_0) = \tfrac{1}{4}(N^2 - \textstyle\sum_i n_i^2) = \tfrac{1}{4}(N^{[2]} - \textstyle\sum_i n_i^{[2]})$$

and

$$\operatorname{var}(T|H_0) = (36\,N^{[3]})^{-1}(N^{[3]} - \textstyle\sum_i n_i^{[3]})(N^{[3]} - \textstyle\sum_j m_j^{[3]}) + \\ + (8\,N^{[2]})^{-1}(N^{[2]} - \textstyle\sum_i n_i^{[2]})(N^{[2]} - \textstyle\sum_j m_j^{[2]}),$$

where

$$N = \textstyle\sum_i n_i = \sum_j m_j,$$

whereas $x^{[k]}$ for any natural k is defined by $x^{[k]} = x(x-1)\ldots(x-k+1)$.

In this case the asymptotic normality of T and S under condition (4. 3) has not yet been shown.

Finally I want to thank Prof. Dr D. VAN DANTZIG, whose suggestions helped me to give the paper its final form.

Publication of the Statistical Department
of the "Mathematisch Centrum",
Amsterdam.

REFERENCES

1. DANTZIG, D. VAN, On the consistency and the power of WILCOXON's two-sample test, Proc. Kon. Ned. Akad. van Wetensch. **54**, 1—8 (1951); also Indagationes Mathematicae **13**, 1—8 (1951).
2. ————, Kadercursus Mathematische Statistiek 1947—1950, Hoofdstuk VI, § 3. (Mathematisch Centrum, Amsterdam).
3. HEMELRIJK, J., Note on WILCOXON's two-sample test, when ties are present, Ann. Math. Stat. **23**, no. 2 (1952).
4. KENDALL, M. G., A new measure of rank correlation, Biometrika **30**, 81 (1938).
5. ————, Rank correlation methods (London, 1948).
6. LEHMANN, E. L., Consistency and unbiasedness of certain non-parametric tests, Ann. Math. Stat. **22**, 165—180 (1951).
7. MANN, H. B., Non-parametric tests against trend, Econometrica **13**, 254—259 (1945).
8. ————, and D. R. WHITNEY, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. **18**, 50—60 (1947).
9. WILCOXON, F., Individual comparisons by ranking methods, Biometrics Bull. **1**, 80—83 (1945).

8250