STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

SD 7

The exact probability distribution of the T statistics for testing against trend and its nomal approximation.

T.J.Terpstra.

1953

KONINKL. NEDERL. AKADEMIE VAN WETENSCHAPPEN - AMSTERDAM
Reprinted from Proceedings, Series A, **56**, No. 5 and Indag. Math., **15**, No. 5, 1953

MATHEMATICS

THE EXACT PROBABILITY DISTRIBUTION OF THE T STATISTIC FOR TESTING AGAINST TREND AND ITS NORMAL APPROXIMATION [1]

BY

T. J. TERPSTRA

(Communicated by Prof. D. VAN DANTZIG at the meeting of September 26, 1953)

1. *Introduction and summary*

Let $\mathbf{x}_{i,h}$ [2]), $i = 1, \ldots, l$, $h = 1, \ldots, n_i$, $\sum_i n_i = N$, be N independent, continuously distributed random variables, all $\mathbf{x}_{i,h}$, $h \leqq n_i$, having the same distribution function as a random variable $\mathbf{x}_i$. Let $x_{i,h}$ denote an observation of $\mathbf{x}_{i,h}$. We shall call the values $x_{i,h}$, $h \leqq n_i$, the *sample* taken from $\mathbf{x}_i$. The hypothesis H_0 states that the variables $\mathbf{x}_i$ all have the same distribution function. The alternative hypothesis H_1 states that the variables $\mathbf{x}_i$ possess an upward trend, defined by means of the inequality (cf. [3]),

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

where

$$\varepsilon_{j,i} = 2\,\mathrm{P}\,[\mathbf{x}_i < \mathbf{x}_j] - 1,$$

so that

$$-1 \leqq \varepsilon_{j,i} \leqq +1 \text{ }^{3)}.$$

Then, for testing the hypothesis H_0 against the alternative hypothesis H_1, use can be made of the statistic [6]

$$\mathbf{T} = \sum_{i<j}\sum \mathbf{U}_{i,j} \text{ }^{4)},$$

where $\mathbf{U}_{i,j}$ is WILCOXON's statistic [5], here being defined as the number of pairs (h, k), $h \leqq n_i$, $k \leqq n_j$, with $i < j$ and $\mathbf{x}_{i,h} < \mathbf{x}_{j,k}$ [5]).

If T_α is defined as the smallest integer T with $\mathrm{P}[\mathbf{T} \geqq T \mid H_0] \leqq \alpha$,

---

[1]) After finishing this paper I learned from Mr SALIB of the University of Liverpool that he has independently obtained a part of the results stated here.

[2]) Random variables will be indicated by printing them in bold type.

[3]) Unless mentioned otherwise, i, j, take the values $1, \ldots, l$.

[4]) In [6] it has been shown that $\mathbf{T}$ is connected with KENDALL's rank correlation statistic $\mathbf{S}$ ([1] and [2]) by means of the relation $2\mathbf{T} - \mathbf{S} = \sum_{i<j}\sum n_i n_j$.

[5]) MANN and WHITNEY [4] define $\mathbf{U}_{i,j}$ as the number of pairs (h, k), $h \leqq n_i$, $k \leqq n_j$, with $i < j$ and $\mathbf{x}_{i,h} > \mathbf{x}_{j,k}$.

Remark: In my paper [6] condition b of Theorem IV, stating that $\lambda^{-1} = \sigma((ln)^{1/2})$, must be replaced by $\lambda^{-1} = \sigma(l^2 n^{1/2})$.

the hypothesis H_0 will be rejected against the alternative hypothesis H_1, if the observed $T \geqq T_\alpha$, the chance of a fault of the first kind being at most α.

It has been shown [6] that under H_0 $\boldsymbol{T}$ for large n_i is asymptotically normally distributed with mean μ and variance σ^2, given by

$$(1.1) \qquad \mu = {}^1/_4 \,(N^2 - \sum_i n_i^2)$$
$$\sigma^2 = {}^1/_{72} \,\{N(N+1)\,(2N+1) - \sum_i n_i(n+1)\,(2n_i+1)\}.$$

In the following section a recurrence relation is derived for the exact probability distribution

$$\mathrm{P}_{n_1,\ldots,n_l}(T) = \mathrm{P}\,[\boldsymbol{T} = T \mid H_0\,;\, n_1, \ldots, n_l].$$

In section 3 tables for the exact distribution function $F_{n_1,\ldots,n_l}(T)$ of T are given for the three sample cases $n_1 \leqq n_2 \leqq n_3 \leqq 5$.

From these distribution functions the significance levels T_α have been derived for $\alpha = 0.005$, 0.010, 0.025, 0.050, 0.100. These significance levels are compared with the levels T_α^*, computed by means of the approximating normal distribution functions.

2. *The probability distribution of* $\boldsymbol{T}$

Because of the continuity of the distribution functions of the variables $\boldsymbol{x}_{i,h}$, all observations $x_{i,h}$ may be assumed to be different from each other, this being true with probability 1. The observations can thus be arranged in order of increasing magnitude.

If $K_{n_1,\ldots,n_l}(T)$ is the number of sequences $\{x_{i,h}\}$ with $\boldsymbol{T} = T$, we obtain by isolating the last observation in each of these sequences the recurrence relation

$$K_{n_1,\ldots,n_l}(T) = \sum_{i=1}^{l} K_{n_1,\ldots,n_i-1,\ldots,n_l}(T - n_{i+1} - \ldots - n_l).$$

If H_0 is true, each of the $N!/\Pi n_i!$ different sequences has the same probability, consequently

$$(2.1) \qquad \mathrm{P}_{n_1,\ldots,n_l}(T) = N^{-1} \sum_{i=1}^{l} n_i \, \mathrm{P}_{n_1,\ldots,n_i-1,\ldots,n_l}(T - n_{i+1} - \ldots - n_l)$$

with initial conditions

$$\mathrm{P}_{n_1,\ldots,n_l}(T) = 0, \text{ if } T < 0 \text{ or if an } n_i < 0,$$

and

$$\mathrm{P}_{0,\ldots,0}(T) = \delta_{T,0} \begin{cases} = 0, & \text{if } T \neq 0, \\ = 1, & \text{if } T = 0, \end{cases}$$

whence for all $n_i \geqq 0$

$$\mathrm{P}_{0,\ldots,0,n_i,0,\ldots,0}(T) = \delta_{T,0}.$$

TABLE 1

The distribution function $F_{n_1,n_2,n_3}(T)$ for $n_1 \leqq n_2 \leqq n_3 \leqq 5$.

$F_{n_1,n_2,n_3}(T) = 1 - F_{n_1,n_2,n_3}(\sum\sum_{i<j} n_i n_j - T - 1)$

n_1	n_2	n_3	$\sum\sum_{i<j} n_i n_j$ \ T	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
1	1	1	3	.167	.500																																				
1	1	2	5	.083	.250	.500																																			
1	2	2	8	.033	.100	.233	.400	.600																															
2	2	2	12	.011	.033	.089	.167	.289	.422	.578																													
1	1	3	7	.050	.150	.300	.500																																		
1	2	3	11	.017	.050	.117	.217	.350	.500																														
2	2	3	16	.005	.014	.038	.076	.138	.219	.324	.438	.562																											
1	3	3	15	.007	.021	.050	.100	.171	.264	.379	.500																												
2	3	3	21	.002	.005	.014	.030	.057	.096	.152	.221	.305	.400	.500																									
3	3	3	27	.001	.002	.005	.011	.021	.037	.061	.095	.139	.194	.260	.334	.415	.500																						
1	1	4	9	.033	.100	.200	.333	.500																															
1	2	4	14	.010	.029	.067	.124	.210	.314	.438	.562	.500																											
2	2	4	20	.002	.007	.019	.038	.071	.117	.181	.257	.350	.448	.552																									
1	3	4	19	.004	.011	.025	.050	.089	.143	.214	.300	.396	.500																										
2	3	4	26	.001	.002	.006	.013	.026	.045	.074	.112	.162	.222	.294	.372	.457	.543																						
3	3	4	33	.000	.001	.002	.004	.009	.015	.026	.042	.064	.093	.130	.175	.228	.289	.355	.427	.500																				
1	4	4	24	.002	.005	.011	.022	.041	.068	.106	.156	.217	.289	.370	.456	.544																							
2	4	4	32	.000	.001	.003	.005	.011	.019	.032	.050	.076	.108	.149	.198	.256	.320	.390	.463	.537																				
3	4	4	40	.000	.000	.001	.002	.003	.006	.010	.017	.026	.040	.058	.080	.109	.144	.185	.232	.285	.343	404	.468	.532																	
4	4	4	48	.000	.000	.000	.001	.001	.002	.004	.006	.010	.015	.023	.033	.046	.063	.084	.110	.140	.176	.216	.260	.309	.361	.416	.472	.528													
1	1	5	11	.024	.071	.143	.238	.357	.500																														
1	2	5	17	.006	.018	.042	.077	.131	.202	.292	.393	.500																											
2	2	5	24	.001	.004	.011	.021	.040	.066	.104	.153	.216	.287	.369	.455	.545																							
1	3	5	23	.002	.006	.014	.028	.050	.081	.125	.181	.248	.325	.411	.500																								
2	3	5	31	.000	.001	.003	.007	.013	.023	.038	.059	.088	.124	.169	.223	.285	.352	.425	.500																					
3	3	5	39	.000	.000	.001	.002	.004	.007	.012	.020	.031	.046	.066	.092	.123	.161	.206	.256	.312	.372	.435	.500																		
1	4	5	29	.001	.002	.006	.011	.021	.035	.056	.083	.120	.165	.219	.281	.350	.424	.500																						
2	4	5	38	.000	.000	.001	.002	.005	.009	.015	.024	.037	.054	.077	.105	.140	.181	.229	.282	.341	.403	.467	.533																		
3	4	5	47	.000	.000	.000	.001	.001	.002	.004	.007	.012	.018	.026	.038	.053	.072	.095	.123	.156	.193	.236	.283	.333	.387	.443	.500														
4	4	5	56	.000	.000	.000	.000	.000	.001	.001	.002	.004	.006	.010	.014	.020	.028	.039	.052	.068	.087	.111	.137	.168	.203	.241	.283	.328	.375	.424	.475	.525									
1	5	5	35	.000	.001	.003	.005	.009	.016	.026	.040	.060	.084	.116	.154	.198	.250	.307	.368	.433	.500																				
2	5	5	45	.000	.000	.000	.001	.002	.004	.006	.010	.016	.025	.036	.050	.069	.092	.119	.152	.190	.233	.280	.331	.386	.442	.500															
3	5	5	55	.000	.000	.000	.000	.001	.001	.002	.003	.005	.007	.011	.016	.023	.032	.044	.058	.076	.097	.122	.151	.184	.220	.261	.304	.350	.399	.449	.500										
4	5	5	65	.000	.000	.000	.000	.000	.000	.001	.001	.002	.002	.004	.006	.008	.012	.016	.022	.030	.039	.051	.065	.082	.101	.124	.150	.179	.211	.246	.283	.324	.366	.410	.455	.500					
5	5	5	75	.000	.000	.000	.000	.000	.000	.000	.000	.001	.001	.001	.002	.003	.004	.006	.009	.012	.016	.021	.028	.036	.046	.057	.071	.087	.105	.126	.149	.175	.203	.234	.267	.302	.340	.378	.418	.459	.500

3. *The distribution function of* $\boldsymbol{T}$ *for* $n_1 \leqq n_2 \leqq n_3 \leqq 5$ *and its comparison with the approximating normal distribution function*

Using recurrence relation (2.1), the COMPUTATION DEPARTMENT of the MATHEMATICAL CENTRE has computed the exact distribution functions $F_{n_1,n_2,n_3}(T)$ for the three sample cases with $n_1 \leqq n_2 \leqq n_3 \leqq 5$.

Because of the symmetric character of the probability distribution of $\boldsymbol{T}$, we can write

$$F_{n_1,\ldots,n_l}(T) = 1 - F_{n_1,\ldots,n_l}(\sum_{i<j}\sum n_i n_j - T - 1).$$

Consequently, the exact distribution function of $\boldsymbol{T}$ has been tabulated only for $T = 0, 1, 2, \ldots, [\mu]$. (Cf. Table 1.)

Denoting by T_α the smallest integer T for which the inequality

$$1 - F_{n_1,\ldots,n_l}(T-1) = \sum_{m \geqq T} \mathrm{P}_{n_1,\ldots,n_l}(m) \leqq \alpha$$

holds, the exact one-sided significance levels T_α corresponding with $\alpha = 0.005$, 0.010, 0.025, 0.050 and 0.100 have been derived from table 1 (cf. table 2).

For large values of n_1, n_2, n_3 $\boldsymbol{T}$ is asymptotically normally distributed with mean μ and variance σ^2 (cf. (1.1) and [6]).

To investigate for which values of n_1, n_2, n_3 this normal distribution function may be considered as a good approximation of the exact distribution function of $\boldsymbol{T}$, the significance levels T'_α, given by this approximating normal distribution function, have been calculated also.

These significance levels T'_α are given by

$$T'_\alpha = \mu + 0{,}5 + k_\alpha \cdot \sigma,$$

where k_α is defined by the equation

$$\alpha = \frac{1}{\sqrt{2\pi}} \int_{k_\alpha}^{\infty} e^{-\frac{1}{2}t^2}\, dt.$$

The constant 0.5 is the so-called correction for continuity. In table 2 the integers T^*_α with $T'_\alpha \leqq T^*_\alpha < T'_\alpha + 1$ are tabulated. From this table we may conclude that for $n_i \geqq 5$ $(i = 1, 2, 3)$ the normal distribution function with mean μ and variance σ^2 is a good approximation of the exact distribution function. The table shows that also for smaller values of n_i the integer T^*_α usually coincides with T_α, sometimes (in particular for $\alpha = 0.005$ and 0.010) is one unit larger and only exceptionally (2, 2, 5; 4, 5, 5; $\alpha = 0.100$) is a unit smaller.

In [6] it is proved that

$$\boldsymbol{T} = \sum_{i<j}\sum \boldsymbol{U}_{i,j} = \sum_{j=2}^{l} \boldsymbol{U}_{(1,\ldots,j-1),j},$$

where $\boldsymbol{U}_{(1,\ldots,j-1),j}$ is defined as the number of pairs $(h, k), h \leqq n_i$, $i = 1, 2, \ldots, j-1$, $k \leqq n_j$, with $x_{i,h} < x_{j,k}$. As the variables $\boldsymbol{U}_{(1,\ldots,j-1),j}$ are independently distributed, we may conclude on basis of the Central Limit

TABLE 2

The exact and approximate significance levels T_α and T^*_α [1])

				$\alpha = .005$		.010		.025		.050		.100	
n_1	n_2	n_3	$\sum\sum_{i<j} n_i n_j$	T_α	T^*_α	T_α	T^*_α	T_α	T^*_α	T_α	T^*_α	T_α	T^*_α
1	1	1	3										
1	1	2	5									5	5
1	2	2	8							8	8	7	7
2	2	2	12					12	12	11	11	10	10
1	1	3	7							7	7	6	7
1	2	3	11					11	11	10	10	10	10
2	2	3	16	16		16	16	15	15	14	14	13	13
1	3	3	15			15		14	14	13	13	12	12
2	3	3	21	20	21	20	20	19	19	18	18	16	16
3	3	3	27	25	26	25	25	23	23	22	22	20	20
1	1	4	9							9	9	8	8
1	2	4	14			14		14	14	13	13	12	12
2	2	4	20	20		19	20	18	18	17	17	16	16
1	3	4	19	19		19	19	17	18	16	16	15	15
2	3	4	26	25	25	24	24	23	23	21	21	20	20
3	3	4	33	30	31	29	30	28	28	26	26	24	24
1	4	4	24	23	24	23	23	21	21	20	20	19	19
2	4	4	32	29	30	29	29	27	27	25	25	24	24
3	4	4	40	36	36	34	35	33	33	31	31	29	29
4	4	4	48	42	43	40	41	38	38	36	36	34	34
1	1	5	11					11		11	11	10	10
1	2	5	17			17		16	16	15	15	14	14
2	2	5	24	23	24	23	23	21	21	20	20	19	18
1	3	5	23	23	23	22	22	21	21	19	19	18	18
2	3	5	31	29	30	28	28	26	26	25	25	23	23
3	3	5	39	35	36	34	34	32	32	30	30	28	28
1	4	5	29	28	28	27	27	25	25	24	24	22	22
2	4	5	38	34	35	33	34	31	32	30	30	28	28
3	4	5	47	41	42	40	40	38	38	36	36	33	33
4	4	5	56	48	49	46	47	44	44	42	42	39	39
1	5	5	35	32	33	31	32	30	30	28	28	26	26
2	5	5	45	40	41	38	39	36	37	34	34	32	32
3	5	5	55	47	48	46	46	43	43	41	41	38	38
4	5	5	65	55	55	53	53	50	50	48	48	45	44
5	5	5	75	62	63	60	61	57	57	54	54	51	51

[1]) For testing against a downward trend the left-sided significance levels are found by subtracting T_α and T^*_α from $\sum\sum_{i<j} n_i n_j$.

Theorem that for more than three samples with $n_i \geqq 5$ the normal distribution function will be also a good approximation of the exact distribution function of $\boldsymbol{T}$.

REFERENCES

1. KENDALL, M. G., A new measure of rank correlation, Biometrika **30**, 81 (1938).
2. ————, Rank correlation methods (London, 1948).
3. MANN, H. B., Non-parametric tests against trend, Econometrica **13**, 254–259 (1945).
4. ———— 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).
5. WILCOXON, F., Individual comparisons by ranking methods, Biometrics Bull. **1**, 80–83 (1945).
6. TERPSTRA, T. J., The asymptotic normality and consistency of KENDALL's test against trend, when ties are present in one ranking, Indagationes Mathematicae **14**, no. 3, 327–333 (1952).