MATHEMATISCH CENTRUM
2e Boerhaavestraat 49
A m s t e r d a m -O.

S190 M50 B

Statistische Afdeling
Memorandum S190 (M50B)

Toets van Terpstra voor het probleem van m [2] rangschikkingen

Gegeven zijn m waarnemers die ieder n objecten $O_1, \ldots, O_n$ naar opklimmende grootte van een of ander gemeenschappelijk kenmerk rangschikken door rangnummers aan de objecten toe te kennen. Het rangnummer dat de i^e waarnemer aan O_j toekent zullen we aanduiden met x_{ij}.
De nulhypothese H_0 die we wensen te toetsen houdt nu in, dat de m rangschikkingen onafhankelijk van elkaar zijn en dat voor iedere waarnemer elke rangschikking van de n objecten dezelfde waarschijnlijkheid heeft.
Deze hypothese wordt getoetst tegen de alternatieve hypothese H, dat er tussen de waarnemers een overeenstemming bestaat in het rangschikken der objecten.

De toetsingsgrootheid $\underline{S}$ waarmee we H_0 tegen H toetsen wordt gedefinieerd door:

$$\underline{S} = \sum_{i<i'} \underline{S}_{i,i'}$$

waarin $\underline{S}_{i,i'}$ de rangcorrelatie-grootheid van Kendall is voor de rangschikkingen i en i' ($1 \leq i < i' \leq m$) (Zie memorandum S47 (M13)).

Voor de berekening van $\underline{S}$ kunnen we ook als volgt te werk gaan:
Laten de rangnummers $x_{i1}, \ldots, x_{in}$ bestaan uit k_i groepen van $t_h^{(i)}$ gelijke rangnummers ($h = 1, 2, \ldots, k_i$; $i = 1, 2 \ldots, m$)
We definieren nu:

$$T_2^{(i)} = 1 - \frac{\sum_1^{k_i} t_h^{(i)} (t_h^{(i)} - 1)}{n(n-1)} \quad \text{voor} \quad n \geq 2$$

$$T_2^{(i)} = 0 \quad \text{voor} \quad n < 2$$

$$T_3^{(i)} = 1 - \frac{\sum_1^{k_i} t_h^{(i)} (t_h^{(i)} - 1)(t_h^{(i)} - 2)}{n(n-1)(n-2)} \quad \text{voor} \quad n \geq 3$$

$$T_3^{(i)} = 0 \quad \text{voor} \quad n < 3$$

We stellen nu een schema op van de volgende vorm:

Rangschikking	Paren Objecten				
	$O_1 O_2$	$O_1 O_3$		$O_{n-2} O_n$	$O_{n-1} O_n$
1 2 ⋮ m					
Totaal	$\underline{S}_1$	$\underline{S}_2$		$\underline{S}_{\binom{n}{2}-2}$	$\underline{S}_{\binom{n}{2}}$

2) Voor m = 2 zie memorandum S 47 (M 13).

In dit schema schrijven we in het vakje behorende bij rangschikking i en het paar objecten $O_j\, O_{j'}$ $(j < j')$:

$+1$ als $\underline{r}_{ij} < \underline{r}_{ij'}$

0 als $\underline{r}_{ij} = \underline{r}_{ij'}$

-1 als $\underline{r}_{ij} > \underline{r}_{ij'}$

Vervolgens bepalen we de kolomtotalen $\underline{s}_e \left(e = 1, 2, \ldots, \binom{n}{2} \right)$
Nu is de toetsingsgrootheid

$$\underline{S} = \frac{1}{2} \sum_{1}^{\binom{n}{2}} \underline{s}_e^2 - \frac{1}{4} n(n-1) \sum_{1}^{m} T_2^{(i)}$$

Een andere methode ter berekening van $\underline{S}$ is de volgende:
We stellen een preferentieschema op van de vorm

	O_1	O_2	-----	O_j	-----	$O_{j'}$	-------	O_n
O_1	X	$\underline{c}_{12}$		$\underline{c}_{1j}$		$\underline{c}_{1j'}$		$\underline{c}_{1n}$
O_2	$\underline{c}_{21}$	X		$\underline{c}_{2j}$		$\underline{c}_{2j'}$		$\underline{c}_{2n}$
⋮								
O_j	$\underline{c}_{j1}$	$\underline{c}_{j2}$		X		$\underline{c}_{jj'}$		$\underline{c}_{jn}$
⋮								
$O_{j'}$	$\underline{c}_{j'1}$	$\underline{c}_{j'2}$		$\underline{c}_{j'j}$		X		$\underline{c}_{j'n}$
⋮								
O_n	$\underline{c}_{n1}$	$\underline{c}_{n2}$		$\underline{c}_{nj}$		$\underline{c}_{nj'}$		X

Hierin is $\underline{c}_{jj'}$ $(j \neq j')$ het aantal malen dat O_j geprefereerd wordt boven $O_{j'}$, d.w.z. het aantal waarden van i waarvoor $\underline{r}_{ij} > \underline{r}_{ij'}$ vermeerderd met de helft van het aantal waarden van i waarvoor $\underline{r}_{ij} = \underline{r}_{ij'}$. We hebben dan:

$$\underline{S} = \sum_{j<j'} \underline{c}_{jj'}^2 - \frac{1}{4} n(n-1) \left\{ m^2 + \sum_{1}^{m} T_2^{(i)} \right\}.$$

Onder de nulhypothese is de verwachting van $\underline{S}$:

$$\mathcal{E}(\underline{S} \mid H_0) = 0$$

en de variantie

$$\mathrm{Var}(\underline{S} \mid H_0) = \sum_{i<i'} \left\{ \frac{1}{9} n(n-1)(n-2)\, T_3^{(i)} \cdot T_3^{(i')} + \frac{1}{2} n(n-1)\, T_2^{(i)} \cdot T_2^{(i')} \right\}$$

Indien $\underline{r}_{i1}, \underline{r}_{i2}, \ldots, \underline{r}_{in}$ verschillend zijn $(i = 1, 2, \ldots, m)$ dan gaat dit over in:

$$\mathrm{Var}(\underline{S} \mid H_0) = \frac{1}{36} m(m-1)\, n(n-1)(2n+5)$$

Bij overeenstemming tussen de waarnemers zal $\underline{S}$ over het algemeen grotere waarden aannemen dan onder de nulhypothese. De kritieke zone krijgt dan dus de vorm $S \geqq S_\alpha$, waarin S_α de kleinste waarde is, die voldoet aan:

$$P\left[\underline{S} \geqq S_\alpha \mid H_o\right] \leqq \alpha .$$

Men vindt de volgende exacte waarden in het geval er geen gelijke rangnummers in één rij voorkomen:

$n=3$ $m=$	$S_{0,05}$	$S_{0,01}$
3	9	-
4	12	18
5	14	22
6	19	27

Voor grote waarden van n is $\underline{S}$ onder H_o bij benadering normaal verdeeld (zie [4]), en dus geldt dan:

$$S_\alpha = \xi_\alpha \sqrt{Var(\underline{S}/H_o)}$$

waarin ξ_α gedefinieerd wordt door

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

zodat ξ_α in een tabel van de normale verdeling kan worden opgezocht.

Voor grote waarden van m heeft $\underline{S}$ onder H_o bij benadering een verdeling die gevonden kan worden door samenstelling van twee χ^2-verdelingen. De grootheid $\underline{S}$ kan namelijk geschreven worden in de volgende vorm:

$$\underline{S} = \tfrac{1}{6}\left\{3\sum_1^m T_2^{(i)} + (n-2)\sum_1^m T_3^{(i)}\right\}\underline{X}_1 + \tfrac{1}{6}\left\{3\sum_1^m T_2^{(i)} - 2\sum_1^m T_3^{(i)}\right\}\underline{X}_2 - \tfrac{1}{4}n(n-1)\sum_1^m T_2^{(i)} .$$

waarin $\underline{X}_1$ en $\underline{X}_2$ voor grote m bij benadering onafhankelijk zijn en χ^2-verdelingen hebben met respectievelijk $n-1$ en $\frac{1}{2}(n-1)(n-2)$ vrijheidsgraden. Als er geen gelijke rangnummers in één rij voorkomen, wordt deze uitdrukking van $\underline{S}$ in $\underline{X}_1$ en $\underline{X}_2$:

$$\underline{S} = \tfrac{m}{6}\left\{(n+1)\,\underline{X}_1 + \underline{X}_2\right\} - \tfrac{1}{4}n(n-1)\,m$$

Voor $n = 3$ volgt hieruit b.v. dat voor voldoende grote waarden van m de overschrijdingskans $P[S \geq S]$ in het geval dat er geen gelijke rangnummers voorkomen als volgt kan worden benaderd:

$$P[S \geq S] \simeq P[\chi_1^2 \geq X] + \frac{2}{\sqrt{3}} e^{-\frac{X}{8}} \cdot P[\chi_1^2 \leq \frac{3}{4}X]$$

$$\left(\simeq \frac{2}{\sqrt{3}} e^{-\frac{X}{8}} \text{ voor } X \geq 11 \text{ in drie decimalen nauwkeurig} \right)$$

waarin $X = \frac{6}{m} S + 9$

en χ_1^2 een χ^2-variabele met 1 vrijheidsgraad is.

Literatuur:

[1] M.G. Kendall, Rank Correlation Methods, London, 1952.

[2] W.J. Dixon and F.J. Massey Jr, Introduction to statistical analysis, Mc Graw Hill Book Comp., 1951.

[3] T.J. Terpstra, A generalization of Kendall's rank correlation statistic. I, Proc. Kon. Ned. Akad. v. Wet. A 58, 690-696; Indagationes Mathematicae 17, 690-696.

[4] T.J. Terpstra, A generalization of Kendall's rank correlation statistic II, Proc. Kon. Ned. Akad. v. Wet. A 59, 59-66; Indagationes Mathematicae 18, 59-66.

[5] Mathematisch Centrum, Methode der m rangschikkingen, rapport S 47 (M 14), een generalisatie van Kendall's rangcorrelatietoets, rapport S 190 (M 49). Een parametervrije toets tegen verloop van groepen waarnemingen, rapport S 168 (M 61).