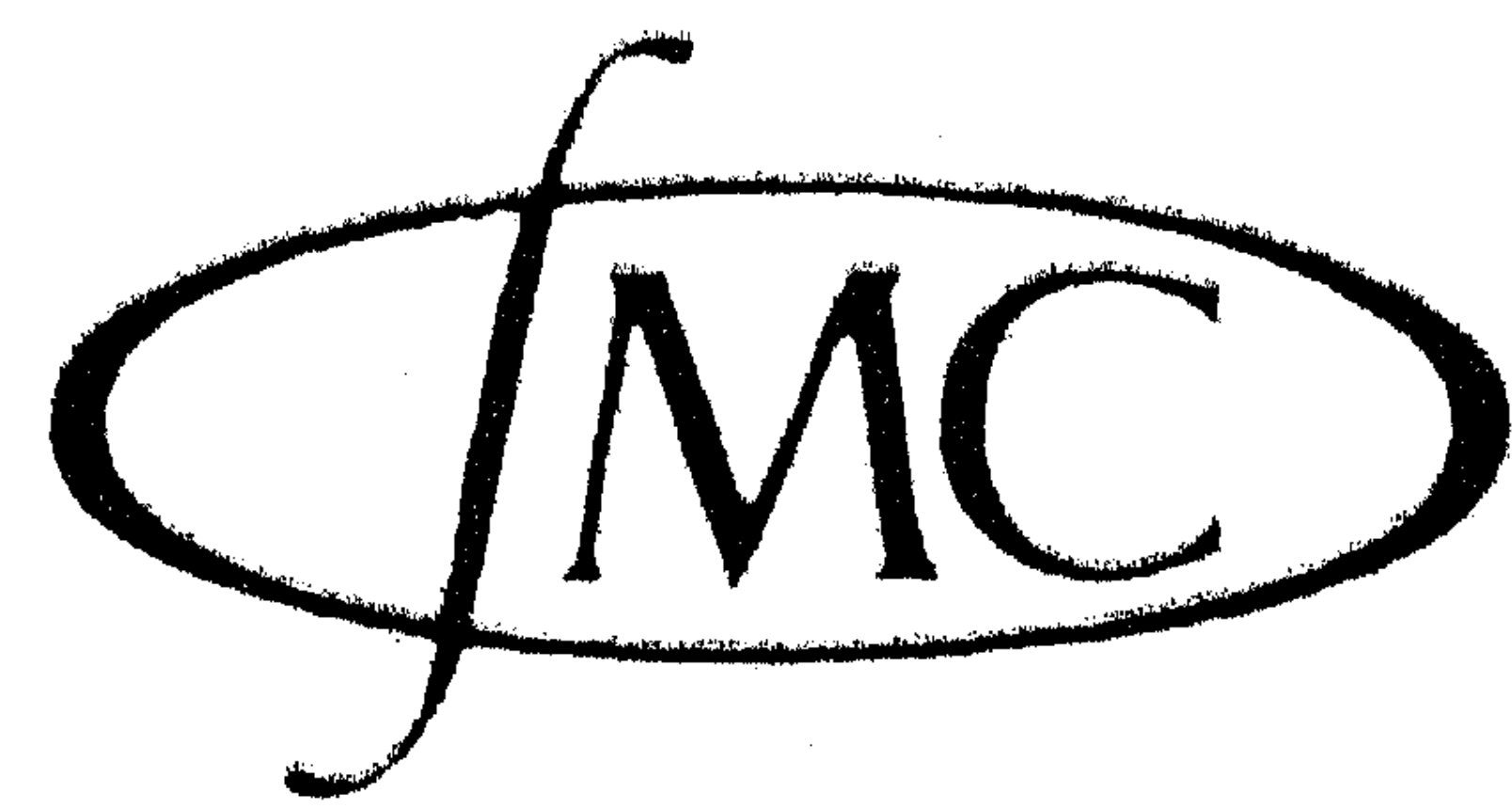


STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 145 (M 49)

Een generalisatie van Kendall's rangcorrelatie-toets

T.J. Terpstra



Een generalisatie van Kendall's rangcorrelatie-toets ¹⁾.
 door T.J.Terpstra.

Gegeven zijn 2 waarnemers (dit kunnen ook meetinstrumenten zijn, of i.d.), die onafhankelijk steekproeven nemen van elk der k stochastische grootheden $x_1, \dots, x_k$ resp. $y_1, \dots, y_k$. De steekproeven voor de eerste waarnemer zijn :

$x_{1,1}, x_{1,2}, \dots, x_{1,n_1}$	van de stochastische grootheid	$x_1,$
$x_{2,1}, x_{2,2}, \dots, x_{2,n_2}$	" " " "	$x_2,$
.....		
$x_{k,1}, x_{k,2}, \dots, x_{k,n_k}$	" " " "	$x_k,$

De uitgebreidheden van de steekproeven voor de eerste waarnemer zijn dus $n_1, n_2, \dots, n_k$; de eerste index van de waarneming $x_{i,j}$ geeft aan, uit welke steekproef deze waarneming afkomstig is, terwijl de tweede index het nummer der waarneming binnen die steekproef aangeeft.

De steekproeven voor de tweede waarnemer zijn:

$y_{1,1}, y_{1,2}, \dots, y_{1,m_1}$	van de stochastische grootheid	$y_1,$
$y_{2,1}, y_{2,2}, \dots, y_{2,m_2}$	" " " "	$y_2,$
.....		
$y_{k,1}, y_{k,2}, \dots, y_{k,m_k}$	" " " "	$y_k,$

De hypothese H_0 houdt in, dat voor elk van beide waarnemers de k steekproeven uit eenzelfde universum (collectie) komen.

Deze hypothese H_0 wensen wij te toetsen tegen hypothesen, die inhouden dat de grootheden $x_1, \dots, x_k$ resp. $y_1, \dots, y_k$ niet dezelfde verdeling hebben, doch verdelingen, die ten opzichte van elkaar verschoven zijn en wel zo, dat de gemiddelden van de y -grootheden gecorreleerd zijn met de gemiddelden van de x -grootheden. Indien men tegelijk tegen positieve en negatieve correlatie wil toetsen, past men de tweezijdige toets toe; indien men alleen tegen positieve resp. negatieve correlatie wil toetsen past men de rechts- resp. linkseenzijdige toets toe.

 1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.

De statistische grootheid $\underline{S}$, waarmee de hypothese H_0 tegen de alternatieve hypothese H , wordt getoetst, wordt op de volgende wijze gedefiniëerd.

Voor de eerste waarnemer wordt elk paar van steekproeven

$x_{k,1}, \dots, x_{k,n_k}$ en $x_{j,1}, \dots, x_{j,n_j}$ ($1 \leq k < j \leq k$) vergeleken, door de $n_k + n_j$ waarnemingen uit de twee steekproeven naar opklimmende grootte te rangschikken en van een rangnummer te voorzien. Indien alle waarnemingen verschillend zijn, zijn dit de rangnummers 1 tot en met $n_k + n_j$. Indien t waarnemingen aan elkaar gelijk zijn, wordt aan elk der t waarnemingen eenzelfde rangnummer toegekend. Dit rangnummer is dan gelijk aan het gemiddelde der rangnummers, welke de waarnemingen gekregen zouden hebben, indien ze alle verschillend waren.

Voor de k^{de} steekproef wordt vervolgens de som van de rangnummers bepaald, welk aantal we $R_{k,j}^{(1)}$ noemen. Uit deze grootheid vormen we de grootheid $\tilde{U}_{k,j}^{(1)}$ volgens $\tilde{U}_{k,j}^{(1)} = R_{k,j}^{(1)} - \frac{1}{2} n_k (n_k + n_j + 1)$. Op dezelfde wijze vormen we voor de tweede waarnemer uit de tweede reeks van steekproeven de grootheid $\tilde{U}_{k,j}^{(2)}$ ($1 \leq k < j \leq k$). De toetsingsgrootheid $\underline{S}$ wordt nu gedefiniëerd door

$$\underline{S} = 4 \sum_{k < j} \tilde{U}_{k,j}^{(1)} \cdot \tilde{U}_{k,j}^{(2)}$$

Indien de hypothese H_0 juist is, bezit de toetsingsgrootheid $\underline{S}$ voor grote waarden van k bij benadering een normale verdeling met gemiddelde en variantie gegeven door

$$\begin{aligned} E(\underline{S} | H_0) &= 0, \\ \text{Var}(\underline{S} | H_0) &= \frac{1}{9} \left\{ \sum_{k < j} n_k n_j (n_k + n_j + 1) \cdot m_k m_j (m_k + m_j + 1) + \right. \\ &\quad \left. + 6 \sum_{k < i < j} n_k n_i n_j \cdot m_k m_i m_j \right\}. \end{aligned}$$

Bij aanwezigheid van correlatie tussen de gemiddelden van de x en y -grootheden neemt $\underline{S}$ in absolute zin over het algemeen grotere waarden aan dan onder de hypothese H_0 en wel grotere positieve waarden bij positieve en grotere negatieve waarden bij negatieve correlatie. De kritieke zônes worden dus van de volgende vorm:

$$\begin{aligned} \text{bij de tweezijdige toets:} & \quad |S| \geq S_{\frac{1}{2}\alpha}, \\ \text{bij de rechts-eenzijdige toets:} & \quad S \geq S_{\alpha}, \\ \text{bij de links-eenzijdige toets:} & \quad S \leq -S_{\alpha}, \end{aligned}$$

waarin S_{α} de kleinste waarde is die $\underline{S}$ kan aannemen, zodanig dat $P[\underline{S} \geq S_{\alpha} | H_0] \leq \alpha$

Voor grote waarden van k geldt bij benadering

$$S_{\alpha} = \xi_{\alpha} \sqrt{\text{Var}(\underline{S} | H_0)},$$

waarbij ξ_{α} wordt gedefiniëerd door

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

en gemakkelijk kan worden gevonden uit een tabel van de normale verdelingsfunctie.

De overschrijdingskans k^* van de uit de steekproeven bepaalde waarde van S wordt gedefinieerd door $k^* = P[\underline{S} \geq S | H_0]$ en kan voor grote k bij benadering worden bepaald met behulp van een tabel van de normale verdelingsfunctie volgens

$$k^* = \frac{1}{\sqrt{2\pi}} \int_u^{\infty} e^{-\frac{1}{2}t^2} dt$$

met

$$u = \frac{S}{\sqrt{\text{Var}(\underline{S} | H_0)}}.$$

Litteratuur:

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.