STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 145 (M 49)

Een generalisatie van Kendall's rangcorrelatie-toets

T.J. Terpstra

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.
S 145 (M 49)

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 $\underline{x}_1, \ldots, \underline{x}_k$ resp. $\underline{y}_1, \ldots, \underline{y}_k$.
De steekproeven voor de eerste waarnemer zijn :

$\underline{x}_{1,1}, \underline{x}_{1,2}, \ldots, \underline{x}_{1,n_1}$ van de stochastische grootheid $\underline{x}_1$,
$\underline{x}_{2,1}, \underline{x}_{2,2}, \ldots, \underline{x}_{2,n_2}$ " " " " $\underline{x}_2$,
. . . . . .
$\underline{x}_{k,1}, \underline{x}_{k,2}, \ldots, \underline{x}_{k,n_k}$ " " " " $\underline{x}_k$,

De uitgebreidheden van de steekproeven voor de eerste waarnemer zijn dus $n_1, n_2, \ldots, n_k$; de eerste index van de waarneming $\underline{x}_{i,j}$ 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:

$\underline{y}_{1,1}, \underline{y}_{1,2}, \ldots, \underline{y}_{1,m_1}$ van de stochastische grootheid $\underline{y}_1$,
$\underline{y}_{2,1}, \underline{y}_{2,2}, \ldots, \underline{y}_{2,m_2}$ " " " " $\underline{y}_2$,
. . . . . .
$\underline{y}_{k,1}, \underline{y}_{k,2}, \ldots, \underline{y}_{k,m_k}$ " " " " $\underline{y}_k$.

De hypothese H_o houdt in, dat voor elk van beide waarnemers de k steekproeven uit eenzelfde universum (collectie) komen.
Deze hypothese H_o wensen wij te toetsen tegen hypothesen, die inhouden dat de grootheden $\underline{x}_1, \ldots, \underline{x}_k$ resp. $\underline{y}_1, \ldots, \underline{y}_k$ niet dezelfde verdeling hebben, doch verdelingen, die ten opzichte van elkaar verschoven zijn en wel zo, dat de gemiddelden van de $\underline{y}$-grootheden gecorreleerd zijn met de gemiddelden van de $\underline{x}$-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_o tegen de alternatieve hypothese H_1 wordt getoetst, wordt op de volgende wijze gedefiniëerd.
Voor de eerste waarnemer wordt elk paar van steekproeven

$$\underline{x}_{h,1}, \ldots\ldots, \underline{x}_{h,n_h} \quad en \quad \underline{x}_{j,1}, \ldots\ldots, \underline{x}_{j,n_j} \qquad (1 \leqq h < j \leqq k)$$

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

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

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

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

$$Var(\underline{S} \mid H_o) = \frac{1}{9} \left\{ \sum_{h<j} n_h n_j (n_h + n_j + 1) \cdot m_h m_j (m_h + m_j + 1) + 6 \sum_{h<i<j} n_h n_i n_j \cdot m_h m_i m_j \right\} .$$

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

bij de tweezijdige toets: $|S| \geqq S_{\frac{1}{2}\alpha}$,
bij de rechts-eenzijdige toets: $S \geqq S_\alpha$,
bij de links-eenzijdige toets: $S \leqq -S_\alpha$,

waarin S_α de kleinste waarde is die $\underline{S}$ kan aannemen, zodanig dat $P[\underline{S} \geqq S_\alpha \mid H_o] \leqq \alpha$
Voor grote waarden van k geldt bij benadering

$$S_\alpha = \Xi_\alpha \sqrt{Var(\underline{S} \mid H_o)} ,$$

waarbij Ξ_α wordt gedefinieerd door

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

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 $k^* = P[\underline{S} \geqq S | H_o]$ 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

$$u = \frac{S}{\sqrt{Var\,(\underline{S}|H_o)}} \ .$$

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.