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

Statistische Afdeling

Memorandum S 190(M49)

49a

Een generalisatie van Kendall's rangcorrelatie-toets *)

Probleemstelling

Gegeven zijn twee waarnemers, die onafhankelijk steekproeven nemen van elk der k stochastische grootheden $\underline{x}_1$, $\underline{x}_2,\ldots,\underline{x}_k$ resp. $\underline{y}_1,\underline{y}_2,\ldots,\underline{y}_k$.

Het j^e element uit de i^e steekproef van de eerste resp. tweede waarnemer duiden wij aan met $\underline{x}_{ij}$ $(1\leqq i\leqq k;\ 1\leqq j\leqq n_i)$ resp. $\underline{y}_{ij}$ $(1\leqq i\leqq k;\ 1\leqq j\leqq m_i)$.

De hypothese H_o houdt in, dat voor elk van beide waarnemers de k steekproeven uit eenzelfde universum (dat voor de twee waarnemers verschillend kan zijn) afkomstig zijn. Deze hypothese wensen wij te toetsen tegen alternatieve 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 $\underline{x}_1,\ldots,\underline{x}_k$ en de gemiddelden van $\underline{y}_1,\ldots,\underline{y}_k$ met elkaar gecorreleerd zijn. Het verdient de voorkeur bij het opzetten van het experiment $n_i = \frac{\sum_i n_i}{k}$ en $m_i = \frac{\sum_i m_i}{k}$ $(i = 1,2,\ldots,k)$ te nemen. Dit komt het onderscheidingsvermogen van de toets ten goede en vereenvoudigt bovendien het rekenwerk.

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 de rechts- resp. linkseenzijdige toets.

Toetsingsgrootheid

De toetsingsgrootheid $\underline{S}$ wordt nu op de volgende wijze gevormd:

Wij vergelijken bij de eerste waarnemer de i^e en de j^e steekproef $(1\leqq i< j\leqq k)$ door de grootheid $\underline{\tilde{W}}_{i,j}^{(1)}$ te berekenen, welke gedefinieerd wordt als het aantal paren (λ,ν) $(1\leqq\lambda\leqq n_i;\ 1\leqq\nu\leqq n_j)$ waarvoor $\underline{x}_{i\lambda} > \underline{x}_{j\nu}$ verminderd met het aantal paren (λ,ν) $(1\leqq\lambda\leqq n_i;\ 1\leqq\nu\leqq n_j)$ waarvoor $\underline{x}_{i\lambda} < \underline{x}_{j\nu}$.
$\underline{\tilde{W}}_{i,j}^{(1)}$ is dus de gereduceerde Wilcoxon-grootheid voor de i^e en de j^e steekproef van de eerste waarnemer.

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

Indien de steekproeven groot zijn, is het sneller om $\underline{\tilde{W}}^{(1)}_{i,j}$ als volgt te berekenen: Wij kennen aan de n_i+n_j waarnemingen uit de i^e en de j^e steekproef van de eerste waarnemer de rangnummers $1,2,\ldots,(n_i+n_j)$ toe naar opklimmende grootte van de waarnemingen.

Indien een aantal waarnemingen aan elkaar gelijk zijn, wordt aan elk van deze waarnemingen hetzelfde rangnummer toegekend, en wel het gemiddelde van de rangnummers, die deze waarnemingen gekregen zouden hebben, als zij alle verschillend waren.

Noemen wij nu de som der rangnummers van de waarnemingen uit de i^e bij vergelijking met de j^e steekproef $\underline{R}^{(1)}_{ij}$, dan is

$$\underline{\tilde{W}}^{(1)}_{ij} = 2.\underline{R}^{(1)}_{ij} - n_i(n_i+n_j+1)$$

Evenzo gaan wij te werk bij de steekproeven van de tweede waarnemer, waar wij krijgen $\underline{\tilde{W}}^{(2)}_{ij}$ $(1 \leqq i \leqq j \leqq k)$.
De toetsingsgrootheid $\underline{S}$ wordt nu gedefinieerd door

$$\underline{S} = \sum_{i<j} \sum_{j} \frac{\underline{\tilde{W}}^{(1)}_{ij} \cdot \underline{\tilde{W}}^{(2)}_{ij}}{n_i \cdot n_j \cdot m_i \cdot m_j}$$

Verwachting en spreiding van $\underline{S}$ onder H_o

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

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

Laten nu de k steekproeven van de eerste waarnemer uit r_1 groepen van $t_h^{(1)}$ $(h = 1,2,\ldots,r_1)$ gelijke waarnemingen en de k steekproeven van de tweede waarnemer uit r_2 groepen van $t_h^{(2)}$ $(h = 1,\ldots,r_2)$ gelijke waarnemingen bestaan.*)

Wij definiëren nu de volgende grootheden:

$$T_2^{(1)} = 1 - \frac{\sum_1^{r_1} t_h^{(1)}(t_h^{(1)}-1)}{n(n-1)}, \text{ voor } n = \sum_i n_i \geqq 2$$

$$T_2^{(1)} = 0, \text{ voor } n = \sum_i n_i < 2$$

$$T_3^{(1)} = 1 - \frac{\sum_1^{r_1} t_h^{(1)}(t_h^{(1)}-1)(t_h^{(2)}-2)}{n(n-1)(n-2)}, \text{ voor } n = \sum_i n_i \geqq 3$$

---

*) De volgorde van de getallen $t_h^{(1)}$ bij rangschikking naar opklimmende grootte der waarnemingen uit de groepen heeft invloed op de exacte verdeling onder H_o van $\underline{S}$, doch niet op de verwachting en de variantie van $\underline{S}$.

$T_3^{(1)} = 0$, voor $n = \sum_i n_i < 3$

$T_{2,3}^{(1)} = \frac{1}{2}T_2^{(1)} - \frac{1}{3}T_3^{(1)}$

voor de eerste waarnemer, en analoog $T_2^{(2)}$, $T_3^{(2)}$, en $T_{2,3}^{(2)}$ voor de tweede waarnemer.

De variantie van $\underline{S}$ onder H_o wordt nu gegeven door:

$$\mathrm{Var}(\underline{S}|H_o) = \left\{2T_{2,3}^{(1)}.T_{2,3}^{(2)} \sum_i (n_i.m_i)^{-1} + \frac{2}{3} T_3^{(1)}.T_{2,3}^{(2)} \sum_i (m_i)^{-1} + \frac{2}{3} T_{2,3}^{(1)}.T_3^{(2)} \sum_i (n_i)^{-1} + \frac{1}{9}k(k-2)T_3^{(1)}.T_3^{(2)}\right\}. \sum_i (n_i m_i)^{-1} +$$

$$\frac{1}{9} T_3^{(1)}.T_3^{(2)} \sum_i (n_i)^{-1} \sum_i (m_i)^{-1} - \frac{2}{3}\sum_i \left\{T_3^{(1)}.T_{2,3}^{(2)}.n_i + T_{2,3}^{(1)}.T_3^{(2)} m_i + 3T_{2,3}^{(1)}.T_{2,3}^{(2)}\right\} (n_i m_i)^{-2}$$

Indien alle waarnemingen van de eerste waarnemer, en ook die van de tweede waarnemer verschillend zijn, gaat dit over in:

$$\mathrm{Var}(\underline{S}|H_o) = \left\{\frac{1}{18}\sum_i (n_i m_i)^{-1} + \frac{1}{9}\sum_i (n_i)^{-1} + \frac{1}{9}\sum_i (m_i)^{-1} + \frac{1}{9}k(k-2)\right\} \times$$

$$\times \sum_i (n_i m_i)^{-1} + \frac{1}{9}\sum_i (n_i)^{-1} \sum_i (m_i)^{-1} - \frac{1}{9}\sum_i (n_i + m_i + \tfrac{1}{2})(n_i m_i)^{-2}$$

voor $n = \sum_i n_i \geqq 3$ èn $m = \sum_i m_i \geqq 3$.

Als $k = 2$ en $n_1 = n_2 = 1$, en als de twee waarnemingen van de eerste waarnemer niet gelijk zijn, dan is $m_1 m_2 \underline{S} = \underline{\tilde{W}}$, de gereduceerde Wilcoxon grootheid voor de twee steekproeven van de tweede waarnemer.

Uit de boven gegeven uitdrukking voor $\mathrm{Var}(\underline{S}|H_o)$ volgt dan

$$\mathrm{Var}(\underline{\tilde{W}}|H_o) = \frac{m_1 m_2 \left\{(m_1+m_2)^3 - \sum_1^{r_2} (t_h^{(2)})^3\right\}}{3(m_1+m_2)(m_1+m_2-1)} .$$

Als de grootheden $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_k$ niet stochastisch zijn, maar de natuurlijke getallen $1,2,\ldots,k$ in een vaste volgorde zijn, dan is de verdeling van $\underline{S}$ onder H_o een andere dan in het hier behandelde geval. Zie hiervoor memorandum S 168(M61). Slechts als bovendien $m_i = \frac{\sum i m_i}{k}$ $(i = 1,2,\ldots,k)$ heeft $\underline{S}$ in beide gevallen onder de nulhypothese dezelfde verdeling. Als $n_i = m_i = 1$ $(i = 1,\ldots,k)$, dan is de toets identiek met Kendall's rangcorrelatietoets.
Wij krijgen dan

$$\mathrm{Var}(\underline{S}\ H_o) = \tfrac{1}{2}k(k-1).T_2^{(1)}T_2^{(2)} + \frac{1}{9}k(k-1)(k-2)T_3^{(1)}.T_3^{(2)}$$

of, als alle waarnemingen van de eerste waarnemer en ook die van de tweede waarnemer verschillend zijn:

$$\mathrm{Var}(\underline{S}|H_o) = \frac{1}{18}k(k-1)(2k+5)$$

Asymptotisch verdeling van $\underline{S}$ onder H_o

Voor grote waarden van k is $\underline{S}$ onder H_o bij benadering normaal verdeeld met gemiddelde 0 en variantie $\mathrm{Var}(S|H_o)$. Bij aanwezigheid van correlatie tussen de gemiddelden van de $\underline{x}$ en de $\underline{y}$ grootheden neemt $\underline{S}$ in absolute zin in het algemeen grotere waarden aan dan onder de nulhypothese, en wel grotere positieve waarden bij positieve en grotere negatieve waarden bij negatieve correlatie. De kritieke zones 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 $S \leq -S_\alpha$

waarin S_α de kleinste waarde is die voldoet aan

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

Voor grote waarden van k geldt bij benadering $S_\alpha = \xi_\alpha \sqrt{\mathrm{Var}(\underline{S}\ H_o)}$ waarin ξ_α gedefiniëerd 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 verdelingsfunctie kan worden gevonden.

De rechtseenzijdige overschrijdingskans k^* van de gevonden waarde S van $\underline{S}$ wordt gedefiniëerd door $k^* = P\left[S \geqq S|H_o\right]$ en voor grote waarden van k geldt bij benadering

$$k^* = \frac{1}{\sqrt{2\pi}} \int_{u}^{\infty} e^{-\frac{1}{2}t^2} . \, dt$$

waarin $u = \dfrac{S}{\mathrm{Var}(\underline{S}|H_o)}$

Op overeenkomstige wijze worden de links-eenzijdige en de tweezijdige overschrijdingskansen gedefinieerd en benaderd.

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. Ak.v.Wet. A 59, 59-66; Indagationes Mathematicae 18, 59-66.

[5] Mathematisch Centrum, Een parametervrije toets tegen verloop voor groepen van waarnemingen, Rapport S 168(M 61).

[6] Mathematisch Centrum, De toets van Wilcoxon, Rapport S 47 (M 7)