STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 47 (M13)

Rangcorrelatie

1952

MATHEMATISCH CENTRUM
2de Boerhaavestr. 49,
A m s t e r d a m -O.
Statistische Afdeling
S47(M13)

Rangcorrelatie[1)]

1. Beschrijving van de methode.

De door M.G. Kendall ontwikkelde methode der rangcorrelatie is toepasbaar op de volgende situatie:

De stochastische grootheden x en y bezitten een simultane verdeling. Over deze verdeling zelf behoeft niets ondersteld te worden.

(x_i,y_i) $(i = 1,\ldots,n)$, zijn onafhankelijke waarnemingsparen van deze stochastische grootheden

Voorbeeld:

i =	1	2	3	4	5	6
x_i	0,11	0,12	0,10	0,11	0,15	0,13
y_i	3,4	3,0	3,2	3,5	3,5	3,5

Wij zeggen dat de waarnemingsparen (x_i,y_i) en (x_j,y_j) positief gecorreleerd zijn, als de volgorde van x_i en x_j hetzelfde is als die van y_i en y_j (bv. $x_i < x_j$ en $y_i < y_j$); zij zijn negatief gecorreleerd als de volgorde van x_i en x_j tegengesteld is aan de volgorde van y_i en y_j (bv. $x_i > x_j$ en $y_i < y_j$) en zij zijn niet gecorreleerd als $x_i = x_j$ of $y_i = y_j$.

In tabel 1 hebben wij van alle tweetallen (x_i,y_i) en (x_j,y_j) uit ons voorbeeld nagegaan of zij positief, negatief dan wel niet gecorreleerd zijn. Een positieve correlatie is aangeduid met +1, een negatieve met -1, terwijl het ontbreken van correlatie wordt aangegeven door een 0.

De toetsingsgrootheid van de methode van rangcorrelatie is nu het aantal positief gecorreleerde tweetallen verminderd met het aantal negatief gecorreleerde, of wel de som van de getallen, die in tabel 1 in de kolom "correlatie" voorkomen.

De verdeling van S voor het geval dat x en y onafhankelijk zijn is bekend (zie § 2). De hypothese dat x en y onafhankelijk

---

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

Tabel 1

Berekening van S voor het voorbeeld

i	j	Correlatie
1	2	-1
1	3	+1
1	4	0
1	5	+1
1	6	+1
2	3	-1
2	4	-1
2	5	+1
2	6	+1
3	4	+1
3	5	+1
3	6	+1
4	5	0
4	6	0
5	6	0
		S = +5

zijn, kan dus getoetst worden.

Is de hypothese van onafhankelijkheid niet vervuld, dan is de waarschijnlijkheid van grote positieve of grote negatieve waarden van S groter, dan wanneer dit wel het geval is. De kritieke zône is daarom van de vorm $|S| \geqq S_o$, en bij éénzijdige toetsing van de vorm $S \geqq S'_o$ (rechtszijdige toetsing) of $S \leqq S''_o$ (linkszijdige toetsing).

2. Verdeling van S als x en y onafhankelijk zijn.

Als er noch bij de x_i noch bij de y_j gelijke waarden voorkomen kunnen wij gebruik te maken van exacte tabellen, die voorkomen in [1] pg 141 (n = 4 t/m,10) en in [2] (tables I and II, n = 4 t/m 40). Bovendien vindt men in [2] table III de kleinste waarden van S, waarvan de overschrijdingskansen onder de hypothese van onafhankelijkheid hoogstens gelijk zijn aan α voor α = 0,005; 0,01; 0,025; 0,05 en 0,10 en n = 4,5,6,...,40.

Als er bij de x_i óf bij de y_i, doch niet bij beide tweetallen of drietallen gelijken voorkomen, kan men voor $n \leqq 10$ gebruik maken van de tabel van Sillitto [4].

Voor grote waarden van n is de verdeling van $\frac{S}{\sigma_S}$ (waarin σ_S de spreiding van S is, die uit een hieronder op te geven formule berekend kan worden) bij benadering normaal met gemiddelde 0 en spreiding 1. Hiervan kunnen we gebruik maken om de hypothese van onafhankelijkheid te toetsen in de gevallen waar de exacte verdeling niet getabelleerd is. Dit geschiedt dan, door in een tabel van de normale verdeling de

overschrijdingskans op te zoeken, die behoort bij de gevonden waarde van $\frac{S}{\sigma_S}$.

Om σ_S te berekenen, nemen wij in de rij der waarnemingen x_i de gelijke waarnemingen in groepen bij elkaar. De aantallen waarnemingen in die groepen duiden wij aan met t_h, waarin $h = 1,2,\ldots,k_1$. Evenzo doet men in de rij der waarnemingen y_j, waar we de overeenkomstige aantallen aanduiden met u_l waarin $l = 1,2,\ldots,k_2$. σ_S kan dan gevonden worden uit de volgende formule:

$$(1)\quad \sigma^2_S = \frac{1}{18}\Big\{ n(n-1)(2n+5) - \sum_{h=1}^{k_1} t_h(t_h-1)(2t_h+5) - \sum_{l=1}^{k_2} u_l(u_l-1)(2u_l+5)\Big\} + \frac{1}{9n(n-1)(n-2)}\sum_{h=1}^{k_1} t_h(t_h-1)(t_h-2)\sum_{l=1}^{k_2} u_l(u_l-1)(u_l-2) + \frac{1}{2n(n-1)}\sum_{h=1}^{k_1} t_h(t_h-1)\sum_{l=1}^{k_2} u_l(u_l-1).$$

In ons voorbeeld van § 1 komt in de rij x_i één tweetal gelijken (dus $k_1=1$ en $t_1=2$) en in de rij y_j één drietal gelijken ($k_2= 1$, $u_1 = 3$) voor. Dus geldt:

$t_1(t_1-1)(2t_1+5) = 2.1.9 = 18$
$u_1(u_1-1)(2u_1+5) = 3.2.11 = 66$
$t_1(t_1-1)(t_1-2) = 0$
$t_1(t_1-1) = 2.1 = 2$
$u_1(u_1-1) = 3.2 = 6$
$n(n-1)(2n+5) = 6.5.17 = 510$
$n(n-1) = 6.5 = 30$

zodat:

$$\sigma^2_S = \frac{1}{18}\{510-18-66\} + \frac{1}{60} \times 2 \times 6 = 23,87$$

en $\sigma_S = 4,89$ is.

Als alle t_h en alle u_l gelijk zijn aan 1 en er dus in geen van beide rijen gelijken voorkomen, gaat formule (2) over in:

$$(2)\quad \sigma_S = \sqrt{\frac{1}{18}\, n(n-1)(2n+5)}$$

Een tabel van deze functie voor n = 40,41,...,100 vindt men in [2] (table IV).

3. Rangcorrelatiecoëfficiënt τ

Als maat voor de correlatie in de rij van waarnemingsparen $(x_1,y_1),\ldots,(x_n,y_n)$ heeft Kendall de coëfficiënt τ gedefinieerd, die + 1 is als de volgorden der waarnemingen in beide rijen $x_1,\ldots,x_n$ en $y_1,\ldots,y_n$ volledig overeenstemmen en -1 is, als deze volgorden volkomen tegengesteld zijn. De definitie van τ is:

$$(3)\quad \tau = \frac{2S}{\left\{n(n-1)-\sum_{h=1}^{k_1} t_h(t_h-1)\right\}^{\frac{1}{2}}\left\{n(n-1)-\sum_{l=1}^{k_2} u_l(u_l-1)\right\}^{\frac{1}{2}}}.$$

Als er in geen van beide rijen gelijke waarnemingen voorkomen wordt deze formule:

$$(4)\quad \tau = \frac{2S}{n(n-1)}.$$

Literatuur:

[1] M.G. Kendall. Rank correlation Methods London 1948, Hoofdstuk 1.

[2] L. Kaarsemaker en A. van Wijngaarden. Tables for use in rank correlation . (1952) Report R 73 of the Computation Department of the Mathematical Centre.

[3] J. Hemelrijk. Kendall's rangcorrelatie-coëfficiënt τ . Hoofdstuk I der cursus "Parametervrije Methoden" Rapport S 59 (1951) Mathematisch Centrum, blz. 1-17.

[4] G.P. Sillitto. "The Distribution of Kendall's coefficient of rankcorrelation in rankings containing ties. Biometrica 34 (1947) p. 36-40.