STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 190 (M 130

De rangcorrelatietoets van KENDALL

Mei 1956

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

Mei 1956

Statistische Afdeling
Rapport S 190 (M 13)

De rangcorrelatietoets van KENDALL[1)]

1. Doel en toepasbaarheid

De door M.G. KENDALL ontwikkelde rangcorrelatietoets is toepasbaar op de volgende situatie:

De stochastische grootheden $\underline{x}$[2)] en $\underline{y}$ bezitten een simultane verdeling en (x_i, y_i) $(i = 1, 2, \ldots, n)$ zijn n onderling onafhankelijke waarnemingsparen van deze stochastische grootheden.

In dit memorandum wordt een methode beschreven om op grond van dit waarnemingsmateriaal de hypothese H_0 te toetsen dat $\underline{x}$ en $\underline{y}$ onderling onafhankelijk verdeeld zijn. Over de vorm van de verdeling van $\underline{x}$ en $\underline{y}$ behoeft hierbij niets ondersteld te worden.

2. De toetsingsgrootheid

KENDALL's toetsingsgrootheid $\underline{S}$ wordt als volgt gedefiniëerd. De waarnemingsparen (x_i, y_i) en (x_j, y_j) leveren ieder tot S een bijdrage

$$\begin{array}{rll} +1 & \text{als} & (x_i - x_j)(y_i - y_j) > 0, \\ 0 & \text{als} & (x_i - x_j)(y_i - y_j) = 0, \\ -1 & \text{als} & (x_i - x_j)(y_i - y_j) < 0. \end{array}$$

S is nu de algebraïsche som van de bijdragen van alle paren (x_i, y_i) en (x_j, y_j), waarvoor $i < j$ is.
Wij zullen de berekening van de toetsingsgrootheid $\underline{S}$ aan de hand van twee voorbeelden nader toelichten.

Voorbeeld 1:

Stel wij hebben de volgende zes waarnemingsparen (x_i, y_i)

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

De grootheid S berekenen wij als volgt: Het waarnemingspaar (x_1, y_1) levert met de paren $(x_2, y_2), \ldots, (x_6, y_6)$ resp. de bijdragen -1, +1, 0, +1, +1. De totale bijdrage van het paar (x_1, y_1) tot S is dus -1+1+1+1 = +2. Op dezelfde manier vinden wij voor de bijdragen

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.
2) Stochastische grootheden worden onderscheiden van getallen (b.v. van de waarden die zij bij een experiment aannemen) door hun symbolen te onderstrepen.

van de paren $(x_1, y_1), \ldots, (x_6, y_6)$ tot S:

(x_1, y_1): -1-1+1+1 = 0
(x_3, y_3): +1+1+1 = 3
(x_4, y_4): 0 + 0 = 0
(x_5, y_5): 0

De grootheid S is dus 2+0+3+0+0 = +5.

Voorbeeld 2:

Stel dat x een grootheid is die slechts de drie waarden 1, 2 en 3 aanneemt en y een die slechts de vier waarden 1, 2, 3 en 4 aanneemt. Bij 30 onderling onafhankelijke waarnemingsparen van deze grootheden wordt het volgende resultaat verkregen:

		x		
		1	2	3	totaal
y	1	6	2	0	8
	2	1	4	2	7
	3	1	3	2	6
	4	1	1	7	9
totaal		9	10	11	30

In deze tabel staan aantallen waarnemingen vermeld. Er zijn b.v. 6 paren (x_i, y_i) met $x_i = y_i = 1$, 3 paren (x_i, y_i) met $x_i = 2$, $y_i = 3$ enz. Verder zien wij dat er onder de 30 x-waarnemingen 9x een 1, 10x een 2 en 11x een 3 voorkomt; onder de 30 y-waarnemingen komt 8x een 1, 7x een 2, 6x een 3 en 9x een 4 voor.

In een geval zoals dit (dus een geval met veel gelijke waarnemingen) kan men het eenvoudigst de grootheid $_2S$ berekenen, dus de som van de bijdragen van alle paren (x_i, y_i) en (x_j, y_j) zowel voor $i < j$ als voor $i > j$.

De bijdrage tot $_2S$ voor ieder der 4 waarnemingsparen (x_i, y_i) met $x_i = y_i = 2$ b.v. vinden wij als volgt: er zijn 2+7+6 paren (x_j, y_j) met $(x_i - x_j)(y_i - y_j) > 0$, nl. de 2 paren met $x_j = y_j = 3$, de 7 paren met $x_j = 3$, $y_j = 4$ en de 6 paren met $x_i = y_i = 1$. Verder zijn er 2 paren (x_j, y_j) met $(x_i - x_j)(y_i - y_j) < 0$, nl. het paar met $x_j = 1$, $y_j = 3$ en het paar met $x_j = 1$, $y_j = 4$. De bijdrage van ieder van deze 4 paren tot $_2S$ is dus 15-2 = 13; tezamen geven zij dus een bijdrage 4 × 13 = 52.

Op deze wijze kan men voor ieder der vakjes in de bovenstaande tabel de bijdrage tot $_2S$ berekenen. Dit geeft:

	1	2	3
1	114	16	0
2	11	52	4
3	0	33	22
4	-13	4	119

Dus

$$2S = 114 + 11 - 13 + 16 + 52 + 33 + 4 + 4 + 22 + 119 = 362$$

$$S = 181.$$

3. Kritieke zônes en overschrijdingskansen

Als de hypothese H_o juist is, dan is de verwachting van $\underline{S}$ gelijk aan 0. Is H_o onjuist, dan zal $\underline{S}$ in het algemeen grote positieve waarden aannemen als $\underline{x}$ en $\underline{y}$ positief gecorreleerd zijn en grote negatieve waarden als $\underline{x}$ en $\underline{y}$ negatief gecorreleerd zijn. De tweezijdige kritieke zône bestaat daarom uit grote waarden van $|S|$, de linkseenzijdige uit grote negatieve en de rechtseenzijdige uit grote positieve waarden van S.

Tabellen van overschrijdingskansen kan men vinden in [1] (pag. 141) voor $n = 4$ t/m 10 en in [2] (tabel I en II) voor $n = 4$ t/m 40. Bovendien vindt men in [2] (tabel III) de linker kritieke waarden van S voor de onbetrouwbaarheidsdrempels $\alpha = 0{,}005$; 0,01; 0,025; 0,05 en 0,10 en $n = 4$ t/m 40. Deze tabellen gelden strikt genomen alleen voor het geval dat er noch bij de x_i noch bij de y_i gelijke waarnemingen voorkomen, maar zij geven een goede benadering voor het geval van weinig gelijke waarnemingen.

Voor grote waarden van n kan men veelal gebruik maken van de benadering met de normale verdeling. Men berekent daartoe de variantie van $\underline{S}$ onder de hypothese H_o als volgt: stel dat de waarnemingen van $\underline{x}$ (resp van $\underline{y}$) uiteenvallen in g (resp. h) groepen gelijke waarnemingen en dat de aantallen waarnemingen in deze groepen $t_1, t_2, \ldots, t_g$ (resp. $u_1, u_2, \ldots, u_h$) zijn. Dan is

$$\sum_{i=1}^{g} t_i = \sum_{j=1}^{h} u_j = n.$$

In voorbeeld 1 is $g=5$, $h=4$, $t_1 = t_2 = t_3 = t_4 = 1$, $t_5 = 2$ en $u_1 = u_2 = u_3 = 1$, $u_4 = 3$; in voorbeeld 2 is: $g=3$, $h=4$, $t_1 = 9$, $t_2 = 10$, $t_3 = 11$ en $u_1 = 8$, $u_2 = 7$, $u_3 = 6$, $u_4 = 9$.

De variantie van $\underline{S}$ wordt dan gevonden uit de formule

$$\sigma^2\{\underline{S}\} = \frac{2\{n^3-\sum_{i=1}^{g} t_i^3 - 3(n^2-\sum_{i=1}^{g} t_i^2)\}\{n^3-\sum_{j=1}^{h} u_j^3 - 3(n^2-\sum_{j=1}^{h} u_j^2)\} + 9(n-2)(n^2-\sum_{i=1}^{g} t_i^2)(n^2-\sum_{j=1}^{h} u_j^2)}{18\,n(n-1)(n-2)} .$$

Als er noch onder de x_i noch onder de y_i gelijke waarnemingen voorkomen dan is

$$t_i = 1 \quad \text{voor iedere } i$$
$$u_j = 1 \quad \text{voor iedere } j$$
$$g = h = n .$$

De variantie van $\underline{S}$ wordt dan

$$\sigma^2\{\underline{S}\} = \frac{1}{18}\, n(n-1)(2n+5) .$$

Een tabel van $\sigma\{\underline{S}\}$ voor dit geval vindt men voor $n = 40$ t/m 100 in [2] (tabel IV).

De grootheid $\frac{S}{\sigma\{\underline{S}\}}$ is nu, voor grote waarden van n en als, zowel bij de x_i als bij de y_i, de groepen gelijke waarnemingen niet te veel in grootte verschillen, bij benadering normaal verdeeld met gemiddelde 0 en spreiding 1. De overschrijdingskans kan dan dus in een tabel der normale verdeling worden opgezocht.

Hierbij past men gewoonlijk een continuïteitscorrectie 1 toe, d.w.z. als rechteroverschrijdingskans neemt men het oppervlak de normale verdeling rechts van $\frac{S-1}{\sigma\{\underline{S}\}}$, als linkeroverschrijdingskans het oppervlak links van $-\frac{S+1}{\sigma\{\underline{S}\}}$ en als tweezijdige overschrijdingskans het oppervlak links van $-\frac{|S|-1}{\sigma\{\underline{S}\}}$ plus het oppervlak rechts van $\frac{|S|-1}{\sigma\{\underline{S}\}}$. Als er bij de x_i óf bij de y_i, doch niet bij beide, tweetallen of drietallen gelijken voorkomen, dus als b.v.

$$t_i = 1 \quad \text{voor iedere } i$$
$$u_j \leq 3 \quad \text{voor iedere } j ,$$

dan kan men als $n \leq 10$ is gebruik maken van exacte tabellen van SILLITTO [4].

In gevallen waarin men geen gebruik kan maken van de exacte tabellen en waar de normale benadering niet van toepassing is, moet men de exacte verdeling zelf berekenen. Hierop zullen wij niet verder ingaan. Voorbeeld 1 is zo'n geval. Bij voorbeeld 2 kan men echter gebruik maken van de normale benadering. Hier was

$$S = 181$$
$$n = 30$$
$$n^3 - \sum_{i=1}^{g} t_i^3 = 22146 \qquad n^2 - \sum_{i=1}^{g} t_i^2 = 598$$
$$n^3 - \sum_{j=1}^{h} u_j^3 = 23190 \qquad n^2 - \sum_{j=1}^{h} u_j^2 = 670 .$$

dus

$$\sigma^2\{S\} = \frac{2.22146.23190 + 9.28.598.670}{18.30.29.28} = \frac{1128097800}{438480} = 2572,75$$

$$\sigma\{S\} = 50,72$$

$$\frac{S-1}{\sigma\{S\}} = 3,55 .$$

In een tabel van de normale verdeling vindt men voor de tweezijdige overschrijdingskans 0,0004.

4. De rangcorrelatiecoëfficiënt τ

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

$$\tau = \frac{2S}{\left\{n^2 - \sum_{i=1}^{k} t_i^2\right\}^{\frac{1}{2}} \left\{n^2 - \sum_{j=1}^{l} u_j^2\right\}^{\frac{1}{2}}} .$$

Als er in geen van beide rijen gelijken voorkomen, dan wordt deze formule

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

Literatuur

[1] KENDALL, M.G., Rank Correlation Methods, London 1948.

[2] KAARSEMAKER, L. en A. VAN WIJNGAARDEN, Tables for use in Rank Correlation, Report R 73 of the Computation Department of the Mathematical Centre (1952) en Statistica 7 (1953), p. 41-54.

[3] HEMELRIJK, J., Kendall's rangcorrelatiecoëfficiënt τ, Hoofdstuk I van de Cursus "Parametervrije Methoden", Rapport S 59 van het Mathematisch Centrum (1951).

[4] SILLITTO, G.P., The distribution of Kendall coefficient of rank correlation in rankings containing ties Biometrika 34 (1947), p. 36-40.