STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 73(M 13b)

Spearman's correlatie-coëfficiënt, wanneer er geen gelijke waarnemingen optreden.

1950.

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

S 73 (M 13b)

Spearman's correlatie-coëfficiënt, wanneer er geen gelijke waarnemingen optreden[1].

We beschouwen het geval dat twee stochastische grootheden x en y[2] een simultane verdeling bezitten. Over deze verdeling zelf behoeft niets ondersteld te worden. (x_1,y_1), $(x_2,y_2)\ldots(x_n,y_n)$ zijn n onafhankelijke waarnemingsparen van deze stochastische grootheden.

Voorbeeld.

i =	1	2	3	4	5	6
x_i	0,107	0,123	0,101	0,111	0,154	0,132
y_i	3,39	3,01	3,17	3,54	3,47	3,49

Om op grond van deze waarnemingen de hypothese H_o te toetsen, dat x en y onafhankelijk verdeeld zijn heeft C.Spearman de volgende parametervrije toetsingsgrootheid ingevoerd:

Nummer de waarnemingen x_i $(i = 1,2,\ldots,n)$ naar opklimmende grootte, evenals de y_i.

Voorbeeld.

i =	1	2	3	4	5	6
rangnummer van x_i	2	4	1	3	6	5
rangnummer van y_i	3	1	2	6	4	5
verschillen der rangnummers, d:	1	3	1	3	2	0
d^2:	1	9	1	9	4	0

Van de paren rangnummers wordt het verschil bepaald en deze verschillen worden gekwadrateerd en opgeteld. Hun som is $S(d^2)$. In dit voorbeeld dus 24.

$S(d^2)$ varieert van 0, bij volkomen overeenstemming, tot $1/3(n^3-n)$, bij volkomen tegengestelde nummering der beide rijen.

We verwerpen de hypothese H_o van onafhankelijkheid ten gunste van de alternatieve hypothese: x en y zijn positief resp. negatief gecorreleerd, als $S(d^2)$ een zeer

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.
2) De onderstreping geeft aan, dat de toetsingsgrootheid stochastisch is, d.w.z. een waarschijnlijkheidsverdeling bezit.

lage resp. zeer hoge waarde aanneemt.

Tabellen van de kans, dat een zekere waarde van $\underline{S}(d^2)$ bereikt of overschreden wordt voor n = 4 t/m 8 zijn te vinden in Kendall [1] p. 142[3]. Voor n = 9 en 10 in Biometrika [2] p. 133[3].

Voor n > 10 kan men gebruik maken van het feit, dat

$$\underline{t} = \left\{n^3-n-6\underline{S}(d^2)\right\} \sqrt{\frac{n-2}{12\underline{S}(d^2)\left\{n^3-n-3\underline{S}(d^2)\right\}}}$$

bij benadering een Studentverdeling heeft met n-2 vrijheidsgraden.

Voor nog grotere n kan men ook van de in dit geval minder nauwkeurige normale benadering gebruik maken. Daarbij geldt dan voor het gemiddelde $\mathcal{E}\,\underline{S}(d^2)$ van $\underline{S}(d^2)$

$$\mathcal{E}\,\underline{S}(d^2) = \frac{1}{6}(n^3-n) = \binom{n+1}{3},$$

en voor de variantie

$$\sigma^2\{\underline{S}(d^2)\} = \frac{1}{36}(n^5+n^4-n^3-n^2) = \frac{1}{3}\binom{n+1}{3}\binom{n+1}{2}.$$

Om de mate van afhankelijkheid van $\underline{x}$ en $\underline{y}$ als een getal tussen -1 en +1 uit te drukken, gebruikt men in dit verband de grootheid

$$\rho = 1 - \frac{6\underline{S}(d^2)}{n^3-n}.$$

De boven ingevoerde grootheid $\underline{t}$ kan dan als volgt in ρ worden uitgedrukt;

$$\underline{t} = \rho\sqrt{\frac{n-2}{1-\rho^2}}.$$

---

3) Zie literatuuropgave achteraan.

Literatuur:

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

[2] S.T.David, M.G.Kendall and A.Stuart, Some questions of distribution in the theory of rank correlation, Biometrika 38 (1951) p. 131-140.