

**stichting
mathematisch
centrum**



AFDELING MATHEMATISCHE STATISTIEK

SW 42/76

APRIL

H. ELFFERS

GRENZEN VOOR KORRELATIEKOEFFICIËNTEN IN EEN
TRIVARIATE VERDELING

2e boerhaavestraat 49 amsterdam

BIBLIOTHEEK MATHEMATISCH CENTRUM
—AMSTERDAM—

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

Grenzen voor korrelatiecoëfficiënten in een trivariate verdeling

door

H. Elffers

SAMENVATTING

Laat $(\underline{x}, \underline{y}, \underline{z})$ een trivariate verdeling bezitten, en zij α een korrelatiecoëfficiënt. Onder- en bovengrenzen voor $\alpha(\underline{x}, \underline{z})$ in termen van $\alpha(\underline{x}, \underline{y})$ en $\alpha(\underline{y}, \underline{z})$ worden besproken voor verscheidene keuzen van α .

TREFWOORD: *korrelatiecoëfficiënten*

1. PROBLEEMSTELLING EN RESULTATEN

Probleemstelling

Als van een drietal variabelen de eerste met de tweede en de tweede met de derde samenhangt, volgt dan dat ook de eerste met de derde samenhangt?

In dit rapport wordt deze vraag bestudeerd voor die samenhang, welke wordt uitgedrukt in een korrelatiecoëfficiënt. Men kan het probleem dan als volgt formuleren: Als van een drietal variabelen twee der korrelatiecoëfficiënten bekend zijn, wat kan dan worden gezegd over de derde korrelatiecoëfficiënt?

Dit wordt onderzocht voor de produktmomentkorrelatiecoëfficiënt, Kendalls (populatie) rangkorrelatiecoëfficiënt, en Spearmans (populatie) rangkorrelatiecoëfficiënt ofwel de "grade" korrelatiecoëfficiënt.

Resultaten

In elk van deze gevallen is het mogelijk onder- en bovengrenzen voor de derde korrelatiecoëfficiënt aan te geven. Voor de produktmomentkorrelatiecoëfficiënt is dit een bekend resultaat, zie bijvoorbeeld KENDALL & STUART [1958], opgave 15.2, dat voor de volledigheid wordt opgenomen.

Laat voor elk tweetal stochastische variabelen $(\underline{x}, \underline{y})$ met een simultane verdeling

- $\rho(\underline{x}, \underline{y})$ de produktmomentkorrelatiecoëfficiënt (definitie zie §3)
- $\tau(\underline{x}, \underline{y})$ Kendalls rangkorrelatiecoëfficiënt (definitie zie §4)
- $\zeta(\underline{x}, \underline{y})$ Spearmans rangkorrelatiecoëfficiënt (definitie zie §5)

aangeven.

Laat $(\underline{x}, \underline{y}, \underline{z})$ een trivariate verdeling hebben, en laten de korrelatiecoëfficiënten van $\underline{x}$ en $\underline{y}$, resp. $\underline{x}$ en $\underline{z}$ gegeven zijn. Dan geldt

$$(1.1) \quad \rho(\underline{x}, \underline{y}) \cdot \rho(\underline{y}, \underline{z}) - \sqrt{(1 - \rho^2(\underline{x}, \underline{y}))(1 - \rho^2(\underline{y}, \underline{z}))} \leq \rho(\underline{x}, \underline{z})$$

$$\leq \rho(\underline{x}, \underline{y}) \cdot \rho(\underline{y}, \underline{z}) + \sqrt{(1 - \rho^2(\underline{x}, \underline{y}))(1 - \rho^2(\underline{y}, \underline{z}))},$$

$$(1.2) \quad |\tau(\underline{x}, \underline{y}) + \tau(\underline{y}, \underline{z})| - 1 \leq \tau(\underline{x}, \underline{z}) \leq 1 - |\tau(\underline{x}, \underline{y}) - \tau(\underline{y}, \underline{z})|,$$

$$(1.3) \quad \zeta(\underline{x}, \underline{y}) \cdot \zeta(\underline{y}, \underline{z}) - \sqrt{(1 - \zeta^2(\underline{x}, \underline{y}))(1 - \zeta^2(\underline{y}, \underline{z}))} \leq \zeta(\underline{x}, \underline{z})$$

$$\leq \zeta(\underline{x}, \underline{y}) \cdot \zeta(\underline{y}, \underline{z}) + \sqrt{(1 - \zeta^2(\underline{x}, \underline{y}))(1 - \zeta^2(\underline{y}, \underline{z}))}.$$

De grenzen (1.1) en (1.2) zijn de scherpst mogelijke; de grenzen (1.3) mogelijk niet. Deze resultaten gelden ook voor de steekproefanaloga van deze koëfficiënten.

Deze grenzen zijn in het algemeen vrij ruim, tenzij de beide gegeven korrelatiekoëfficiënten in absolute waarde zeer hoog zijn. Zie fig. 1 en 2 (pp. 16 en 17). De konklusie luidt dus dat over samenhang tussen $\underline{x}$ en $\underline{z}$ i.h.a. weinig gezegd kan worden op grond van samenhang tussen $\underline{x}$ en $\underline{y}$ en tussen $\underline{y}$ en $\underline{z}$.

2. KONVENTIES EN NOTATIE

In het volgende wordt stilzwijgend verondersteld dat elke voorkomende stochastische variabele een verdeling met eindig tweede moment bezit.

De verwachting van een stochastische variabele $\underline{x}$ wordt aangegeven met $E\underline{x}$, de variantie met $\sigma^2(\underline{x})$, de standaardafwijking met $\sigma(\underline{x})$. De kovariantie van een bivariate stochastische variabele wordt genoteerd met $\text{kov}(\underline{x}, \underline{y})$.

Met $A \subset B$ wordt bedoeld dat A een *echte* deelverzameling van B is, met $A \subseteq B$ dat $A \subset B$ of $A = B$ geldt.

3. GRENZEN VOOR DE PRODUKTMOMENTKORRELATIEKOËFFICIËNT

Zij $\underline{x} \stackrel{\text{def}}{=} (\underline{x}^{(1)}, \underline{x}^{(2)}, \dots, \underline{x}^{(n)})'$ een n-dimensionale stochastische variabele, $n \geq 2$. Noteer de kovariantiematrix van $\underline{x}$ met

$$\text{kov}(\underline{x}) \stackrel{\text{def}}{=} (\text{kov}(\underline{x}^{(i)}, \underline{x}^{(j)}))_{i=1, j=1}^n.$$

We zullen gebruik maken van het welbekende lemma:

LEMMA 3.1. *Er bestaat een stochastische variabele met kovariantiematrix Σ dan en slechts dan als Σ een niet-negatief definitie symmetrische matrix is. $\square$*

Een principale minor van een vierkante matrix V is de determinant van een submatrix van V die uit V ontstaat door nul of meer rijen en overeenkomstige kolommen weg te laten.

Een bekend resultaat uit de matrixrekening luidt:

LEMMA 3.2. Een vierkante matrix Σ is niet-negatief definitief d.e.s.d.a. alle principale minoren van Σ niet negatief zijn.

BEWIJS. Zie bijvoorbeeld MIRSKY [1955], Th. 13.3.5. $\square$

De produktmomentkorrelatiecoëfficiënt ρ van een bivariate stochast $(\underline{x}, \underline{y})$ is gedefinieerd als

$$\rho(\underline{x}, \underline{y}) \stackrel{\text{def}}{=} E((\underline{x} - E\underline{x})(\underline{y} - E\underline{y})) \cdot (\sigma(\underline{x})\sigma(\underline{y}))^{-1}.$$

Voor ρ geldt onder meer:

$$\begin{aligned}\rho(\underline{x}, \underline{y}) &= \rho(\underline{y}, \underline{x}), \\ \rho(\underline{x}, \underline{x}) &= 1.\end{aligned}$$

De korrelatiematrix van een n-dimensionale stochast $\underline{x} \stackrel{\text{def}}{=} (\underline{x}^{(1)}, \underline{x}^{(2)}, \dots, \underline{x}^{(n)})'$ is gedefinieerd als

$$\text{rho}(\underline{x}) \stackrel{\text{def}}{=} (\rho(\underline{x}^{(i)}, \underline{x}^{(j)}))_{i=1, j=1}^n.$$

Op grond van de definitie van ρ geldt:

$$\text{rho}(\underline{x}) = \text{kov}((\underline{x}^{(1)}/\sigma(\underline{x}^{(1)}), \dots, \underline{x}^{(n)}/\sigma(\underline{x}^{(n)}))'),$$

d.w.z. $\text{rho}(\underline{x})$ is een kovariantiematrix met éenen op de diagonaal.

Dit feit en de Lemma's 1 en 2 leiden tot de volgende:

STELLING. Er bestaat een stochastische variabele $\underline{x} \stackrel{\text{def}}{=} (\underline{x}^{(1)}, \dots, \underline{x}^{(n)})'$ met korrelatiematrix P d.e.s.d.a. P een symmetrische $n \times n$ matrix is met éenen op de diagonaal, waarvan alle principale minoren niet negatief zijn.

Toepassing van deze stelling op een bivariate stochast geeft het bekende resultaat dat

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

slechts een korrelatiematrix kan zijn wanneer $1 \geq 0$ en $1 - \rho^2 \geq 0$.

Wanneer in het vervolg het symbool ρ wordt gebruikt, zal niet steeds de restrictie $|\rho| \leq 1$ vermeld worden, doch wel impliciet bedoeld zijn.

De vraagstelling van dit rapport is nu het bepalen van de verzameling

$D_\rho = \{(\rho_{12}, \rho_{23}, \rho_{13}) \in [-1, +1]^3 \mid (\rho_{12}, \rho_{23}, \rho_{13}) \text{ zijn de drie korrelatiekoefficiënten van een trivariate verdeling}\}.$

Op grond van de stelling geldt

$$\begin{aligned}
 (\rho_{12}, \rho_{23}, \rho_{13}) \in D_\rho &\Leftrightarrow \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix} \text{ heeft slechts niet-negatieve minoren} \\
 &\Leftrightarrow 1 \geq 0, 1 - \rho_{12}^2 \geq 0, 1 - \rho_{23}^2 \geq 0, 1 - \rho_{13}^2 \geq 0, \\
 &\quad 1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{23}\rho_{13} \geq 0 \\
 &\Leftrightarrow (\text{in verband met de impliciete beperking tot } [-1, +1]^3) \\
 &\quad 1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{23}\rho_{13} \geq 0 \\
 &\Leftrightarrow \rho_{12}\rho_{23} - \sqrt{(1-\rho_{12}^2)(1-\rho_{23}^2)} \leq \rho_{13} \\
 &\quad \leq \rho_{12}\rho_{23} + \sqrt{(1-\rho_{12}^2)(1-\rho_{23}^2)}. \tag{3.1}
 \end{aligned}$$

Daarmee is (1.1) bewezen.

Merk op dat de linkergrens in (3.1) ≥ -1 en de rechtergrens $\leq +1$ is, voor alle $(\rho_{12}, \rho_{23}) \in [-1, +1]^2$.

(3.1) impliceert dat ρ_{12} , ρ_{13} en ρ_{23} voldoen aan

$$\frac{(\rho_{23} - \rho_{13})^2}{2(1-\rho_{12})} + \frac{(\rho_{23} + \rho_{13})^2}{2(1+\rho_{12})} \leq 1,$$

d.w.z. dat voor vaste ρ_{12} de mogelijke waarden ρ_{13} en ρ_{23} binnen of op de rand van een ellips liggen. Zie Fig. 1 (p. 16).

In het bijzondere geval $\rho_{12} = \rho_{23} = \rho$ zeg, gaat (3.1) over in

$$2\rho^2 - 1 \leq \rho_{13} \leq 1.$$

De grenzen in (3.1) worden d.e.s.d.a. aangenomen als de korrelatiematrix van de beschouwde verdeling singulier is, d.w.z. dat er een lineaire relatie tussen de beschouwde variabelen bestaat.

VOORBEELD van het aannemen van de ondergrens in (3.1): Zij $(\underline{x}, \underline{y})$ stochastisch onafhankelijk met gelijke variantie

$$\underline{x}^{(1)} \stackrel{\text{def}}{=} \underline{x}, \quad \underline{x}^{(2)} \stackrel{\text{def}}{=} \underline{x} + \underline{y}, \quad \underline{x}^{(3)} \stackrel{\text{def}}{=} \underline{y}, \quad \rho_{ij} \stackrel{\text{def}}{=} \rho(\underline{x}^{(i)}, \underline{x}^{(j)}).$$

Dan geldt

$$\rho_{12} = \rho_{23} = \frac{1}{2} \sqrt{2} \approx 0,71 \quad \text{en} \quad \rho_{13} = 0.$$

Dit voorbeeld laat tevens zien hoe zwak de overdracht van samenhang van variabelen is.

Steekproeven

Daar men de steekproefproduktmomentkorrelatiecoëfficiënt kan beschouwen als een populatieproduktmomentkorrelatiecoëfficiënt, door de steekproef te beschouwen als een populatie waarop een stochastische variabele is gedefinieerd die elk der steekproefwaarden met gelijke kans aanneemt, geldt (1.1) ook wanneer men voor de ρ 's de steekproefcoëfficiënten substitueert.

Dit resultaat is echter niet van praktisch nut, daar men, over een tri-variate steekproef beschikkende, beter alle coëfficiënten direkt kan uitrekenen.

4. GRENZEN VOOR KENDALLS RANGKORRELATIEKOËFFICIËNT

Laat $(\underline{x}, \underline{y})$ een bivariate verdeling bezitten, en laten $(\underline{x}_1, \underline{y}_1)$ en $(\underline{x}_2, \underline{y}_2)$ twee stochastisch onafhankelijke trekkingen van $(\underline{x}, \underline{y})$ zijn.

$(\underline{x}_1, \underline{y}_1)$ en $(\underline{x}_2, \underline{y}_2)$ heten

konkordant als $\text{sign}((- \underline{x}_1, \underline{y}_1))$

diskordant als $\text{sign}((- \underline{x}_1, \underline{y}_2))$

geknoopt (tied) als $\text{sign}((- \underline{x}_1, \underline{y}_1))$

Laten p_k , p_d en p_t de kansen op eekon geknoopt paar zijn, dan geldt

$$p_k + p_d + p_t = 1.$$

Kendalls (populatie) rangkorrelatieën als

$$\tau(-\underline{x}, \underline{y}) = -\tau(\underline{x}, \underline{y})$$

en derhalve geldt

$$\tau(\underline{x}, \underline{y}) = 2p_k - (1 - p_t).$$

Voor τ geldt onder meer

$$\tau(-\underline{x}, \underline{y}) = -\tau(\underline{x}, \underline{y})$$

en

$$|\tau(\underline{x}, \underline{y})| \leq 1.$$

Wanneer het symbool τ wordt gebruikt, wordt de restrikte $|\tau| \leq 1$ vermeld worden, dochel

We willen de verzameling

$$D_\tau \stackrel{\text{def}}{=} \{(\tau_{12}, \tau_{23}, \tau_{13}) \in [-1, 1]^3\}$$

Kendall rangkorrelatiekoëffie

bepalen.

Zij nu $(\underline{x}^{(1)}, \underline{x}^{(2)}, \underline{x}^{(3)})$ een trivariate

laten $(\underline{x}_1^{(1)}, \underline{x}_1^{(2)}, \underline{x}_1^{(3)})$ en $(\underline{x}_2^{(1)}, \underline{x}_2^{(2)}, \underline{x}_2^{(3)})$ twee s.o. trekkingen uit deze verdeling zijn.

Definieer de volgende eventualiteiten

$$\left. \begin{aligned} K_{ij} &\stackrel{\text{def}}{=} \text{sign}((\underline{x}_1^{(i)} - \underline{x}_2^{(i)})(\underline{x}_1^{(j)} - \underline{x}_2^{(j)})) = 1 \\ D_{ij} &\stackrel{\text{def}}{=} \text{sign}((\underline{x}_1^{(i)} - \underline{x}_2^{(i)})(\underline{x}_1^{(j)} - \underline{x}_2^{(j)})) = -1 \end{aligned} \right\} \quad i, j = 1, 2, 3,$$

d.w.z. de eventualiteit dat de i-de en j-de komponent van de twee trekkingen konkordant, resp. diskordant zijn.

Definieer de eventualiteiten

$$\begin{aligned} T_i &\stackrel{\text{def}}{=} \underline{x}_1^{(i)} = \underline{x}_2^{(i)} \quad i = 1, 2, 3, \\ T_{ij} &\stackrel{\text{def}}{=} T_i \wedge T_j \quad i, j = 1, 2, 3, \\ T_{123} &\stackrel{\text{def}}{=} T_1 \wedge T_2 \wedge T_3, \end{aligned}$$

d.w.z. de eventualiteit dat de i-de komponent een knoop oplevert etc.

Zij

$$\begin{aligned} p_{kk} &\stackrel{\text{def}}{=} P(K_{12} \wedge K_{23}), \\ p_{kd} &\stackrel{\text{def}}{=} P(K_{12} \wedge D_{23}), \\ p_{dk} &\stackrel{\text{def}}{=} P(D_{12} \wedge K_{23}), \\ p_{dd} &\stackrel{\text{def}}{=} P(D_{12} \wedge D_{23}), \\ t_i &\stackrel{\text{def}}{=} P(T_i) \quad i = 1, 2, 3, \\ t_{ij} &\stackrel{\text{def}}{=} P(T_{ij}) \quad i, j = 1, 2, 3, \\ t_{123} &\stackrel{\text{def}}{=} P(T_{123}), \end{aligned}$$

$$\tau_{ij} \stackrel{\text{def}}{=} \tau(\underline{x}^{(i)}, \underline{x}^{(j)}) \quad i, j = 1, 2, 3,$$

zodat geldt

$$\begin{aligned} \tau_{12} &= 2P(K_{12}) - 1 + P(\text{sign}((\underline{x}_1^{(1)} - \underline{x}_2^{(1)})(\underline{x}_1^{(2)} - \underline{x}_2^{(2)})) = 0) \\ &= 2P(K_{12} \wedge K_{23}) + 2P(K_{12} \wedge D_{23}) + 2P(K_{12} \wedge T_3) - 1 + P(T_1 \vee T_2) \\ &= 2p_{kk} + 2p_{kd} + 2P(K_{12} \wedge T_3) - 1 + t_1 + t_2 - t_{12}. \end{aligned}$$

Evenzo geldt

$$\tau_{23} = 2p_{kk} + 2p_{dk} + 2P(K_{23} \wedge T_1) - 1 + t_2 + t_3 - t_{23}.$$

Gebruikmakend van

$$p_{kk} + p_{kd} + p_{dk} + p_{dd} + t_1 + t_2 + t_3 - t_{12} - t_{23} - t_{13} + t_{123} = 1$$

volgt

$$\begin{aligned} (4.0) \quad \tau_{12} + \tau_{23} - 1 + 4p_{dd} - 2P(K_{12} \wedge T_3) - 2P(K_{23} \wedge T_1) + 2t_1 + 2t_3 + \\ - t_{12} - t_{23} - 3t_{13} + 2t_{123} = 2p_{kk} + 2p_{dd} - 1 + t_1 + t_3 - t_{13}. \end{aligned}$$

Merk op dat

$$K_{13} \iff (K_{12} \wedge K_{23}) \vee (D_{12} \wedge D_{23}) \vee (K_{13} \wedge T_2)$$

waarbij de leden van de disjunctie in het rechterlid wederzijds uitsluitend zijn, zodat geldt

$$\begin{aligned} \tau_{13} &= 2P(K_{13}) - 1 + t_1 + t_3 - t_{13} \\ &= 2p_{kk} + 2p_{dd} + 2P(K_{13} \wedge T_2) - 1 + t_1 + t_3 - t_{13}. \end{aligned}$$

Derhalve is het rechterlid van (4.0) gelijk aan $\tau_{13} - 2P(K_{13} \wedge T_2)$,

zodat volgt

$$\begin{aligned}\tau_{13} = & \tau_{12} + \tau_{23} - 1 + 4p_{dd} + 2P(K_{13} \wedge T_2) - 2P(K_{12} \wedge T_3) + \\ & - 2P(K_{23} \wedge T_1) + 2t_1 + 2t_3 - t_{12} - t_{23} - 3t_{13} + 2t_{123}\end{aligned}$$

In deze uitdrukking zullen we een aantal termen afschatten en wel m.b.v.

$$\begin{aligned}p_{dd} & \geq 0, \\ 2P(K_{13} \wedge T_2) & \geq 0, \\ P(K_{12} \wedge T_3) & \leq P(\neg T_1 \wedge \neg T_2 \wedge T_3) = t_3 - t_{13} - t_{23} + t_{123}, \\ P(K_{23} \wedge T_1) & \leq t_1 - t_{12} - t_{13} + t_{123},\end{aligned}$$

waarmee volgt

$$\begin{aligned}\tau_{13} & \geq (\tau_{12} + \tau_{23}) - 1 + t_{12} + t_{23} + t_{13} - 2t_{123} \\ & = (\tau_{12} + \tau_{23}) - 1 + P(T_{12} \vee T_{23} \vee T_{13}).\end{aligned}$$

Kortheidshalve noteren we

$$T^{(2)} \stackrel{\text{def}}{=} T_{12} \vee T_{23} \vee T_{13},$$

d.w.z. de eventualiteit van minstens twee geknoopte componenten in de twee waarnemingen, en

$$t^{(2)} \stackrel{\text{def}}{=} P(T^{(2)}) = t_{12} + t_{23} + t_{13} - 2t_{123}.$$

Het resultaat kan dan worden uitgedrukt als

$$(4.1) \quad \tau_{13} \geq (\tau_{12} + \tau_{23}) - 1 + t^{(2)}.$$

De ongelijkheid (4.1) geldt voor elke trivariate verdeling. Substitueren we $\underline{x}^{(1)}$ voor $\underline{x}^{(1)}$ in het bovenstaand argument, dan gaat (4.1) over in

$$\tau_{13} \leq 1 - (\tau_{23} - \tau_{12}) - t^{(2)}.$$

Door evenzo $-\underline{x}^{(2)}$ voor $\underline{x}^{(2)}$ en $-\underline{x}^{(3)}$ voor $\underline{x}^{(3)}$ te substitueren volgt nog een tweetal ongelijkheden, welke samen te vatten zijn in:

$$(4.2) \quad |\tau_{12} + \tau_{23}| - (1-t^{(2)}) \leq \tau_{13} \leq (1-t^{(2)}) - |\tau_{12} - \tau_{23}|.$$

Daar $t^{(2)} \geq 0$ geldt, tevens

$$(4.3) \quad |\tau_{12} + \tau_{23}| - 1 \leq \tau_{13} \leq 1 - |\tau_{12} - \tau_{23}|.$$

Daarmee is bewezen dat

$$D_{\tau} \subseteq D'_{\tau} \stackrel{\text{def}}{=} \left\{ (\tau_{12}, \tau_{23}, \tau_{13}) \in [-1, +1]^3 \mid |\tau_{12} + \tau_{23}| - 1 \leq \tau_{13} \leq 1 - |\tau_{12} - \tau_{23}| \right\}.$$

Dat tevens geldt $D_{\tau} = D'_{\tau}$ zullen we aantonen door te laten zien dat voor

$$N_{\tau} \stackrel{\text{def}}{=} \left\{ (\tau_{12}, \tau_{23}, \tau_{13}) \in [-1, +1]^3 \mid (\tau_{12}, \tau_{23}, \tau_{13}) \text{ zijn de Kendall-korrelatiecoëfficiënten van een trivariaat normale verdeling} \right\}$$

geldt

$$N_{\tau} = D'_{\tau}.$$

Dan immers volgt $D'_{\tau} = N_{\tau} \subseteq D_{\tau} \subseteq D'_{\tau}$, dus $D_{\tau} = D'_{\tau}$ d.w.z. (1.2).

LEMMA 4.1. Als $(\underline{x}, \underline{y})$ een bivariaat normale verdeling heeft dan geldt

$$\tau(\underline{x}, \underline{y}) = \frac{2}{\pi} \arcsin(\rho(\underline{x}, \underline{y})).$$

BEWIJS. Zie bijvoorbeeld GIBBONS [1971]. $\square$

Laat $\underline{x} = (\underline{x}^{(1)}, \underline{x}^{(2)}, \underline{x}^{(3)})'$ een normale verdeling hebben met

$$\rho(\underline{x}) = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix} \text{ z.d.d. (3.2) geldt met gelijkheidsteken}$$

aan één van beide zijden, zeg bij de ondergrens, dus

$$\rho_{12}\rho_{23} - \sqrt{(1-\rho_{12}^2)(1-\rho_{23}^2)} = \rho_{13}.$$

Substitutie van $\sin(\frac{\pi}{2} \tau_{ij})$ voor ρ_{ij} geeft na enige manipulatie

$$\sin(\frac{\pi}{2}((\tau_{12} + \tau_{23}) - 1)) = \sin(\frac{\pi}{2}\tau_{13})$$

of

$$\sin(\frac{\pi}{2}((\tau_{12} + \tau_{23}) - 1)) = \sin(\frac{\pi}{2}\tau_{13}),$$

hetgeen, voor $(\tau_{12}, \tau_{13}, \tau_{23}) \in [-1, +1]^3$, impliceert

$$|\tau_{12} + \tau_{23}| - 1 = \tau_{13}.$$

Dit is juist de ondergrens van (4.3).

Op analoge wijze demonstreert men dat de bovengrens kan worden aangenomen. Daar de transformatie $\tau_{13} = \frac{2}{\pi} \arcsin(\rho_{13})$ een continue monotone transformatie is, worden ook alle waarden van τ_{13} tussen beide grenzen door normale verdelingen gerealiseerd, d.w.z. $N_{\tau} = D'_{\tau}$.

Voor vaste τ_{12} liggen volgens (4.3) de mogelijke τ_{23} en τ_{13} binnen of op de rand van een rechthoek, vgl. Fig. 2 (p. 17).

Voor $\tau_{12} = \tau_{23} = \tau$ zeg, gaat (4.2) over in

$$|2\tau| - 1 \leq \tau_{13} \leq 1.$$

Enige algebra laat zien dat $D_{\tau} \subset D_{\rho}$, vgl. Fig. 3 (p. 18).

Voor continue stochastische variabelen geldt dat een der grenzen in (4.3) wordt aangenomen d.e.s.d.a. minstens een der p_{kk} , p_{kd} , p_{dk} of p_{dd} nul is. Dit volgt direkt uit de afschatting, en de gebruikte symmetrie-

argumenten.

Voor niet continue variabelen moet dan nog een aantal weinig doorzichtige voorwaarden over de kansen op knopen etc. zijn vervuld.

Eindige populaties en steekproeven

Bij diskrete stochastische variabelen kan men iets aan scherpte winnen door i.p.v. (4.3) (4.2) te gebruiken, mits men over een geschikte afschatting van $t^{(2)}$ beschikt. Daar het resultaat, in de ogen van de schrijver, voornamelijk gebruikt kan worden om te demonstreren dat kennis van τ_{12} en τ_{23} alléén nog erg weinig zegt over τ_{13} , is bovenstaande niet zo interessant daar men dan ook niet over kennis omtrent $t^{(2)}$ zal beschikken.

Ook kan men zich afvragen of de ongelijkheden (4.2) en (4.3) ook voor het steekproefanalogon van τ geldt. Vanzelfsprekend kan men beter, wanneer men over een steekproef van een trivariate stochast beschikt, alle drie de korrelatiecoëfficiënten berekenen, dan slechts twee en grenzen voor de derde. Toch geven de volgende resultaten inzicht in de onderlinge relatie van Kendallcoëfficiënten.

Zij $D = \{d_1, \dots, d_N\}$ een eindige populatie, van omvang N en zij $\underline{d}$ een aselekte trekking uit D , en $(x^{(1)}, x^{(2)}, x^{(3)}): D \rightarrow \mathbb{R}^3$. Zij $(\underline{x}^{(1)}, \underline{x}^{(2)}, \underline{x}^{(3)}) \stackrel{\text{def}}{=} (x^{(1)}(\underline{d}), x^{(2)}(\underline{d}), x^{(3)}(\underline{d}))$. Dan geldt:

$$P(T_{123}) \geq \frac{1}{N}$$

en dus

$$t^{(2)} = P(T^{(2)}) \geq P(T_{123}) \geq \frac{1}{N}$$

en

$$1 - t^{(2)} \leq \frac{N-1}{N}.$$

In dit geval impliceren de ongelijkheden (4.2)

(4.4) (aselekte trekking uit een eindige populatie van omvang N)

$$|\tau_{12} + \tau_{23}| - \frac{N-1}{N} \leq \tau_{13} \leq \frac{N-1}{N} - |\tau_{12} - \tau_{23}|.$$

In het bijzonder kan men dit voor steekproeven gebruiken: Zij gegeven een steekproef $d_1, \dots, d_N$ met een drietal kenmerken $(x^{(1)}(d_i), x^{(2)}(d_i), x^{(3)}(d_i))$, $i = 1, \dots, N$. Kendalls steekproefrangkorrelatiecoëfficiënt $\tau_a(x^{(1)}, x^{(2)})$ is gedefinieerd als (vgl. KENDALL [1955])

$$\begin{aligned} \tau_a(x^{(1)}, x^{(2)}) &\stackrel{\text{def}}{=} \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \text{sign}((x^{(1)}(d_i) - x^{(1)}(d_j)) \\ &\quad (x^{(2)}(d_i) - x^{(2)}(d_j))) \\ &= \frac{N}{N-1} \cdot \frac{1}{N^2} \sum_{i,j=1}^N \text{sign}((x^{(1)}(d_i) - x^{(1)}(d_j)) \\ &\quad (x^{(2)}(d_i) - x^{(2)}(d_j))). \end{aligned}$$

Beschouw de gegeven steekproef als een populatie waarop een stochast $\underline{d}$ is gedefinieerd met $P(\underline{d}=d_i) = \frac{1}{N}$, $i = 1, \dots, N$. Dan geldt:

$$\begin{aligned} \tau(x^{(1)}(\underline{d}), x^{(2)}(\underline{d})) &= \frac{1}{N^2} \sum_{i,j=1}^N \text{sign}((x^{(1)}(d_i) - x^{(1)}(d_j)) \\ &\quad (x^{(2)}(d_i) - x^{(2)}(d_j))) \end{aligned}$$

en derhalve

$$\tau_a(x^{(1)}, x^{(2)}) = \frac{N}{N-1} \tau(x^{(1)}(\underline{d}), x^{(2)}(\underline{d})).$$

Substitutie van deze relatie in (4.4) geeft, met de notatie

$$\tau_{ij}^a \stackrel{\text{def}}{=} \tau_a(x^{(i)}, x^{(j)}) \quad i, j = 1, 2, 3,$$

$$(4.5) \quad |\tau_{12}^a + \tau_{23}^a| - 1 \leq \tau_{13}^a \leq 1 - |\tau_{12}^a - \tau_{23}^a|,$$

d.w.z. de relatie (1.2) geldt ook voor de τ_a coëfficiënt.

5. GRENZEN VOOR SPEARMANS RANGKORRELATIEKOEFFICIENT

Laat $(\underline{x}, \underline{y})$ een bivariate verdeling bezitten, en laat F de marginale verdelingsfunctie van $\underline{x}$, G die van $\underline{y}$ zijn.

Spearman's rangkorrelatiecoëfficiënt of "grade" korrelatiecoëfficiënt $\zeta(\underline{x}, \underline{y})$ is gedefinieerd als

$$\zeta(\underline{x}, \underline{y}) \stackrel{\text{def}}{=} \rho(F(\underline{x}), G(\underline{y})).$$

We zijn geïnteresseerd in de verzameling

$$D_{\zeta} = \{(\zeta_{12}, \zeta_{23}, \zeta_{13}) \in [-1, +1]^3 \mid (\zeta_{12}, \zeta_{23}, \zeta_{13}) \text{ zijn de grade korrelatie-} \\ \text{coëfficiënten van een tri-} \\ \text{variate verdeling}\}.$$

Een onmiddellijk gevolg van de definitie van ζ is

$$(5.1) \quad D_{\zeta} \subseteq D_{\rho}.$$

Op grond van (3.1) bewijst dit (1.3).

Wanneer we op dezelfde wijze als bij de Kendallkorrelatiecoëfficiënt te werk gaan, en

$$N_{\zeta} \stackrel{\text{def}}{=} \{(\zeta_{12}, \zeta_{23}, \zeta_{13}) \in [-1, +1]^3 \mid (\zeta_{12}, \zeta_{23}, \zeta_{13}) \text{ de grade korrelatie-} \\ \text{coëfficiënten van een} \\ \text{trivariate normale ver-} \\ \text{deling}\}$$

bepalen, blijkt N_{ζ} een echt deel van D_{ρ} te zijn:

LEMMA 5.1. *Als $(\underline{x}, \underline{y})$ bivariaat normaal verdeeld is, dan geldt*

$$\zeta(\underline{x}, \underline{y}) = \frac{6}{\pi} \arcsin(\rho(\underline{x}, \underline{y})).$$

BEWIJS. Zie bijvoorbeeld MORAN [1948]. $\square$

Substitutie van $\sin\left(\frac{\pi}{6}\zeta_{ij}\right)$ voor ρ_{ij} in de grenzen (3.2) geeft

$$(5.2) \quad (\zeta_{12}, \zeta_{13}, \zeta_{23}) \in N_\zeta \iff$$

$$\frac{6}{\pi} \arcsin\left(2\sin\left(\frac{\pi}{6}\zeta_{12}\right)\sin\left(\frac{\pi}{6}\zeta_{23}\right) - \frac{1}{2}\sqrt{\left(1-4\sin^2\left(\frac{\pi}{6}\zeta_{12}\right)\right)\left(1-4\sin^2\left(\frac{\pi}{6}\zeta_{23}\right)\right)}\right) \leq \zeta_{13} \leq$$

$$\frac{6}{\pi} \arcsin\left(2\sin\left(\frac{\pi}{6}\zeta_{12}\right)\sin\left(\frac{\pi}{6}\zeta_{23}\right) + \frac{1}{2}\sqrt{\left(1-4\sin^2\left(\frac{\pi}{6}\zeta_{12}\right)\right)\left(1-4\sin^2\left(\frac{\pi}{6}\zeta_{23}\right)\right)}\right).$$

Deze grenzen zijn iets scherper dan die uit (1.3), alhoewel het verschil gering is, maximaal in absolute waarde 0,055. Vergelijk Fig. 3 (p. 18). Het blijft daarmee een open vraag of $D_\zeta = D_\rho$, dan wel $D_\zeta \subset D_\rho$. Het verschil tussen N_ζ en D_ρ is zo gering, dat gebruik van de grenzen (1.3) zeker verantwoord is.

Dat punten op de rand van D_ρ kunnen worden aangenomen blijkt uit het voorbeeld van de eindige populatie $Z = \{(1,2,1), (2,1,3), (3,4,2), (4,3,4)\}$, waarvan de drie gradecoefficiënten van de variabelen geïnduceerd door aselect trekken van een element uit Z gelijk zijn aan 0,6, 0 en 0,8.

Daar ook de steekproefversie van Spearmans koëfficiënt geschreven kan worden als de steekproef-produktmomentkorrelatiekoëfficiënt van zekere functies van x en y (nl. de rangen, waarbij het er niet toe doet hoe men het knopenprobleem oplost), geldt het resultaat eveneens voor de steekproef-koëfficiënt.

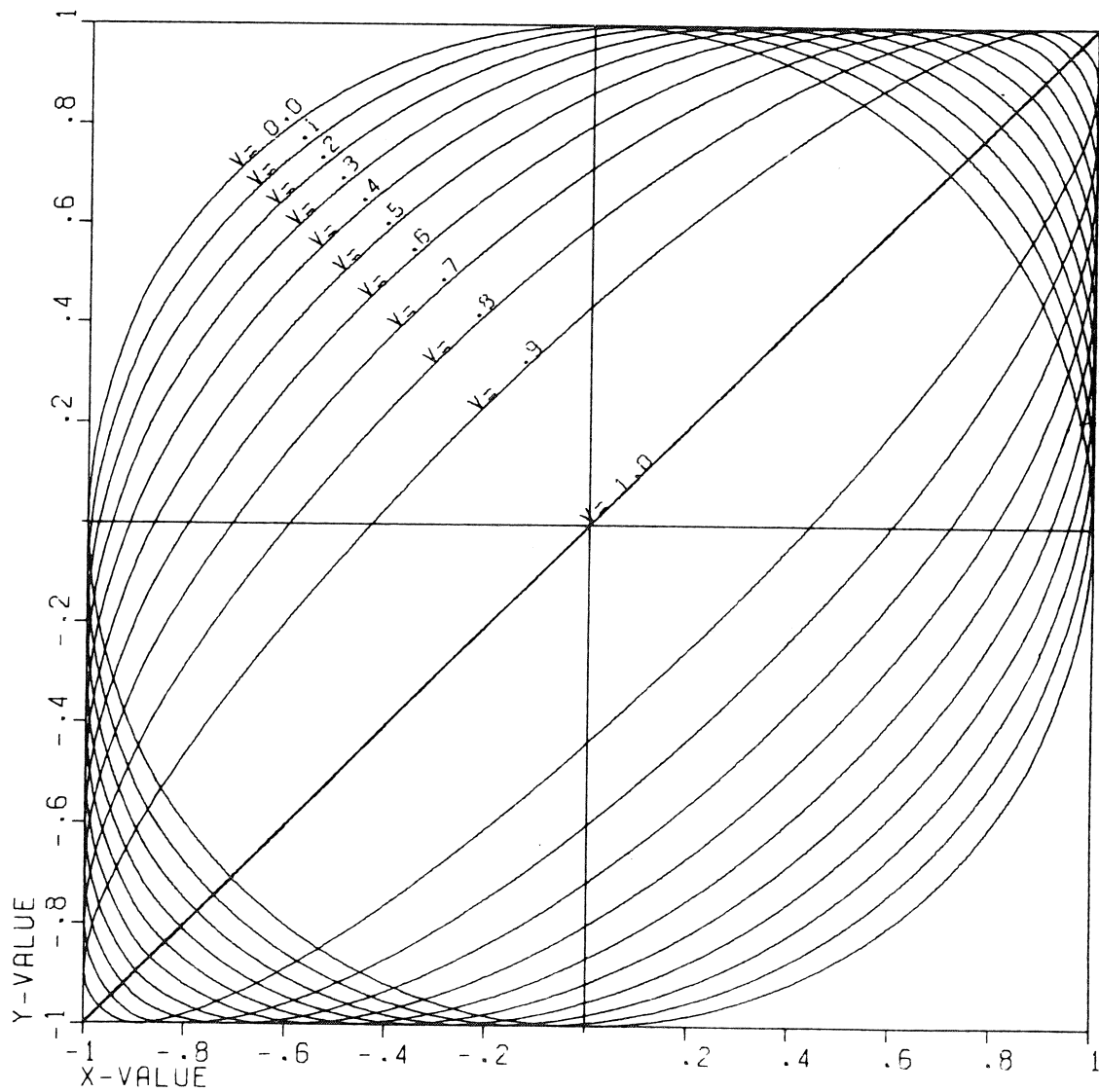


FIG. 1

De ellipsen $\frac{(x-y)^2}{2(1-v)} + \frac{(x+y)^2}{2(1+v)} = 1$ voor verschillende waarden van v .
 Wanneer $\rho_{12} = v$, dan kunnen ρ_{13} en ρ_{23} alle x - en y -waarden op of binnen de v -ellips aannemen.

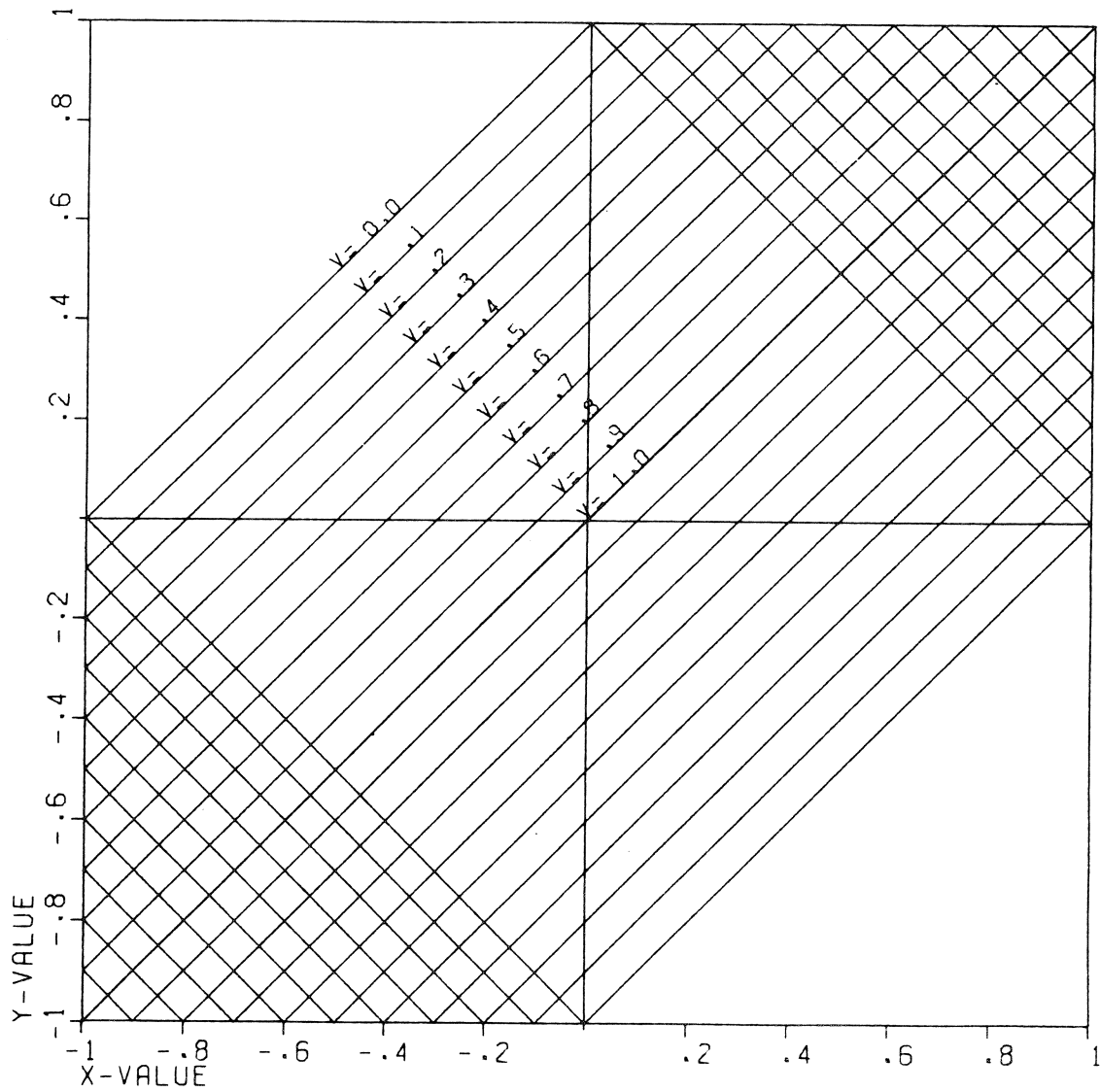


FIG. 2

De rechthoeken $|v+x| - 1 = y$ of $|v-x| + 1 = y$ voor verschillende waarden van v .

Wanneer $\tau_{12} = v$, dan kunnen τ_{13} en τ_{23} alle x - en y -waarden op of binnen de v -rechthoek aannemen.

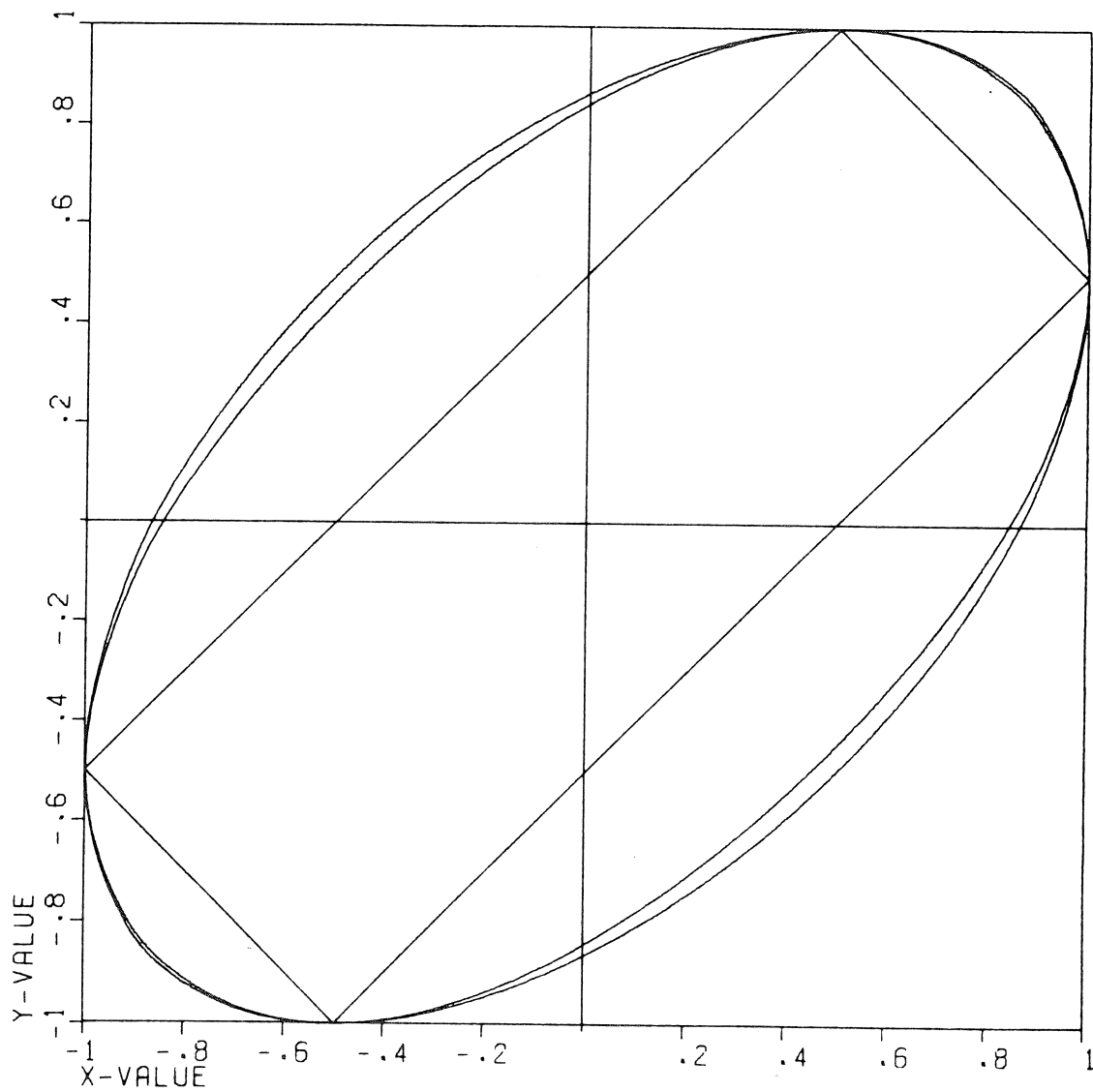


FIG. 3 .

Doorsneden van (van binnen naar buiten) D_τ , N_ζ en D_ρ voor respectievelijk $\tau_{13} = 0,5$, $\zeta_{13} = 0,5$ en $\rho_{13} = 0,5$. Het (x,y) vlak staat voor het (τ_{12}, τ_{23}) -, (ζ_{12}, ζ_{23}) - en (ρ_{12}, ρ_{23}) -vlak.

LITERATUUR

GIBBONS, J.D., *Nonparametric Statistical Inference* (McGraw-Hill, New York, 1971).

KENDALL, M.G., *Rank Correlation Methods*, 2nd ed. (Charles Griffin, London, 1955).

KENDALL, M.G. & A. STUART, *The Advanced Theory of Statistics*, vol. 1 (3 vol. ed.) (Charles Griffin, London, 1958).

MIRSKY, L., *An introduction To Linear Algebra* (Clarendon, Oxford, 1955).

MORAN, P.A.P., *Rank correlation and product moment correlation*, *Biometrika* 35 (1948) 203-207.

ONTVANGEN 13 MEI 1976