MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

STATISTISCHE AFDELING

Leiding: Prof. Dr D. van Dantzig
Chef van de Statistische Consultatie: Prof. Dr J. Hemelrijk

Rapport S 105 (M 41)

Uittreksel en uitwerking van het artikel "A class of statistics with asymptotically normal distributions", door W.HOEFDING, Ann. Math. Stat. 19 (1948), pp. 293-325.

door

D.J.Stoker

1953

Een klasse van statistische grootheden met asymptotische normale verdeling met speciale verwijzing naar parametervrije toetsingsmethoden voor onafhankelijkheid.

1. Definitie van een U-statistische grootheid.

Stel X_ν, $\nu = 1, \ldots, n$ is een steekproef van n vectoren

$$(1) \qquad X_\nu = (X_\nu^{(1)}, \ldots, X_\nu^{(r)}),$$

en $\Phi(x_1, \ldots, x_m)$ een functie van m ($\leq n$) vectorargumenten. De functie van de steekproef

$$(2) \qquad U = U(X_1, \ldots, X_n) = \frac{1}{n(n-1)\cdots(n-m+1)} \sum'' \Phi(X_{\alpha_1}, \ldots, X_{\alpha_m}),$$

waar $\sum''$ een sommatie over alle permutaties $(\alpha_1, \ldots, \alpha_m)$ van gehele getallen aangeeft zodanig dat

$$(3) \qquad 1 \leq \alpha_i \leq n, \quad \alpha_i \neq \alpha_j \text{ as } i \neq j, \quad (i,j = 1, \ldots, n),$$

een U-statistische grootheid wordt genoemd. U is het gemiddelde van de waarden van Φ in de verzameling van gerangschikte deelverzamelingen van m elementen van de steekproef (1). U is symmetrisch in $X_1, \ldots, X_n$ (een functie $f = f(x_1, \ldots, x_n)$ wordt symmetrisch genoemd indien deze invariant is ten aanzien van alle permutaties van zijn argumenten).
Een functie $\Phi(x_1, \ldots, x_m)$ die aan (2) voldoet, wordt een kern van de statistische grootheid U genoemd.

Stel $\mathfrak{D}$ is een deelverzameling van alle verdelingsfuncties (v.f.'s) in de r-dimensionale ruimte waartoe $F = F(x) = F(x^{(1)}, \ldots, x^{(r)})$ behoort. $\Theta(F)$ wordt een functioneel van F gedefinieerd op $\mathfrak{D}$, genoemd indien aan iedere F die tot $\mathfrak{D}$ behoort, een grootheid $\Theta(F)$ wordt toegekend. Stel verder dat voor een functioneel $\Theta = \Theta(F)$ van $F(x)$ een steekproefgrootte n bestaat waarvoor een zuivere schatting van Θ voor iedere v.f. F in $\mathfrak{D}$ bestaat Dit wil zeggen, als $\underline{X}_1, \ldots, \underline{X}_n$ n onafhankelijke stochastische vectoren met dezelfde v.f. F zijn, dan bestaat er een functie $\Phi(x_1, \ldots, x_n)$ van n vectorargumenten (1) zodanig dat

$$(4) \qquad \mathcal{E}\, \Phi(\underline{X}_1, \ldots, \underline{X}_n) = \int_{R_n} \cdots \int \Phi(x_1, \ldots, x_n)\, dF(x_1) \cdots dF(x_n) = \Theta(F)$$

voor iedere F in $\mathfrak{D}$. $\Phi(x_1, \ldots, x_n)$ is dus een zuivere schatting van Θ op $\mathfrak{D}$. Een functioneel $\Theta(F)$ van de vorm (4) noemen wij regelmatig op $\mathfrak{D}$.

Stel $m (\leq n)$ is de kleinste steekproefgrootte waarvoor een zuivere schatting $\Phi(x_1, \ldots, x_m)$ van de regelmatige functioneel Θ op $\mathfrak{D}$ bestaat, d.w.z.

$$(5) \qquad \Theta(F) = \int \cdots \int \Phi(x_1, \ldots, x_m)\, dF(x_1) \cdots dF(x_m)$$

voor iedere F in $\mathcal{D}$, m wordt de <u>graad</u> van de regelmatige functioneel op $\mathcal{D}$ genoemd.
Iedere functie $\Phi(x_1,\ldots,x_m)$ die aan (5) voldoet, wordt de kern van de regelmatige functioneel $\Theta(F)$ genoemd.

$\Phi_0(x_1,\ldots,x_m)$ is een <u>symmetrische kern</u> in $x_1,\ldots,x_m$ van $\Theta(F)$ indien

$$(6)\qquad \Phi_0(x_1,\ldots,x_m)=\frac{1}{m!}\sum{}''\Phi(x_{\alpha_1},\ldots,x_{\alpha_m}),$$

waar de som Σ'' zich over alle permutaties $(\alpha_1,\ldots,\alpha_m)$ van $(1,\ldots,m)$ uitstrekt.

Voor $n=m$, reduceert de statistische grootheid U zich dus tot de symmetrische kern (6) van $\Theta(F)$.

Halmos [3] heeft zekere optimale eigenschappen van U-statistische grootheden als zuivere schattingen van regelmatige functionelen bewezen, n l. dat, wanneer $\Theta(F)$ een regelmatige functioneel van de m^e graad op een verzameling $\mathcal{D}$, die alle geheel en al discontinue verdelingsfuncties bevat, is, U dan de enige zuivere schatting op $\mathcal{D}$ is die symmetrisch in $x_1,\ldots,x_n$ is en U heeft de kleinste variantie van alle zuivere schattingen op $\mathcal{D}$, met andere woorden, U is de doeltreffendste schatting van $\Theta(F)$ op $\mathcal{D}$.

Uit (2) en (6) volgt nu dat wij een U-statistische grootheid in de volgende vorm kunnen schrijven:

$$(7)\qquad U(x_1,\ldots,x_n)=\binom{n}{m}^{-1}\sum{}'\Phi_0(x_{\alpha_1},\ldots,x_{\alpha_m})$$

daar de kern Φ_0 symmetrisch in zijn m vectorargumenten is en de som Σ' zich over alle indices α uitstrekt, zodanig dat

$$1\le\alpha_1<\alpha_2<\cdots<\alpha_m\le n.$$

2. <u>De variantie van een U-statistische grootheid.</u>

Stel $\underline{x}_1,\ldots,\underline{x}_n$ zijn n onafhankelijke stochastische vectoren met dezelfde v.f. $F(x)=F(x^{(1)},\ldots,x^{(r)})$ en stel

$$(8)\qquad \underline{U}=U(\underline{x}_1,\ldots,\underline{x}_n)=\binom{n}{m}^{-1}\Sigma'\Phi(\underline{x}_{\alpha_1},\ldots,\underline{x}_{\alpha_m})$$

daar $\Phi(x_1,\ldots,x_m)$ symmetrisch in $x_1,\ldots,x_m$ is en Σ' dezelfde betekenis als in (7) heeft. Stel dat de functie Φ n niet bevat. Nu is $\mathcal{E}\,\underline{U}=\mathcal{E}\,\Phi(\underline{x}_1,\ldots,\underline{x}_m)=\Theta=\Theta(F)$.

Stel

$$(9)\qquad \Phi_c(x_1,\ldots,x_c)=\mathcal{E}\,\Phi(x_1,\ldots,x_c,\underline{x}_{c+1},\ldots,\underline{x}_m)\qquad,(c=1,\ldots,m)$$

daar $x_1,\ldots,x_c$ willekeurig vaste vectoren zijn en de verwachting ten opzichte van de stochastische vectoren $\underline{x}_{c+1},\ldots,\underline{x}_m$ genomen

wordt.
Dan is

$$(10)\qquad \Phi_{c-1}(x_1,\cdots x_{c-1}) = \xi\ \Phi_c(x_1,\cdots,x_{c-1},\underline{x}_c),$$

en

$$(11)\qquad \xi\ \Phi_c(\underline{x}_1,\cdots,\underline{x}_c) = \Theta \qquad ,\ (c=1,\cdots,m)$$

Definieer nu

$$(12)\qquad \psi(x_1,\cdots,x_m) = \Phi(x_1,\cdots,x_m) - \Theta$$

$$(13)\qquad \psi_c(x_1,\cdots,x_c) = \Phi_c(x_1,\cdots,x_c) - \Theta \quad ,\ (c=1,\cdots,m).$$

Nu is ook

$$(14)\qquad \psi_{c-1}(x_1,\cdots,x_{c-1}) = \xi\ \psi_c(x_1,\cdots,x_{c-1},\underline{x}_c),$$

$$(15)\qquad \xi\ \psi_c(\underline{x}_1,\cdots,\underline{x}_c) = \xi\ \psi(\underline{x}_1,\cdots,\underline{x}_m) = 0 \quad ,\ (c=1,\cdots,m).$$

Veronderstel dat de variantie van $\psi_c(\underline{x}_1,\cdots,\underline{x}_c)$ bestaat en stel

$$(16)\qquad \jmath_0 = 0\ ,\quad \jmath_c = \xi\ \psi_c^2(\underline{x}_1,\cdots,\underline{x}_c) \quad ,\ (c=1,\cdots,m)$$

Nu is

$$\jmath_c = \xi\ \Phi_c^2(\underline{x}_1,\cdots,\underline{x}_c) - \Theta^2.$$

Als voor een zekere oorspronkelijke verdeling $F=F_0$ en een zeker geheel getal d $\jmath_d(F_0)=0$ is, betekent dit dat $\psi_d(\underline{x}_1,\cdots,\underline{x}_d)=0$ is met een waarschijnlijkheid 1. Uit (14) en (16) volgt dat $\jmath_d=0$ betekent dat $\jmath_1=\cdots=\jmath_{d-1}=0$ is.

Als $\jmath_1(F_0)=0$ is, is de regelmatige functioneel $\Theta(F)$ stationnair voor $F=F_0$. Als

$$(17)\qquad \jmath_1(F_0)=\cdots=\jmath_d(F_0)=0\ ,\quad \jmath_{d+1}(F_0)>0\ ,\quad (1\le d\le m)$$

is $\Theta(F)$ stationnair van orde d voor $F=F_0$.

Als $(\alpha_1,\cdots,\alpha_m)$ en $(\beta_1,\cdots,\beta_m)$ twee verzamelingen zijn bestaande uit m verschillende gehele getallen $1\le\alpha_i,\beta_i\le n$ en c het aantal gehele getallen is dat beide verzamelingen gemeen hebben, volgt uit de symmetrie van ψ dat

(18) $$\xi\{\psi(\underline{x}_{\alpha_1},\ldots,\underline{x}_{\alpha_m})\,\psi(\underline{x}_{\beta_1},\ldots,\underline{x}_{\beta_m})\} = \zeta_c$$

Als de variantie van $\underline{U}$ bestaat, dan is deze gelijk aan

$$\begin{aligned}\sigma^2(\underline{U}) &= \xi(\underline{U}^2) - \xi^2(\underline{U})\\ &= \binom{n}{m}^{-2}\xi\{\Sigma'\Phi(\underline{x}_{\alpha_1},\ldots,\underline{x}_{\alpha_m})\}^2 - \Theta^2\\ &= \binom{n}{m}^{-2}\xi\{\Sigma'\psi(\underline{x}_{\alpha_1},\ldots,\underline{x}_{\alpha_m})\}^2\\ &= \binom{n}{m}^{-2}\sum_{c=0}^{m}\sum^{(c)}\xi\{\psi(\underline{x}_{\alpha_1},\ldots,\underline{x}_{\alpha_m})\,\psi(\underline{x}_{\beta_1},\ldots,\underline{x}_{\beta_m})\}\end{aligned}$$

waar $\sum^{(c)}$ de sommatie over alle indices aanduidt, zodanig dat

$$1\le\alpha_1<\alpha_2<\cdots<\alpha_m\le n,\quad 1\le\beta_1<\cdots<\beta_m\le n$$

en precies aan c vergelijkingen

$$\alpha_i=\beta_j$$

wordt voldaan. Uit (18) volgt dat elke term in $\sum^{(c)}$ gelijk is aan ζ_c . Het aantal termen in $\sum^{(c)}$ is gelijk aan

$$\binom{m}{c}\binom{n-m}{m-c}\binom{n}{m} = \frac{n(n-1)\cdots\cdots(n-2m+c+1)}{c!\;(m-c)!\;(m-c)!}$$

(De m α's kunnen op $\binom{n}{m}$ manieren uit de getallen $1,2,\ldots,n$ gekozen worden. c van de m α's kunnen op $\binom{m}{c}$ manieren gekozen worden. Nu zijn c waarden van de β's gekozen en de overblijvende $m-c$ β's mogen niet gelijk zijn aan de m α's. Dus kunnen de overblijvende β's op $\binom{n-m}{m-c}$ manieren gekozen worden. Let goed op dat de α's en β's alle onderling verschillend zijn.)

Omdat $\zeta_0 = 0$ is, is

(19) $$\sigma^2(\underline{U}) = \binom{n}{m}^{-1}\sum_{c=1}^{m}\binom{m}{c}\binom{n-m}{m-c}\zeta_c$$

Een overeenkomstige formulering kan gemakkelijk afgeleid worden voor het geval dat de verdelingen $F_\nu(x)$ van $\underline{x}_\nu$, $\nu=1,\ldots,n$, verschillend zijn.

Hoeffding [5] bewijst vervolgens de volgende twee stellingen aangaande de grootheden $\zeta_1,\ldots,\zeta_m$ en $\sigma^2(\underline{U})$.

Stelling A.

De grootheden $\zeta_1,\ldots,\zeta_m$ zoals gedefinieerd in (16) voldoen aan de ongelijkheden

(20) $$0\le\frac{\zeta_c}{c}\le\frac{\zeta_d}{d}\qquad \text{als}\quad 1\le c<d\le m.$$

Stelling B.

De variantie $\sigma^2(\underline{U}_n)$ van een U-statistische grootheid $\underline{U}_n = U(\underline{x}_1,\ldots,\underline{x}_n)$, waar $\underline{x}_1,\ldots,\underline{x}_n$ onafhankelijk en identiek verdeeld is, voldoet aan de ongelijkheden

(21) $$\frac{m^2}{n}\zeta_1 \leq \sigma^2(\underline{U}_n) \leq \frac{m}{n}\zeta_m .$$

$n\sigma^2(\underline{U}_n)$ is een afnemende functie van n,

(22) $$(n+1)\sigma^2(\underline{U}_{n+1}) \leq n\sigma^2(\underline{U}_n) ,$$

die voor $n=m$ zijn bovenste grens $m\zeta_m$ aanneemt en naar zijn onderste grens $m^2\zeta_1$ nadert als n toeneemt:

(23) $$\sigma^2(\underline{U}_m) = \zeta_m ,$$

(24) $$\lim_{n\to\infty} n\sigma^2(\underline{U}_n) = m^2\zeta_1 .$$

Als $\xi\,\underline{U}_n = \theta(F)$ stationnair is van de orde $\geq d-1$ voor de v.f. van $\underline{x}_\alpha$, kan (21) vervangen worden door

(25) $$\frac{m}{d}\,K_n(m,d)\,\zeta_d \leq \sigma^2(\underline{U}_n) \leq K_n(m,d)\,\zeta_m$$

waarbij

(26) $$K_n(m,d) = \binom{n}{m}^{-1} \sum_{c=d}^{m} \binom{m-1}{c-1}\binom{n-m}{m-c} .$$

<u>Opmerking</u>: Uit (19) en (20) volgt dat een nodige en voldoende voorwaarde voor het bestaan van $\sigma^2(\underline{U})$ het bestaan van

(27) $$\zeta_m = \xi\left\{\Phi(\underline{x}_1,\cdots,\underline{x}_m)\right\}^2 - \theta^2$$

of van $\xi\left\{\Phi^2(\underline{x}_1,\cdots,\underline{x}_m)\right\}$ is.

Voor $\zeta_1 > 0$ is $\sigma^2(\underline{U}) = O(n^{-1})$ (zie o.a. verg. (21)).

Als $\theta(F)$ stationnair is van de d^{e} orde voor $F=F_0$, d.w.z. als aan (17) voldaan is, dan is $\sigma^2(\underline{U})$ van de orde $n^{-(d+1)}$. Alleen wanneer, voor een zekere $F=F_0$, $\theta(F)$ stationnair is van de m^{e} orde, waarbij m de graad van $\theta(F)$ is, krijgen wij $\sigma^2(\underline{U})=0$ en $\underline{U}$ is dan gelijk aan een constante (spr 0).

3. De covariantie van twee U-statistische grootheden.

Beschouw een verzameling van g U-statistische grootheden

$$\underline{U}^{(\gamma)} = \binom{n}{m(\gamma)}^{-1} \Sigma' \Phi^{(\gamma)}(\underline{x}_{\alpha_1},\cdots,\underline{x}_{\alpha_{m(\gamma)}}) , \quad (\gamma=1,\cdots,g),$$

waar $\underline{U}^{(\gamma)}$ een functie van dezelfde n onafhankelijke gelijk verdeelde stochastische vectoren $\underline{x}_1,\cdots,\underline{x}_n$ is. Uit de veronderstelling volgt dat $\Phi^{(\gamma)}$ symmetrisch in zijn $m(\gamma)$ argumenten $(\gamma=1,\ldots,g)$ is.

Stel

$$\xi\left\{\underline{U}^{(\gamma)}\right\} = \xi\left\{\Phi^{(\gamma)}(\underline{x}_1,\ldots,\underline{x}_{m(\gamma)})\right\} = \theta^{(\gamma)} \qquad (\gamma=1,\ldots,g)$$

(28) $$\psi^{(\gamma)}(x_1,\ldots,x_{m(\gamma)}) = \Phi^{(\gamma)}(x_1,\ldots,x_{m(\gamma)}) - \theta^{(\gamma)} \qquad (\gamma=1,\ldots,g)$$

(29) $$\psi_c^{(\gamma)}(x_1,\ldots x_c) = \mathcal{E}\{\psi^{(\gamma)}(x_1,\ldots,x_c,\underline{x}_{c+1},\ldots,\underline{x}_m)\}$$

$$(c=1,\ldots,m(\gamma);\ \gamma=1,\ldots,g)$$

(30) $$\zeta_c^{(\gamma,\delta)} = \mathcal{E}\{\psi_c^{(\gamma)}(\underline{x}_1,\ldots,\underline{x}_c)\,\psi_c^{(\delta)}(\underline{x}_1,\ldots,\underline{x}_c)\} \qquad (\gamma,\delta=1,\ldots,g)$$

Voor $\gamma=\delta$ is

(31) $$\zeta_c^{(\gamma)} = \zeta_c^{(\gamma\gamma)} = \mathcal{E}\{\psi_c^{(\gamma)}(\underline{x}_1,\ldots,\underline{x}_c)\}^2$$

Stel $\sigma(\underline{U}^{(\gamma)},\underline{U}^{(\delta)}) = \mathcal{E}\{(\underline{U}^{(\gamma)}-\theta^{(\gamma)})(\underline{U}^{(\delta)}-\theta^{(\delta)})\}$

is de covariantie van $\underline{U}^{(\gamma)}$ en $\underline{U}^{(\delta)}$.

Soortgelijk als voor de variantie, wordt voor $m(\gamma)\leq m(\delta)$ gevonden

(32) $$\sigma(\underline{U}^{(\gamma)},\underline{U}^{(\delta)}) = \binom{n}{m(\gamma)}^{-1}\sum_{c=1}^{m(\gamma)}\binom{m(\delta)}{c}\binom{n-m(\delta)}{m(\gamma)-c}\zeta_c^{(\gamma,\delta)}$$

De rechterkant is symmetrisch in γ en δ.

Voor $\gamma=\delta$ is (32) de variantie van $\underline{U}^{(\gamma)}$ (zie (19)).

Uit (24) en (32) volgt dat

$$\lim_{n\to\infty} n\sigma^2(\underline{U}^{(\gamma)}) = m^2(\gamma)\,\zeta_1^{(\gamma)}$$

$$\lim_{n\to\infty} n\sigma(\underline{U}^{(\gamma)},\underline{U}^{(\delta)}) = m(\gamma)m(\delta)\,\zeta_1^{(\gamma,\delta)}$$

Als dus $\zeta_1^{(\gamma)}\neq 0$ en $\zeta_1^{(\delta)}\neq 0$, nadert de productmoment-correlatie $\rho(\underline{U}^{(\gamma)},\underline{U}^{(\delta)})$ tussen $\underline{U}^{(\gamma)}$ en $\underline{U}^{(\delta)}$ naar de limiet

(33) $$\lim_{n\to\infty}\rho(\underline{U}^{(\gamma)},\underline{U}^{(\delta)}) = \frac{\zeta_1^{(\gamma,\delta)}}{\sqrt{\zeta_1^{(\gamma)}\,\zeta_1^{(\delta)}}}$$

4. Limietstellingen voor het geval van gelijk verdeelde $\underline{x}_\nu$'s.

Hoeffding bewijst de volgende vijf stellingen in verband met de asymptotische verdeling van $\underline{U}$-statistische grootheden en zekere aanverwante functies.

Stelling 1.1:

Stel $\underline{x}_1,\ldots,\underline{x}_n$ zijn n onafhankelijke, gelijk verdeelde stochastische vectoren

$$\underline{x}_\alpha = \left(\underline{x}_\alpha^{(1)}, \ldots, \underline{x}_\alpha^{(k)}\right) \qquad (\alpha = 1, \ldots, n)$$

Stel verder

$$\Phi^{(\gamma)}(x_1, \ldots, x_{m(\gamma)}) \qquad (\gamma = 1, \ldots, g)$$

zijn g reële-waarde functies die n niet bevatten en waarbij $\Phi^{(\gamma)}$ symmetrisch in zijn $m(\gamma)$ $(\leqq n)$ vectorargumenten $x_\alpha = (x_\alpha^{(1)}, \ldots, x_\alpha^{(\gamma)})$ $(\alpha = 1, \ldots, m(\gamma);\ \gamma = 1, \ldots, g)$ is. Definieer nu

$$\underline{U}^{(\gamma)} = \binom{n}{m(\gamma)}^{-1} \sum{}' \Phi^{(\gamma)}(\underline{x}_{\alpha_1}, \ldots, \underline{x}_{\alpha_{m(\gamma)}}), \quad (\gamma = 1, \ldots, g)$$

waarbij de sommatie zich over alle indices uitstrekt zodanig dat $1 \leqq \alpha_1 < \ldots < \alpha_{m(\gamma)} \leqq n$. Indien de verwachtingen

$$\theta^{(\gamma)} = \mathcal{E}\left\{\Phi^{(\gamma)}(\underline{x}_1, \ldots, \underline{x}_{m(\gamma)})\right\} \qquad (\gamma = 1, \ldots, g)$$

en

$$\mathcal{E}\left\{\Phi^{(\gamma)}(\underline{x}_1, \ldots, \underline{x}_{m(\gamma)})\right\}^2 \qquad (\gamma = 1, \ldots, g)$$

bestaan, dan nadert de simultane verdelingsfunctie van

$$\sqrt{n}\,(\underline{U}^{(1)} - \theta^{(1)}), \ldots, \sqrt{n}\,(\underline{U}^{(g)} - \theta^{(g)})$$

voor $n \to \infty$, tot de g-dimensionale normale verdelingsfunctie met gemiddelde 0 en $\left[m(\gamma)m(\delta)\,\zeta_1^{(\gamma,\delta)}\right]$, waarbij $\zeta_1^{(\gamma,\delta)}$ door (30) gedefinieerd is, als covariantiematrix. De limietverdeling is niet singulier als de determinant $\left|\zeta_1^{(\gamma,\delta)}\right| > 0$.

Opmerkingen: Voor $g = 1$ volgt uit de stelling dat de verdeling van een U-statistische grootheid onder zekere voorwaarden tot de normale vorm nadert. Voor $m = 1$ is U de som van n onafhankelijke stochastische veranderlijken en is stelling 1.1 dus de Centrale Limietstelling voor zulke sommen. Voor $m > 1$ is $\underline{U}$ de som van stochastische veranderlijken die in het algemeen niet onafhankelijk zijn.

De stelling toont aan dat in het geval van onderling onafhankelijke en gelijk verdeelde $\underline{x}_\alpha$'s, het bestaan van $\mathcal{E}\left\{\Phi^2(\underline{x}_1, \ldots, \underline{x}_m)\right\}$ voldoende voor de asymptotische normaliteit van $\underline{U}$ is. Aan de functie Φ worden geen regelmatigheidsvoorwaarden opgelegd.

Volgens stelling B overschrijdt $\sigma^2(\underline{U})$ zijn asymptotische waarde $\frac{m^2}{n}\zeta_1$ voor iedere eindige n . Als wij dus stelling 1.1 gebruiken als een benadering van de verdeling van $\underline{U}$ voor n groot maar eindig, dan onderschatten wij de variantie van $\underline{U}$. In vele toepassingen is dit ongewenst en moeten wij liever de volgende stelling gebruiken:

Stelling 1.2:

Onder de voorwaarden van stelling 1.1 en als

$$\zeta_1^{(\gamma)} > 0 \quad , \qquad (\gamma = 1, \ldots, g)$$

nadert de simultane v.f. van

$$\frac{U^{(1)} - \theta^{(1)}}{\sigma(U^{(1)})} , \ldots\ldots , \frac{U^{(g)} - \theta^{(g)}}{\sigma(U^{(g)})}$$

voor $n \to \infty$ tot de g-dimensionale normale v.f. met gemiddelde 0 en $[\rho^{(\gamma,\delta)}]$ als covariantiematrix, waarbij

$$\rho^{(\gamma,\delta)} = \lim_{n\to\infty} \frac{\sigma(U^{(\gamma)}, U^{(\delta)})}{\sigma(U^{(\gamma)})\,\sigma(U^{(\delta)})} = \frac{\zeta_1^{(\gamma,\delta)}}{\sqrt{\zeta_1^{(\gamma)} \cdot \zeta_1^{(\delta)}}} ,$$

$$(\gamma, \delta = 1, \ldots\ldots, g)$$

Opmerking: De stellingen (1.1) en (1.2) betreffen dus de asymptotische verdeling van $U^{(1)}, \ldots\ldots, U^{(g)}$ die zuivere schattingen zijn van $\theta^{(1)}, \ldots\ldots, \theta^{(g)}$. De zuiverheid van een statistische grootheid speelt natuurlijk geen rol bij het asymptotisch gedrag. Stelling (1.1) kan uitgebreid worden tot een grotere klasse van statistische grootheden, vandaar

Stelling 1.3:

Stel $$U^{(\gamma)'} = U^{(\gamma)} + \frac{b_n^{(\gamma)}}{\sqrt{n}} , \qquad (\gamma = 1, \ldots, g)$$

waarbij $U^{(\gamma)}$ in stelling (1.1) gedefinieerd is en $b_n^{(\gamma)}$ een aselecte veranderlijke is. Indien aan de voorwaarden van stelling (1.1) is voldaan en $\lim_{n\to\infty} \mathcal{E}\{b_n^{(\gamma)2}\} = 0$, $(\gamma = 1, \ldots, g)$ is,

dan nadert de simultane verdeling van

$$\sqrt{n}\,(U^{(1)'} - \theta^{(1)}), \ldots\ldots, \sqrt{n}\,(U^{(g)'} - \theta^{(g)})$$

tot de normale verdeling met gemiddelde 0 en $\{m(\gamma)\,m(\delta)\,\zeta_1^{(\gamma,\delta)}\}$ als covariantiematrix.

Stelling 1.4:

Stel $x_1, \ldots, x_n$ is een aselecte steekproef uit een $\varkappa$-dimensionale verzameling met $F(x) = F(x^{(1)}, \ldots\ldots\ x^{(\varkappa)})$ als v.f. en stel

$$\theta^{(\gamma)}(F) = \int \ldots\ldots \int \Phi^{(\gamma)}(x_1, \ldots\ldots, x_{m(\gamma)})\, dF(x_1) \ldots\ldots dF(x_{m(\gamma)})$$

$$(\gamma = 1, \ldots\ldots, g)$$

zijn regelmatige functionels van F, waarbij $\Phi^{(\gamma)}(x_1, \ldots\ldots, x_{m(\gamma)})$ symmetrisch in de vectoren $x_1, \ldots\ldots, x_{m(\gamma)}$ is en n niet bevat.
Als nu $S(x)$ de v.f. van de aselecte steekproef is, en indien de variantie van

$$\theta^{(\gamma)}(\underline{S}) = \frac{1}{n^m} \sum_{\alpha_1=1}^{n} \cdots \sum_{\alpha_{m(\gamma)}=1}^{n} \phi^{(\gamma)}(\underline{x}_{\alpha_1}, \ldots, \underline{x}_{\alpha_{m(\gamma)}})$$

bestaat, dan nadert de simultane v.f. van

$$\sqrt{n}\{\theta^{(1)}(\underline{S}) - \theta^{(1)}(F)\}, \ldots, \sqrt{n}\{\theta^{(g)}(\underline{S}) - \theta^{(g)}(F)\}$$

tot de g-dimensionale normale v.f. met gemiddelde 0 en $\{m(\gamma)\, m(\delta)\, \zeta_1^{(\gamma,\delta)}\}$ als covariantiematrix.

Stelling 1.5:

Stel $(\underline{U}') = (\underline{U}^{(1)'}, \ldots, \underline{U}^{(g)'})$ is een stochastische vector, waarbij $\underline{U}^{(\gamma)'}$ in stelling (1.3) gedefinieerd is en veronderstel dat aan de voorwaarden van stelling (1.3) voldaan is. Indien de functie $h(y) = h(y^{(1)}, \ldots, y^{(g)})$, nniet bevat en evenals de partiële afgeleiden van de 2^e orde in de omgeving van het punt $(y) = (\theta) = (\theta^{(1)}, \ldots, \theta^{(g)})$ continu is, dan nadert de verdeling van de stochastische veranderlijke $\sqrt{n}\{h(\underline{U}') - h(\theta)\}$ tot de normale verdeling met gemiddelde 0 en

$$\sum_{\gamma=1}^{g} \sum_{\delta=1}^{g} m(\gamma)\, m(\delta) \left(\frac{\partial h(y)}{\partial y^{(\gamma)}}\right)_{y=\theta} \left(\frac{\partial h(y)}{\partial y^{(\delta)}}\right)_{y=\theta} \zeta_1^{(\gamma,\delta)}$$

als variantie.

Opmerking: Stelling (1.5) is dus een uitbreiding van de vorige stellingen en betreft de asymptotische verdeling van een functie van $\underline{U}$- of $\underline{U}'$-statistische grootheden.

Aangezien ieder moment waaruit de variantie bestaat in de vorm $\underline{U}' = \theta(\underline{S})$ geschreven kan worden (zie par. 5 A en stelling (1.4)), is stelling (1.5) een uitbreiding van de stelling over een functie van momenten.

De hierbovengenoemde limietstellingen kunnen uitgebreid worden tot het geval waarbij de $\underline{x}_\alpha$'s verschillende verdelingen hebben. Hoeffding [5] zet uiteen hoe de uitbreidingen kunnen geschieden.

5. Speciale gevallen van U-statistische grootheden.

A. Momenten en functies van momenten.

Stel $S = S(x)$ is de verdelingsfunctie van de steekproef (1). Substitutie van S voor F in (5) geeft

$$\theta(S) = \frac{1}{n^m} \sum_{\alpha_1=1}^{n} \cdots \sum_{\alpha_m=1}^{n} \phi(x_{\alpha_1}, \ldots, x_{\alpha_m}) \tag{A.1}$$

Als $m=1$ is $\theta(S) = U$. Als $m=2$, dan is

$$\theta(S) = \frac{1}{n^2} \sum_{\alpha_1=1}^{n} \sum_{\alpha_2=1}^{n} \Phi(x_{\alpha_1}, x_{\alpha_2})$$

$$= \frac{1}{n^2} \sum_{\substack{\alpha_1=1 \\ \alpha_1 \neq \alpha_2}}^{n} \sum_{\alpha_2=1}^{n} \Phi(x_{\alpha_1}, x_{\alpha_2}) + \frac{1}{n} \left[\frac{1}{n} \sum_{\alpha=1}^{n} \Phi(x_\alpha, x_\alpha) \right]$$

$$= \frac{n-1}{n} U + \frac{1}{n} \left\{ \frac{1}{n} \sum_{\alpha=1}^{n} \Phi(x_\alpha, x_\alpha) \right\},$$

en $\theta(S)$ is een lineaire functie van U-statistische grootheden met coëfficiënten die van n afhangen. Dit geldt voor iedere m. In het algemeen is $\theta(S)$ echter geen zuivere schatting van $\theta(F)$. Indien de verwachting van $\theta(S)$ echter bestaat voor iedere F in $\mathcal{D}$, dan is

$$\mathcal{E}\{\theta(S)\} = \theta(F) + O(n^{-1}) \qquad \text{(zie ook (4))}$$

en dus kan de schatting $\theta(S)$ van $\theta(F)$ asymptotisch zuiver over $\mathcal{D}$ genoemd worden.

De steekproefmomenten hebben de vorm (A.1). Hieruit volgt dus dat de zuivere schattingen van de momenten U-statistische grootheden zijn.

Uit de stellingen (1.1), (1.2) en (1.4) volgt dat de steekproefmomenten asymptotisch normaal verdeeld zijn en uit stelling (1.5) volgt hetzelfde voor een functie van die momenten indien aan de vereiste voorwaarden voldaan is. Dit resultaat is bekend (zie Cramér: Math. Methods of Statistics).

De steekproefmomenten hebben de vorm (A.1). Hun kern Φ kan als volgt verkregen worden:

Momenten t.o.v. de oorsprong:

$$\mu'_{\nu_1,\ldots,\nu_r} = \int \cdots \int (x^{(1)})^{\nu_1} \cdots (x^{(r)})^{\nu_r} \, dF(x_1^{(1)}, \ldots, x^{(r)})$$

Momenten t.o.v. het gemiddelde: Een moment t.o.v. het gemiddelde is een veelterm in momenten μ' t.o.v. de oorsprong. Voor $r=1$ (d.w.z. in het geval van een ééndimensionale verdelingsfunctie) is

$$\sigma^2 = \int (x-\mu)^2 \, dF(x)$$

waarbij μ het gemiddelde van de oorspronkelijke verzameling is. Geef met $\hat{\sigma}^2$ de geschatte waarde uit de steekproef van σ^2 aan. Dan is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \text{waarbij} \qquad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Nu is $\mathcal{E}\left\{\frac{n}{n-1}\hat{\sigma}^2\right\}=\sigma^2$ zoals bekend is.

Het is gemakkelijk om k-statistische grootheden te verkrijgen, die parametervrije zuivere schattingen van de momenten t.o.v. het gemiddelde van enige orde zijn (zie Halmos [3]) [1]. $\Phi(x_1,\ldots,x_n)$ wordt een parametervrije zuivere schatting van $\theta(F)$ genoemd indien $\mathcal{E}\left(\Phi(\underline{x}_1,\ldots,\underline{x}_n)\right)=\theta(F)$ wanneer $\Phi(\underline{x}_1,\ldots,\underline{x}_n)$ ook al bestaan, waarbij $\underline{x}_1,\ldots,\underline{x}_n$ n onafhankelijke stochastische vectoren zijn.

Ook de k-statistische grootheden van Fisher zijn k-statistische grootheden. Dit volgt ook uit hun definitie als zuivere schattingen van de cumulanten, symmetrisch in de steekproefwaarden (Kendall I, blz. 256). Zij zijn dus ook asymptotisch normaal verdeeld.

B. Gemiddeld verschil en concentratie-coëfficiënt.

Als $y_1,\ldots,y_n$ n onafhankelijke, reële, aselecte veranderlijken zijn, dan wordt het gemiddelde verschil van Gini (zonder herhaling) gedefinieerd door

$$\underline{d}=\frac{1}{n(n-1)}\sum_{\alpha\neq\beta}|y_\alpha-y_\beta|$$

Indien de y_α's allemaal dezelfde verdeling F hebben, is het gemiddelde van $\underline{d}$

$$\delta=\mathcal{E}(\underline{d})=\int\int|y_1-y_2|\,dF(y_1)\,dF(y_2)$$

---

1) Als de steekproefmomenten van de m^e orde gedefinieerd worden door

$$g_m(x_1,\ldots,x_n)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^m$$

waar $\bar{x}=\frac{1}{n}\sum_i x_i$, dan bewijst Halmos dat, als $f_m(x_1,\ldots,x_n)$ zuivere schattingen van de momenten t.o.v. het gemiddelde van de oorspronkelijke verzameling zijn, deze gegeven worden door

$$f_1(x_1,\ldots,x_n)=g_1(x_1,\ldots,x_n)$$
$$f_2(x_1,\ldots,x_n)=\frac{n}{n-1}g_2(x_1,\ldots,x_n)$$
$$f_3(x_1,\ldots,x_n)=\frac{n^2}{(n-1)(n-2)}g_3(x_1,\ldots,x_n)$$

en voor $m>3$ kan f_m nog in termen van de g_m's uitgedrukt worden, maar niet meer als een constant veelvoud van de g_m's. In het algemeen is f_m een lineaire combinatie van $g_1,\ldots,g_m$ met rationale coëfficiënten, waarvan de noemers $(n-1)(n-2)\ldots\ldots\ldots(n-m+1)$ zijn.

en de variantie van $\underline{d}$ is

$$\sigma^2(\underline{d}) = \binom{n}{m}^{-1} \sum_{l=1}^{m} \binom{m}{l}\binom{n-m}{m-l} \gamma_l \quad \text{waarbij } m=2 \text{ is,}$$

$$= \binom{n}{2}^{-1}\left\{2(n-2)\gamma_1(\delta) + \gamma_2(\delta)\right\}$$

$$= \frac{2}{n(n-1)}\left\{2(n-2)\gamma_1(\delta) + \gamma_2(\delta)\right\}$$

waar

$$\gamma_1(\delta) = \mathcal{E}\left\{\Psi_1^2(\underline{y}_1|\delta)\right\}$$

(B.1)
$$= \mathcal{E}\left\{\Phi_1^2(\underline{y}_1|\delta)\right\} - \delta^2$$

$$= \int\left\{\int|y_1-y_2|\,\alpha F(y_2)\right\}^2 \alpha F(y_1) - \delta^2$$

en

$$\gamma_2(\delta) = \iint(|y_1-y_2|)^2\,\alpha F(y_1)\,\alpha F(y_2) - \delta^2$$

(B.2)
$$= \iint(y_1-y_2)^2\,\alpha F(y_1)\,\alpha F(y_2) - \delta^2 = 2\sigma^2(\underline{y}) - \delta^2.$$

De notatie $\gamma_1(\delta)$, $\gamma_2(\delta)$ dient om het verband tussen de functionels van F en de functionel $\delta(F)$ aan te duiden; δ is alleen het symbool van de functionel en niet een bepaalde waarde daarvan. Soortgelijk is $\Phi(y_1, y_2|\delta) = |y_1 - y_2|$.

Nair [10] heeft $\sigma^2(\underline{d})$ voor verscheidene speciale verdelingen berekend.

Volgens stelling (1.1) is $\sqrt{n}(\underline{d}-\delta)$ asymptotisch normaal indien $\gamma_2(\delta)$ bestaat.

De concentratiecoëfficiënt van Gini wordt voor niet-negatieve waarden van $\underline{y}_1, \ldots, \underline{y}_n$ gedefinieerd door

$$\underline{G} = \frac{\underline{d}}{2\underline{\bar{y}}}$$

waarbij $\underline{\bar{y}} = \frac{\sum \underline{y}_\alpha}{n}$. $\underline{G}$ is hier een functie van twee $\underline{U}$-statistische grootheden. Indien de $\underline{y}_\alpha$'s gelijk verdeeld zijn, indien $\mathcal{E}\{\underline{y}^2\}$ bestaat en indien $\mu = \mathcal{E}\{\underline{y}\} > 0$, dan volgt uit stelling (1.5) dat $\sqrt{n}\left(\underline{G} - \frac{\delta}{2\mu}\right)$ asymptotisch normaal is met gemiddelde 0 en variantie

$$\sum_{\gamma=\delta,\mu}\sum_{\lambda=\delta,\mu} m(\gamma)m(\lambda)\left(\frac{\partial\left(\frac{\alpha}{2\bar{y}}\right)}{\partial\gamma}\Big/_{\substack{\alpha=\delta\\ \bar{y}=\mu}}\right)\left(\frac{\partial\left(\frac{\alpha}{2\bar{y}}\right)}{\partial\lambda}\Big/_{\substack{\alpha=\delta\\ \bar{y}=\mu}}\right)\gamma_1(\gamma,\lambda)$$

waarbij $m(\delta) = 2$ en $m(\mu) = 1$,

$$= \frac{1}{\mu^2}\gamma_1(\delta) - \frac{\delta}{\mu^3}\gamma_1(\delta,\mu) + \frac{\delta^2}{4\mu^4}\gamma_1(\mu)$$

waarbij $\gamma_1(\delta)$ gegeven wordt door (B.1) en

$$\gamma_1(\mu) = \int y^2 \, dF(y) - \mu^2 = \sigma^2(y)$$

$$\gamma_1(\mu, \delta) = \iint y_1 |y_1 - y_2| \, dF(y_1) \, dF(y_2) - \mu\delta.$$

C. Functies van rangen en tekens van verschillen tussen veranderlijken.

Definieer $S(u)$ door

$$S(u) = \begin{cases} -1 & < \\ 0 & \text{als } u = 0 \\ 1 & > \end{cases} \tag{C.1}$$

en stel

$$C(u) = \tfrac{1}{2}\{1 + S(u)\} = \begin{cases} 0 & < \\ \tfrac{1}{2} & \text{als } u = 0 \\ 1 & > \end{cases} \tag{C.2}$$

Als

$$x_\alpha = \left(x_\alpha^{(1)}, \ldots, x_\alpha^{(r)} \right), \qquad (\alpha = 1, \ldots, n)$$

een steekproef van n vectoren met r componenten elk is, definiëren wij de rang $R_\alpha^{(i)}$ van $x_\alpha^{(i)}$ door

$$\begin{aligned} R_\alpha^{(i)} &= \tfrac{1}{2} + \sum_{\beta=1}^{n} C\left(x_\alpha^{(i)} - x_\beta^{(i)}\right) \qquad (i = 1, \ldots, r) \\ &= \frac{n+1}{2} + \tfrac{1}{2} \sum_{\beta=1}^{n} S\left(x_\alpha^{(i)} - x_\beta^{(i)}\right). \end{aligned} \tag{C.3}$$

Indien de getallen $x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}$ allemaal verschillend zijn, heeft het kleinste de rang 1, of op één na het kleinste de rang 2, enz. Als sommige gelijk zijn, wordt de gemiddelde rang door (C.3) bepaald.

Iedere functie van de rang is een functie van uitdrukkingen $C\left(x_\alpha^{(i)} - x_\beta^{(i)}\right)$ of $S\left(x_\alpha^{(i)} - x_\beta^{(i)}\right)$.

Omgekeerd, aangezien $S\left(x_\alpha^{(i)} - x_\beta^{(i)}\right) = S\left(R_\alpha^{(i)} - R_\beta^{(i)}\right)$, is iedere functie van uitdrukkingen $S\left(x_\alpha^{(i)} - x_\beta^{(i)}\right)$ of $C\left(x_\alpha^{(i)} - x_\beta^{(i)}\right)$ een functie van de rang.

Beschouw een regelmatige functionel $\theta(F)$ waarvan de kern $\phi(x_1, \ldots, x_m)$ alleen afhangt van de tekens van de verschillen tussen de veranderlijken

$$S\left(x_\alpha^{(i)} - x_\beta^{(i)}\right) \qquad (\alpha, \beta = 1, \ldots, m;\ i = 1, \ldots, r) \tag{C.4}$$

De overeenkomstige $\underline{C}$-statistische grootheid is een functie van de rang van de steekproefveranderlijken.

De functie Φ kan alleen een eindig aantal waarden $C_1, \ldots, C_N$ aannemen. Als $\pi_i = P\{\Phi = C_i\}$, $(i = 1, \ldots, N)$, krijgen wij

$$\theta = C_1 \pi_1 + \ldots + C_N \pi_N \quad , \quad \sum_{i=1}^{N} \pi_i = 1 \; .$$

π_i is een regelmatige functionel waarvan de kern $\Phi_i(x_1, \ldots, x_m)$ gelijk is aan 1 of 0 al naar gelang $\Phi = C_i$ of $\neq C_i$. Dus is

$$\Phi = C_1 \Phi_1 + \ldots + C_N \Phi_N \; .$$

Wil $\theta(F)$ bestaan, dan moet C_i eindig zijn en dus is Φ begrensd. Dus bestaat $\mathcal{E}\{\Phi^2\}$ en als $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ gelijk verdeeld zijn, nadert de v.f. van $\sqrt{n}(\underline{U} - \theta)$ volgens stelling (1.1) tot een normale v.f. die niet-singulier is voor $\zeta > 0$.

Een voorbeeld van een zodanig functionel is

D. De correlatie tussen tekens van verschillen.

Beschouw de tweedimensionale steekproef

(D.1) $$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

$$\left[\text{of } (x_1^{(1)}, x_1^{(2)}), (x_2^{(1)}, x_2^{(2)}), \ldots, (x_n^{(1)}, x_n^{(2)})\right] .$$

Met elke twee waarnemingen van deze steekproef komen er een paar tekens van verschillen tussen de desbetreffende veranderlijken overeen:

(D.2) $$S(x_\alpha - x_\beta) \; , \; S(y_\alpha - y_\beta) \; . \quad (\alpha \neq \beta \; ; \alpha, \beta = 1, \ldots, n)$$

(D.2) is een verzameling van $n(n-1)$ paren van tekens van verschillen.

Omdat $$\sum_{\alpha \neq \beta} S(x_\alpha - x_\beta) = 0 = \sum_{\alpha \neq \beta} S(y_\alpha - y_\beta)$$

is de covariantie t van de tekens van de verschillen (D.2)

(D.3) $$t = \frac{1}{n(n-1)} \sum_{\alpha \neq \beta} S(x_\alpha - x_\beta) \, S(y_\alpha - y_\beta)$$

of $t = \frac{S}{n(n-1)}$, waarbij S de S van Kendall [7] is.

We noemen t de verschiltekencovariantie van de steekproef (D.1).

Als alle x's en alle y's verschillend zijn, dan is

$$\sum_{\alpha \neq \beta} S^2(x_\alpha - x_\beta) = n(n-1) = \sum_{\alpha \neq \beta} S^2(y_\alpha - y_\beta)$$

en is t de productmomentcorrelatie van de tekens van de verschillen. t is een lineaire functie van het (minimum) aantal

inversies in de permutatie van de rang van x en y (Kendall [7] blz. 8).

$\underline{t}$ is een $\underline{U}$-statistische grootheid. $\underline{t}$ is een functie van een aselecte steekproef uit een tweedimensionale verzameling en is dus een zuivere schatting van de regelmatige functionel van de tweede graad

$$\tau = \iiiint S(x_1 - x_2)\, S(y_1 - y_2)\, \alpha F(x_1, y_1)\, \alpha F(x_2, y_2) \tag{D.4}$$

τ is de covariantie van de tekens van de verschillen van de overeenkomstige componenten van $(\underline{x}_1, \underline{y}_1)$ en $(\underline{x}_2, \underline{y}_2)$ in de verzameling van paren van onafhankelijke vectoren $(\underline{x}_1, \underline{y}_1), (\underline{x}_2, \underline{y}_2)$ met gelijke v.f. $F(x, y)$. Als $F(x, y)$ continu is, is τ de product-momentcorrelatie van de tekens van de verschillen.

Stel nu

$$\overline{F}(x, y) = \tfrac{1}{4}\{F(x-0, y-0) + F(x-0, y+0) + F(x+0, y-0) + F(x+0, y+0)\}, \tag{D.5}$$

dan is

$$\Phi_1(x, y \mid \tau) = \mathcal{E}\{\Phi_1(x_1, y_1; \underline{x}_2, \underline{y}_2)\} \tag{D.6}$$

$$= \iint S(x_1 - x_2)\, S(y_1 - y_2)\, \alpha F(x_2, y_2)$$

$$= \int_{\substack{x_2 < x_1 \\ y_2 < y_1}} \alpha F - \int_{\substack{x_2 > x_1 \\ y_2 < y_1}} \alpha F - \int_{\substack{x_2 < x_1 \\ y_2 > y_1}} \alpha F + \int_{\substack{x_2 > x_1 \\ y_2 > y_1}} \alpha F$$

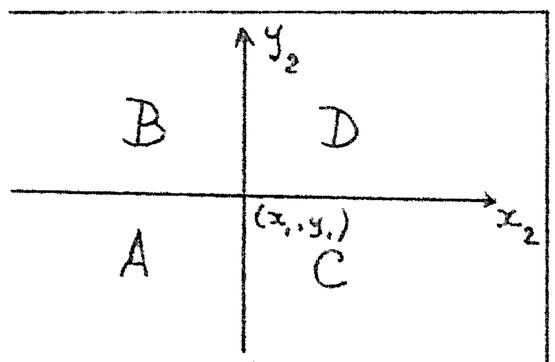
$$= F(x_1 - 0, y_1 - 0) - [F(\infty, y_1 - 0) - F(x_1 + 0, y_1 - 0)] +$$

$$- [F(x_1 - 0, \infty) - F(x_1 - 0, y_1 + 0)] +$$

$$+ [1 - F(\infty, y_1 + 0) - F(x_1 + 0, \infty) + F(x_1 + 0, y_1 + 0)]$$

$$= 1 - 2\,\overline{F}(x, \infty) - 2\,\overline{F}(\infty, y) + 4\,\overline{F}(x, y)$$

(uit D.5)



$A = F(x_1, y_1)$

of $F(x_1 - 0, y_1 - 0)$ in het discontinue geval.

$A + B = F(x_1, \infty)$

$(A + C) = F(\infty, y_1)$

$A + B + C + D = 1.$

Verder is $\tau = \mathcal{E}\{\Phi_1(\underline{x}_1, \underline{y}_1 \mid \tau)\}$

Nu is de variantie van $\underline{t}$

$$\sigma^2(t) = \binom{n}{2}^{-1} \sum_{c=1}^{2} \binom{2}{c}\binom{n-2}{2-c} \zeta_c(\tau) = \frac{2}{n(n-1)}\left[2(n-2)\zeta_1(\tau) + \zeta_2(\tau)\right] \tag{D.7}$$

waarbij

$$\zeta_1(\tau) = \mathcal{E}\{\Phi_1^2(\underline{x}_1, \underline{y}_1 \mid \tau)\} - \tau^2 \tag{D.8}$$

(D.9) $$\zeta_2(\tau) = \mathcal{E}\left\{S^2(\underline{x}_1 - \underline{x}_2)\, S^2(\underline{y}_1 - \underline{y}_2)\right\} - \tau^2$$

Als $F(x, y)$ <u>continu</u> is, dan is

$$\zeta_2(\tau) = 1 - \tau^2$$

en $\overline{F}(x, y)$ in (D 6) kan vervangen worden door $F(x, y)$.

Als $\underline{x}$ en $\underline{y}$ <u>onafhankelijk</u> zijn en een <u>continue</u> verdelingsfunctie hebben, dan is $(\tau = 0)$

$$\zeta_1(\tau) = \zeta_1 = \mathcal{E}\left\{\Phi_1^2(\underline{x}_1, \underline{y}_1)\right\}$$

$$= \iint\left\{\iint S(x_1 - x_2)\, S(y_1 - y_2)\, dF(x_2)\, dF(y_2)\right\}^2 dF(x_1)\, dF(y_1)$$

$$= \iint\left[2F(x_1) - 1\right]^2\left[2F(y_1) - 1\right]^2 dF(x_1)\, dF(y_1)$$

zoals volgt uit (D.6) voor het onafhankelijke en continue geval

Nu is $$\int (2u - 1)^2\, du = \left[\tfrac{4}{3}u^3 - \tfrac{4}{2}u^2 + u\right]_0^1 = \tfrac{4}{3} - 2 + 1 = \tfrac{1}{3}$$

$$\therefore \quad \zeta_1 = \tfrac{1}{9}.$$

Ook is $$\zeta_2(\tau) = \zeta_2 = 1$$

Dus

(D.10) $$\sigma^2(\underline{t}) = \frac{2}{n(n-1)}\left[\frac{2(n-2)}{9} + 1\right] = \frac{2(2n+5)}{9n(n-1)}$$

Dit stemt overeen met Kendall [7], waar $\sigma^2(S) = \frac{1}{15}n(n-2)(2n+5)$ en $t = \frac{2S}{n(n-1)}$.

In het geval (D.10) is de verdeling van $\underline{t}$ onafhankelijk van de ééndimensionale verdelingen van $\underline{x}$ en $\underline{y}$. Dit is echter niet meer het geval als de onafhankelijke veranderlijken discontinu zijn. Dan hangt $\sigma^2(\underline{t})$ af van $P\{\underline{x}_1 = \underline{x}_2\}$, $P\{\underline{y}_1 = \underline{y}_2\}$ en $P\{\underline{x}_1 = \underline{x}_2 = \underline{x}_3\}$, $P\{\underline{y}_1 = \underline{y}_2 = \underline{y}_3\}$.

Uit stelling (1.1) volgt dat de v.f van $\sqrt{n}(\underline{t} - \tau)$ naar de normale vorm nadert. Vergelijk ook Kendall [11] voor het speciale geval dat alle permutaties van de rang van $\underline{x}$ en $\underline{y}$ even waarschijnlijk zijn, hetgeen overeenkomt met de onafhankelijkheid van de continue (stochastische) aselecte veranderlijken $\underline{x}$ en $\underline{y}$.

In het algemene geval is de asymptotische normaliteit van $\underline{t}$ bewezen door Daniels en Kendall [2] en door Hoeffding [4].

De functionel $\tau(F)$ is stationnair (en dus is de normale limietverdeling van $\sqrt{n}(\underline{t} - \tau)$ dan singulier) als $\zeta_1 < 0$ is In het continue geval betekent dit dat $\Phi_1(\underline{x}, \underline{y}\,|\,\tau) = \tau$ of dat aan

(D.11) $$4F(\underline{x}, \underline{y}) = 2F(\underline{x}, \infty) + 2F(\infty, \underline{y}) - 1 + \tau.$$

wordt voldaan met een waarschijnlijkheid 1. Dit is b.v. het geval wanneer $\underline{y}$ een toenemende functie van $\underline{x}$ is. Dan is $t = \tau = 1$ met waarschijnlijkheid 1 en $\sigma^2(\underline{t}) = 0$.

Een voorbeeld van een geval waar aan (D.11) voldaan is en $\sigma^2(\underline{t}) > 0$ is, is het volgende: $\underline{x}$ is homogeen verdeeld over het interval (0,1) en

(D.12) $\underline{y} = \underline{x} + \frac{1}{2}$ voor $0 \leqq \underline{x} < \frac{1}{2}$, $\underline{y} = \underline{x} - \frac{1}{2}$ voor $\frac{1}{2} \leqq \underline{x} \leqq 1$.

Hier is $\tau = 0$, $\gamma_2 = 1$ en $\sigma^2(\underline{t}) = \frac{2}{n(n-1)}$ uit (D.7) met $\gamma_1 = 0$.

E De toets van Mann tegen "verloop" ("trend").

Stel $\underline{y}_1, \ldots, \underline{y}_n$ zijn n onafhankelijke, reële, stochastische veranderlijken, waarbij $\underline{y}_\alpha$ de continue v.f. $F_\alpha(y)$, $(\alpha = 1, \ldots, n)$ heeft. Wij willen de hypothese H_0 toetsen dat de $\underline{y}_\alpha$'s $(\alpha = 1, \ldots, n)$ alle dezelfde verdeling hebben, dus

$$H_0 : \quad F_1(y) = \ldots = F_n(y)$$

tegen de alternatieve hypothese H_1 van een "stijgend verloop", dus

$$H_1 : \quad F_1(y) < F_2(y) < \ldots < F_n(y).$$

Mann [9] heeft een toets van H_0 tegenover H_1 voorgesteld die gebaseerd is op het aantal T ongelijkheden $\underline{y}_\alpha < \underline{y}_\beta$, waarbij $\alpha < \beta$ is. Nu is

$$2\underline{T} - \frac{n(n-1)}{2} = \sum_{\alpha<\beta} s(\underline{y}_\beta - \underline{y}_\alpha) \qquad \text{(Kendall[7], verg. (17))}$$

$$= \sum_{\alpha<\beta} s(\alpha-\beta)\, s(\underline{y}_\alpha - \underline{y}_\beta) .$$

De $\underline{U}$-statistische grootheid

$$\underline{t} = \frac{4\underline{T}}{n(n-1)} - 1 = \frac{2}{n(n-1)} \sum_{\alpha<\beta} s(\alpha-\beta)\, s(\underline{y}_\alpha - \underline{y}_\beta)$$

is dezelfde als (D 3) voor het speciale geval waar één component geen stochastische veranderlijke is.

Stel

$$\tau_{\alpha\beta} = s(\alpha-\beta) \iint s(y_1 - y_2)\, dF_\alpha(y_1)\, dF_\beta(y_2)$$

$$= s(\alpha-\beta) \left\{ 2 \int F_\beta(y)\, dF_\alpha(y) - 1 \right\} .$$

Indien H_0 geldt, is $\tau_{\alpha\beta} = s(\alpha-\beta)\left\{2\int F(y)\, dF(y) - 1\right\} = 0$

Indien H_1 geldt, is $\tau_{\alpha\beta} < 0$, want $s(\alpha-\beta) = -1$ en $\{\ldots\} > 0$.

Aangezien $\mathcal{E}(\underline{t}) = \tau_n = \frac{2}{n(n-1)} \sum_{\alpha<\beta} \tau_{\alpha\beta}$

volgt dat

(E.1) $\mathcal{E}(\underline{t}) = 0$ onder H_0 en $\mathcal{E}(\underline{t}) < 0$ onder H_1.

Veronderstel $P(\underline{t} < a_n | H_o) \leqq \varepsilon$, waarbij a_n het grootste getal is dat hieraan voldoet, dan wordt ε de onbetrouwbaarheidsdrempel genoemd.

Dit betekent dus dat de waarschijnlijkheid dat een aselecte steekproef een waarde van t zal geven kleiner dan a_n, indien H_o juist is, ten hoogste gelijk is aan ε, de onbetrouwbaarheidsdrempel. ε is dus de kans op het ten onrechte verwerpen van de juiste hypothese (fout van de eerste soort). Maar wij kunnen ook een fout van de tweede soort maken, d.w.z. H_o wordt niet verworpen, terwijl een alternatieve hypothese $H_1 \neq H_o$ geldt. Dit wordt gegeven door

$$P[\underline{t} \geqq a_n | H_1] = \beta \qquad \text{waarbij} \quad H_1 \neq H_o \quad \text{geldt.} \tag{E.2}$$

Het is uiteraard ook gewenst dat β zo klein mogelijk zal zijn. Het onderscheidingsvermogen van de toets van de hypothese H_o tegen de alternatieve hypothese H_1 wordt gedefinieerd door

$$P[\underline{t} < a_n | H_1] = 1 - \beta , \tag{E.3}$$

dat is de waarschijnlijkheid dat H_o terecht verworpen zal worden wanneer $H_1 \neq H_o$ geldt.

Uit de definitie van a_n volgt dat $a_n \to o$ voor $n \to \infty$ en omdat $\sigma^2(\underline{t}) = O(\frac{1}{n})$ (want $\underline{t}$ is een $\underline{U}$-statistische grootheid), volgt uit de ongelijkheid van Tchebycheff (Cramér [1], blz. 182) dat $P[\underline{t} < a_n | H_1] \to 1$, d.w.z. dat de toets bruikbaar is en dientengevolge asymptotisch zuiver is.

(Let wel dat onder H_o: $\tau_n = \mathcal{E}(\underline{t}) = o$

en onder H_1: $\tau_n = \mathcal{E}(\underline{t}) < o$.

Volgens de ongelijkheid van Tchebycheff is

$$P\{|\underline{t} - \tau_n| \geqq K\sigma\} \leqq \frac{1}{K^2} \qquad \text{waarbij } K > o$$

of $$P\{|\underline{t} - \tau_n| \geqq K\} \leqq \frac{\sigma^2}{K^2} = \frac{\sigma^2(\underline{t})}{K^2}$$

d.i. $$P\{(\underline{t} - \tau_n)^2 \geqq K^2\} \leqq \frac{\sigma^2(\underline{t})}{K^2}$$

$$\therefore \beta = P\{\underline{t} \geqq a_n | H_1\} = P\{\underline{t} - \tau_n \geqq a_n - \tau_n | H_1\} \leqq \frac{\sigma^2(\underline{t})}{(a_n - \tau_n)^2} = \frac{O(\frac{1}{n})}{(a_n - \tau_n)^2} .$$

$\left(\tau_n < o \text{ onder } H_1 \text{ en } |\tau_n| > |a_n| \text{ voor voldoende grote } n \;\therefore\; a_n - \tau_n > o \right)$

Hierin is $\tau_n \neq o$ en $a_n \to o$ voor $n \to \infty$, $\therefore \beta \to o$, waaruit volgt dat $P\{t < a_n | H_1\} \to 1$.)

Mann heeft ook aangetoond dat de toets asymptotisch zuiver is en heeft bovendien voldoende voorwaarden gegeven waaronder de toets zuiver is voor eindige n .

Uit de algemene theorie van de $\underline{U}$-statistische grootheden voor het geval waar de $\underline{x}_\alpha$'s verschillende v.f.'s hebben, volgt dat de verdeling van $\frac{\underline{t} - \tau_n}{\sigma(\underline{t})}$ asymptotisch normaal is

(Hoeffding [5], stellingen 8.1 en 8.2).

F. Parametervrije toets voor onafhankelijkheid.

Veronderstel dat de stochastische veranderlijken $\underline{x}$, $\underline{y}$ een continue simultane v.f. $F(x,y)$ hebben. Wij willen de hypothese H_o toetsen dat $\underline{x}$ en $\underline{y}$ onafhankelijk zijn, d.w.z. dat

$$F(x,y) = F(x,\infty).\ F(\infty,y)\ .$$

Omdat de verdeling van iedere statistische grootheid, die alleen de rang van de veranderlijken bevat onder de hypothese H_o niet van de v.f. van de verzameling afhangt, is gesuggereerd verscheidene "order statistics", o.a. ook de verschiltekencorrelatie $\underline{t}$, te gebruiken om voor onafhankelijkheid te toetsen

Uit de voorgaande resultaten kunnen wij het asymptotisch onderscheidingsvermogen van de toets voor onafhankelijkheid, dat op $\underline{t}$ gebaseerd is, verkrijgen. Indien H_o geldt, is $\mathcal{E}(\underline{t}) = \tau = o$ (vgl. par. D) en de kritieke zone van grootte ε van de t-toets kan gedefinieerd worden door $|t| > c_n$, waarbij c_n het kleinste getal is dat voldoet aan de ongelijkheid

$$P\{|\underline{t}| > c_n \,|\, H_o\} \leqq \varepsilon \qquad \text{(F.1)}$$

ε wordt de onbetrouwbaarheidsdrempel genoemd. Volgens stelling (1.2) en (D.10) kunnen wij schrijven $c_n = \frac{2\lambda_n}{3\sqrt{n}}$, waarbij λ_n tot een positieve constante λ nadert, die van ε afhangt.

Aangezien $\sigma^2(\underline{t}) = O\left(\frac{1}{n}\right)$, nadert het onderscheidingsvermogen

$$P\left\{|\underline{t}| \geqq \frac{2\lambda_n}{3\sqrt{n}} \,\middle|\, H\right\}$$

tot 1 als $n \to \infty$ in de alternatieve hypothese H met $\tau(F) \neq o$.

Als $\tau = o$, is $\lim_{n\to\infty} P\left\{|\underline{t}| \geqq \frac{2\lambda_n}{3\sqrt{n}} \,\middle|\, H\right\} < 1$ (uit ongelijkheid van Tchebycheff)

Als $\tau = o$ en $\gamma(\tau) < \frac{1}{9}$, krijgen wij zelfs

$$\lim_{n\to\infty} P\left\{|\underline{t}| \geqq \frac{2\lambda_n}{3\sqrt{n}} \,\middle|\, H\right\} < \varepsilon$$

en ten opzichte van deze alternatieven is de toets asymptotisch onzuiver. In het geval van de verdeling (D.12) geldt b.v. dat

$$P\left\{|\underline{t}| \geqq \frac{2\lambda_n}{3\sqrt{n}} \,\middle|\, H\right\} \longrightarrow o\ .$$

In dit geval is er een functionele betrekking tussen de veranderlijken, en de verdeling verschilt dus enorm van het geval van onafhankelijkheid.

De vraag is nu of er parametervrije toetsingsmethoden voor onafhankelijkheid bestaan die zuiver zijn of asymptotisch zuiver zijn.

Verdere analyse van parametervrije toetsingsmethoden in het geval van onafhankelijkheid.

In een parametervrije toetsingsmethode van een statistische hypothese worden geen onderstellingen omtrent de functionele vorm van de verdeling van de verzameling gemaakt. Een algemene theorie van parametervrije toetsingsmethoden is nog niet ontwikkeld en het ziet er naar uit of er geen bevredigende definitie van een "beste" parametervrije toetsingsmethode verkregen kan worden. Wenselijke eigenschappen van een "goede" parametervrije toetsingsmethode zijn zuiverheid en bruikbaarheid. Een toets van een hypothese H_o wordt bruikbaar ten opzichte van een bepaalde klasse toelaatbare hypothesen genoemd, indien de waarschijnlijkheid van het niet verwerpen van H_o wanneer een hypothese $\neq H_o$ van die klasse geldt, tot nul nadert met toenemende grootte van de steekproef.

Een aselecte steekproef van grootte n is gegeven en daaruit willen wij bepalen of de twee stochastische veranderlijken $\underline{x}$ en $\underline{y}$ onafhankelijk zijn. Wij nemen aan dat de v.f. $F(x,y)$ van $(\underline{x},\underline{y})$ continu is. Stel Ω' is de klasse van continue v.f.'s $F(x,y)$ en Ω'' is de klasse van v.f.'s die continue simultane en marginale waarschijnlijkheden hebben,

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x\, \partial y} \quad . \quad f_1(x) = \int f(x,y)\, dy \quad . \quad f_2(y) = \int f(x,y)\, dx$$

H_o: $F(x,y) = F(x,\infty).\, F(\infty,y)$ is de hypothese die getoetst moet worden

$F(x,y)$ is dus een tweedimensionale v.f. en wij schrijven

$$D(x,y) = F(x,y) - F(x,\infty).\, F(\infty,y)$$

en

(F.2) $$\Delta = \Delta(F) = \int D^2(x,y)\, dF(x,y)$$, waarbij de integraal zich over R_2 (zie Cramér [1]) uitstrekt.

De stochastische veranderlijken $\underline{x}, \underline{y}$ met de v.f. $F(x,y)$ is dan en slechts dan onafhankelijk als $D(x,y) \equiv 0$.

Stelling (F.1):

Als $F(x,y)$ tot Ω'' behoort, is $\Delta(F) = 0$ dan en slechts dan als $D(x,y) \equiv 0$.

Bewijs:

Als $D(x,y) \equiv 0$, dan is $\Delta(F) = 0$.

Omgekeerd, veronderstel $D(x,y) \not\equiv 0$. Aangezien $F(x,y)$ in Ω'' is, is de functie $f(x,y) - f_1(x)\, f_2(y)$ continu. Nu is

$$D(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} d(u,v)\, du\, dv \quad .$$

Voor $D(x,y) \not\equiv 0$ is $\alpha(x,y) \not\equiv 0$ en aangezien

$$\iint \alpha(x,y)\,dx\,dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{f(x,y) - f_1(x) f_2(y)\}\,dx\,dy$$

$$= \iint f(x,y)\,dx\,dy - \int f_1(x)dx \int f_2(y)\,dy$$

$$= 1 - 1 = 0$$

bestaat er een rechthoek Q in R_2 zodanig dat $\alpha(x,y) > 0$ indien (x,y) in Q is. Dientengevolge is $D(x,y)$ bijna overal in $Q \neq 0$ en $f(x,y) > 0$ in Q. Dus is

$$\Delta(F) \geqq \iint_Q D^2(x,y)\, f(x,y)\,dx\,dy > 0$$

$$\neq 0 \quad ,$$ hetgeen de stelling bewijst.

Als $F(x,y)$ discontinu is, kan $\Delta(F) = 0$ zijn terwijl $D(x,y) \not\equiv 0$. Dit is b.v. het geval voor de verdeling

$$P\{\underline{x} = 0, \underline{y} = 1\} = P\{\underline{x} = 1, \underline{y} = 0\} = \tfrac{1}{2}.$$

Het is nog onbekend of $D(x,y) \equiv 0$ is als $\Delta(F) = 0$ is voor continue of absoluut continue $F(x,y)$.

Uit stelling (F.1) zien wij dus dat alleen wanneer de stochastische veranderlijken $\underline{x}$, $\underline{y}$ onafhankelijk zijn $\Delta(F) = 0$ is.

Veronderstel nu

(F.3) $$C(u) = \begin{cases} 1 & \text{voor } u \geqq 0 \\ 0 & \text{voor } u < 0 \end{cases}$$

$$\Psi(x_1, x_2, x_3) = C(x_1 - x_2) - C(x_1 - x_3)$$

$$\Phi(x_1, y_1; \ldots; x_5, y_5) = \tfrac{1}{4}\Psi(x_1, x_2, x_3)\Psi(x_1, x_4, x_5)\Psi(y_1, y_2, y_3)\Psi(y_1, y_4, y_5)$$

Wij kunnen schrijven

(F.4) $$\Delta = \int \cdots \int \Phi(x_1, y_1; \ldots; x_5, y_5)\, dF(x_1, y_1) \cdots dF(x_5, y_5).$$

(I) <u>De statistische grootheid D</u>. Stel $(\underline{x}_1, \underline{y}_1), \ldots, (\underline{x}_n, \underline{y}_n)$ is een aselecte steekproef uit de verzameling met de v.f. $F(x,y)$, $n \geqq 5$ en stel

(F.5) $$\underline{D} = \underline{D}_n = \frac{1}{n(n-1)\cdots(n-4)} \sum{}'' \Phi(\underline{x}_{\alpha_1}, \underline{y}_{\alpha_1}; \ldots; \underline{x}_{\alpha_5}, \underline{y}_{\alpha_5})$$

waarbij $\sum''$ de sommatie over alle α aanduidt zodanig dat

$$\alpha_i = 1, \ldots, n \; ; \; \alpha_i \neq \alpha_j \text{ als } i \neq j \quad (i,j = 1, \ldots, 5).$$

Omdat het aantal termen in $\sum''$ $n(n-1)\ldots(n-4)$ is volgt uit (F.4) dat

(F.6) $$\mathcal{E}\,\underline{D} = \mathcal{E}(\underline{D}) = \Delta$$

In het geval van onafhankelijkheid is $\mathcal{E}\,\underline{D} = 0$ (want $\Delta = 0$ voor onafhankelijkheid zoals volgt uit (F.3) en (F.4)) en dientengevolge kan $\underline{D}$ zowel negatieve als positieve waarden aannemen. Het zal bewezen worden dat $-\frac{1}{60} \leqq \underline{D}_n \leqq \frac{1}{30}$, waarbij de bovenste grens $\frac{1}{30}$ voor iedere n bereikt kan worden, terwijl het minimum van $\underline{D}_n$, $\frac{1}{60}$, blijkbaar toeneemt met n.

De stochastische veranderlijke $\underline{D}$ zoals gedefinieerd in (F.5) behoort tot de klasse van U-statistische grootheden. Hieruit volgt dus onmiddellijk dat:

$$\Phi(x_1, y_1; \ldots; x_5, y_5) = D_5 = \frac{1}{5!} \sum'' \phi(x_{\alpha_1}, y_{\alpha_1}; \ldots; x_{\alpha_5}, y_{\alpha_5})$$

$$\Phi_k(x_1, y_1; \ldots; x_k, y_k) = \int \cdots \int \phi(x_1, y_1; \ldots; x_k, y_k; x_{k+1}, y_{k+1}; \ldots; x_5, y_5)$$

$$dF(x_{k+1}, y_{k+1}) \ldots dF(x_5, y_5) \quad , \quad (k = 1, \ldots, 5)$$

$$\zeta_k = \int \cdots \int \left\{ \Phi_k(x_1, y_1; \ldots; x_k, y_k) - \Delta \right\}^2 dF(x_1, y_1) \ldots dF(x_k, y_k)$$

Dan is de variantie van $\underline{D}_n$

(F.7) $$\mathrm{Var}\, \underline{D}_n = \binom{n}{5}^{-1} \sum_{k=1}^{5} \binom{5}{k} \binom{n-5}{5-k} \zeta_k$$

Verder is $$25\,\zeta_1 \leqq n\, \mathrm{Var}\, \underline{D}_n \leqq 5\,\zeta_5$$

$n\, \mathrm{Var}\, \underline{D}_n$ is een afnemende functie van n en

(F.8) $$\lim_{n \to \infty} n\, \mathrm{Var}\, \underline{D}_n = 25\,\zeta_1 .$$

Volgens stelling (1.1) is de stochastische veranderlijke $\sqrt{n}\,(\underline{D}_n - \Delta)$ asymptotisch normaal met gemiddelde 0 en variantie $25\,\zeta_1$.

Het zal ook bewezen worden dat in het geval van onafhankelijkheid $\zeta_1 = 0$ is zodat $\sqrt{n}\,\underline{D}_n$ een ontaarde normale limietverdeling heeft. ($\Delta = 0$ voor het geval van onafhankelijkheid.) Hoeffding heeft ook aangetoond dat $n\,\underline{D}_n$ een niet-normale limietverdeling heeft, maar aangezien het bewijs daarvan tamelijk gecompliceerd is, wordt dit hier niet gegeven.

(II) De berekening van $\underline{D}$. De rang van $\underline{x}_\alpha$ wordt gegeven door

$$\sum_{\beta=1}^{n} c(\underline{x}_\alpha - \underline{x}_\beta) .$$

Schrijf nu

$$\underline{a}_\alpha = \sum_{\beta=1}^{n} c(\underline{x}_\alpha - \underline{x}_\beta) - 1$$

$$\underline{b}_\alpha = \sum_{\beta=1}^{n} c(\underline{y}_\alpha - \underline{y}_\beta) - 1$$

en $$\underline{c}_\alpha = \sum_{\beta=1}^{n} c(\underline{x}_\alpha - \underline{x}_\beta) \cdot c(\underline{y}_\alpha - \underline{y}_\beta) - 1 .$$

Hier is $a_\alpha + 1$ en $b_\alpha + 1$ dus de rang van resp. $\underline{x}_\alpha$ en y_α. c_α is het aantal elementen uit de steekproef waarvoor beide $\underline{x}_\beta < \underline{x}_\alpha$ en $y_\beta < y_\alpha$. (Omdat $F(x,y)$ continu is, kunnen wij aannemen dat $\underline{x}_\alpha \neq \underline{x}_\beta$ en $y_\alpha \neq y_\beta$ als $\alpha \neq \beta$.) Schrijf verder

$$A = \sum_{\alpha=1}^{n} a_\alpha (a_\alpha - 1) b_\alpha (b_\alpha - 1)$$

(F.9)
$$B = \sum_{\alpha=1}^{n} (a_\alpha - 1)(b_\alpha - 1) c_\alpha$$

$$C = \sum_{\alpha=1}^{n} c_\alpha (c_\alpha - 1)$$

dan is $\Sigma'' \Phi(\underline{x}_{\alpha_1}, y_{\alpha_1}; \ldots; \underline{x}_{\alpha_5}, y_{\alpha_5}) =$

$$= \Sigma'' \Big[\{ C(\underline{x}_{\alpha_1} - \underline{x}_{\alpha_2}) - C(\underline{x}_{\alpha_1} - \underline{x}_{\alpha_3}) \} \{ C(\underline{x}_{\alpha_1} - \underline{x}_{\alpha_4}) - C(\underline{x}_{\alpha_1} - \underline{x}_{\alpha_5}) \} \{ C(y_{\alpha_1} - y_{\alpha_2}) +$$

$$- C(y_{\alpha_1} - y_{\alpha_3}) \} \{ C(y_{\alpha_1} - y_{\alpha_4}) - C(y_{\alpha_1} - y_{\alpha_5}) \} \Big]$$

uit (F.3) en uit (F.5) en geeft na rangschikking in verschillende gevallen

$$= A - 2(n-2)B + (n-2)(n-3)C .$$

(F.10)
$$\therefore \quad \underline{D} = \frac{A - 2(n-2)B + (n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)}$$

Om dus D te berekenen voor een gegeven steekproef, moeten eerst de getallen a_α, b_α en c_α voor elk element van de steekproef berekend worden, dan A, B, en C uit (F.9) en dit substitueren in (F.10).

(III) <u>De variantie van $\underline{D}$ in het geval van onafhankelijkheid</u>.

Omdat $F(x,y)$ volgens aanname continu is, is ook $F(x,\infty)$ en $F(\infty,y)$ continu. Dan zijn de ongelijkheden $x_1 < x_2$ en $F(x_1,\infty) < F(x_2,\infty)$ equivalent behalve als $F(x_1,\infty) = F(x_2,\infty)$. Hetzelfde geldt voor $y_1 < y_2$ en $F(\infty,y_1) < F(\infty,y_2)$. Dit toont dus aan dat de functie Φ, (F.3), niet van waarde verandert indien x_i, y_i door $F(x_i,\infty)$, $F(\infty,y_i)$ vervangen wordt (op o, behoudens een verzameling met waarschijnlijkheid nul). Δ en D zijn dus invariant voor de transformatie

$$u = F(x,\infty) \ , \ v = F(\infty,y) \ ; \ \underline{u} = F(\underline{x},\infty) \ , \ \underline{v} = F(\infty,\underline{y}).$$

In het geval van <u>onafhankelijkheid</u> is dus $F(x,y) = uv$ en

$$\gamma_k = \int_0^1 \cdots \int_0^1 \{ \Phi_k'(u_1,v_1;\ldots;u_k,v_k) \}^2 \, du_1 \, dv_1 \cdots du_k \, dv_k$$

waarbij Φ_k' gedefinieerd wordt door Φ_k met x_i, y_i en $F(x_i,y_i)$ door u_i, v_i en $u_i v_i$ resp. (Voor onafhankelijkheid is $\Delta = 0$.)

Hier is dus

$$\Phi_K' = \int \cdots \int \Phi'(u_1, v_1; \ldots; u_5, v_5)\, du_{K+1}\, dv_{K+1} \ldots du_5\, dv_5$$

$$= \int \cdots \int \tfrac{1}{5!} \Sigma'' \Phi(u_{\alpha_1}, v_{\alpha_1}; \ldots; u_{\alpha_5}, v_{\alpha_5})\, du_{K+1}\, dv_{K+1} \ldots du_5\, dv_5$$

en uit (F 3) is

$$= \int_0^1 \cdots \int_0^1 \tfrac{1}{4.5!} \Sigma'' \{C(u_{\alpha_1} - u_{\alpha_2}) - C(u_{\alpha_1} - u_{\alpha_3})\}\{C(u_{\alpha_1} - u_{\alpha_4}) - C(u_{\alpha_1} - u_{\alpha_5})\} \times$$

$$\times \{C(v_{\alpha_1} - v_{\alpha_2}) - C(v_{\alpha_1} - v_{\alpha_3})\}\{C(v_{\alpha_1} - v_{\alpha_4}) - C(v_{\alpha_1} - v_{\alpha_5})\} \times$$

$$\times du_{K+1}\, dv_{K+1} \ldots du_5\, dv_5$$

$$\therefore (4.5!)^2 \gamma_K = \int_0^1 \cdots \int_0^1 \left\{ \int_0^1 \cdots \int_0^1 \Sigma'' [C(u_{\alpha_1} - u_{\alpha_2}) - C(u_{\alpha_1} - u_{\alpha_3})][C(u_{\alpha_1} - u_{\alpha_4}) - C(u_{\alpha_1} - u_{\alpha_5})] \times \right.$$

$$\left. \times [C(v_{\alpha_1} - v_{\alpha_2}) - C(v_{\alpha_1} - v_{\alpha_3})][C(v_{\alpha_1} - v_{\alpha_4}) - C(v_{\alpha_1} - v_{\alpha_5})]\, du_{K+1}\, dv_{K+1} \ldots du_5\, dv_5 \right\}^2 \times$$

$$\times du_1\, dv_1 \ldots du_K\, dv_K$$

De som Σ'' strekt zich uit over alle permutaties $(\alpha_1, \ldots, \alpha_5)$ van de getallen 1, ...,5 en $C(z) = \begin{cases} z \\ 0 \end{cases}$ voor $z \gtrless 0$.

Hoeffding vindt de volgende waarden voor γ_K :

$$\gamma_1 = 0 \quad , \qquad 200.30^2\, \gamma_2 = \frac{2}{9} \quad , \qquad 600.30^2\, \gamma_3 = \frac{14}{3} \, ,$$

$$600.30^2\, \gamma_4 = \frac{164}{9} \quad , \qquad 120.30^2\, \gamma_5 = 12 \, .$$

(Ik kan er niet in slagen deze bepaalde integralen te verifiëren).

Uit (F.7) volgt nu

$$\mathrm{Var}\, \underline{D}_n = \frac{5!}{n(n-1)(n-2)(n-3)(n-4)} \left\{ 5\binom{n-5}{4}\gamma_1 + \frac{20}{2}\binom{n-5}{3}\gamma_2 + \frac{60}{6}\binom{n-5}{2}\gamma_3 + \frac{120}{24}(n-5)\gamma_4 + \gamma_5 \right\}$$

$$\therefore \mathrm{Var}\,(30\,\underline{D}_n) = \frac{120\left\{\frac{10}{6}(n-5)(n-6)(n-7)\frac{1}{100} + \frac{10}{2}(n-5)(n-6)\frac{14.3}{600} + \frac{5(n-5)164}{609} + \frac{9}{10}\right\}}{9\,n(n-1)(n-2)(n-3)(n-4)}$$

$$(F.10') \qquad = \frac{2n^3 + 6n^2 - 84n + 128}{9n(n-1)(n-2)(n-3)(n-4)} = \frac{2(n^2 + 5n - 32)}{9n(n-1)(n-3)(n-4)}$$

Een andere methode om de coëfficiënten γ_K in het geval van onafhankelijkheid te bepalen is om $\mathrm{Var}\, \underline{D}_n$ voor $n = 5, 6, 7$ uit de exacte verdelingen, als gegeven in de volgende paragraaf, te berekenen en $\lim_{n\to\infty} n^2\, \mathrm{Var}\, \underline{D}_n$ uit de asymptotische verdeling van $n\underline{D}_n$ (zie Hoeffding [6], blz. 552, par. 8)

(IV) <u>De exacte verdeling van $\underline{D}$ in het geval van onafhankelijkheid voor $n = 5, 6, 7$.</u>

Stel $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is een steekproef uit een verzameling met een continue verdelingsfunctie. Wij beperken

ons dus tot steekproeven met $x_i \neq x_j$ en $y_i \neq y_j$ als $i \neq j$. Stel $(x_1', y_{\beta_1}'), \ldots\ldots, (x_n', y_{\beta_n}')$ is een herrangschikking van $(x_1, y_1), \ldots\ldots, (x_n, y_n)$ zodanig dat $x_1' < x_2' < \ldots < x_n'$ en $y_1' < y_2' < \ldots < y_n'$. De permutatie $\Pi = (\beta_1, \ldots\ldots, \beta_n)$ van $(1, \ldots, n)$ wordt de rangschikking van de steekproef S genoemd.

D_n hangt slechts van de rangschikking van de steekproef af Wij schrijven dus $D_n = D_n(\Pi) = D_n(\beta_1, \ldots\ldots, \beta_n)$. Als $(\beta_{\alpha_1}', \ldots\ldots, \beta_{\alpha_m}')$ een permutatie van $m\ (< n)$ van de gehele getallen $1, \ldots, n$ is zodanig dat $\beta_1' < \beta_2' < \ldots\ldots < \beta_m'$, dan definiëren wij $D_m(\beta_{\alpha_1}', \ldots, \beta_{\alpha_m}') = D_m(\alpha_1, \ldots, \alpha_m)$. Als wij in (F.5) (x_α, y_α) vervangen door (α, β_α) krijgen wij

$$(F.11) \qquad D_n = D_n(\beta_1, \ldots\ldots, \beta_n) = \binom{n}{5}^{-1} \Sigma' D_5(\beta_{\alpha_1}, \ldots\ldots, \beta_{\alpha_5})$$

waarbij Σ' de sommatie over alle α aanduidt zodanig dat $1 \leqq \alpha_1 < \alpha_2 < \ldots\ldots < \alpha_5 \leqq n$.

Als we met $\Pi^{(i)}$ de permutatie aanduiden, die verkregen wordt uit $\Pi = (\beta_1, \ldots\ldots, \beta_n)$ door weglating van β_i, dan krijgen wij de herleidingsformule

$$D_n(\Pi) = \binom{n}{5}^{-1} \Sigma' D_5(\beta_{\alpha_1}, \ldots\ldots, \beta_{\alpha_5}) \qquad 1 \leqq \alpha_1 < \ldots < \alpha_5 \leqq n$$

$$D_{n-1}(\Pi) = \binom{n-1}{5}^{-1} \Sigma' D_5(\beta_{\alpha_1}, \ldots\ldots, \beta_{\alpha_5}) \qquad 1 \leqq \alpha_1 < \ldots < \alpha_5 \leqq n-1$$

$$(F.12) \qquad \text{d i.} \qquad n\, D_n(\Pi) = (n-5) \sum_{i=1}^{n} D_{n-1}(\Pi^{(i)})$$

Uit (F.5) en (F.3) volgt

$$D_5(\beta_1, \ldots, \beta_5) = \frac{1}{5!} \Sigma'' \Phi(\alpha_1, \beta_{\alpha_1}; \ldots\ldots; \alpha_5, \beta_{\alpha_5})$$

Dus $$60\, D_5(\beta_1, \ldots, \beta_5) = \frac{1}{2} \cdot \frac{1}{4} \Sigma'' \psi(\alpha_1, \alpha_2, \alpha_3) \cdot \psi(\alpha_1, \alpha_4, \alpha_5) \cdot \psi(\beta_{\alpha_1}, \beta_{\alpha_2}, \beta_{\alpha_3}) \cdot \psi(\beta_{\alpha_1}, \beta_{\alpha_4}, \beta_{\alpha_5})$$

waarbij Σ'' uitgestrekt is over alle permutaties van $\alpha_1, \ldots, \alpha_5$. Dus

$$60\, D_5(\beta_1, \ldots, \beta_5) = \Sigma'' \frac{1}{8} \left[c(\alpha_1 - \alpha_2) - c(\alpha_1 - \alpha_3)\right]\left[c(\alpha_1 - \alpha_4) - c(\alpha_1 - \alpha_5)\right] \psi(\beta_{\alpha_1}, \beta_{\alpha_2}, \beta_{\alpha_3}) \cdot \psi(\beta_{\alpha_1}, \beta_{\alpha_4}, \beta_{\alpha_5})$$

$\alpha_1, \ldots, \alpha_5$ zijn dus alle mogelijke permutaties van de getallen $1, \ldots, 5$. Allen wanneer $\alpha_1 = 3$, is

$$\left[c(\alpha_1 - \alpha_2) - c(\alpha_1 - \alpha_3)\right]\left[c(\alpha_1 - \alpha_4) - c(\alpha_1 - \alpha_5)\right] \neq 0$$

en wel met de volgende permutaties:

3, 1, 4, 2, 5 (+1)	3, 2, 4, 1, 5 (+1)	3, 1, 5, 2, 4 (+1)
3, 1, 4, 5, 2 (-1)	3, 2, 4, 5, 1 (-1)	3, 1, 5, 4, 2 (-1)
3, 4, 1, 2, 5 (-1)	3, 4, 2, 1, 5 (-1)	3, 5, 1, 2, 4 (-1)
3, 4, 1, 5, 2 (+1)	3, 4, 2, 5, 1 (+1)	3, 5, 1, 4, 2 (+1)

3, 2, 5, 1, 4 (+1)
3, 2, 5, 4, 1 (-1)
3, 5, 2, 1, 4 (-1)
3, 5 2, 4, 1 (+1)

Nu is $\psi(\beta_3 . \beta_1 . \beta_4) . \psi(\beta_3 . \beta_5 . \beta_2) = -\psi(\beta_3 . \beta_1 . \beta_4) . \psi(\beta_3 . \beta_2 . \beta_5)$ enz.,
zodat

$$60\, D_5(\beta_1, \ldots, \beta_5) = \tfrac{1}{8}\left\{\delta\psi(\beta_3 . \beta_1 . \beta_4) . \psi(\beta_3 . \beta_2 . \beta_5) + \delta\psi(\beta_3 . \beta_1 . \beta_5) . \psi(\beta_3 . \beta_2 . \beta_4)\right\}$$
$$= \psi(\beta_3 . \beta_1 . \beta_4) . \psi(\beta_3 . \beta_2 . \beta_5) + \psi(\beta_3 . \beta_1 . \beta_5) . \psi(\beta_3 . \beta_2 . \beta_4)$$

of

$$60\, D_5(\beta_1, \ldots, \beta_5) = \begin{cases} 0 & \text{as } \beta_3 \neq 3 \\ 2 & \text{as } \beta_3 = 3 \quad \text{en } \beta_1 . \beta_2 < 3 \quad \text{of } \beta_1 . \beta_2 > 3 \\ -1 & \text{as } \beta_3 = 3 \quad \text{en } \beta_1 < 3, \beta_2 > 3 \text{ of } \beta_1 > 3, \beta_2 < 3 \end{cases} \tag{F 13}$$

Nu is

$$D_n(\beta_1, \ldots, \beta_n) = D_n(\beta_2 . \beta_1 . \beta_3 \ldots . \beta_n) = D_n(\beta_1, \ldots, \beta_{n-2} . \beta_n . \beta_{n-1})$$
$$= D_n(\beta_n . \beta_{n-1}, \ldots, \beta_1) \tag{F 14}$$

Voor $n = 5$ volgt dit uit (F 13) en voor algemene n uit (F 11)
Uit de symmetrie van D_n t.o.v. x en y volgt ook dat D_n niet van waarde verandert, indien in de permutatie $(\beta_1, \ldots, \beta_n)$ de getallen 1, 2 of de getallen $n-1, n$ verwisseld worden of als de permutatie door zijn inverse vervangen wordt.

In het geval van onafhankelijkheid hebben alle $n!$ mogelijke rangschikkingen dezelfde waarschijnlijkheid $\frac{1}{n!}$. Om dus de verdeling van $\underline{D}_n$ te vinden, moet het aantal rangschikkingen bepaald worden dat een bijzondere waarden van D_n geeft.

Voor $n = 5$ zijn er $5! = 120$ rangschikkingen. Volgens (F 14) behoeven wij slechts de rangschikkingen te beschouwen met $\beta_1 < \beta_2$, $\beta_4 < \beta_5$, $\beta_1 < \beta_4$. Hun aantal is $\frac{120}{2.2.2} = 15$. Die rangschikkingen onder deze 15, waarbij $\beta_3 \neq 3$, geven $D_5 = 0$. Er blijven dus de volgende drie permutaties over

(1, 2, 3, 4, 5), (1, 4, 3, 2, 5), (1, 4, 3, 5, 2)

Uit (F.13) vinden we als respectievelijke waarden van $60\, D_5$ 2, -1, -1. Dus is

$$P\{60\,\underline{D}_5 = 2\} = \tfrac{1}{15} \; ; \quad P\{60\,\underline{D}_5 = -1\} = \tfrac{2}{15} \; ; \quad P\{60\,\underline{D}_5 = 0\} = \tfrac{12}{15}$$

$\therefore$ $n = 5$

x	$15\,P\{60\,\underline{D}_5 = x\}$	$P\{60\,\underline{D}_5 \geqq x\}$
-1	2	1.0000
0	12	0.8667
2	1	0.0667

Analoog kunnen de verdelingen van $\underline{D}_6$, $\underline{D}_7$, ... berekend worden door gebruik te maken van de betrekkingen (F.11) en (F.14) Zo is voor $n = 6$:

x	$90\, P\{180\, \underline{D}_6 = x\}$	$P\{180\, \underline{D}_6 \geqq x\}$
-2	4	1 0000
-1	23	0 9556
0	36	0 6444
1	16	0 2444
2	1	0 0667
3	4	0 0556
6	1	0 0111

Hoeffding geeft ook een tabel voor het geval $n = 7$.

Uit (F 13) en (F 11) volgt dat $-\frac{1}{60} \leqq \underline{D}_n \leqq \frac{2}{60} = \frac{1}{30}$ voor $n = 5,6,\ldots$

De bovenste grens $\frac{1}{30}$ wordt bereikt voor $\Pi = (1, \ldots, n)$ en iedere n. Volgens de gevallen $n = 5,6,7$ lijkt het alsof het minimum van $\underline{D}_n$ schijnbaar toeneemt met n, $(-\frac{1}{60}, -\frac{1}{90}, -\frac{1}{115})$.

Uit $\mathcal{E}\, \underline{D}_n = \Delta$, volgt dat $\Delta \leqq \frac{1}{30}$.

(V) Uitvoering van de D-toets om te toetsen op onafhankelijkheid.

Als een aselecte steekproef uit een tweedimensionale verdeling met een continue v.f gegeven is, dan wordt als volgt op onafhankelijkheid getoetst:

$$H_o : \quad F(x,y) = F(x,\infty).F(\infty,y)$$

Stel ζ_n is het kleinste getal dat voldoet aan

(F.15) $$P\{\underline{D}_n > \zeta_n \mid H_o\} \leqq \alpha ,$$

waarbij α $(0 < \alpha < 1)$ de gekozen onbetrouwbaarheidsdrempel is.

Bereken D_n zoals in par. F(II) is aangetoond. Als $D_n > \zeta_n$ zoals uit de exacte verdeling verkregen is, dan wordt de nulhypothese H_o, dat de twee stochastische grootheden $\underline{x}$ en $\underline{y}$ onafhankelijk zijn, verworpen. Voor $n = 5,6,7$ kan ζ_n uit de berekende tabellen verkregen worden.

Uit de ongelijkheid van Tchebycheff en (F.10') volgt dat

$$P\left\{|30\,\underline{D}_n - 30\,\Delta| \geqq \sqrt{Var\,(30\,\underline{D})}\cdot\frac{1}{\sqrt{\alpha}}\right\} \leqq (\sqrt{\alpha})^2$$

maar $\Delta = 0$ bij onafhankelijkheid, dus

$$P\left\{30\,\underline{D}_n > \sqrt{\frac{2(n^2+5n-32)}{n(n-1)(n-3)(n-4)\alpha}}\right\} \leqq \alpha$$

waaruit volgt

$$30\,\varsigma_n \leqq \sqrt{\frac{2(n^2+5n-32)}{n(n-1)(n-3)(n-4)\alpha}}$$

omdat ς_n het grootste getal is dat aan (F.15) voldoet. Dus is $\varsigma_n = O(\frac{1}{n})$.

Als $\Delta > 0$, is $\Delta - \varsigma_n > 0$ voor voldoend grote n. Dan is

$$P\{D_n > \varsigma_n \mid H\} \geqq P\{|D_n - \Delta| \leqq \Delta - \varsigma_n\} \geqq 1 - \frac{Var\, D_n}{(\Delta - \varsigma_n)^2}$$

$$\longrightarrow 1 \quad als \quad n \longrightarrow \infty \quad (uit\ F.8),$$

waarbij H de alternatieve hypothese $\Delta > 0$ voorstelt.

Dus volgt hieruit en uit stelling (F.1) dat de D-toets bruikbaar is ten opzichte van de klasse Ω'', en dus asymptotisch zuiver is.

(VI) Enkele opmerkingen.

Volgens Hoeffding lijkt het alsof het onderscheidingvermogen van de D-toets in vergelijking met de productmomentcorrelatietoets voor een normale verdeling met een correlatie ς, enigszins laag is voor kleine waarden van $|\varsigma|$ en voor $n \longrightarrow \infty$. Het resultaat is misschien niet geheel bevredigend voor waarden van n waarin wij geïnteresseerd zijn. Aan de andere kant is te verwachten dat een toets, die bruikbaar is ten opzichte van een grote klasse van alternatieven, een lager onderscheidingsvermogen zal hebben met betrekking tot een deelklasse van alternatieven dan een toets, die optimale kenmerken ten opzichte van de deelklasse heeft. Dit werpt dus het probleem op om uit een gegeven klasse van parametervrije toetsingsmethoden (zoals die, welke bruikbaar zijn ten opzichte van Ω'') een toetsingsmethode te kiezen, die het grootste onderscheidingsvermogen heeft ten opzichte van zekere parametrische alternatieven (b.v. normale verdelingen). (Zie Hoeffding: Optimum Nonparametric Tests).

(VII) Een paar verdere stellingen.

Stel dat $\Pi_{n\nu}$ $(\nu = 1,2,\ldots,n!)$ de $n!$ mogelijke rangschikkingen van steekproeven (grootte n) uit een tweedimensionale verdeling met continue v.f. $F(x,y)$ zijn (vergelijk par. F(IV)).

Als $F(x,y) = F(x,\infty)\,F(\infty,y)$, dan is

$$P\{\Pi_{n\nu}\} = \frac{1}{n!} \qquad (\nu = 1,\ldots,n!) \tag{F.16}$$

voor iedere n.

Als (F.16) gegeven is, volgt dan daaruit dat, voor een zekere n, de veranderlijken $\underline{x}, \underline{y}$ onafhankelijk zijn? Dit geldt niet voor $n = 2$, want dan is (F.16) equivalent met $P\{(1,2)\} = \frac{1}{2}$. Indien de verdeling een waarschijnlijkheidsdichtheid $f(x,y)$ heeft, is

$$P\{(1,2)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Big[\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,du\,dv + \int_{x}^{\infty}\int_{y}^{\infty} f(u,v)\,du\,dv\Big] f(x,y)\,dx\,dy$$

hetgeen $= \frac{1}{2}$ is wanneer $f(x,y) = f(-x,y)$.

Nu hebben wij echter

<u>Stelling A</u>. Als $F(x,y)$ tot de klasse Ω'' behoort en (F.16) voor een zekere $n \geqq 5$ geldt, dan is

(F.17) $$F(x,y) = F(x,\infty).\,F(\infty,y)$$

<u>Bewijs</u>: Omdat $\sum \underline{D}_n = \Delta$, kunnen wij schrijven

(F.18) $$\sum_{\nu=1}^{n!} D_n(\pi_{n\nu})\,P\{\pi_{n\nu}\} = \Delta .$$

Als (F.16) geldt, heeft het linkerlid van (F.18) dezelfde waarde als wanneer (F.17) geldt. Maar wanneer (F.17) geldt, is $\Delta = 0$.

Uit (F.16) volgt dus dat $\Delta = 0$ en uit stelling (F.1) volgt dat dit voldoende is voor (F.17). Hiermee is de stelling bewezen. Als dus de oorspronkelijke verdeling tot Ω'' behoort en als voor een zekere $n \geqq 5$ de waarschijnlijkheden van de $n!$ rangpermutaties gelijk zijn, zijn de stochastische veranderlijken onafhankelijk.

<u>Stelling B</u>. Er bestaan geen toetsingsmethoden voor onafhankelijkheid, die alleen van de rangvolgorde van de waarnemingen afhangen en zuiver zijn bij elke onbetrouwbaarheidsdrempel ten opzichte van de klassen Ω' of Ω''.

<u>Bewijs</u>: Elke critieke zone van een rangordetoets (d.i. een toets die alleen van de rangvolgorde van de waarnemingen afhangt) voor onafhankelijkheid is een verzameling $S_m = \{\pi_{n\nu_1}, \ldots\ldots, \pi_{n\nu_m}\}$ van m rangschikkingen. In het geval van onafhankelijkheid is

$$P[S_m] = P[\underline{\pi}_{n\nu} \in S_m] = \frac{m}{n!}$$

Wij kunnen ons dus beperken tot onbetrouwbaarheidsdrempels $\frac{m}{n!}$, $m = 1,2,\ldots,n!-1$. Teneinde de stelling te bewijzen, zal het voldoende zijn om aan te tonen dat voor iedere $n = 2,3,\ldots,$ voor een bepaalde m $(1 \leqq m \leqq n!-1)$ en voor iedere S_m, een v.f. F in Ω'' bestaat, zodanig dat

$$P[S_m \mid F] < \frac{m}{n!}$$

Wij bewijzen echter een iets meer algemene stelling dat dit geldt voor $m = 1,2,3$.

Wij definiëren de tweedimensionale verdeling A_n zodanig dat de waarschijnlijkheidsmassa gelijkmatig (uniform) verdeeld is over de $n-1$ segmenten

(F.19) $\qquad \frac{k-1}{n-1} < x \leqq \frac{k}{n-1} \quad ; \qquad y - x = \frac{n-2k}{n-1}$

$(k = 1, 2, \ldots, n-1)$

en nul is in elk gebied dat geen deel van deze segmenten bevat.

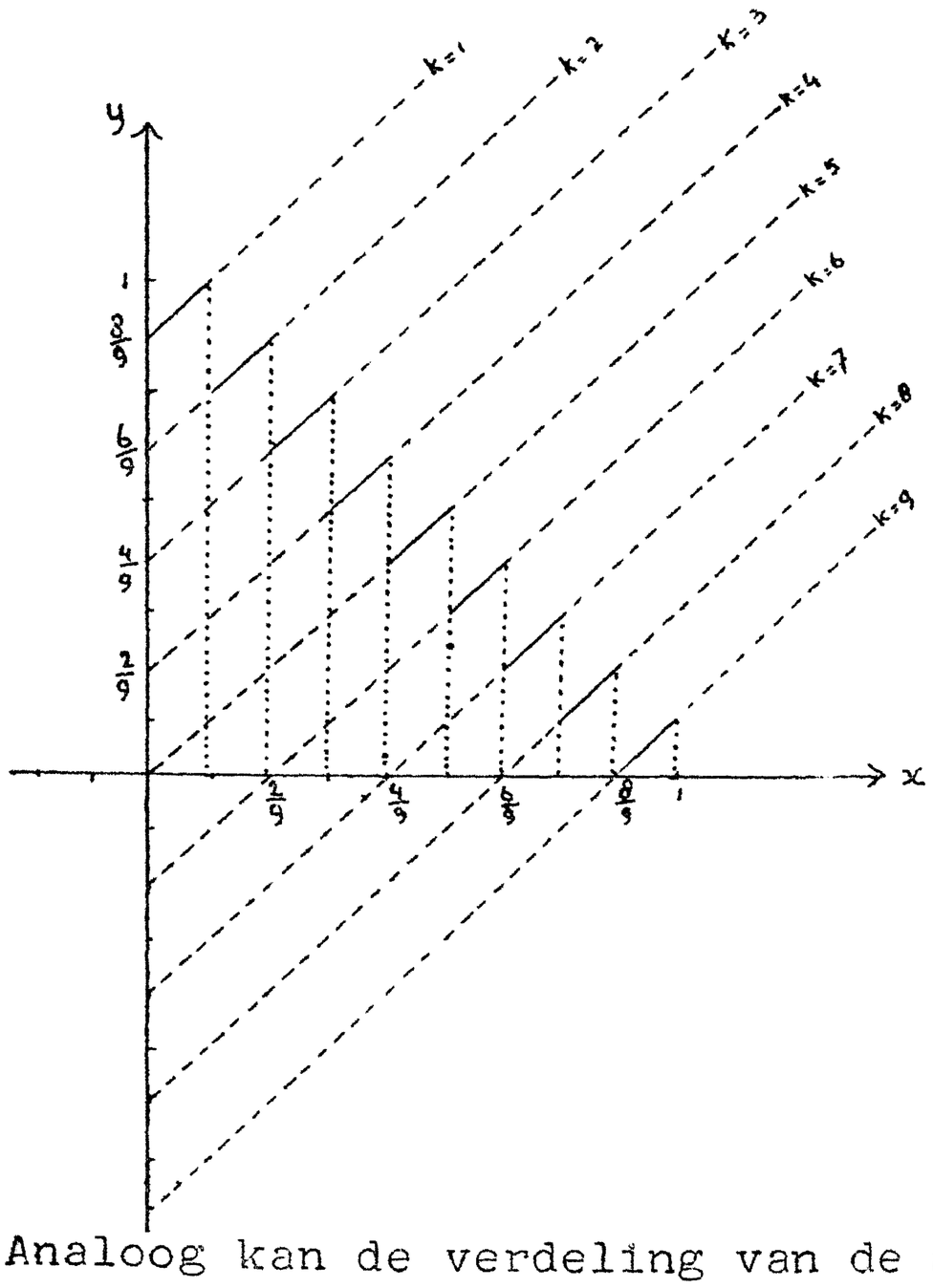
Stel B_n is een verdeling die gelijkmatig verdeeld is over de $n-1$ segmenten

(F.20) $\qquad \frac{k-1}{n-1} < x \leqq \frac{k}{n-1} \quad ; \qquad x + y = \frac{2k-1}{n-1} \qquad (k = 1, 2, \ldots, n-1)$

en een verdelingsdichtheid gelijk nul buiten deze segmenten.

De verdeling van de segmenten (F.19) ziet er dus als volgt uit:

Stel $n = 10$:



K	x	y
1	$0 < x \leqq \frac{1}{9}$	$x + \frac{8}{9}$
2	$\frac{1}{9} < x \leqq \frac{2}{9}$	$x + \frac{6}{9}$
3	$\frac{2}{9} < x \leqq \frac{3}{9}$	$x + \frac{4}{9}$
4	$\frac{3}{9} < x \leqq \frac{4}{9}$	$x + \frac{2}{9}$
5	$\frac{4}{9} < x \leqq \frac{5}{9}$	x
6	$\frac{5}{9} < x \leqq \frac{6}{9}$	$x - \frac{2}{9}$
7	$\frac{6}{9} < x \leqq \frac{7}{9}$	$x - \frac{4}{9}$
8	$\frac{7}{9} < x \leqq \frac{8}{9}$	$x - \frac{6}{9}$
9	$\frac{8}{9} < x \leqq 1$	$x - \frac{8}{9}$

Analoog kan de verdeling van de segmenten (F.20) overzichtelijk gemaakt worden.

De v.f.'s van A_n en B_n is continu, met

$$F(x, \infty) = F(\infty, x) = x \qquad (0 \leqq x \leqq 1)$$

De waarschijnlijkheid dat $(\underline{x}, \underline{y})$ in elk van de segmenten (F.19) of (F.20) ligt, is $\frac{1}{n-1}$. De waarschijnlijkheden $P(\pi | A_n)$ en $P(\pi | B_n)$ kunnen gemakkelijk verkregen worden in termen van de multinomiale verdeling met $(n-1)$ gelijke waarschijnlijkheden.

In het bijzonder is

(F.21) $\qquad P(1, 2, \ldots, n | A_2) = 1 \qquad$ en $\qquad P(n, n-1, \ldots, 1 | B_2) = 1$

(Want voor $n = 2$ in (F.19) is $0 < x \leqq 1$ en $y = x$. $x_1, \ldots, x_n$ liggen dus tussen 0 en 1. Als de x-waarden in volgorde $1, \ldots, n$ vol-

gens rangnummers geranschikt zijn, zullen de y -waarden ook in volgorde staan $(y = x)$ Let wel dat π de rangschikking is van de rangnummers van de y -waarden als de x -waarden volgens hun rangnummers gerangschikt zijn. Dus $P(1,2,\ldots,n|A_2)$ is de waarschijnlijkheid dat $y_1 < y_2 < \ldots\ldots < y_n$ wanneer $x_1 < x_2 < \ldots < x_n$ voor de verdeling A_2. Voor B_2 is $0 < x \leqq 1$ en $y = 1 - x$.)

Ook is

(F 22) $$P(1,2,\ldots,n|A_n) = P(n,n-1,\ldots,1|B_n) = (n-1)\left(\frac{1}{n-1}\right)^n = \left(\frac{1}{n-1}\right)^{n-1}$$

(Er zijn $(n-1)$ segmenten waarover de n punten $(\underline{x},\underline{y})$ verdeeld kunnen worden. De waarschijnlijkheid dat het punt $(\underline{x},\underline{y})$ in een bepaald segment voor x valt is $\frac{1}{n-1}$. De waarschijnlijkheid dat alle n punten $(\underline{x},\underline{y})$ in **hetzelfde** segment voor x valt, is $\left(\frac{1}{n-1}\right)^n$. Omdat er $(n-1)$ segmenten zijn, is de waarschijnlijkheid dat alle n punten $(\underline{x},\underline{y})$ in eenzelfde segment voor x vallen, $(n-1)\left(\frac{1}{n-1}\right)^{n-1}$. Een rangschikking $(1,2,\ldots,n)$, d.w.z. $y_1 < y_2 < \ldots\ldots < y_n$ als $x_1 < x_2 < \ldots\ldots < x_n$, is alleen dan mogelijk als alle n punten $(\underline{x},\underline{y})$ in hetzelfde segment voor x vallen — dit volgt onmiddellijk uit de tabel op de vorige bladzijde voor $n = 10$. Dus $P(1,2,\ldots,n|A_n) = \left(\frac{1}{n-1}\right)^{n-1}$. Evenzo $P(n,n-1,\ldots,1|B_n) = \left(\frac{1}{n-1}\right)^{n-1}$.)

Verder is $P(n,n-1,\ldots,1|A_n) = P(1,2,\ldots,n|B_n) = 0$.

(Want n punten $(\underline{x},\underline{y})$ (neem b.v. $n = 10$) kunnen als A_n gegeven is, niet zodanig over $n-1$ (d.i. 9) segmenten worden verdeeld dat de rangnummers van de y -waarden in een dalende volgorde staan bij stijgende volgorde van de rangnummers van de x -waarden)

In het algemeen krijgen wij dus: als π_n alle permutaties van $1,2,\ldots,n$ voorstelt, is òf $P(\pi_n|A_n) = 0$, òf $P(\pi_n|B_n) = 0$. Want elke π_n met $P(\pi_n|A_n) \neq 0$ bevat ten minste een stijging in rangvolgorde van twee of meer opeenvolgende getallen (algemeen $i, i+1, \ldots, i+k$) die niet voorafgegaan worden door kleinere getallen of gevolgd worden door grotere getallen. Aan de andere kant, als een π_n' met $P(\pi_n'|B_n) \neq 0$ een stijging in rangvolgorde bevat, wordt die of voorafgegaan door kleinere getallen of gevolgd door grotere getallen. Als dus $P(\pi_n|A_n) \neq 0$, is $P(\pi_n|B_n) = 0$ en soortgelijk als $P(\pi_n|B_n) \neq 0$, moet $P(\pi_n|A_n) = 0$ zijn.

Uit (F.21) volgt dat voor **iedere** verzameling S_m van m rangschikkingen die niet $(1,2,\ldots,n)$ of $(n,n-1,\ldots,1)$ bevat, is $P(S_m|A_2) = 0$ of $P(S_m|B_2) = 0$. Wij behoeven dus alleen kritieke zones te beschouwen die zowel $(1,2,\ldots,n)$ als $(n,n-1,\ldots,1)$ bevatten. Voor $m = 1$ bestaan er geen zodanige zones. Voor $m = 2$ is er één. Uit (F.22) volgt dat voor $n > 2$

$$P(1,2,\ldots,n \mid A_n) + P(n,n-1,\ldots,1 \mid A_n) = \left(\frac{1}{n-1}\right)^{n-1}$$

$$< \frac{2}{n}\left(\frac{1}{n-1}\right)^{n-2} < \frac{2}{n!}$$

Ten slotte, als Π_n iedere permutatie met uitzondering van $(1,2,\ldots,n)$ of $(n,n-1,\ldots,1)$ voorstelt, krijgen wij uit **het voorgaande** betoog voor A_n of ook voor B_n

$$P(1,2,\ldots,n) + P(n,n-1,\ldots,1) + P(\Pi_n) = \left(\frac{1}{n-1}\right)^{n-1} < \frac{3}{n!}$$

Dus is ook $P(S_3 \mid F) < \frac{3}{n!}$. Hiermee is dus het bewijs voor de v.f.'s in Ω' voltooid.

Teneinde de stelling te bewijzen voor v.f.'s in Ω'', kunnen wij de verdelingen A_n en B_n vervangen door verdelingen A_n' en B_n' die continue marginale en gezamenlijke dichtheden hebben en zodanig zijn dat de waarschijnlijkheden $P(\pi \mid A_n')$ en $P(\pi \mid B_n')$ willekeurig weinig verschillen van $P(\pi \mid A_n)$ respectievelijk $P(\pi \mid B_n)$. Zo b.v. kan A_2' gedefinieerd worden door de continue dichtheid

$$\begin{aligned}
f(x,y) &= K(\varepsilon - y + x) && \text{voor } 0 \leqq y-x \leqq \varepsilon,\ x \leqq 1-\varepsilon,\ y \geqq \varepsilon;\\
&= K(\varepsilon - x + y) && \text{voor } -\varepsilon \leqq y-x \leqq 0,\ x \geqq \varepsilon,\ y \leqq 1-\varepsilon;\\
&= K(x+y-\varepsilon) && \text{voor } x+y \geqq \varepsilon,\ x \leqq \varepsilon,\ y \leqq \varepsilon;\\
&= K(2-\varepsilon-x-y) && \text{voor } x+y \leqq 2-\varepsilon,\ x \geqq 1-\varepsilon,\ y \geqq 1-\varepsilon;\\
&= 0 && \text{overal elders,}
\end{aligned}$$

waarbij $K = \frac{3}{3\varepsilon^2 - 4\varepsilon^3}$ en $0 < \varepsilon \leqq \frac{1}{2}$. Voor voldoend kleine ε, voldoet de verdeling aan de eisen.

Het bewijs van de stelling voor Ω'' toont ook aan dat er geen zuivere toetsingsmethode voor onafhankelijkheid, die alleen van de rangvolgorde van de waarnemingen afhangt, voor $n = 2$ en iedere betrouwbaarheidsdrempel (wij behoeven slechts $\varepsilon = \frac{1}{2}$ te nemen) bestaat. Volgens Hoeffding kan ook bewezen worden dat voor $n = 3$ iedere $m = 1,2,\ldots,5$ en iedere S_m, de ongelijkheid $P(S_m) < \frac{m}{3!}$ ten minste voor een van de verdelingen A_2, A_3, B_2, B_3 geldt. Het is nog onbekend of er rangvolgorde toetsingsmethoden voor onafhankelijkheid bestaan die zuiver zijn voor bepaalde steekproefgrootten n en bepaalde onbetrouwbaarheidsdrempels $\frac{m}{n!}$, ten opzichte van de klasse Ω'' van verdelingsfuncties F.

(VIII) Enkele opmerkingen.

Er bestaan dus geen toetsingsmethoden voor onafhankelijkheid die alleen van de rangvolgorde van de waarnemingen afhangen en die zuiver zijn op iedere onbetrouwbaarheidsdrempel ten

opzichte van de klasse Ω' of Ω'' van verdelingsfuncties (zie par. F(VII)).

Er bestaan echter wel toetsingsmethoden voor onafhankelijkheid die alleen van de rangvolgorde van de waarnemingen afhangen (dit zijn parametervrije toetsingsmethoden) die bruikbaar zijn en dus asymptotisch zuiver, ten minste ten opzichte van de klasse Ω'' van verdelingsfuncties (zie einde van par. F(V)).

G. Rangcorrelatie en graadcorrelatie.

Hoeffding [5] is er verder in geslaagd om de rangcorrelatiecoëfficiënt van Spearman (ρ) (zie Kendall [7]) met behulp van de graadcorrelatie K (zie Yule+Kendall [12]) en de rangcorrelatiecoëfficiënt τ van Kendall als een U-statistische grootheid uit te drukken.

Als een steekproef $(x_\alpha^{(1)}, x_\alpha^{(2)})$, $(\alpha = 1, \ldots, n)$ gegeven is met alle $x_\alpha^{(1)}$'s en alle $x_\alpha^{(2)}$'s verschillend, dan definieert Hoeffding

$$K' = \frac{12}{(n^3-n)} \sum_{\alpha=1}^{n} \left(R_\alpha^{(1)} - \frac{n+1}{2}\right)\left(R_\alpha^{(2)} - \frac{n+1}{2}\right) \quad \text{(zie ook par. 5.6)}$$

hetgeen equivalent is met Spearman's ρ.

Substitutie van (C.3) hierin geeft

$$K' = \frac{3}{n^3-n} \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \sum_{\gamma=1}^{n} S\left(x_\alpha^{(1)} - x_\beta^{(1)}\right) S\left(x_\alpha^{(2)} - x_\gamma^{(2)}\right)$$

hetgeen geschreven kan worden als

$$K' = \frac{(n-2)K + 3t}{n+1} \tag{G.1}$$

waarbij t de rangcorrelatiecoëfficiënt is (verg. D.3) en

$$K = \frac{3}{n(n-1)(n-2)} \sum'' S\left(x_\alpha^{(1)} - x_\beta^{(1)}\right) S\left(x_\alpha^{(2)} - x_\gamma^{(2)}\right)$$

en de sommatie zich over alle onderling verschillende α, β, γ uitstrekt.

$\underline{K}$ is een $\underline{U}$-statistische grootheid, en als een functie van een aselecte steekproef uit een verdeling met v.f. F, is K een zuivere schatting van de 3e graad regelmatige functioneel

$$K = 3 \int \cdots \int S\left(x_1^{(1)} - x_2^{(1)}\right) S\left(x_1^{(2)} - x_3^{(2)}\right) dF(x_1)\, dF(x_2)\, dF(x_3) \tag{G.2}$$

Hoeffding vindt voor continue en onafhankelijke veranderlijken $\underline{x}^{(1)}$, $\underline{x}^{(2)}$

$$\sigma^2(\underline{K}) = \frac{n^2-3}{n(n-1)(n-2)} \tag{G.3}$$

(G.4) $$\sigma^2(\underline{t},\underline{K}) = \frac{2(n+2)}{3n(n-1)}$$

waaruit volgt

$$\sigma^2(\underline{K}') = \frac{(n-2)^2\sigma^2(\underline{K}) + 6(n-2)\sigma(\underline{t},\underline{K}) + 9\sigma^2(\underline{t})}{(n+1)^2}$$

$$= \frac{1}{n-1}$$

hetgeen overeenstemt met het resultaat door Kendall [7] afgeleid voor Spearman's ρ.

Volgens stelling 1.1 is $\sqrt{n}(\underline{K}-K)$ asymptotisch normaal verdeeld met gemiddelde nul en $9\zeta_1(K)$ als variantie. Hetzelfde geldt voor de verdeling van de rangcorrelatiecoëfficiënt $\underline{K}'$, zoals volgt uit stelling 1.3 tezamen met (G.1).

Uit stelling 1.3 volgt ook dat de gezamenlijke verdeling van $\sqrt{n}(\underline{t}-\tau)$ en $\sqrt{n}(\underline{K}-K)$ (of $\sqrt{n}(\underline{K}'-K)$) tot de normale verdeling nader met varianties $4\zeta_1(\tau)$ en $9\zeta_1(K)$ en covariantie $6\zeta_1(K,\tau)$.

Literatuur.

[1] H.Cramér, Mathematical methods of statistics, Princeton University Press, 1946.

[2] H.E.Daniels and M.G.Kendall, The significance of rank correlations where parental correlation exists, Biometrika, vol 34, (1947), pp. 197-208.

[3] P.R.Halmos, The theory of unbiased estimation, Annals of Math Stat., vol. 17 (1946), pp. 34-43.

[4] W.Hoeffding, On the distribution of the rank correlation coefficient τ, when the variates are not independent, Biometrika, vol. 34 (1947), pp. 183-196.

[5] W.Hoeffding, A class of statistics with asymptotically normal distribution, Annals of Math. Stat. vol. 19 (1948), pp. 293-325.

[6] W.Hoeffding, A non-parametric test of independence, Annals of Math. Stat., vol. 19 (1948), pp. 546-557

[7] M.G.Kendall, Rank correlation methods, Charles Griffin and Co. Ltd., London (1948).

[8] M.G.Kendall, Advanced statistics, Charles Griffin and Co Ltd., London (1948), volume I.

[9] H.B.Mann, Non-parametric tests against trend, Econometrica, vol. 13 (1945), pp. 245-259.

[10] U.Nair, The standard error of Gini's mean difference, Biometrika, vol. 28 (1936), pp. 428-436.

[11] M.G.Kendall, A new measure of rank correlation, Biometrika, vol. 30 (1938), pp. 81-93.

[12] G.U.Yule and M.G.Kendall, Introduction to the theory of statistics, Charles Griffin and Co. Ltd., (1950), pp. 268-270.