STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

AFDELING MATHEMATISCHE STATISTIEK

Rapport S 311

Simultane verdelingen en correlatiecoëfficiënten bij gegeven marginale verdelingen

door

.F.W. Steutel

April 1963

Inleiding en samenvatting

Hoewel de vraag, in hoeverre een simultane verdeling bepaald is, als de marginale verdelingen gegeven zijn, in de practijk niet zo vaak voorkomt, bestaan hierover toch verbreide misverstanden. Zo wordt vrij algemeen gedacht, dat iedere twee-dimensionale simultane verdeling, waarvan de beide marginale verdelingen normaal zijn, normaal is. Tegenvoorbeelden hiervan zijn gegeven door Fréchet [3] en [4].

Het algemenere probleem: welke n-dimensionale verdelingsfuncties zijn mogelijk, als de één-dimensionale marginale verdelingsfuncties voorgeschreven zijn, is voor n=2 behandeld door Fréchet [4] en [5], voor n=3 door Féron [2] en Dall' Aglio [1]. In dit rapport vermelden we een aantal van hun nog weinig bekende resultaten, waarbij we ons in hoofdzaak bezig houden met het geval n=2. In §1 geven we na ~~het~~ een overzicht van het probleem voor willekeurige n een afleiding van Fréchet's resultaten, die slechts weinig van die in [4] verschilt. Met behulp van deze resultaten wordt in §2 onderzocht in hoeverre de correlatiecoëfficient wordt beperkt door het voorschrijven van de marginale verdelingen. In §3 geven we een aantal voorbeelden.

1. Simultane verdelingen

Gegeven zijn n verdelingsfuncties $F_1(x_1),\ldots,F_n(x_n)$. We definiëren $\mathscr{H}_n$ als de verzameling van alle n-dimensionale verdelingsfuncties $H(x_1,\ldots,x_n)$, die $F_1(x_1),\ldots,F_n(x_n)$ als marginale verdelingen hebben, dus met de eigenschap

$$(1) \qquad H(\infty,\ldots,\infty,x_j,\infty,\ldots,\infty) = F_j(x_j) \quad (j=1,2,\ldots,n).$$

$\mathscr{H}_n$ is niet leeg, immers de verdelingsfunctie $\prod_{j=1}^{n} F_j(x_j)$ is een element van $\mathscr{H}_n$. Voor alle $H \in \mathscr{H}_n$ geldt

$H(x_1,\ldots,x_n) = P\{\underline{x}_1 \leq x_1,\ldots,\underline{x}_n \leq x_n\} \leq P\{\underline{x}_j \leq x_j\} = F_j(x_j)$ voor alle j dus ook $H(x_1,\ldots,x_n) \leq \min[F_1(x_1),\ldots,F_n(x_n)]$. Anderzijds is $1-H(x_1,\ldots,x_n) = P\{\text{niet } (\underline{x}_1 \leq x_1,\ldots,\underline{x}_n \leq x_n)\} \leq P\{\underline{x}_1 > x_1\}+\ldots+P\{\underline{x}_n > x_n\} = n-\{F_1(x_1)+\ldots+F_n(x_n)\}$, dus $H(x_1,\ldots,x_n) \geq F_1(x_1)+\ldots+F_n(x_n) - n+1$.

Omdat ook steeds $H(x_1,\ldots,x_n) \geq 0$ is, geldt $H(x_1,\ldots,x_n) \geq \max[F_1(x_1)+\ldots+F_n(x_n)-n+1,0]$. Samenvattend hebben we dus voor alle $H \in \mathcal{H}_n$.

(2) $\max[F_1(x_1)+\ldots+F_n(x_n)-n+1,0] \leq H(x_1,\ldots,x_n) \leq [\min F_1(x_1),\ldots,F_n(x_n)]$.

Met enige moeite toont men aan, dat $\min[F_1(x_1),\ldots,F_n(x_n)]$ voor alle n een verdelingsfunctie [1] is (zie [2] voor n=3) en dus het grootste element van $\mathcal{H}_n$, d.w.z.: voor iedere $H \in \mathcal{H}_n$ geldt $H(x_1,\ldots,x_n) \leq \min[F_1(x_1),\ldots,F_n(x_n)]$ voor alle $(x_1,\ldots,x_n)$. Voor $n \geq 3$ zijn er voorbeelden (zie [2]), waarbij $\max[F_1(x_1)+\ldots+F_n(x_n)-n+1,0]$ geen verdelingsfunctie en dus geen element van $\mathcal{H}_n$ is, terwijl bij ieder punt $(x_1',\ldots,x_n')$ een element $H \in \mathcal{H}_n$ bestaat met $H(x_1',\ldots,x_n') = \max[F_1+\ldots+F_n-n+1,0]$.
Voor $n \geq 3$ bevat $\mathcal{H}_n$ dus in het algemeen geen kleinste element.

In [1] wordt een noodzakelijke en voldoende voorwaarde gegeven, opdat $\mathcal{H}_3$ een kleinste element heeft.

We beperken ons verder tot n=2 en schrijven x,y,F,G en $\mathcal{H}$ in plaats van x_1,x_2, F_1,F_2 en $\mathcal{H}_2$.

Definieert men

(3)
$$\begin{cases} H_0(x,y) = F(x)G(y) \\ H_1(x,y) = \max[F(x)+G(y)-1,0] \\ H_2(x,y) = \min[F(x),G(y)] \end{cases},$$

dan geldt dus voor iedere $H \in \mathcal{H}$ en alle x en y

(2') $H_1(x,y) \leq H(x,y) \leq H_2(x,y)$.

---

1) Zie Appendix a).

H_1 en H_2 zijn elementen van $\mathcal{H}$: zij voldoen aan (1) en men gaat gemakkelijk na, dat zij verdelingsfuncties zijn. H_1 en H_2 zijn dus respectievelijk het kleinste en het grootste element van $\mathcal{H}$.

Ieder element van $\mathcal{H}$ voldoet aan (2'). Omgekeerd is iedere verdelingsfunctie, die aan (2') voldoet een element van $\mathcal{H}$, omdat uit (2') volgt, dat H aan (1) voldoet. We hebben dus

Stelling 1: $\mathcal{H}$ bestaat uit alle verdelingsfuncties H(x,y), die aan de ongelijkheid (2') voldoen.

Als $(\underline{x}, \underline{y})$ de verdelingsfunctie H(x,y) heeft, dan is $P\{x_1 < \underline{x} \leqslant x_2 \text{ en } y_1 < \underline{y} \leqslant y_2\} = H(x_2,y_2)-H(x_2,y_1)-H(x_1,y_2)+ + H(x_1,y_1)$. We beschouwen nu $H_2(x,y)$ en kiezen $(x_j,y_k) \in \{(x,y) \mid F(x) < G(y)\}$ $(j=1,2;\ k=1,2)$. Nu is de kans, dat $(\underline{x}_2,\underline{y}_2)$ - met verdelingsfunctie H_2- in de rechthoek met hoekpunten (x_j,y_k) ligt gelijk aan $H_2(x_2,y_2)-H_2(x_2,y_1)-H_2(x_1,y_2)+H_2(x_1,y_1)=F(x_2)-F(x_2)-F(x_1)+ +F(x_1) = 0$.
Hetzelfde geldt voor een rechthoek met hoekpunten in $\{(x,y) \mid F(x) > G(y)\}$. Als F en G continu zijn en stijgend [2], dan worden de gebieden waar $F(x) < G(y)$ resp. $F(x) > G(y)$ is gescheiden door de continue kromme [2] $F(x)=G(y)$. Uit het bovenstaande volgt nu, dat in dit geval het punt $(\underline{x}_2,\underline{y}_2)$ met kans 1 op de kromme $F(x)=G(y)$ ligt. Een analoge redenering geldt voor het punt $(\underline{x}_1,\underline{y}_1)$ met verdelingsfunctie H_1. We vinden zo

Stelling 2: zijn F en G continu en stijgend, dan ligt het punt $(\underline{x}_1,\underline{y}_1)$ met kans 1 op de kromme $F(x)=G(y)= 1$ en het punt $(\underline{x}_2,\underline{y}_2)$ met kans 1 op de kromme $F(x)=G(y)$.

---

[2] We beperken ons hierbij tot het gebied waar $0 < F < 1$ en $0 < G < 1$.

Opm.: voor algemenere verdelingsfuncties geldt een analoge eigenschap; de formulering wordt dan echter veel minder eenvoudig.

Zoals we al zagen bevat $\mathcal{H}$ altijd het element $H_o(x,y)= F(x).G(y)$. Als $\underline{x}$ of $\underline{y}$ univalent is d.w.z. met kans 1 constant is, dan is H_o het enige element van $\mathcal{H}$, immers als (bijv.) $P\{\underline{x} = x_o\} = 1$ is, dan is

$$\max [F(x)+G(y)-1,0]= \min [F(x),G(y)] = F(x)G(y) =$$

$$= \begin{cases} 0 & \text{als } x < x_o \\ G(y) & \text{als } x \geqslant x_o \end{cases},$$

dus $H_1=H_2=H_o$, in overeenstemming met het feit, dat een paar stochastische variabelen onafhankelijk is, als één van beiden univalent is.

Is omgekeerd bijv. $H_1=H_o$ en is $\underline{y}$ niet univalent, dus is er een y_o met $0 < G(y_o) < 1$, dan geldt

$$F(x) < 1-G(y_o) \Rightarrow H_1(x,y_o)=0 = F(x)G(y_o) \Rightarrow F(x) = 0$$

$$F(x) \geq 1-G(y_o) \Rightarrow F(x)+G(y_o)-1=F(x)\ G(y_o) \Rightarrow F(x)= 1\ ,$$

dus $\underline{x}$ is univalent. Een analoge redenering geldt als $H_2=H_o$ is, dus geldt

<u>Stelling 3</u> : $\underline{x}$ of $\underline{y}$ is univalent $\Leftrightarrow H_1=H_o \Leftrightarrow H_2 = H_o \Leftrightarrow H_1= H_2 = H_o$.

2 <u>Correlatiecoëfficienten</u> [3]

Analoog aan de formule

$$(4) \qquad \mathcal{E}\underline{x} = -\int_{-\infty}^{0} F(x)dx + \int_{0}^{\infty} \{1-F(x)\}dx \qquad [4]$$

geldt voor $\mathcal{E}\,\underline{x}\,\underline{y}$ de relatie

---

3) We beperken ons in het volgende tot verdelingen met een eindig tweede moment.

4) Zie Appendix 6).

$$(5)\quad \mathcal{E}_{\underline{x}\,\underline{y}} = \int_{-\infty}^{0}\int_{-\infty}^{0} H(x,y)\,dx\,dy - \int_{0}^{\infty}\Big[\int_{-\infty}^{0}\{F(x)-H(x,y)\}\,dx\Big]dy +$$

$$- \int_{-\infty}^{0}\Big[\int_{0}^{\infty}\{G(y)-H(x,y)\}\,dx\Big]dy + \int_{0}^{\infty}\int_{0}^{\infty}\{1-F(x)-G(y)+H(x,y)\}\,dxdy$$

en dus met (4)

$$(6)\quad \mathcal{E}_{\underline{x}\,\underline{y}} - \mathcal{E}_{\underline{x}}\,\mathcal{E}_{\underline{y}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{H(x,y)-F(x)G(y)\}\,dxdy$$

Uit (6) volgt, dat de correlatiecoëfficient ρ gedefinieerd (voor positieve $\sigma(\underline{x})$ en $\sigma(\underline{y})$) door

$$\rho = \frac{\mathcal{E}_{\underline{x}\,\underline{y}} - \mathcal{E}_{\underline{x}}\,\mathcal{E}_{\underline{y}}}{\sigma(\underline{x})\quad\sigma(\underline{y})}$$

groter (kleiner) wordt, als H(x,y) wordt vervangen door een verdelingsfunctie, die voor alle x en y groter (kleiner) is. Het is duidelijk, dat H_1 de kleinste correlatiecoëfficient ρ_1 levert en H_2 de grootste ρ_2. Zijn H^* en H^{**} elementen van $\mathcal{H}$ met correlatiecoëfficienten ρ^* en ρ^{**} dan is $H=\lambda H^* + (1-\lambda)\,H^{**} \in \mathcal{H}$ voor alle $0 \leq \lambda \leq 1$, terwijl voor de correlatiecoëfficient ρ van H geldt $\rho = \lambda\rho^* + (1-\lambda)\rho^{**}$. Zo leveren de verdelingsfuncties $\lambda H_1+(1-\lambda)H_0$ alle negatieve correlatiecoëfficienten ($\geq \rho_1$) en de verdelingsfuncties $\lambda H_2+(1-\lambda)H_0$ alle positieve ($\leq \rho_2$). Resumerend hebben we

<u>Stelling 4</u> : van de elementen van $\mathcal{H}$ heeft H_1 de kleinste correlatiecoëfficient ρ_1 en H_2 de grootste ρ_2. Is een ρ gegeven met $\rho_1 \leq \rho \leq \rho_2$ dan is er een element van $\mathcal{H}$ met correlatiecoëfficient ρ .

Opm.: als een der marginale verdelingen de verdeling van een univalente variabele is, zijn $\underline{x}$ en $\underline{y}$ onafhankelijk en zou men $\rho_1 = \rho_2 = 0$ kunnen definiëren. Er zijn echter, zoals uit het volgende zal blijken, minder triviale voorbeelden waarbij $\rho_1 > -1$ en/of $\rho_2 < 1$ is. In het algemeen wordt dus de correlatiecoëfficient door het voorschrijven van de marginale verdelingen wezenlijk beperkt.

Het is bekend, dat de correlatiecoëfficient dan en alleen dan 1 of -1 is, als $(\underline{x},\underline{y})$ met kans 1 op een rechte lijn (niet evenwijdig aan één der coördinaatassen) ligt. Het triviale geval, dat $\underline{x}$ en $\underline{y}$ beide univalent zijn moeten we hierbij uitsluiten. Opdat $\rho_1=-1$ is, is het dus noodzakelijk en voldoende dat F en G zodanig zijn, dat voor een reële $a<0$ en een reële b een simultane verdeling mogelijk is, waarbij $\underline{y}=a\,\underline{x}+b$ is met kans 1, dus dat $P\{\underline{x}\leq x\}=P\{\underline{y}\geq ax+b\}$ is voor alle x, d.w.z. $F(x)=1-G^-(ax+b)$ [5]). Als bijv. F=G en F symmetrisch is, dan is aan deze voorwaarde voldaan. Analoog vindt men voor $\rho_2=1$ de voorwaarde $F(x)=G(\alpha x+\beta)$ $(\alpha>0)$.
In het speciale geval, dat F en G continu en stijgend zijn vinden we in overeenstemming met het bovenstaande, dat $F(x)+G(y)=1$ en $F(x)=G(y)$ rechte lijnen moeten voorstellen (zie St. 2). We formuleren tenslotte

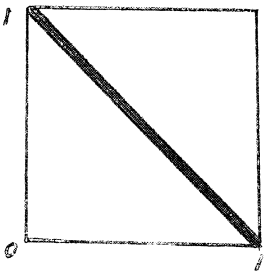
Stelling 5 : als F(x) en G(y) niet beide verdelingsfuncties zijn van een univalente variabele, dan geldt: Opdat $\rho_1=-1$ is, is het noodzakelijk en voldoende dat er reële $a<0$ en b bestaan met $F(x)=1-G^-(ax+b)$. Opdat $\rho_2=1$ is, is het noodzakelijk en voldoende, dat er reële $\alpha>0$ en β bestaan met $F(x)=G(\alpha x+\beta)$.

3 Voorbeelden.

1) $F(x)=x \quad (0\leq x\leq 1)$, $G(y)=y \quad (0\leq y\leq 1)$.

$H_1(x,y)=\max(x+y-1,0)$

$\rho_1=-1$



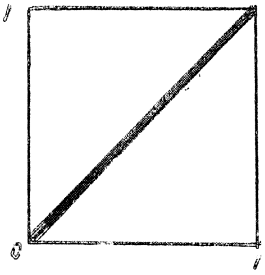
$(\underline{x},\underline{y})$ ligt met kans 1 op dik aangegeven lijn.

---

5) $G^-(x)=\lim_{\varepsilon\downarrow 0} G(x-\varepsilon)$.

$H_2(x,y)=\min(x,y)$

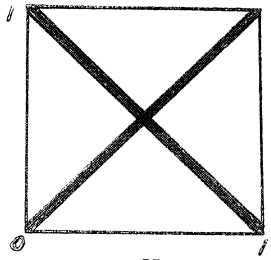
$\rho_2=1$



($\underline{x}$,$\underline{y}$) ligt met kans 1 op dik aangegeven lijn.

$H(x,y)=\frac{H_1(x,y)+H_2(x,y)}{2}$

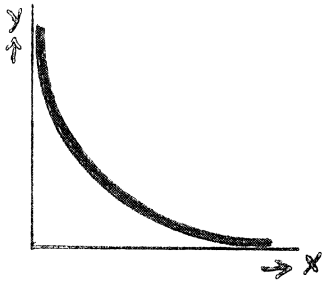
$\rho=0$



2) $F(x)=1-e^{-x}$ $(x\geq 0)$, $G(y)=1-e^{-y}$ $(y\geq 0)$.

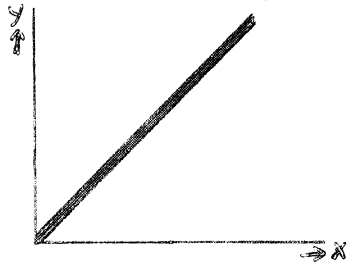
$H_1(x,y)=\max(1-e^{-x}-e^{-y},0)$

$\rho_1=1-\frac{\pi^2}{6}=-0{,}645$



$H_2(x,y)=1-\max(e^{-x},e^{-y})$

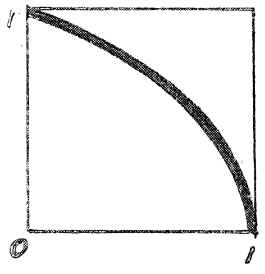
$\rho_2=1$



3) $F(x)=x$ $(0\leq x\leq 1)$, $G(y)=y^2$ $(0\leq y\leq 1)$.

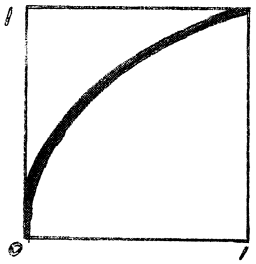
$H_1(x,y)=\max(x+y^2-1,0)$

$\rho_1=-\frac{2}{5}\sqrt{6}=-0{,}98$



$H_2(x,y) = \min(x, y^2)$

$\rho_2 = \frac{2}{5}\sqrt{6} = 0{,}98$



$(\underline{x},\underline{y})$ ligt met kans 1 op dik aangegeven lijn.

4) Gumbel [6] geeft de volgende klasse van verdelingsfuncties (elementen van $\mathcal{H}$) aan:

$$(7)\quad H_a(x,y) = F(x)\,G(y)\left[1+a\left\{1-F(x)\right\}\left\{1-G(y)\right\}\right] \quad (-1 \leq a \leq 1).$$

Uit (6) en (7) volgt voor de correlatiecoefficient ρ_a

$$\rho_a = \frac{a}{\sigma(\underline{x})\sigma(\underline{y})} \int_{-\infty}^{\infty} F(x)\left\{1-F(x)\right\}dx \int_{-\infty}^{\infty} G(y)\left\{1-G(y)\right\}dy,$$

waaruit men kan afleiden dat $|\rho_a| \leq \frac{1}{3}$ is. De waarden $\frac{1}{3}$ en $-\frac{1}{3}$ worden bereikt, als F en G homogene verdelingen voorstellen.

Zijn F en G gestandaardizeerde normale verdelingsfuncties, dan geeft differentiatie van (7)

$$(8)\qquad h(x,y) = \frac{\partial^2}{\partial x \partial y} H(x,y) = \frac{1}{2\pi} e^{-\frac{x^2+y^2}{2}} \left[1+a\left\{2F(x)-1\right\}\cdot\left\{2F(y)-1\right\}\right]$$

met $\rho = \frac{a}{\pi}$. De verdelingsdichtheid (8) geeft een voorbeeld van een niet-normale verdeling met normale marginale verdelingen. Een ander voorbeeld hiervan levert de verdelingsfunctie $\frac{F+G}{2}$ waarbij het punt $(\underline{x},\underline{y})$ met kans 1 op het lijnenpaar $x^2-y^2=0$ ligt.

$\rho_{\frac{H_1+H_2}{2}}$

5) Door Runnenburg en Steutel [7] worden voor het geval F=G verdelingsfuncties van de gedaante

$$(9)\qquad H(x,y) = A(x)\,B(y) + C(x)\,D(y)$$

beschouwd.

Omdat de verdelingsfuncties (7) een deelklasse vormen van de klasse, die door (9) wordt aangegeven is het duidelijk, dat met de functies (9) hogere correlaties kunnen worden bereikt dan met de functies van Gumbel. Zo vindt men voor de functies (9) als uiterste correlatiecoëfficienten bij de marginale verdelingen uit de voorbeelden 1) en 2) $-3/4$ en $3/4$ resp. 0,648 en -0,480. De functies van Gumbel hebben het voordeel van hun bijzonder eenvoudige gedaante.

Appendix

a) Volledigheidshalve geven we hier een definitie van het begrip verdelingsfunctie (in n variabelen):

een reële functie $H(x_1,\ldots,x_n)$ is dan en slechts dan een verdelingsfunctie, als de volgende voorwaarden vervuld zijn:

1) H is continu van rechts, d.w.z. $\lim_{\varepsilon_1 \downarrow 0 \ldots \varepsilon_n \downarrow 0} H(x_1+\varepsilon_1,\ldots,x_n+\varepsilon_n) = H(x_1,\ldots,x_n)$.

2) $\lim_{x_i \to -\infty} H(x_1,\ldots,x_n)=0 \quad (i=1,2,\ldots,n)$

3) $\lim_{x_1 \to \infty \ldots x_n \to \infty} H(x_1,\ldots,x_n)=1$.

4) voor alle $-\infty < a_i \leqslant b_i < \infty$ geldt

$$H(b_1,\ldots,b_n)-\{H(a_1,b_2,\ldots b_n)+\ldots+H(b_1,\ldots,b_{n-1},a_n)\}$$
$$+\ldots+(-)^n H(a_1,\ldots,a_n) \geqslant 0.$$

Voor n=2 wordt voorwaarde 4) dus $H(b_1,b_2)-H(a_1,b_2)-H(b_1,a_2)+H(a_1,a_2) \geqslant 0$.

b) Voor differentieerbare F(x) geldt $\mathcal{E}\underline{x} = I_1 + I_2$, waarin

$I_1 = \int_{-\infty}^{0} x\, F'(x)\, dx$ en $I_2 = \int_{0}^{\infty} x\, F'(x)dx$. Nu is

$$I_1 = \lim_{T\to\infty}\left\{\int_{-T}^{0} x\, F'(x)\, dx\right\} = \lim_{T\to\infty}\left\{\left[x\, F(x)\right]_{-T}^{0} - \int_{-T}^{0} F(x)dx\right\} =$$

$$= -\int_{-\infty}^{0} F(x)dx,$$

omdat $\lim_{T\to\infty} T\, F(-T) = \lim_{T\to\infty} T\int_{-\infty}^{-T} F'(x)dx \leq \lim_{T\to\infty} -\int_{-\infty}^{-T} x\, F'(x) = 0$ is, als $\mathcal{E}\underline{x}$ bestaat. Analoog bewijst men dat $I_2 = \int_{0}^{\infty}\{1-F(x)\}\, dx$, zodat $\mathcal{E}\underline{x} = -\int_{-\infty}^{0} F(x)dx + \int_{0}^{\infty}\{1-F(x)\}\, dx$.

In het algemene geval bewijst men (4) evenals (5) met behulp van de stelling van Fubini bijv.: $I_2 = \int_{0}^{\infty} x\, dF(x) = \int_{0}^{\infty}\int_{0}^{x} dy\, dF(x) =$

$$= \int_{0}^{\infty}\int_{y}^{\infty} dF(x)\, dy = \int_{0}^{\infty}\{1-F(y)\}\, dy.$$

Literatuur

[1] Dall' Aglio, G, Les fonctions extrêmes de la classe de Fréchet à 3 dimensions, Publications de l'Inst. de Stat. de l'Univ. de Paris. Vol IX, fasc. 2 , 1960.

[2] Féron, R, Sur les tableaux de corrélation dont les marges sont données (cas de l'espace à trois dimensions), publications de l'Inst. de Stat. de Paris, Vol. V fasc.1, 1956.

[3] Fréchet, M, Sur un essai infondé de sauver le coefficient classique dit de corrélation, Revue de l'Inst. Int. de Stat. 3/4 , 1950.

[4] Fréchet,M, Sur les tableaux de corrélation dont les marges sont données. Ann.Univ. de Lyon III ser. fasc. 14A, 1951.

[5] Fréchet, M, Sur les tableaux de corrélation dont les marges sont données, Comptes Rendus Académie des Sciences, vol. 242, Paris, 1956.

[6] Gumbel,E,J, Multivariate distributions with given margins and analytical examples, Bull. de l'Inst. Int. de Stat., XXXVII, 3, 1960.

[7] Runnenburg, J.Th. en F.W. Steutel, On Markov chains, the transition function of which is a finite sum of products of functions of one variable, Report S 304 (VP 18) of the Math. Centre, Amsterdam, 1962.