STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM
AFDELING ZUIVERE WISKUNDE

ZW 1970-003

Voordracht in de serie
"Elementaire onderwerpen vanuit hoger standpunt belicht"

door

Prof.Dr. R. Doornbos

Afhankelijkheid

0. Samenvatting

Er worden vijf opeenvolgend sterkere definities besproken van positieve (of negatieve) afhankelijkheid tussen twee stochastische variabelen x en y. Hierbij wordt onder positieve afhankelijkheid verstaan dat grote waarden van x in het algemeen samen voorkomen met grote waarden van y. Het zwakste van de genoemde kenmerken is dat de correlatiecoefficient van nul verschilt. Bij elk op elkaar volgend tweetal definities wordt een voorbeeld gegeven van een paar stochastische variabelen dat wel aan het zwakkere en niet aan het sterkere kenmerk voldoet. Als toepassing wordt een uitschieterprobleem behandeld.

1. Inleiding

De onafhankelijkheid van twee stochastische variabelen $\underline{x}$ en $\underline{y}$ is ondubbelzinnig gedefinieerd door de volgende relatie tussen de cumulatieve verdelingsfuncties

$$(1.1) \qquad F(x,y) = F_1(x).\ F_2(y).$$

waarin

$$(1.2) \qquad F(x,y_2) \ := \ P(\underline{x} \leq x \wedge \underline{y} \leq y)$$

de simultane verdelingsfunctie van $\underline{x}$ en $\underline{y}$ is en F_1 en F_2 de marginale verdelingsfuncties van $\underline{x}$ en $\underline{y}$ voorstellen.

Afhankelijkheid daarentegen treedt op in alle gevallen waarin niet aan (1.1) is voldaan en kan zich op verschillende wijzen manifesteren. In deze voordracht worden alleen die vormen van afhankelijkheid beschouwd waarbij globaal gesproken kan worden van positieve of negatieve afhankelijkheid. Dat wil zeggen dat grote waarden van $\underline{x}$ vooral samen zullen optreden met grote waarden van $\underline{y}$ en kleine waarden van $\underline{x}$ met kleine waarden van $\underline{y}$. Bij negatieve afhankelijkheid is juist het tegendeel het geval. Er is in deze gevallen dus sprake van positieve of negatieve correlatie. De vijf opeenvolgende strengere definities die wij geven hebben betrekking op negatieve afhankelijkheid. Door omkering van het teken van de betreffende ongelijkheid verkrijgt men telkens een definitie voor positieve afhankelijkheid. Deze definities zijn grotendeels te vinden in E.L. Lehmann, Some concepts of dependence, Annals of Mathematical Statistics, vol. 37 (1966) pp. 1137-1153. Het verband met uitschieterproblemen dat ook in deze voordracht wordt besproken wordt behandeld in R. Doornbos, Slippage Tests, Mathematical Centre Tracts no. 15 (1966).

Achtereenvolgens bespreken wij de volgende definities

$$(A) \qquad E(\underline{x}\,\underline{y}) < E(\underline{x})\,E(\underline{y}) \quad \text{(correlatie)}$$

$$(B) \qquad P(\underline{x} \leq x \wedge \underline{y} \leq y) \leq P(\underline{x} \leq x)\ P(\underline{y} \leq y),$$

voor alle x en y, waarbij in minstens één punt (x,y) het < teken geldt (kwadrant-afhankelijkheid).
Gelijkwaardig hiermee is vanzelfsprekend

(B') $$P(\underline{x} \leq x \mid \underline{y} \leq y) \leq P(\underline{x} \leq x).$$

(C) $$P(\underline{x} \leq x \mid \underline{y} \leq y_1) \leq P(\underline{x} \leq x \mid \underline{y} \leq y_2),$$

als $y_1 < y_2$.

(D) $$P(\underline{x} \leq x \mid \underline{y} = y_1) \leq P(\underline{x} \leq x \mid \underline{y} = y_2)$$

voor $y_1 < y_2$ (regressie-afhankelijheid)

(E) $$f(x_1,y_1)\, f(x_2,y_2) \leq f(x_1,y_2)\, f(x_2,y_1)$$

als $x_1 \leq x_2$ en $y_1 \leq y_2$, waarbij f(x,y) de simultane kansdichtheid is in het geval van continu verdeelde variabelen en de kans op het paar (x,y) in het discrete geval (likelihood ratio-afhankelijkheid). Met (E) equivalent is

(E') $$\frac{\partial^2 \log f(x,y)}{\partial x\, \partial y} \leq 0,$$

zo de uitdrukking in het linkerlid bestaat.

2. Betrekkingen tussen de definities A - E.

Eerst willen wij bewijzen dat (A) uit (B) volgt. Dit volgt direct uit een lemma afkomstig van Hoeffding dat als volgt luidt:

Als de verwachtingen $E(\underline{x}\,\underline{y})$, $E(\underline{x})$ en $E(\underline{y})$ bestaan, dan geldt:

(2.1) $$E(\underline{x}\,\underline{y}) - E(\underline{x})\, E(\underline{y}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[F(x,y) - F_1(x)\, F_2(y)\right] dx\, dy.$$

Bewijs:
Definieer

$$I(u,x) := \begin{cases} 1 \text{ als } u \geq x \\ 0 \text{ als } u < x \end{cases}$$

Beschouw twee paren stochatische variabelen $(\underline{x}_1,\underline{y}_1)$ en $(\underline{x}_2,\underline{y}_2)$, beide met verdelingsfunctie $F(x,y)$ en onafhankelijk van elkaar.
Dan geldt:

$$(2.2) \qquad 2\left[E(\underline{x}_1\underline{y}_1) - E(\underline{x}_1)\,E(\underline{y}_1)\right] = E\left[(\underline{x}_1-\underline{x}_2)\,(\underline{y}_1-\underline{y}_2)\right] =$$

$$= E\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[I(u,\underline{x}_2) - I(u,\underline{x}_1)\right]\left[I(v,\underline{y}_2) - I(v,\underline{y}_1)\right]\,du\,dv.$$

Immers

$$I(u,x_2) - I(u,x_1) = \begin{cases} +1 \text{ op } [x_2,x_1) \text{ als } x_1 > x_2 \\ -1 \text{ op } [x_1,x_2) \text{ als } x_1 < x_2, \end{cases}$$

dus

$$\int_{-\infty}^{\infty}\left[I(u,x_2) - I(u,x_1)\right]\,du = \int_{x_2}^{x_1} du \text{ of } -\int_{x_1}^{x_2} du = (x_1-x_2).$$

Als overeenkomstig de veronderstelling de verwachtingen $E|\underline{x}\,\underline{y}|$, $E|\underline{x}|$ en $E|y|$ bestaan, dan mogen in het laatste lid van (2.2) de verwachtingen onder de integraaltekens worden genomen.
Maar er geldt b.v.

$$E\left[I(u,\underline{x}_2) - I(v,\underline{y}_2)\right] = P\left[\underline{x} \leq u \wedge \underline{y} \leq v\right] = F(u,v).$$

Dus het laatste lid van (2.2) gaat over in

$$2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[F(u,v) - F_1(u)\,F_2(v)\right]\,du\,dv,$$

waarmee het lemma is bewezen.

Uit (2.1) volgt dan onmiddelijk dat definitie (B) definitie (A) impliceert.

Voorbeeld 2.1

De verdeling van $\underline{x}$ en $\underline{y}$ wordt gegeven door de volgende tabel:

y \ x	-1	0	1
+1	0	$\frac{1}{5}$	0
0	$\frac{1}{5}$	0	$\frac{1}{5}$
-1	0	$\frac{1}{5}$	$\frac{1}{5}$

Neem $x = -\frac{1}{2}$, $y = +\frac{1}{2}$, dan geldt:

$$P(\underline{x} \leq x \wedge \underline{y} \leq y) = \frac{1}{5}$$
$$P(\underline{x} \leq x) = \frac{1}{5}$$
$$P(\underline{y} \leq y) = \frac{4}{5}$$

Er is dus in $(-\frac{1}{2},\frac{1}{2})$ niet voldaan aan (B), terwijl

$$E(\underline{x}\,\underline{y}) = -\frac{1}{5},\ E(\underline{x}) = \frac{1}{5},\ E(\underline{y}) = -\frac{1}{5},$$

zodat (A) wel vervuld is.
Dat uit (C) (B') en dus ook (B) volgt is direct in te zien door in (C) in te vullen $y_1 = y$ en $y_2 = \infty$.

Voorbeeld 2.2

Geef de verdeling van $\underline{x}$ en $\underline{y}$ door

y \ x	-1	+1
+1	0,22	0,15
0	0,10	0,15
-1	0,18	0,20

Dan geldt

$$0{,}47 = P(\underline{x} \leq -1 \mid \underline{y} \leq -1) > P(\underline{x} \leq -1 \mid \underline{y} \leq 0) = 0{,}44,$$

waaruit blijkt dat (C) niet geldt, terwijl (B) wel geldt.

Nu het bewijs dat (C) uit (D) volgt. (D) wil zeggen

$$h(y) := P(\underline{x} \leq x \mid \underline{y} = y) \tag{2.3}$$

is niet afnemend in y. Het linkerlid van voorwaarde (C) is

$$P(\underline{x} \leq x \mid \underline{y} \leq y_1) = \frac{P(\underline{x} \leq x \wedge \underline{y} \leq y_1)}{P(\underline{y} \leq y_1)} = \frac{\int_{-\infty}^{y_1} h(y)\,dF_2(y)}{P(\underline{y} \leq y_1)}$$

$$= E\left[h(\underline{y}) \mid \underline{y} \leq y_1\right].$$

Evenzo is het rechterlid gelijk aan

$$E\left[h(\underline{y}) \mid \underline{y} \leq y_2\right].$$

Omdat h(y) niet afnemend is volgt hieruit (C).

Voorbeeld 2.3

Geef de verdeling van $\underline{x}$ en $\underline{y}$ door

y \ x	-1	+1
+1	0,20	0,18
0	0,20	0,17
-1	0,10	0,15

Er geldt nu het volgende:

$$P(\underline{x} \leq -1 \mid y \leq 1) = 0{,}50 \quad P(\underline{x} \leq -1 \mid y = 1) = 0{,}53$$
$$P(\underline{x} \leq -1 \mid y \leq 0) = 0{,}48 \quad P(\underline{x} \leq -1 \mid y = 0) = 0{,}54$$
$$P(\underline{x} \leq -1 \mid y \leq -1) = 0{,}40 \quad P(\underline{x} \leq -1 \mid y = -1) = 0{,}40.$$

Dus voor $x = -1$, $y_1 = 0$ en $y_2 = 1$ geldt (D) niet, terwijl (C) wel vervuld is.

Opmerking: De definities (C) en (D) zijn niet symmetrisch in $\underline{x}$ en $\underline{y}$. De symmetrie kan eventueel worden hersteld door te eisen dat naast de genoemde voorwaarden ook de voorwaarde is vervuld die wordt verkregen door x en y te verwisselen, hetgeen uiteraard in beide gevallen leidt tot een sterkere eis.

Het bewijs dat (E) (D) impliceert geven we alleen voor het continue geval aan, het discrete geval gaat precies zo.
Volgens (E) is

$$f(x_1,y_1)\, f(x_2,y_2) - f(x_1,y_2)\, f(x_2,y_1) \leq 0 \text{ als}$$

$x_1 < x_2$ en $y_1 < y_2$.
Dus is ook

$$\int_{-\infty}^{x} dx_1 \int_{x}^{\infty} dx_2 \{f(x_1,y_1)\, f(x_2,y_2) - f(x_1,y_2)\, f(x_2,y_1)\} \leq 0.$$

Of

$$P(\underline{x} \leq x \mid \underline{y} = y_1)\, P(\underline{x} > x \mid y = y_2) \leq$$
$$P(\underline{x} \leq x \mid \underline{y} = y_2)\, P(\underline{x} > x \mid \underline{y} = y_1).$$

Of

$$P(\underline{x} \leq x \mid \underline{y} = y_1)\, [1 - P(\underline{x} \leq x \mid \underline{y} = y_2)] \leq$$
$$P(\underline{x} \leq x \mid \underline{y} = y_2)\, [1 - P(\underline{x} \leq x \mid \underline{y} = y_1)],$$

waaruit onmiddelijk (D) volgt.

Voorbeeld 2.4

Neem voor $f(x|y)$ een Cauchy verdeling met parameter $-y$:

$$f(x|y) = \frac{1}{\pi} \frac{1}{1+(x+y)^2} ,$$

met voor y een willekeurige verdeling. Als y toeneemt verschuift de verdeling naar links, dus (D) geldt. Maar aan (E) is slechts voldaan als geldt:

$$\frac{f(x_1,y_1)}{f(x_1,y_2)} \leq \frac{f(x_2,y_1)}{f(x_2,y_2)} , \tag{2.4}$$

met andere woorden als de likelihood ratio

$$\lambda = \frac{f(x,y_1)}{f(x,y_2)}$$

monotoon stijgend is in x, of

$$\frac{f(x|y_1)}{f(x|y_2)}$$

monotoon in x. Zoals bekend is dit niet het geval, want λ neemt tussen $-\infty$ en een waarde x_0 af (die van y_1 en y_2 afhangt), stijgt dan tussen x_0 en een tweede waarde x_0' om daarna weer te dalen voor $x > x_0'$.

3. Het verband met het uitschieterprobleem

Dit verband wordt met het volgende voorbeeld toegelicht. Stel de variabelen

$$\underline{u}_i \quad (i=1,\ldots,k)$$

hebben gamma verdelingen met parameters ε_i en β_i. ($\varepsilon_i > -1, \beta_i > 0$)
De kansdichtheid van $\underline{u}_i$ is dus

$$(3.1) \qquad f_i(u) = \frac{1}{\Gamma(\varepsilon_i)\beta^{\varepsilon_i}} \, u^{\varepsilon_i - 1} \, e^{-u/\beta_i} \quad (0 \leq u \leq \infty).$$

Wij willen de hypothese toetsen dat

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k,$$

bij bekende ε_i, tegen de alternatieven

$$H_i: \beta_1 = \ldots = \beta_{i-1} = \beta_{i+1} = \ldots = \beta_k > \beta_i,$$

voor een onbekende waarde van i. Dit is het alternatief voor een uitschieter naar links. Het geval van een uitschieter naar rechts verloopt geheel analoog. De volgende toetsingsmethode, die bepaalde optimale eigenschappen heeft als $\varepsilon_1 = \ldots = \varepsilon_k$, kan nu worden toegepast. De quotienten

$$\underline{x}_j = \frac{\underline{u}_j}{\Sigma \underline{u}_i} \quad (j=1,\ldots,k)$$

worden berekend. Men kan aantonen dat $\underline{x}_j$ een beta-verdeling heeft met als dichtheid

$$(3.2) \qquad f_j(x) = \frac{\Gamma(A)}{\Gamma(\varepsilon_j)\Gamma(A-\varepsilon_j)} \, x^{\varepsilon_j - 1} \, (1-x)^{A-\varepsilon_j - 1} \quad (0 \leq x \leq 1)$$

waarin

$$A := \sum_{i=1}^{k} \varepsilon_i.$$

Er worden nu kritieke waarden g_j bepaald, zodanig dat

$$P(\underline{x}_j \leq g_j) = \alpha/k$$

en als één van de $\underline{x}_i$ de bijbehorende kritieke waarde onderschrijdt, wordt H_0 verworpen. De onbetrouwbaarheid van deze toets is de kans P dat minstens één van de $\underline{x}_j$ kleiner is dan de bijbehorende waarde g_j.

Volgens een van de ongelijkheden van Bonferroni geldt:

$$(3.3) \qquad \sum_i P(\underline{x}_i \leq g_i) + \sum_{i<j} P(\underline{x}_i \leq g_i \wedge \underline{x}_j \leq g_j) \leq P \leq \sum_i P(\underline{x}_i \leq g_i).$$

Als nu geldt dat

$$(3.4) \qquad P(\underline{x}_i \leq g_i \wedge \underline{x}_j \leq g_j) \leq P(\underline{x}_i \leq g_i)\, p(\underline{x}_j \leq g_j) = \alpha^2/k^2$$

dan gaat (3.3) over in

$$k.\ \alpha/k - \binom{k}{2} \alpha^2/k^2 \leq P \leq k.\ \alpha/k,$$

of

$$\alpha - \tfrac{1}{2}\alpha^2 \frac{k-1}{k} \leq \alpha,$$

dus

$$(3.5) \qquad \alpha - \tfrac{1}{2}\alpha^2 < P \leq \alpha.$$

Mits dus (3.4) waar is geeft deze methode een toets waarvan de onbetrouwbaarheid in goede benadering bekend is (α zal meestal in de buurt van 0,05 of 0,01 liggen). Maar (3.4) is juist voorwaarde (B) voor kwadrantafhankelijkheid. Om dit in ons geval te bewijzen moet de simultane kansdichtheid van $\underline{x}_i$ en $\underline{x}_j$ worden afgeleid. Deze is

$$(3.6) \qquad f(x_i,x_j) = \frac{\Gamma(A)}{\Gamma(\varepsilon_i)\Gamma(\varepsilon_j)\Gamma(A-\varepsilon_i-\varepsilon_j)}\, x_i^{\varepsilon_i-1}\, x_j^{\varepsilon_j-1}\, (1-x_i-x_j)^{A-\varepsilon_i-\varepsilon_j-1},$$

$$x_i \geq 0,\ x_j \geq 0,\ x_i + x_j \leq 1.$$

Een partieel bewijs geeft voorwaarde (E'), want

$$\frac{\partial^2 \log f(x_i,x_j)}{\partial x_i\, \partial x_j} = -\,C.\, \frac{A-\varepsilon_i-\varepsilon_j-1}{(1-x_i-x_j)^2},$$

waarin C een positieve constante is. Dus (E') geldt als

$$A - \varepsilon_i - \varepsilon_j - 1 \geq 0 \text{ voor alle i en j,}$$

wat in de meeste practische toepassingen wel het geval is. Het is echter ook mogelijk om in dit geval (B) en dus (3.4) algemeen te bewijzen. Hiervoor verwijzen we naar de eerder vermelde Tract over Slippage Tests pag. 31, waarin nog verschillende andere toepassingen te vinden zijn.