Associatiematen : enkele keuzecriteria

THE-RC 59481

Eindhoven University of Technology
Computing Centre Note 22

ASSOCIATIEMATEN: ENKELE KEUZECRITERIA

J. Praagman

november 1984

ASSOCIATIEMATEN: ENKELE KEUZECRITERIA

Voor het beschrijven van de samenhang tussen twee variabelen A en B op grond van de in een kruistabel $\{n_{ij}\}$ weergegeven observaties is een groot aantal associatiematen beschikbaar. In het onderstaande wordt een globale indeling van die maten gegeven en worden een aantal eigenschappen en overwegingen aangestipt die de keuze van een bepaalde maat in een gegeven situatie kunnen vergemakkelijken. We beperken ons daarbij tot associatie tussen op nominaal of ordinaal nivo gemeten variabelen.
Voor wat betreft de notatie: n_{ij} is het aantal onderzoekseenheden, dat op variabele A de waarde i en op variabele B de waarde j aanneemt. ($i = 1,\ldots, r$; $j = 1,\ldots, k$). Oftewel n_{ij} is de frequentie van cel (i,j) van de kruistabel van A en B, met A als rij- en B als kolomvariabele. Verder geven we met $n_{i.}$, $n_{.j}$ en $n_{..}$, resp. de rijsommen, kolomsommen en totaalsom aan.
Dus bv.

$$n_{i.} = \sum_{j=1}^{k} n_{ij}.$$

Associatie

We spreken in de meest ruime zin van associatie tussen twee variabelen als er sprake is van samenhang of een of ander verband tussen die variabelen. Hiermee is dus nog geen uitspraak gedaan over de vorm of de richting van een dergelijk verband.
In vergelijking met het begrip statistische onafhankelijkheid:

> er is sprake van associatie zolang de variabelen niet statistisch onafhankelijk zijn.

Merk op dat we hier te maken hebben met een "asymmetrie". Aan de ene kant de precies gedefinieerde toestand van statistische onafhankelijkheid, aan de andere kant associatie als zijnde al het andere. Nagaan of er associatie bestaat kan dus door vast te stellen (b.v. met de χ^2-toets) of de variabelen statistisch onafhankelijk zijn of niet [1]. Meestal zijn we echter niet alleen geïnteresseerd in de vraag of er al dan niet een verband bestaat maar meer nog in de sterkte van een eventueel verband.

Associatiematen

Voor het meten van die sterkte dienen de associatiematen. Daarbij kunnen we in ieder geval twee mogelijkheden onderscheiden:

i) maten die eigenlijk alleen uitgaan van het "nulpunt" van statistische onafhankelijkheid en op een of andere wijze bepalen hoever een gegeven resultaat daarvan afwijkt.
ii) maten waarbij ook het alternatief duidelijk omschreven is of anders gezegd, waarbij duidelijk omschreven is wanneer er maximale associatie is.
Nu wordt nagegaan in hoeverre het verkregen resultaat afwijkt van het nulpunt in de richting van dit omschreven alternatief.

Tot de eerste kategorie horen met name alle op de χ^2 statistic gebaseerde maten, zoals de coëfficiënten van Cramer, Pearson en Tschuprow, tot de tweede kategorie o.a. de coëfficiënten van Goodman en Kruskal, van Yule, van Somer en van Kendall.
Om dergelijke maten zinvol te kunnen gebruiken, zijn een aantal eigenschappen gewenst (zie bv. Kendall en Stuart, 1961, p. 538, of Galtung, 1967, p. 207 e.v.).
De belangrijkste van deze eigenschappen zijn:

1. a) de coëfficiënt is nul als er geen associatie tussen de variabelen bestaat ("voldoende" formulering);
   b) de coëfficiënt is nul als er geen associatie tussen de variabelen bestaat, en alleen dan ("nodig" formulering);
2. a) de coëfficiënt neemt de maximale waarde aan als de variabelen maximaal samenhangen[2)] ("voldoende" formulering";
   b) de coëfficiënt neemt de maximale waarde aan als de variabelen maximaal samenhangen en alleen dan ("nodige" formulering);
3. de coëfficiënt moet, indien van toepassing, de richting van het verband aangeven;
4. de coëfficiënt moet genormeerd zijn;
5. de waarden van de coëfficiënt moeten onderling vergelijkbaar zijn;
6. de coëfficiënt moet onafhankelijk zijn van het totaal aantal waarnemingen;
7. de coëfficiënt moet onafhankelijk zijn van het aantal klassen van de variabelen;

8. de coëfficiënt moet interpreteerbaar zijn.

We zullen nu mede aan de hand van deze lijst nader ingaan op de genoemde twee groepen associatiematen.

Op χ^2 gebaseerde maten

Deze maten gaan zoals gezegd uit van het begrip statistische onafhankelijkheid. Voor χ^2, gedefinieerd door

$$\chi^2 = n_{..} \sum_{i,j} \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1 = n\left(\sum \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1\right) \qquad (1)$$

en de erop gebaseerde coëfficiënten geldt dan ook als belangrijkste voordeel, dat de waarde nul 1-1 duidig correspondeert met statistische onafhankelijkheid, dus met de situatie waarin er geen enkel verband tussen A en B bestaat. Er is dus voldaan aan eigenschap 1, zelfs aan de sterkste vorm van 1b.
De χ^2-waarde zelf is als associatiemaat echter ongeschikt. Aan veel van de in de vorige paragraaf genoemde eisen is nl. niet voldaan.
De belangrijkste bezwaren zijn, dat de χ^2 waarde:

a) afhankelijk is van het totaal aantal waarnemingen ($n_{..}$)
b) afhankelijk is van het aantal klassen van de verdelingen van de beide variabelen (dus van r en k)
c) afhankelijk is van de verdeling van de beide variabelen over deze klassen, de zogenaamde marginale verdelingen
d) niet genormeerd is.

Ook de tweede eigenschap levert problemen. Behalve de in voetnoot 2 genoemde moeilijkheid om vast te stellen wat precies maximale samenhang is, is als gevolg van de net genoemde bezwaren a t/m c ook de maximale χ^2 waarde afhankelijk van $n_{..}$, r, k en de marginale verdelingen.
Zo geldt dat voor een rxk tabel de maximale waarde die χ^2 kan aannemen, gelijk is aan

$$\chi^2_{max} = n_{..} (\min(r, k) - 1).$$

Bovendien kan deze maximale waarde alleen bereikt worden als de marginale verdelingen van beide variabelen gelijk zijn, of door het samenvoegen van klassen aan elkaar gelijk gemaakt kunnen worden.

Aangezien dit laatste punt van algemener belang is, staan we er wat langer bij stil en wel aan de hand van een vergelijking van de tabellen 1a en 1b.

		B 1	B 2	
A	1	40	0	40
	2	0	60	60
		40	60	100

$\chi^2 = \chi^2_{max} = 100$

Tabel 1a.

		B 1	B 2	
A	1	40	0	40
	2	10	50	60
		50	50	100

$\chi^2 = 66.67$

Tabel 1b.

Voor beide tabellen geldt dat de χ^2 waarde maximaal is *bij de gegeven marginale verdelingen*. Wat kunnen we nu zeggen op grond van een *vergelijking* van de χ^2 waarden? Is de samenhang tussen A en B in tabel 1a sterker dan in tabel 1b? Het is duidelijk, dat deze vraag pas kan worden beantwoord als precieser is omschreven wat in dit geval onder samenhang wordt verstaan. Kendall en Stuart (1961, p. 540) maken in dit verband voor 2x2 tabellen onderscheid tussen wat zij noemen *"complete association"* en *"absolute association"*.
Onder "absolute association" verstaan ze een sitatie zoals in tabel 1a is gegeven, waar met iedere waarde voor A 1-1 duidig een waarde voor B correspondeert. Tabel 1b geeft een voorbeeld van het zwakkere begrip "complete association": iedere eenheid met op variabele A de waarde 1 heeft ook de waarde 1 op variabele B, maar omgekeerd geldt niet voor iedere eenheid met B = 1 dat ook voor A = 1. Desgewenst kunnen we deze begrippen generaliseren naar het algemenere geval van een rxk tabel.
De vraag in hoeverre het verschil in χ^2 waarde tussen de tabellen 1a en 1b relevant is, kunnen we dus ook beschouwen als de vraag of we in een vorm van "absolute association" of juist in een vorm van "complete association" geïnteresseerd zijn. In het eerste geval is het verschil wel, in het tweede niet relevant. Het gebruik van χ^2 is dus alleen zinvol in gevallen waarin we in "absolute association" zijn geïnteresseerd, en ook alleen dan is aan eigenschap 2b voldaan.

De verschillende op χ^2 gebaseerde coëfficiënten zijn voorgesteld, om met behoud van eigenschap 1 en 2 een of meer van de bovengenoemde bezwaren te ondervangen.

$$\phi = \sqrt{\frac{\chi^2}{n_{..}}} \qquad (2)$$

Deze coëfficiënt is door de deling door $n_{..}$ niet meer afhankelijk van het totaal aantal onderzoekseenheden, maar de bezwaren b t/m d, blijven evenals de opmerkingen t.a.v. de maximaal bereikbare waarde onverminderd van kracht. De ϕ is dan ook minder geschikt als associatiemaat.

$$V^2 = \frac{\chi^2}{n_{..}\ (\min(r, k) - 1)} \qquad \text{Cramers } V^2 \qquad (3)$$

In feite dus $V^2 = \chi^2/\chi^2_{max}$, m.a.w. V^2 is een genormeerde χ^2, met maximale waarde 1.
De bezwaren a, b en d vervallen nu dus, maar (c) blijft gelden. Zodat ook voor V^2 weer geldt dat het van de marginale verdelingen afhangt of de maximale waarde bereikt kan worden.
Tenslotte noemen we nog de door Pearson voorgestelde contingency coëfficiënt

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n_{..}}} \qquad (4)$$

Aangezien deze weer de bezwaren b, c en d heeft, is een iets gewijzigde vorm, nl. de genormeerde contingency coëfficiënt beter hanteerbaar.

$$C^1 = \frac{C}{C_{max}} = \sqrt{\frac{\chi^2}{\chi^2 + n_{..}} \cdot \frac{\min(r, k)}{\min(r, k) - 1}} \qquad (5)$$

Waarvoor weer alleen bezwaar (c) overblijft.

Samenvattend kan dus worden gesteld, dat Cramers V^2 en de genormeerde C^1 van Pearson voldoen aan de meeste voor associatiematen gewenste eigenschappen. Belangrijkste bezwaar blijft de afhankelijkheid van de marginale verdeling, waardoor de onderlinge vergelijkbaarheid van op verschillende tabellen gebaseerde waarden wordt bemoeilijkt (eigenschap 5).
Belangrijkste voordeel van deze coëfficiënten is dat ze doordat ze op χ^2 zijn gebaseerd, gevoelig zijn voor alle vormen van samenhang (eigenschap 1b).
In tabel 2 zijn de bevindingen van deze paragraaf nog eens schematisch weergegeven.

	eigenschap									
maat	1a	1b	2a	2b	3	4	5	6	7	8
χ^2	+	+	+	1)		-	-	-	-	-
ϕ	+	+	+	1)		-	-	+	-	-
V	+	+	+	1)		+	-	+	+	-
C	+	+	+	1)		-	-	+	-	-
C^1	+	+	+	1)		+	-	+	+	-
λ	+	-	+	1)		+	?	+	+	+
γ	+	-	+	2)	+	+	-	+	+	+

1) alleen voor "absolute association"

2) alleen voor "complete association".

Tabel 2.: Overzicht eigenschappen coëfficiënten.

PRE-maten

Tot de tweede kategorie hoort een groot aantal coëfficiënten van het zogenaamde Proportional Reduction in Error type. De idee hier achter is dat wanneer er sprake is van samenhang tussen twee variabelen A en B, dat dan kennis van de A-score van een objekt ons moet kunnen helpen bij het schatten van zijn B-score, en wel meer naarmate het verband sterker is.

We vergelijken daartoe twee situaties:

a) "voorspel" de B-score van een aselect uit de onderzoeksgroep gekozen objekt.
b) idem als de A-score gegeven is.

De algemene gedaante van de PRE-maat is dan

$$PRE_{B/A} = \frac{\text{kans op fouten bij (a) - kans op fouten bij (b)}}{\text{kans op fouten bij (a)}}$$

Dus inderdaad de relatieve vermindering van de kans op een fout doordat de A-score bekend is.

Een bepaalde coëfficiënt ontstaat nu door

- vast te stellen wat als fout wordt aangemerkt
- vast te stellen volgens welk voorschrift de B-score in beide situaties wordt voorspeld.

Hierbij is vooral de keuze van het voorschrift in situatie (b) belangrijk. Dit legt nl. vast hoe de informatie over de A-score wordt benut voor het voorspellen van de B-score, oftewel welk verband er tussen A en B wordt verondersteld. Hiermee wordt dus in feite gedefinieerd naar welk alternatief we kijken en wat we onder maximale samenhang verstaan.
Kenmerkend voor deze coëfficiënten is, dat:

1) ze asymmetrisch zijn. We kijken naar de voorspelling van B op grond van A. Analoog kan natuurlijk ook $PRE_{A/B}$ worden gedefinieerd.
2) de waarde nul niet meer 1-1 duidig met statistische onafhankelijkheid correspondeert. Nu betekent de waarde nul alleen het ontbreken van het verband, zoals dat impliciet is verondersteld door de keuze van het voorschrift volgens welk de B-score in situatie (b) wordt voorspeld. In dit geval zouden we van preditieve onafhankelijkheid kunnen spreken. Er is nu dus wel aan eigenschap 1a, maar niet aan eigenschap 1b voldaan.
3) de maximale waarde 1 1-1 duidig met maximale preditieve afhankelijkheid correspondeert. Kennis van de A-score impliceert dat dan de B-score perfekt kan worden voorspeld.
   Dat desondanks ook voor de tweede eigenschap in doorsnee alleen de zwakke a formulering geldt, hangt samen met de afhankelijkheid van de marginale verdelingen. We komen daar nog op terug.

Een voorbeeld van een dergelijke coëfficiënt is de λ van Goodman en Kruskal. Daarbij wordt in beide situaties gestreefd naar een maximale kans op een goede voorspelling. Dat betekent dat in situatie (a) de voorspelling gelijk aan de modus van de marginale B-verdeling wordt gekozen. En in situatie (b) aan de modus van de B-scores binnen de rij van de kruistabel, die hoort bij de gegeven A-score.
Er kan worden afgeleid dat geldt:

$$\lambda_{B/A} = \frac{\sum_i \max_j n_{ij} - \max_j n_{.j}}{n_{..} - \max_j n_{.j}} \qquad (6)$$

Zoals gezegd geldt voor deze coëfficiënt eigenschap 1 in de zwakkere vorm (1a). Tabel 3 geeft een eenvoudig voorbeeld van een tabel, waarin

wel samenhang bestaat, terwijl $\lambda_{B/A} = 0$ (en ook $\lambda_{A/B} = 0$). Verder is gemakkelijk in te zien dat de eigenschappen 4, 6, 7 en 8 ook gelden. Voor wat betreft de eigenschappen 2 en 5 bestaan er weer een paar beperkingen i.v.m. de marginale verdelingen.

		B			
		1	2	3	
A	1	80	40	30	150
	2	35	20	20	75
	3	35	15	25	75
		150	75	75	300

Tabel 3.: $\lambda_{B/A} = \lambda_{A/B} = 0$

Uit de bovenstaande omschrijving van het model waarop λ is gebaseerd, volgt onmiddellijk dat $\lambda_{B/A} = 1$ (dat is de maximale waarde) wordt als bij iedere A-score, maar precies één B-score voorkomt. (In termen van de kruistabel: als iedere rij maar precies één cel heeft die niet gelijk aan nul is.)
In feite dus een asymmetrische vorm van wat eerder "absolute association" is genoemd. Toen is er ook op gewezen, dat "absolute association" alleen kan worden bereikt bij gelijke marginale verdelingen. M.a.w. de maximale waarde van $\lambda_{B/A}$ kan alleen worden bereikt wanneer de marginale verdelingen gelijk zijn, of wanneer door samenvoegen van klassen van variabele A deze verdelingen gelijk gemaakt kunnen worden.
Net als bij χ^2 is eigenschap 2b ook nu alleen geldig als we kijken naar "absolute association".
De tabellen 4a en 4b illustreren tenslotte de problemen met eigenschap 5.

		B		
		1	2	
A	1	40	10	50
	2	20	30	50
		60	40	100

Tabel 4a.: $\lambda_{B/A} = .25$

		B		
		1	2	
A	1	48	12	60
	2	16	24	40
		64	36	100

Tabel 4b.: $\lambda_{B/A} = .22$

We vinden twee verschillende λ's, m.a.w. de proportie reductie in de foute voorspellingen is voor beide tabellen verschillend. Maar dat is niet het resultaat van een grotere samenhang tussen A en B in tabel 4b,

de conditionele verdelingen van B, gegeven A zijn immers in beide tabellen gelijk! In werkelijkheid wordt het verschil veroorzaakt door het verschil tussen de marginale verdelingen van variabele A in de beide tabellen.
Aan eigenschap 5 is dus niet voldaan, door dat de coëfficiënt niet onafhankelijk is van de marginale verdelingen.

Ordinale samenhang

Nog duidelijker voorbeelden van coëfficiënten uit de tweede kategorie zijn die welke speciaal gevoelig zijn voor ordinale samenhang en strikt genomen alleen toepasbaar zijn, als zowel A als B op ordinaal nivo gemeten is. Er wordt gekeken in hoeverre de ordening van de onderzoekseenheden op grond van variabele A overeenstemt met die op grond van variabele A. Dus in hoeverre er een *monotoon* stijgend of dalend verband is tussen de variabelen.
Als voorbeeld van een dergelijke associatiemaat, kijken we naar de ook door Goodman en Kruskal voorgestelde γ coëfficiënt. (Voor een 2x2 tabel is deze coëfficiënt identiek met de Q van Yule.) Eerst gaan we voor ieder paar onderzoekseenheden x en y na tot welke van de volgende 3 categorieën het behoort. (x_A is de score van x op A, etc.)

I de consistente paren
met $x_A > y_A$ en $x_B > y_B$ òf $x_A < y_A$ en $x_B < y_B$

II de inconsistente paren
met $x_A > y_A$ en $x_B < y_B$ òf $x_A < y_A$ en $x_B > y_B$

III paren met minstens 1 knoop
$x_A = y_A$ en/of $x_B = y_B$

Geven we nu het aantal paren in de 3 categorieën aan met resp. P_I, P_{II}, P_{III}, dan is Goodman en Kruskals γ gedefinieerd door:

$$\gamma = \frac{P_I - P_{II}}{P_I + P_{II}}$$

Dus bij aselecte trekking van twee onderzoekseenheden is γ het verschil tussen de kansen op resp. een consistent en een inconsistent paar, *dit alles gegeven dat in het paar geen knopen voorkomen.*

γ kan dus alle waarden tussen -1 en +1 aannemen, -1 betekent dat voor ieder paar zonder knopen de ordening van de twee eenheden volgens beide variabelen tegengesteld is, +1 juist dat die ordening in beide gevallen dezelfde is. We gaan ook nu weer het lijstje met eigenschappen na. Gemakkelijk is in te zien dat aan de eigenschappen 1a, 3, 4, 6, 7 en 8 is voldaan. Eigenschap 1b geldt niet. Aangezien de γ een maat is voor een speciale vorm van associatie, komt $\gamma = 0$ ook voor in al die gevallen waarin sprake is van een andere vorm van associatie, maar waarbij het aantal consistente en inconsistente paren elkaar in evenwicht houden.

Ook nu leveren de eigenschappen 2 en 5 weer meer problemen.

γ is immers gedefinieerd als het verschil van twee *voorwaardelijke* kansen, de kansen op een consistent en een inconsistent paar onder de voorwaarde dat in het paar geen knopen voorkomen. De *coëfficiënt wordt dus alleen over de niet geknoopte paren berekend.*

Voor tabellen met verschillende fracties ongeknoopte paren (F.O.P.) zijn de γ's dus op ongelijke fracties van het totaal aantal paren gebaseerd en daardoor niet zonder meer vergelijkbaar.

Indirekt spelen de marginale verdelingen hier ook weer een rol, voor iedere tabel geldt nl. dat de maximale F.O.P. wordt bepaald door deze marginale verdelingen.

Zo is bijvoorbeeld voor 2x2 tabellen F.O.P. maximaal 0.50, maar voor een 2x2 tabel met beide marginale verdelingen 20, 80, is de F.O.P. maximaal 0.32! Deze zelfde overwegingen zijn ook van belang i.v.m. eigenschap 2.

In de tabellen 5a en 5b geldt in beide gevallen dat $\gamma = 1$, omdat alle niet geknoopte paren consistent zijn, maar de F.O.P. en daarmee dus het gedeelte van het totale aantal paren, waarop γ is gebaseerd is duidelijk verschillend.

		B 1	B 2	
A	1	50	0	50
	2	0	50	50
		50	50	100

Tabel 5a.: $\gamma = 1$; F.O.P. = .50

		B 1	B 2	
A	1	50	0	50
	2	50	50	100
		100	50	150

Tabel 5b.: $\gamma = 1$; F.O.P. = .22

Moeten we nu toch in beide gevallen van maximale samenhang spreken? In ieder geval geldt eigenschap 2b alleen dan wanneer we maximale samenhang zo definiëren, dat beide gegeven voorbeelden daaronder vallen. In de terminologie van Kendall en Stuart betekent dat voor 2x2 tabellen, dat we ons met γ moeten beperken tot vormen van "complete association". [3] De zwakkere eigenschap van 2a geldt in alle gevallen waarin we samenhang als ordinale samenhang en maximale samenhang, dus als monotoon stijgend of dalend, opvatten.

Vergelijking associatiematen

Tenslotte illustreren we aan de hand van twee voorbeelden een paar van de verschillen en overeenkomsten tussen de besproken associatiematen.

Het eerst voorbeeld betreft de hiernaast in tabel 6 gegeven 2x3 tabel. Voor deze tabel is de waarde van de coëfficiënten V^2, C^1, $\lambda_{B/A}$ en γ bepaald als functie van de celinhouden n_{11} en n_{12}.

		B			
		1	2	3	
A	1	n_{11}	n_{12}	$40-n_{11}-n_{12}$	40
	2	$20-n_{11}$	$30-n_{12}$	$10+n_{11}+n_{12}$	60
		20	30	50	100

Tabel 6.

De resultaten zijn in figuur 1 t/m 4 met behulp van hoogtelijnen weergegeven. De ellipsvormige hoogtelijnen in figuur 1 en 2 voor de op χ^2 gebaseerde V^2 en C^1 illustreren, dat hier alle afwijkingen van het "nulpunt" (statistische onafhankelijkheid): $n_{11} = 8$, $n_{12} = 12$ ongeacht hun richting worden gesignaleerd. Daartegenover Goodman en Kruskal's γ (figuur 4), die alleen verandert bij wijzigingen van n_{11} en n_{12} waardoor het ordinale verband wordt versterkt.
Voor $\lambda_{B/A}$ zien we weer een ander patroon. Het meest opvallend hier is het grote gebied waar $\lambda_{B/A} = 0$. Dit is een gevolg van de hier gehanteerde voorspelregel. Vergelijk ook met het voorbeeld uit tabel 3.

In het tweede voorbeeld willen we de afhankelijkheid van de marginale verdelingen laten zien. Bij alle in het voorgaande besproken coëfficiënten kwam deze afhankelijkheid naar voren.

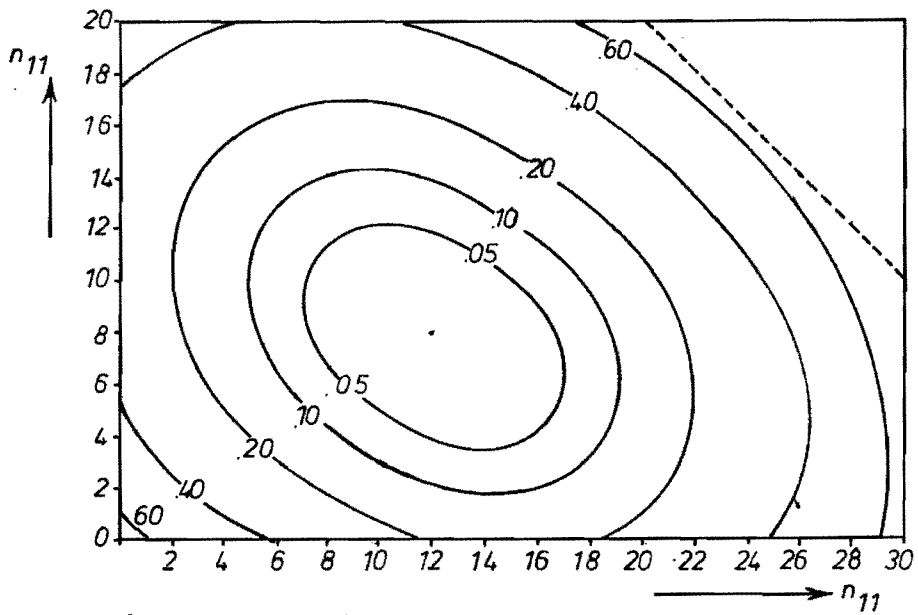


Fig. 1. Hoogtelijnen V^2 voor tabel 6

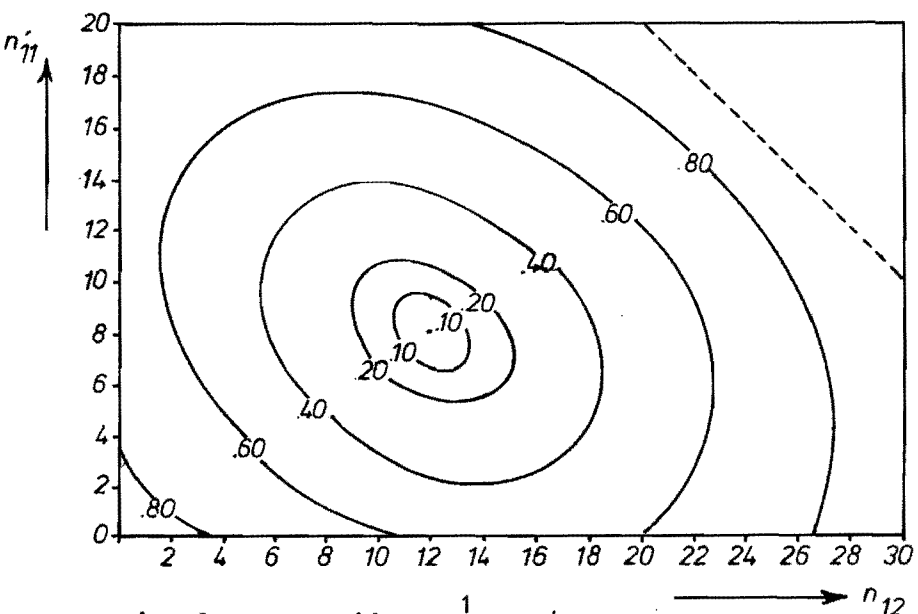


Fig. 2. Hoogtelijnen C^1 voor tabel 6

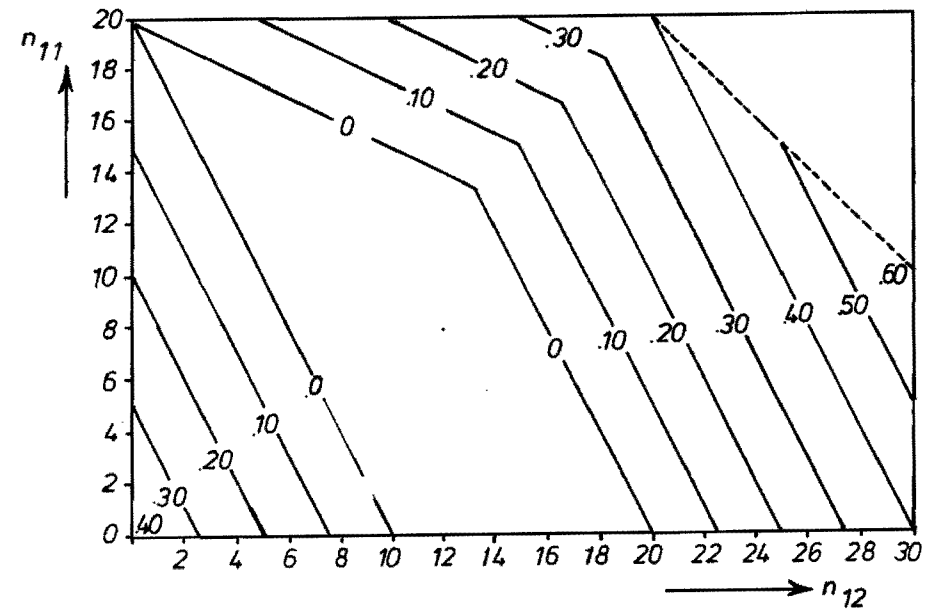


Fig. 3. Hoogtelijnen λ_{BIA} voor tabel 6

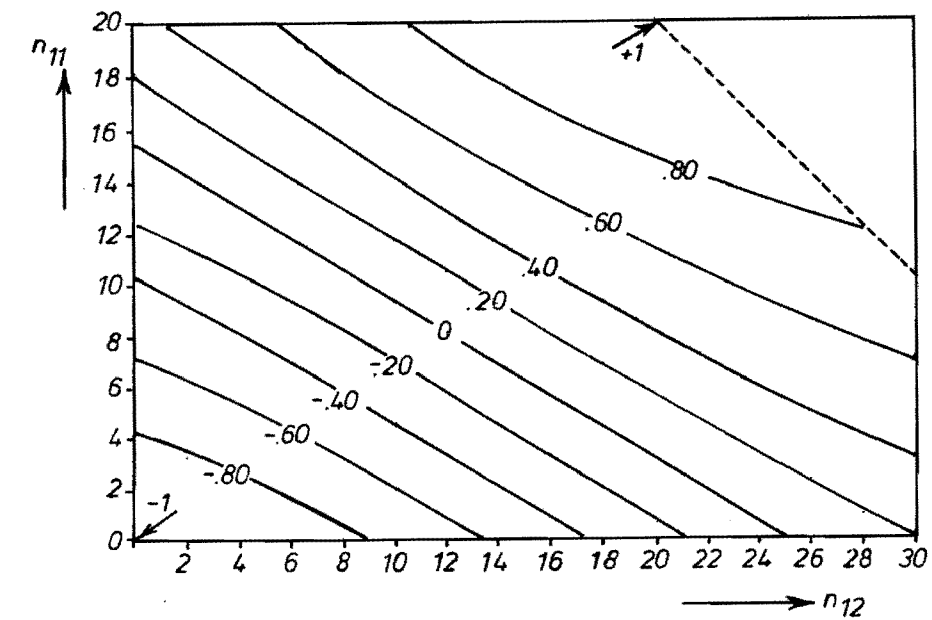


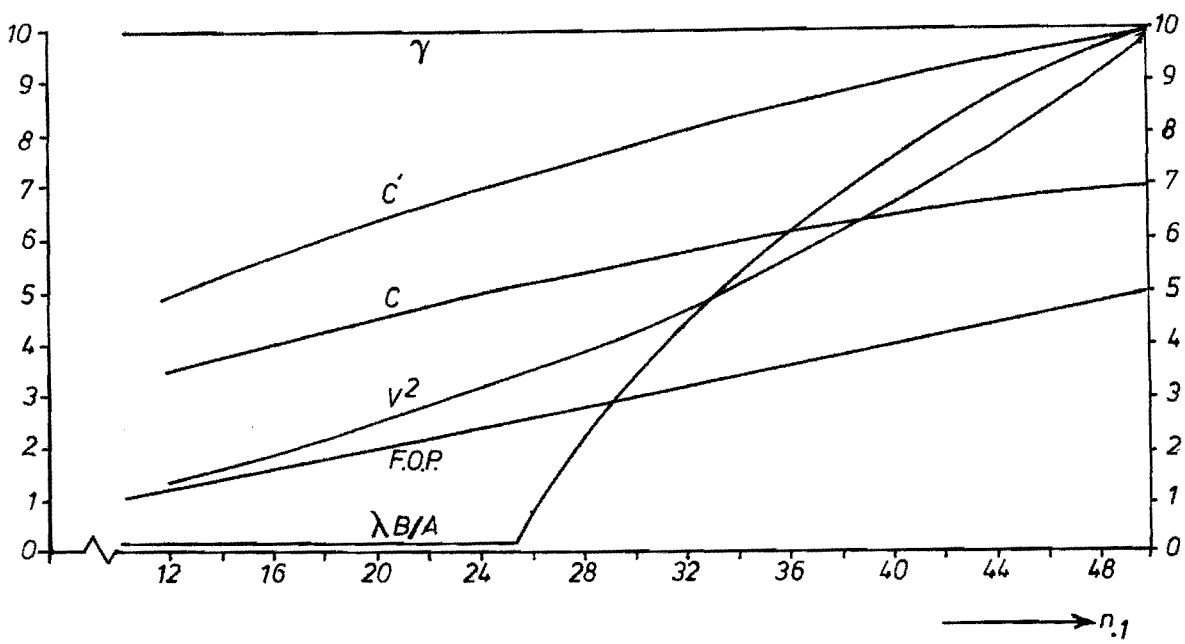
Fig. 4. Hoogtelijnen γ voor tabel 6

Dit aspect is vooral van belang in verband met de onderlinge vergelijkbaarheid van op verschillende tabellen gebaseerde coëfficiënten.
Tabel 7 geeft de 2x2 tabel die in dit voorbeeld wordt gebruikt. Voor ieder van de coëfficiënten is, nu als functie van de marginale verdeling van B ($10 \leq n_{.1} \leq 50$), de maximale waarde bepaald.

		B		
		1	2	
A	1	$n_{.1}$	$50-n_{.1}$	50
	2	0	50	50
		$n_{.1}$	$100-n_{.1}$	100

Tabel 7.

Figuur 5 geeft het resultaat. In verband met γ is ook de maximale F.O.P. in de figuur opgenomen. Opnieuw valt de ongevoeligheid van $\lambda_{B/A}$ op. Voor $10 \leq n_{.1} \leq 25$ is deze steeds nul, ongeacht de verdere verdeling in de tabel. Verder zien we een duidelijk verschil tussen γ enerzijds en de op χ^2 gebaseerde coëfficiënten anderzijds. Zoals al eerder aangestipt kan dit verschil worden verklaard, vanuit het verschil tussen "complete" en "absolute association". "Complete association" kan worden bereikt ongeacht de marginale verdelingen. "Absolute association" alleen als de marginale verdelingen gelijk zijn of door het samenvoegen van klassen gelijk gemaakt kunnen worden.



Figuur 5. Maximale waarde van enkele associatiematen voor tabel 7, als functie van $n_{.1}$

Conclusies

- Het toetsen op onafhankelijkheid levert geen antwoord op de vraag naar sterkte en vorm van een eventueel verband. De toets kijkt alleen hoe groot de likelihood is dat de geobserveerde samenhang ook in de populatie bestaat.
- De op χ^2 gebaseerde associatiematen geven aan hoe groot de afwijking is t.o.v. statistische onafhankelijkheid en niet hoe groot de overeenkomst is met een vorm van maximale samenhang.
- Om een maat te krijgen die dit laatste doet moet de omschrijving van associatie worden toegespitst, wat betekent dat van de vele mogelijke soorten van samenhang er een groot aantal moeten afvallen.
- Wanneer het gebruik waarvoor een associatiemaat moet dienen en daarmee het begrip maximale samenhang scherp kan worden omschreven, dan is het veelal zonder veel moeite mogelijk een passende maat te vinden of te definiëren. Hierbij kan de PRE-aanpak goede diensten bewijzen.
- Alle genoemde coëfficiënten zijn afhankelijk van de marginale verdelingen, dit bemoeilijkt met name onderlinge vergelijking.
- De keuze van een coëfficiënt zal moeten worden gedaan mede op grond van de volgende punten:
  - is het zinvol om te toetsen
  - is er reden om een bepaald verband (een bepaald patroon in de tabel) te veronderstellen
  - zijn de variabelen op nominaal of ordinaal nivo gemeten.

Noten

1) In het algemeen is het niet eenvoudig om statistische onafhankelijkheid vast te stellen. Vooral omdat de onderzoekseenheden veelal niet een aselecte steekproef uit één of ander populatie zijn, kan hierbij immers niet gebruik worden gemaakt van de statistische toetsingstheorie.

2) Deze tweede eigenschap is wat moeilijker vast te stellen dan de eerste. We moeten daarvoor nl. afspreken wat maximale samenhang betekent, en dit zal in z'n algemeenheid, dus zonder de vorm van de samenhang aan te geven, niet goed mogelijk zijn.

3) Voor grotere tabellen kunnen we voor ordinale samenhang het onderscheid tussen "absolute" "complete association" generaliseren, door het onderscheid tussen een strikt monotoon stijgend verband (d.w.z., dat voor ieder tweetal eenheden x en y waarvoor $x_A > y_A$ ook $x_B > y_B$) en een zogenaamd monotoon niet-dalend verband (d.w.z., dat $x_A > y_A$ impliceert dat $x_B \geq y_B$). Goodman en Kruskals γ is nu maximaal zodra het verband tussen A en B aan deze tweede (zwakkere) monotonie voldoet.
De vergelijkbare coëfficiënt d van Somer is alleen maximaal (=1) wanneer er sprake is van strikte monotonie (zie tabel 8).

		B			
		1	2	3	
A	1	50	0	0	50
	2	50	50	0	100
	3	0	50	50	100
		100	100	50	250

Tabel 8a.: $\gamma = 1$; $d = 0.75$

		B			
		1	2	3	
A	1	50	0	0	50
	2	0	50	0	50
	3	0	0	50	50
		50	50	50	150

Tabel 8b.: $\gamma = 1$; $d = 1$

Literatuur

Galtung, J. (1967) - Theory and methods of social research London, George Allen & Anwin Ltd.

Kendall, M.G. en A. Stuart (1961) - The advanced theory of statistics, vol. 2, London; Charles Griffin.