

DE WEERGAVE VAN DE SAMENHANGSSTRUKTUUR BIJ DE ANALYSE VAN $2 \times 2 \times 2$ TABELLEN

G.J. VAN DRIEL *)

Samenvatting

In dit artikel wordt uiteengezet hoe de samenhangsstructuur in een $2 \times 2 \times 2$ tabel, zoals die bij een analyse "op de hand" eenvoudig is vast te stellen, door een drietal formele methoden wordt weergegeven en van additionele informatie wordt voorzien. Deze drie methoden zijn de zeta-analyse, de log-lineaire analyse en de logitanalyse. De zeta-analyse geeft, naast de samenhangsstructuur, kengetallen voor de intensiteit van de onderscheiden samenhangen. De beide andere methoden zijn onderling zeer nauw verwant, mits bij de logitanalyse een bepaalde interactie-term als verklarende variabele wordt gehanteerd. Bij de log-lineaire analyse worden tabellen geproduceerd die een vergelijking met de waargenomen tabel en daarmee een uitspraak over de relevantie of de significantie van de gevonden samenhangen mogelijk maken. De logitanalyse, tenslotte, geeft regressiecoëfficiënten van de samenhangen.

1. Inleiding

Het eenvoudigste categorische systeem met drie variabelen is de tripele dichotomie, die kan worden weergegeven in een $2 \times 2 \times 2$ tabel. In dit artikel worden de drie variabelen achtereenvolgens x , y en z genoemd, en genummerd 1, 2 en 3. De klasse-aanduidingen zijn respectievelijk X en X^C , Y en Y^C , Z en Z^C . Een en ander leidt tot de weergave van de frekwenties w_{ijk} in de acht klassen van het drievoudige categorische systeem conform de tabel links onder. Waar het mogelijk is om de frekwenties met een eenvoudiger symbool aan te geven, worden de letters a tot en met h conform de tabel rechts onder gebruikt.

		in Z		in Z^C	
		X	X^C	X	X^C
Y		w_{111}	w_{211}	w_{112}	w_{212}
Y^C		w_{121}	w_{221}	w_{122}	w_{222}

	in Z		in Z^C	
	X	X^C	X	X^C
Y	a	b	e	f
Y^C	c	d	g	h

*) Vakgroep Statistiek, Faculteit der Economische Wetenschappen, Erasmus Universiteit Rotterdam, Postbus 1738, 3000 DR Rotterdam, tel. 010 - 408 1523.

Bij de analyse van een $2 \times 2 \times 2$ tabel "op de hand" zal men in het algemeen de paarsgewijze samenhangen in de drie (te vormen) marginale 2×2 tabellen en de zes (te vormen) 2×2 tabellen-binnen-een-deelgroep onderzoeken. Bij de marginale tabellen is de samenhang tussen de variabelen eenduidig gedefinieerd als het ontbreken van onafhankelijkheid; we spreken in dit kader verder steeds van (de aanwezigheid van) marginale samenhang. Ook in de afzonderlijke deeltabellen is samenhang op dezelfde wijze eenduidig gedefinieerd. Het is gebruikelijk om van directe samenhang tussen twee variabelen te spreken zodra **niet** in beide relevante deeltabellen sprake is van onafhankelijkheid. Met de zes al of niet aanwezige paarsgewijze samenhangen is een specifieke vorm van samenhangsstructuur vastgelegd. Deze kan in een diagram worden weergegeven. Dit wordt gedaan in paragraaf 2. We noemen een samenhangsstructuur en het korresponderende diagram "kompleet" als geen van de zes samenhangen ontbreekt. Naast deze zes paarsgewijze samenhangen kan ook de aanwezigheid van tweede-orde interactie (zie paragraaf 2) eenvoudig op de hand worden vastgesteld.

Het is daarom goed om zich af te vragen of men bij de analyse van $2 \times 2 \times 2$ tabellen wel verder moet gaan dan deze analyse op de hand. Het antwoord op deze vraag wordt gegeven door de praktijk, waar in empirische tabellen vrijwel nooit een exacte paarsgewijze onafhankelijkheid valt waar te nemen, zodat het diagram van de samenhangsstructuur bijna altijd compleet is en daarmee in zekere zin waardeloos wordt.

Omdat echter de intensiteit van de samenhangen sterk kan verschillen, ontstaat er behoefte aan een grondiger inzicht in de mate van samenhang. Daartoe wordt in dit artikel aandacht geschonken aan een drietal formele of, zo men wil, mechanistische methoden voor de analyse van $2 \times 2 \times 2$ tabellen. Ten behoeve van de lezer die met één of meer van deze methoden niet vertrouwd is, wordt van elke methode een korte uitleg gegeven.

De eerste methode is gebaseerd op de analyse van de ζ -coëfficiënten van een $2 \times 2 \times 2$ tabel, zoals gesystematiseerd door Lammerts van Bueren [6]; verder aan te geven als zeta-analyse. De weergave van de samenhangsstructuur volgens deze methode wordt besproken in paragraaf 3. In het kader van dit artikel past ook een inzicht in alle mogelijke samenhangsstructuren. Dit overzicht kan worden verkregen op basis van een verdieping van de zeta-analyse. Hieraan is paragraaf 4 gewijd.

De tweede methode is gebaseerd op het schatten van de parameters van een log-lineair model ter beschrijving van een $2 \times 2 \times 2$ tabel. De weergave van de samenhangsstructuur volgens deze methode wordt besproken in paragraaf 5.

De derde methode is gebaseerd op de drievoudige logitanalyse van een $2 \times 2 \times 2$ tabel. De weergave van de samenhangsstructuur volgens deze methode wordt besproken in paragraaf 6.

De konklusies van het onderzoek zijn samengevat in paragraaf 7.

2. De analyse van een $2 \times 2 \times 2$ tabel op de hand

Zoals in de inleiding is gezegd, kunnen $2 \times 2 \times 2$ tabellen heel goed op de hand worden geanalyseerd. We illustreren deze uitspraak met een voorbeeld. De gepresenteerde tabel heeft betrekking op 300 propedeusestudenten, onderscheiden naar het tentamenresultaat voor vak x [X = geslaagd voor vak x, X^C = niet geslaagd voor vak x], het tentamenresultaat voor vak y [Y = geslaagd voor vak y, Y^C = niet geslaagd voor vak y] en de geschiktheid voor de studie z [Z = geschikt voor de studie, Z^C = niet geschikt voor de studie]. Aan de vraag hoe men de uitkomsten voor die laatste variabele zou moeten vaststellen, gaan we hier voorbij. De frekwentietabel van de 300 tentamenresultaten is als volgt.

	in Z		in Z^C	
	X	X^C	X	X^C
Y	112	28	6	14
Y^C	48	12	24	56

In de beide deeltabellen voor de geschikte en de niet-geschikte studenten bestaat onafhankelijkheid tussen de variabelen x en y, dat wil zeggen tussen de resultaten voor vak x en vak y. Tellen we de tabellen op, dan is er wel marginale samenhang tussen de variabelen x en y. Men noemt deze marginale samenhang daarom wel geïnduceerde samenhang en spreekt ook wel van schijn-associatie tussen de variabelen x en y.

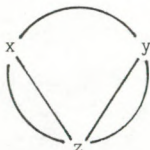
Hoewel niet strikt noodzakelijk, is het soms handig om de tabel te herschikken ten behoeve van het onderzoek naar de samenhangen tussen de variabelen x en z respectievelijk y en z. We doen dit voor de variabelen x en z als volgt.

	in Y		in Y^C	
	X	X^C	X	X^C
Z	112	28	48	12
Z^C	6	14	24	56

In de beide deeltabellen en in het totaal is er een duidelijke positieve samenhang tussen de variabelen x en z waar te nemen; van de geschikte studenten is 80% geslaagd voor vak x en van de niet-geschikte studenten slechts 30%.

Bij het onderzoek naar de samenhang tussen de variabelen y en z worden analoge resultaten gevonden: van de geschikte studenten is 70% geslaagd voor vak y en van de niet-geschikte studenten slechts 20%.

De uitkomsten zijn "ideaal" in de zin dat binnen de deelgroepen Z en Z^C exacte onafhankelijkheid tussen de variabelen x en y is waar te nemen, waardoor een zinvol [= inkompleet] diagram van de samenhangsstructuur is te tekenen. In dit diagram wordt een rechte lijn getrokken tussen twee punten, zodra er directe samenhang bestaat tussen de met die punten korresponderende variabelen. Een boog wordt getrokken tussen twee punten zodra er marginale samenhang bestaat tussen de met die punten korresponderende variabelen. In dit voorbeeld ontbreekt alleen de directe samenhang tussen de variabelen x en y, zodat het diagram van de samenhangsstructuur er als volgt uit komt te zien.



Naast het onderzoek naar de drie directe en de drie marginale paarsgewijze samenhangen is het ook mogelijk om na te gaan of er sprake is van tweede-orde interactie. Dit begrip is in 1935 door Bartlett [3] voor de $2 \times 2 \times 2$ tabel (in negatieve zin) gedefinieerd door te stellen dat er geen tweede-orde interactie in zo'n tabel is indien en slechts indien geldt dat:

$$\frac{w_{111}w_{221}}{w_{121}w_{211}} = \frac{w_{112}w_{222}}{w_{122}w_{212}}.$$

Gemakshalve zal in het vervolg kortweg van "de" interactie worden gesproken, omdat de eerste-orde interactie directe samenhang is genoemd. Uit de definitie volgt onmiddellijk dat het ontbreken van directe samenhang tussen twee variabelen een voldoende voorwaarde is voor het ontbreken van verdere interactie. In het gegeven voorbeeld is aan deze voldoende voorwaarde dus voldaan. Op de betekenis van de aanwezigheid, of beter gezegd de afwezigheid, van interactie in een $2 \times 2 \times 2$ tabel komen we in de volgende paragrafen nog terug.

3. De zeta-analyse van een $2 \times 2 \times 2$ tabel

Uitgaande van de acht frekwenties w_{ijk} [$i, j, k = 1, 2$] van de tabel, wordt het frekwentietotaal symbolisch weergegeven als

$$\sum_i \sum_j \sum_k w_{ijk} = w_{+++}.$$

Op soortgelijke wijze noteren we alle mogelijke marginale frekwenties, zoals

$$\sum_i w_{ijk} = w_{+jk} \quad \text{en} \quad \sum_i \sum_j w_{ijk} = w_{++k}.$$

Vervolgens worden de relatieve frekwenties gedefinieerd als $p_{ijk} = w_{ijk}/w_{+++}$ en daarmee liggen p_{+jk} , enzovoorts, vast.

Bij de zeta-analyse van een $2 \times 2 \times 2$ tabel staan de variabiliteiten van de relatieve-frekwentieverdelingen (ook wel entropieën genoemd) centraal. Hier-voor wordt het symbool H gebruikt. Er zijn zeven variabiliteiten:

$$H(x,y,z) = - \sum_i \sum_j \sum_k p_{ijk} \log p_{ijk}$$

$$H(x,y) = - \sum_i \sum_j p_{ij+} \log p_{ij+} \quad \text{en analoog voor } H(x,z) \text{ en } H(y,z)$$

$$H(x) = - \sum_i p_{i++} \log p_{i++} \quad \text{en analoog voor } H(y) \text{ en } H(z).$$

Uit deze variabiliteiten zijn vervolgens vier kovariabiliteiten te bepalen. Deze kovariabiliteiten zijn als volgt gedefinieerd:

$$T(x:y:z) = H(x) + H(y) + H(z) - H(x,y,z)$$

$$T(x:y) = H(x) + H(y) - H(x,y) \quad \text{en analoog voor } T(x:z) \text{ en } T(y:x).$$

Het is te bewijzen dat $T(x:y) = 0$ dan en slechts dan als tussen de variabelen x en y marginale onafhankelijkheid bestaat. De betekenis van $T(x:y:z) = 0$ is soortgelijk. Onder meer door Pannekoek [7] is gewezen op het feit dat de (marginale) kovariabiliteiten proportioneel zijn [met een factor $2n$] met de toetsingsgrootheid G^2 zoals die in de log-lineaire analyse wordt gebruikt voor de vaststelling van de "goodness of fit" van een onafhankelijkheidsmodel.

Uit de zeven variabiliteiten kunnen negen konditionele variabiliteiten worden afgeleid.

Notatie en definitie zijn als volgt:

$$H_Z(x,y) = H(x,y,z) - H(z)$$

$$H_Z(x) = H(x,z) - H(z) \quad \text{en analoog voor } H_Y(\dots) \text{ en } H_X(\dots).$$

$$H_Z(y) = H(y,z) - H(z)$$

Met behulp van deze negen konditionele variabiliteiten worden vervolgens drie konditionele kovariabiliteiten gedefinieerd.

$$T_Z(x:y) = H_Z(x) + H_Z(y) - H_Z(x,y) \quad \text{en analoog voor } T_Y(x:z) \text{ en } T_X(y:z)$$

De waarden van de variabiliteiten en de kovariabiliteiten die behoren bij de tabel uit paragraaf 2 zijn hieronder vermeld:

$H(x,y,z) = 1,748$	$T(x:y:z) = 0,237$	
$H(x,y) = 1,321$	$H(x,z) = 1,174$	$H(y,z) = 1,211$
$H(x) = 0,657$	$H(y) = 0,691$	$H(z) = 0,637$
$H_Z(x,y) = 1,111$	$H_Y(x,z) = 1,057$	$H_X(y,z) = 1,091$
$H_Z(x) = 0,537$	$H_Y(x) = 0,630$	$H_X(y) = 0,664$
$H_Z(y) = 0,574$	$H_Y(z) = 0,520$	$H_X(z) = 0,517$
$T(x:y) = 0,027$	$T(x:z) = 0,120$	$T(y:z) = 0,117$
$T_Z(x:y) = 0,000$	$T_Y(x:z) = 0,093$	$T_X(y:z) = 0,090$

Het is te bewijzen dat, bijvoorbeeld, de term $T_Y(x:z)$ is op te vatten als een gewogen gemiddelde van de kovariabiliteiten tussen x en z in de beide deelgroepen Y en Y^c . Aangezien kovariabiliteiten niet-negatief zijn, kunnen de konditionele kovariabiliteiten slechts gelijk zijn aan nul dan en slechts dan als er in beide deeltabellen sprake is van onafhankelijkheid. Een positieve konditionele kovariabiliteit impliceert dus directe samenhang tussen de in het geding zijnde variabelen in minstens één van beide deeltabellen.

Tenslotte kunnen gestandaardiseerde symmetrische dan wel asymmetrische vormen van de (konditionele) kovariabiliteiten worden bepaald. Wij beperken ons tot de symmetrische variant. Zo geldt bijvoorbeeld dat:

$$\zeta(x:y) = 2 T(x:y) / [H(x) + H(y)]$$

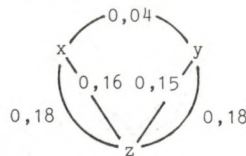
$$\text{en} \quad \zeta_Z(x:y) = 2 T_Z(x:y) / [H_Z(x) + H_Z(y)]$$

Het is te bewijzen dat deze associatiematen tussen 0 en 1 liggen. Voor het gekozen voorbeeld worden de volgende zeta-coëfficiënten gevonden:

$$\begin{array}{lll} \zeta(x:y) = 0,04 & \zeta(x:z) = 0,18 & \zeta(y:z) = 0,18 \\ \zeta_z(x:y) = 0,00 & \zeta_y(x:z) = 0,16 & \zeta_x(y:z) = 0,15 \end{array}$$

Recentelijk is door Abrahamse [2] de asymptotische standaardfout van de zeta-coëfficiënt in een tweedimensionale tabel afgeleid. In een in bewerking zijnde studie wordt hetzelfde gedaan voor de konditionele zeta-coëfficiënt in een drie- of meerdimensionale tabel. Dit opent de mogelijkheid om te kijken naar de significantie van dergelijke zeta-coëfficiënten. In dit artikel wordt hier niet op ingegaan.

Het zal uit het voorgaande duidelijk zijn dat bij een zeta-analyse altijd hetzelfde diagram van de samenhangsstructuur wordt gevonden als bij de analyse op de hand, maar dat de bogen en lijnen nu voorzien kunnen worden van een maat: $\zeta(...)$ voor een marginale samenhang en $\zeta_z(...)$ voor een directe samenhang.



Het enige bezwaar van de zeta-analyse is dat deze geen onmiddellijk inzicht verschaft in het effect en de omvang van de interactie. Wel is door Abrahamse [1] aangetoond dat het verschil tussen de som van $T(x:y)$, $T(x:z)$ en $T(y:z)$ en $T(x:y:z)$ is te splitsen in twee componenten, waarvan de één het effect van de interactie weergeeft en de ander het effect van de zogenaamde "induced association", maar het bezwaar blijft dat de getallen die deze effecten weergeven in absolute zin niet veel betekenis hebben.

4. Twaalf mogelijke samenhangsstructuren

Het is de moeite waard om na te gaan welke samenhangsstructuren binnen een $2 \times 2 \times 2$ tabel kunnen voorkomen. Hiertoe gaan we nog wat dieper in op de zeta-analyse. We definiëren thans de term Q als:

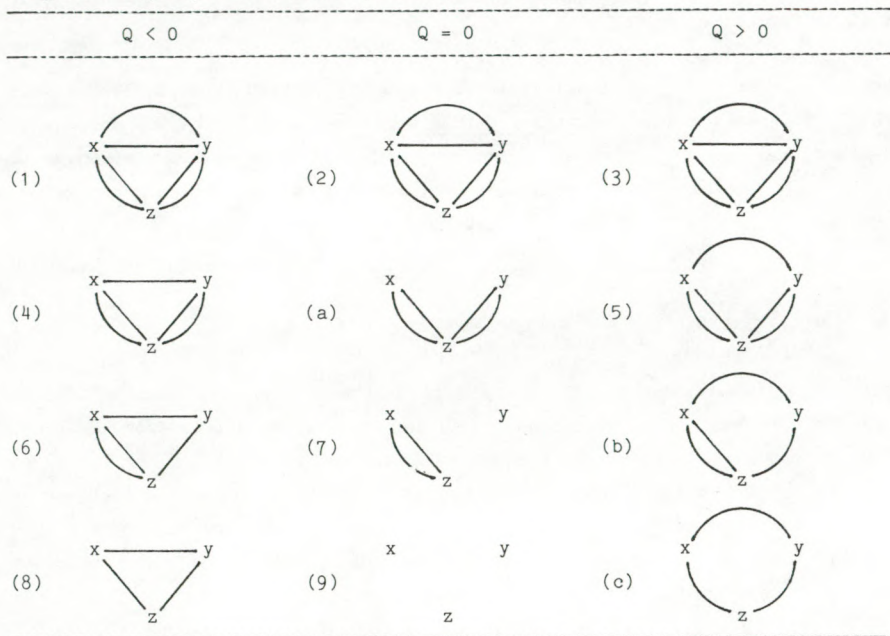
$$\begin{aligned} Q &= T(x:y) + T(x:z) + T(y:z) - T(x:y:z) \\ &= H(x) + H(y) + H(z) - H(x,y) - H(x,z) - H(y,z) + H(x,y,z). \end{aligned}$$

Het is te bewijzen dat Q ook gelijk is aan het verschil tussen $T(x:y)$ en $T_z(x:y)$ en dat dus, als gevolg van de symmetrie in deze relaties, bovendien geldt:

$$T(x:y) - T_z(x:y) = T(x:z) - T_y(x:z) = T(y:z) - T_x(y:z) = Q.$$

Dit laatste resultaat stelt ons in staat om te bewijzen dat het aantal mogelijke samenhangsstructuren in een $2 \times 2 \times 2$ tabel beperkt is en zich laat vangen [behoudens verwisseling van variabelen] in diagram 1, ontleend aan figuur 9.1.1 in Lammerts van Bueren [6].

Diagram 1 Mogelijke en onmogelijke samenhangsstructuren in een $2 \times 2 \times 2$ tabel



Van structuur (a) is aan te tonen dat deze in een $2 \times 2 \times 2$ tabel onbestaanbaar is; het vermoeden bestaat dat deze uitspraak ook voor $r \times s \times t$ tabellen geldt. De structuren (b) en (c) zijn alleen bestaanbaar in gedegenereerde tabellen, dat zijn tabellen met één of meer nullen in de marges.

Aldus blijven er in eerste instantie negen mogelijk samenhangsstructuren over, die in een aantal gevallen echter zowel zonder als met interactie mogelijk zijn. Omdat het ontbreken van een directe samenhang een voldoende voorwaarde

is voor het ontbreken van interactie, bestaan de structuren (5), (7) en (9) slechts in de variant zonder interactie, terwijl van de structuren (6) en (8) kan worden bewezen dat ze alleen mét interactie kunnen bestaan. Tenslotte is structuur (9) voor ons doel niet interessant: in een tabel zonder samenhangsstructuur is natuurlijk ook geen samenhang te ontdekken. Derhalve zijn er uiteindelijk twaalf, verschillende samenhangsstructuren representerende, $2 \times 2 \times 2$ tabellen te onderscheiden: zes zonder en zes met interactie. Deze zijn in diagram 2 gepresenteerd met een numeriek voorbeeld. Het symbool "I" in een plaatje geeft de aanwezigheid van interactie in de bij dat plaatje behorende tabel aan. Het zal uit diagram 1 duidelijk zijn dat alle plaatjes geroteerd kunnen worden. In diagram 2 is dit in feite gedaan bij de plaatjes 10 [= structuur (6) uit diagram 1] en 11 [= structuur (7) uit diagram 1].

5. De log-lineaire analyse van een $2 \times 2 \times 2$ tabel

In zijn meest algemene vorm (zie bijvoorbeeld Fienberg [5]) luidt het log-lineaire model voor drie kwalitatieve variabelen:

$$\log v_{ijk} = u + u_1(i) + u_2(j) + u_3(k) + u_{12}(ij) + u_{13}(ik) + u_{23}(jk) + u_{123}(ijk)$$

waarin v_{ijk} staat voor de volgens dat model te schatten frekwentie in de cel met index $[ijk]$ en waarin alle u -tjes met een index gemiddeld nul zijn over de lopende index c.q. indices. In deze vorm zijn er acht parameters en acht waargenomen frekwenties, zodat de v_{ijk} 's zullen samenvallen met de korresponderende w_{ijk} 's. Het model krijgt dus pas betekenis onder een nader in te voeren hypothese of modelspecificatie, bijvoorbeeld $u_{123}(ijk) = u_{23}(jk) = 0$ [voor alle i, j, k]. Voor drie kwalitatieve variabelen zijn er acht (zogenoemde hiërarchische) log-lineaire model-specificaties mogelijk.

model	omschrijving	betekenis
1.	$[12][13][23]$	$u_{123}(ijk) = 0$
2.	$[12][13]$	$u_{123}(ijk) = u_{23}(jk) = 0$
3.	$[12][23]$	$u_{123}(ijk) = u_{13}(ik) = 0$
4.	$[13][23]$	$u_{123}(ijk) = u_{12}(ij) = 0$
5.	$[12][3]$	$u_{123}(ijk) = u_{13}(ik) = u_{23}(jk) = 0$
6.	$[13][2]$	$u_{123}(ijk) = u_{12}(ij) = u_{23}(jk) = 0$
7.	$[23][1]$	$u_{123}(ijk) = u_{12}(ij) = u_{13}(ik) = 0$
8.	$[1][2][3]$	$u_{123}(ijk) = u_{12}(ij) = u_{13}(ik) = u_{23}(jk) = 0$

Diagram 2. Twaalf specifieke samenhangsstructuren bij 2 x 2 x 2 tabellen



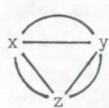
	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	60	0	20	20
Y ^C	20	20	0	60

Q < 0



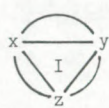
	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	60	0	0	40
Y ^C	20	20	20	40

Q < 0



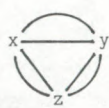
	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	58 ⁵	1 ⁵	18 ⁵	21 ⁵
Y ^C	21 ⁵	18 ⁵	1 ⁵	58 ⁵

Q = 0



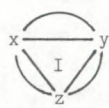
	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	52	8	6	34
Y ^C	28	12	14	46

Q = 0



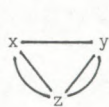
	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	50	10	10	30
Y ^C	30	10	10	50

Q > 0



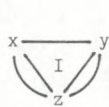
	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	54	6	10	30
Y ^C	26	14	10	50

Q > 0



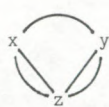
	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	45	15	5	35
Y ^C	35	5	15	45

Q < 0



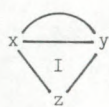
	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	50	10	0	40
Y ^C	30	10	20	40

Q < 0



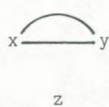
	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	48	12	8	32
Y ^C	32	8	12	48

Q > 0



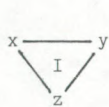
	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	60	0	40	20
Y ^C	20	20	40	0

Q < 0



	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	50	10	50	10
Y ^C	30	10	30	10

Q = 0



	in Z		in Z ^C	
	X	X ^C	X	X ^C
Y	56	4	40	20
Y ^C	24	16	40	0

Q < 0

Zolang de verschillen tussen de geschatte v_{ijk} 's en de waargenomen w_{ijk} 's³⁵ klein genoeg zijn in termen van, bijvoorbeeld, de toetsingsgrootte G^2 , kan men overgaan op een verscherpte specificatie.

Maximale aannemelijkheidsschattingen van de u -tjes en daarmee van de v_{ijk} 's leiden tot een stelsel lineaire vergelijkingen in deze geschatte v_{ijk} 's met als rechterleden bepaalde combinaties van waargenomen randtotalen. De modelspecificatie legt precies vast welke randtotalen dit zijn. Zo geeft bijvoorbeeld de modelspecificatie [12][23] het volgende stelsel:

$$\begin{array}{ll} v_{111} + v_{112} = w_{11+} & v_{111} + v_{211} = w_{+11} \\ v_{121} + v_{122} = w_{12+} & v_{112} + v_{212} = w_{+12} \\ v_{211} + v_{212} = w_{21+} & v_{121} + v_{221} = w_{+21} \\ v_{221} + v_{222} = w_{22+} & v_{122} + v_{222} = w_{+22} \end{array}$$

Het is niet moeilijk in te zien dat, omdat w_{+++} gegeven is, van deze acht vergelijkingen in acht onbekenden er twee afhankelijk zijn. De gegeven modelspecificatie leidt dus tot zes onafhankelijke vergelijkingen in acht onbekenden en heeft daarom twee vrijheidsgraden.

Voor de oplossing van het bij een bepaalde modelspecificatie behorende stelsel vergelijkingen bestaat een (iteratieve) standaard-procedure, die hieronder wordt uiteengezet. Het aantrekkelijke van deze procedure is dat alle hiërarchische log-lineaire modellen ermee zijn te schatten. Bij drie variabelen is deze procedure, behalve bij de modelspecificatie $u_{123(ijk)} = 0$ [voor alle i, j, k], niet-iteratief zolang er geen extra restricties op geschatte v_{ijk} 's in de vorm van gefixeerde nullen zijn.

Expliciet geeft deze methode dus alleen maar de geschatte v_{ijk} 's onder de gekozen modelspecificatie, maar de geschatte u_{ijk} 's zijn hieruit door oplossing van een stelsel lineaire vergelijkingen te bepalen.

Ter illustratie van deze iteratieve methode van de log-lineaire analyse gaan we weer uit van het voorbeeld uit paragraaf 2. De gedachtengang achter de oplossingsmethode is vrij eenvoudig. Ga na in hoeverre de jota-vector [dat is bij de analyse van een $2 \times 2 \times 2$ tabel de vector met acht énen als elementen] voldoet aan alle vergelijkingen. Bij de modelspecificatie [12][13][23] zijn dat er twaalf, waarvan vijf afhankelijke. Beschouw vervolgens de eerste vergelijking en verander de waarden van de v_{ijk} 's die daarin voorkomen proportioneel zodanig dat deze wel aan die vergelijking voldoen. Doe vervolgens hetzelfde bij alle volgende vergelijkingen.

In de praktijk gaat dit met blokken van twee of vier vergelijkingen tegelijk. Zoals blijkt in het voorbeeld worden na het "gladmaken" van het derde blok van vier vergelijkingen, de vergelijkingen in het eerste blok van vier weer geschonden. Daar begint dan een nieuwe iteratie. Het voorbeeld spreekt verder voor zich zelf.

*** Modelspecificatie [12][13][23] ***

Stapnummer en betekenis		schattingen van de v_{ijk} 's na deze stap			
0	opstarten procedure met start-	1	1	1	1
	waarden zonder enige interactie	1	1	1	1
1	$v_{ij+} = w_{ij+}$	59	21	59	21
	i.v.m. specificatie [12]	36	34	36	34
2	$v_{i+k} = w_{i+k}$	99	15	19	27
	i.v.m. specificatie [13]	61	25	11	43
3	$v_{+jk} = w_{+jk}$	121	19	8	12
	i.v.m. specificatie [23]	43	17	17	63
4	noodzakelijke herhaling	111	26	7	16
	van stap 1	52	15	20	53
5 t/m 8	noodzakelijke herhalingen	...	..	..	..
	van de stappen 2 t/m 5	..	..	..	..
9	noodzakelijke herhaling	112	28	6	14
	van stap 6	48	12	24	56

De procedure is gekonvergeerd naar een oplossing die aan alle vergelijkingen voldoet en die een perfecte aansluiting geeft met de waargenomen frekwenties. Dit betekent dus dat er geen (tweede-orde) interactie in deze $2 \times 2 \times 2$ tabel aanwezig is. Het zal duidelijk zijn dat de procedure even goed betekenis heeft, indien er wel interactie aanwezig is, omdat dan het effect ervan door vergelijking van de geschatte en waargenomen frekwenties naar voren komt.

Er kan in dit voorbeeld nog gekeken worden naar scherpere specificaties. In plaats van deze procedure op de voet te volgen, geven we in Tabel 1 de uitkomsten van de volledige log-lineaire analyse.

Tabel 1. Volledige log-lineaire analyse van de 2 x 2 x 2 tabel uit paragraaf 2

schattingen van de v_{ijk} 's onder acht verschillende modelspecificaties

specificatie [12][13][23] [1 vrijheidsgraad]	112 48	28 12	6 24	14 56
specificatie [12][13] [2 vrijheidsgraden]	99 61	15 25	19 11	27 43
specificatie [12][23] [2 vrijheidsgraden]	103 31	37 29	15 41	5 39
specificatie [13][23] [2 vrijheidsgraden]	112 48	28 12	6 24	14 56
specificatie [12][3] [3 vrijheidsgraden]	79 48	28 45	39 24	14 23
specificatie [13][2] [3 vrijheidsgraden]	86 74	21 19	16 14	37 33
specificatie [23][1] [3 vrijheidsgraden]	88 38	52 22	13 51	7 29
specificatie [1][2][3] [4 vrijheidsgraden]	68 59	39 34	34 29	20 17

We zien dat ook onder de modelspecificatie $u_{123(ijk)} = u_{12(ij)} = 0$ [voor alle i, j, k] nog een perfecte aansluiting tussen geschatte en waargenomen frekwenties wordt gevonden. Het is eenvoudig te bewijzen dat in een 2 x 2 x 2 tabel waarin geen directe samenhang bestaat tussen één paar variabelen, de korresponderende log-lineaire modelspecificatie altijd een perfecte aansluiting geeft. Voor twee ontbrekende directe samenhangen geldt hetzelfde.

De konklusie is dus dat ook bij een volledige log-lineaire analyse van een 2 x 2 x 2 tabel altijd dezelfde samenhangsstructuur wordt gevonden als bij een analyse op de hand.

Het voordeel van de log-lineaire analyse is daarbij dat in situaties waarin geen perfecte aansluiting wordt gevonden bij één van de modelspecificaties 2 tot en met 8 en het diagram van de samenhangsstructuur dus compleet is, een indruk wordt verkregen van de verschillen tussen geschatte en waargenomen frekwenties.

Volledigheidshalve laten we ook nog zien hoe de u_{ijk} 's op eenvoudige wijze zijn te bepalen uit de v_{ijk} 's die behoren bij een bepaalde modelspecificatie. Omdat u-tjes met een index gemiddeld nul zijn over de lopende index c.q. indices, is zo'n u-tje hetzij positief hetzij negatief, maar van dezelfde waarde. We kunnen daarom de lopende index weglaten en alleen de juiste tekens plaatsen. Zo hebben we bij de modelspecificatie [13][23] onder meer:

$$\log v_{111} = u + u_1 + u_2 + u_3 + u_{13} + u_{23}$$

en
$$\log v_{211} = u - u_1 + u_2 + u_3 - u_{13} + u_{23}$$

Aldus hebben we in dit voorbeeld zes onafhankelijke lineaire vergelijkingen in zes u-tjes met de logaritmen van de v_{ijk} 's als bekende termen. Voor het gekozen voorbeeld zijn de uitkomsten:

$$u = + 3,255 \quad ; \quad u_1 = + 0,135 \quad ; \quad u_2 = - 0,135$$

$$u_3 = + 0,347 \quad ; \quad u_{13} = + 0,558 \quad ; \quad u_{23} = + 0,558.$$

6. De logitanalyse van een 2 x 2 x 2 tabel

Aangezien een waargenomen dichotomie is te beschouwen als een steekproef uit een alternatieve verdeling, door aan een "Succes" de waarde 1 toe te kennen en aan een "Mislukking" de waarde 0, kan bij een 2 x 2 x 2 tabel een logitanalyse worden uitgevoerd met verklarende variabelen die slechts de waarden 1 en 0 kunnen aannemen. Omdat er sprake is van één te verklaren variabele, moeten voor de complete analyse van de tabel drie logitanalyses worden uitgevoerd. Er is een sterke verwantschap met de multi-pele regressie-analyse, zoals blijkt uit de modelomschrijving van de logitanalyse in het geval dat de variabele z wordt "verklaard" met behulp van de variabelen x en y .

$$\text{logit } z = \log [P(z)/(1 - P(z))] = \gamma_1 + \gamma_2 x + \gamma_3 y = f(x, y).$$

Het symbool $P(z)$ beschrijft de kans dat de te verklaren variabele z de waarde 1 aanneemt. De regressiecoëfficiënt γ_2 beschrijft de **direkte** logit-lineaire samenhang tussen z en x en γ_3 die tussen z en y . De uitwerking van de modelomschrijving leidt tot:

$$P(z)/[1 - P(z)] = \exp [f(x,y)]$$

$$\text{of} \quad P(z) = \exp [f(x,y)] / [1 + \exp [f(x,y)]].$$

Deze schrijfwijze van de kans dat z de waarde 1 aanneemt, laat zien dat een logitmodel altijd leidt tot kansuitspraken die netjes tussen 0 en 1 liggen. Indien een computer ter beschikking staat, kunnen maximale aannemelijkheids-schattingen van de regressiecoëfficiënten via een iteratieve procedure worden berekend. Indien men voor elk van de vier combinaties van de x en y -variabele over enkele tientallen waarnemingen van z beschikt, kunnen de regressiecoëfficiënten van het logitmodel ook worden geschat zonder dat er iteraties aan te pas komen. Deze zogenoemde minimum χ^2 -methode wordt onder meer beschreven door Theil [8].

In beide gevallen kunnen met de schattingen van de regressiecoëfficiënten tevens schattingen van hun asymptotische standaardfouten worden verkregen. In de praktijk zal men deze standaardfouten natuurlijk gebruiken om significantietoetsen uit te voeren op de gevonden regressiecoëfficiënten of, anders gezegd, om na te gaan of de direkte logit-lineaire samenhangen tussen de betreffende variabelen significant zijn. We gaan in dit artikel niet op dit aspect in.

Bij de logitanalyse van een $2 \times 2 \times 2$ tabel is een perfecte aansluiting tussen waargenomen en geschatte relatieve frekwenties altijd mogelijk als naast de konstante term drie verklarende variabelen worden opgenomen. Het ligt voor de hand om hiervoor in de eerste plaats de dichotome variabelen x en y op te nemen en als derde verklarende variabele een (eveneens) dichotome variabele die de interactie tussen z , x en y moet verklaren. We geven deze variabele het symbool r . Het logitmodel dat een perfecte aansluiting geeft met z als dichotome te verklaren variabele is dan te schrijven als:

$$\text{logit } z = \gamma_1 + \gamma_2 x + \gamma_3 y + \gamma_4 r$$

Alle dichotome variabelen in dit model nemen (slechts) de waarden 1 en 0 aan.

Door verwisseling van de variabelen worden de andere twee direkte logit-lineaire samenhangen gedefinieerd. De paarsgewijze marginale samenhangen kunnen worden bepaald door het uitvoeren van zes enkelvoudige logitanalyses op de drie marginale 2×2 tabellen.

Er blijkt (zie bijvoorbeeld Bishop e.a. [4]) een zeer nauwe band tussen log-lineaire analyse en logitanalyse te bestaan, indien de verklarende variabele r op een bepaalde manier wordt gedefinieerd. Om dit te laten zien keren we even terug naar de schrijfwijze van het log-lineaire model uit paragraaf 5 en de verkorte symbolische weergave a t/m h voor de frekwenties van de $2 \times 2 \times 2$ tabel. Voor het volledige log-lineaire model geldt onder meer:

$$\log v_{111} = \log a = u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23} + u_{123}$$

$$\log v_{112} = \log e = u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23} + u_{123}$$

Uit deze twee vergelijkingen en uit drie andere paren van zulke vergelijkingen volgt:

$$\log (a/e) = 2u_3 + 2u_{13} + 2u_{23} + 2u_{123}$$

$$\log (c/g) = 2u_3 + 2u_{13} - 2u_{23} - 2u_{123}$$

$$\log (b/f) = 2u_3 - 2u_{13} + 2u_{23} - 2u_{123}$$

$$\log (d/h) = 2u_3 - 2u_{13} - 2u_{23} + 2u_{123}$$

We introduceren nu tijdelijk x^* , y^* en r^* als verklarende variabelen die slechts de waarden $+1$ en -1 kunnen aannemen en schrijven het logitmodel in verband daarmee als:

$$\text{logit } z = \gamma_1^* + \gamma_2^* x^* + \gamma_3^* y^* + \gamma_4^* r^*.$$

Nu volgt onmiddellijk dat $2u_3$ gelijk is aan γ_1^* , dat $2u_{13}$ gelijk is aan γ_2^* en dat $2u_{23}$ gelijk is aan γ_3^* , terwijl $2u_{123}$ gelijk is aan γ_4^* , mits r^* de waarde $+1$ heeft als $x^* = +1$ én $y^* = +1$ maar ook als $x^* = -1$ én $y^* = -1$. Hieruit volgt dat r^* gelijk is aan het produkt van x^* en y^* . De verklarende variabele r^* staat daarmee inderdaad voor de interactie in de tabel.

De relaties tussen de verklarende variabelen met en zonder ster zijn tamelijk simpel:

$$x = \frac{1}{2}x^* + \frac{1}{2} \quad ; \quad y = \frac{1}{2}y^* + \frac{1}{2} \quad ; \quad r = \frac{1}{2}r^* + \frac{1}{2} = 2xy - x - y + 1.$$

Hetzelfde geldt voor de relaties tussen de γ 's zonder ster en de u-tjes uit het log-lineaire model:

$$\gamma_1 = 2(u_3 - u_{13} - u_{23} - u_{123}) ; \gamma_2 = 4u_{13} ; \gamma_3 = 4u_{23} ; \gamma_4 = 4u_{123}.$$

Gemakshalve schrijven we het logitmodel ook nog uit in de vier vergelijkingen

$$\log(a/e) = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4$$

$$\log(c/g) = \gamma_1 + \gamma_2$$

$$\log(b/f) = \gamma_1 + \gamma_3$$

$$\log(d/h) = \gamma_1 + \gamma_4$$

De numerieke uitkomsten van de drievoudige logitanalyse van het voorbeeld in paragraaf 2 kunnen als volgt worden gepresenteerd, met achtereenvolgens z, y en x als te verklaren variabele.

$$\text{logit } z = -1,540 + 2,234 x + 2,234 y + 0,000 r$$

$$\text{logit } y = -1,386 + 0,000 x + 2,234 z + 0,000 r$$

$$\text{logit } x = -0,847 + 0,000 y + 2,234 z + 0,000 r$$

Indien in een $2 \times 2 \times 2$ tabel géén interactie bestaat, betekent dit dat van de vier vergelijkingen van het logitmodel er één afhankelijk wordt. In dat geval is er ook altijd een perfecte aansluiting mogelijk tussen waargenomen relatieve frekwenties en geschatte kansen als er drie parameters worden geschat. In het gekozen voorbeeld doet zich deze situatie voor. Merk voorts op dat het ontbreken van de direkte samenhang tussen x en y keurig wordt teruggevonden.

Het is heel goed voorstelbaar dat men bij het introduceren van een dichotome variabele voor de interactie zou denken aan de variabele $r = xy$. Deze variabele zou dan de waarde 1 aannemen als $x = y = 1$ en de waarde 0 in de drie overige gevallen. Nog afgezien van het feit dat men op deze manier de direkte band tussen log-lineaire analyse en logitanalyse doorsnijdt, geeft deze definitie van r ook aanleiding tot problemen rond sommige plaatjes uit diagram 2. We hebben dan immers de volgende vier vergelijkingen in vier onbekenden.

$$\log(a/e) = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4$$

$$\log(c/g) = \gamma_1 + \gamma_2$$

$$\log(b/f) = \gamma_1 + \gamma_3$$

$$\log(d/h) = \gamma_1$$

Uit de vier vergelijkingen op de vorige pagina volgt bijvoorbeeld dat=

$$\gamma_2 = \log c + \log h - \log d - \log g.$$

Indien er nu in de "complementaire deelgroep", dat is in Y^C , onafhankelijkheid zou bestaan tussen de variabelen x en $z^{*})$ impliceert dit dat $ch = dg$. De konsekwentie is onmiddellijk duidelijk, γ_2 wordt gelijk aan 0. Hoewel onafhankelijkheid in een complementaire deelgroep, gekombineerd met samenhang in de primaire deelgroep nogal bijzonder is, is het feit dat twee van de regressiecoëfficiënten uitsluitend bepaald worden door de frekwenties in de complementaire tabel beslist een negatief punt voor deze alternatieve definitie van de interactie-variabele r .

Logitanalyse is oorspronkelijk ontworpen om bij een dichotome te verklaren variabele en continue of diskrete verklarende variabelen een regressieberekening te kunnen uitvoeren die niet kan leiden tot geschatte kansen groter dan 1 of kleiner dan 0. Met de invoering van dummy-variabelen kwamen ook de dichotome verklarende variabelen aan de orde. Tegen deze achtergrond is logitanalyse met een interactieterm echter geen vaste regel doch eerder een uitzondering. Daarom is het misschien goed om de aandacht te vestigen op het feit dat als er wel interactie bestaat maar men in de logitanalyse geen interactieterm meeneemt als verklarende variabele, er vreemde dingen kunnen gebeuren bij de weergave van de samenhangsstructuur, met name als er twee of drie van de marginale samenhangen [bogen] in het diagram ontbreken. Zo blijkt bijvoorbeeld dat in dat geval voor plaatje 12 uit diagram 2 een samenhangsstructuur conform het niet in diagram 2 opgenomen plaatje van structuur (9) wordt gevonden.

Hierbij kan nog een tweetal opmerkingen worden gemaakt. In de eerste plaats mag het geen wonder heten dat er juist bij het logitmodel zonder interactieterm problemen kunnen ontstaan. Het logitmodel is [in de logits] een strikt lineair model. In de gewone regressie-analyse is algemeen bekend dat afhankelijkheid gepaard kan gaan met covariantie nul en omgekeerd dat covariantie nul geen onafhankelijkheid impliceert. Voor een logitmodel geldt een analoge redenering.

*) -----
Onafhankelijkheid binnen één deelgroep is alleen mogelijk indien er in de $2 \times 2 \times 2$ tabel (tweede-orde) interactie bestaat.

In de tweede plaats is het van belang om te weten dat de genoemde problemen vrijwel onverkort blijven spelen als er sprake is van een complete samenhangsstructuur die in de buurt ligt van de bedoelde probleemsituaties. Dan worden er bij een logitanalyse zonder interactieterm bijvoorbeeld insignifikante logit-lineaire samenhangen gevonden, terwijl een logitanalyse met interactieterm wel significante samenhangen te zien geeft.

7. Konklusies.

In paragraaf 4 van dit artikel is getoond dat er bij een $2 \times 2 \times 2$ tabel twaalf specifieke samenhangsstructuren zijn te onderscheiden indien tabellen met en zonder interactie apart worden geteld. Deze twaalf structuren zijn weergegeven in diagram 2.

Zeta-analyse, log-lineaire analyse en logitanalyse [met een korrekt gedefinieerde interactie-variabele] van een $2 \times 2 \times 2$ tabel geven de structuur van zo'n tabel altijd korrekt weer. Daarbij geeft de zeta-analyse een maat mee aan de verschillende samenhangen, die in ieder geval onderlinge vergelijking van de intensiteit van de samenhangen mogelijk maakt. Aldus kan ook bij een complete samenhangsstructuur een beeld worden gevormd van de naast aanliggende incompleete structuur.

De log-lineaire analyse produceert tabellen met geschatte frekwenties zoals die behoren bij de verschillende modelspecificaties. Deze geschatte frekwenties laten zich op overzichtelijke wijze vergelijken met de waargenomen frekwenties. Desgewenst kunnen bij een complete samenhangsstructuur de verschillen op significantie worden getoetst. Daarmee is ook de log-lineaire analyse een handige, relatief weinig gekompliceerde, methode om een grondige analyse van een $2 \times 2 \times 2$ tabel uit te voeren.

Indien een logitanalyse van een $2 \times 2 \times 2$ tabel met een korrekte interactieterm wordt uitgevoerd, is de methode in wezen identiek met de log-lineaire analyse voor het volledige model. Laat men echter de interactieterm achterwege, dan ontstaan er bij de logitanalyse in sommige gevallen vreemde problemen rond de korrekte weergave van de samenhangsstructuur. Bij een complete samenhangsstructuur die "in de buurt" ligt van één van de bedoelde bijzondere structuren, stuit men in de praktijk ook op de genoemde problemen. Logitanalyse zonder interactieterm is dus niet de meest geëigende techniek om de structuur van de samenhangen in een $2 \times 2 \times 2$ tabel mee op te sporen.

Literatuur

- [1] Abrahamse, A.P.J. (1982), Second-order interaction and induced association in a three-way contingency table, report 8220/S, Econometrisch Instituut, Erasmus Universiteit Rotterdam
- [2] Abrahamse, A.P.J. (1986), On the sampling behaviour of the covariability coefficient ζ , report 8603/A, Econometrisch Instituut, Erasmus Universiteit Rotterdam
- [3] Bartlett, M.S. (1935), "Contingency table interactions", Suppl. J.R. Statistical Society, 2, blz 248-252
- [4] Bishop, Y.M.M., Fienberg, S.E. en Holland, P.W. (1975), Discrete Multivariate Analysis, M.I.T. Press, Cambridge, Massachusetts
- [5] Fienberg, S.E. (1980), The analysis of cross-classified categorical data, 2nd ed., M.I.T. Press, Cambridge, Massachusetts
- [6] Lammerts van Bueren, W.M. (1982), Measuring association in nominal data, proefschrift, Erasmus Universiteit Rotterdam
- [7] Pannekoek, J. (1983), Informatiematen bij de analyse van categorische gegevens, Kwantitatieve Methoden, 10, blz 77-103
- [8] Theil, H. (1972), Statistical decomposition analysis, North Holland, Amsterdam

Ontvangen: 27-03-86

Geaccepteerd: 11-11-86