



CENTRAAL BUREAU VOOR DE STATISTIEK

Hoofdafdeling Statistische Methoden

Postbus 959, 2270 AZ VOORBURG

INFORMATIEMATEN BIJ DE ANALYSE VAN
CATEGORISCHE GEGEVENS

J. Pannekoek ^{*)}

Samenvatting

In dit rapport wordt een associatiemaat behandeld die gebruikt kan worden bij de analyse van categorische gegevens wanneer er onderscheid gemaakt wordt tussen afhankelijke en onafhankelijke variabelen. Deze maat, de PRU (Proportional Reduction Uncertainty), is te interpreteren als de proportie verklaarde variatie in de afhankelijke variabelen. Een vereenvoudiging in de beschrijving van de data kan verkregen worden door categorieën van onafhankelijke variabelen samen te voegen of door restricties op te leggen aan de associatiestructuur zoals die geformuleerd kunnen worden door middel van log-lineaire modellen. De PRU is te gebruiken om het verlies aan verklaarde variatie te meten dat door deze vereenvoudigingen ontstaat. Een en ander is toegepast op een onderzoek naar de invloed van een aantal achtergrondvariabelen op de schoolkeuze. Een model zonder restricties voor de bij dit onderzoek gehanteerde vierdimensionale kruistabel zou 559 parameters bevatten. Het bleek mogelijk een model met 88 parameters te specificeren zonder veel aan verklaarde variatie te verliezen.

^{*)} De in dit rapport weergegeven opvattingen zijn die van de auteur en komen niet noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek.

1. Inleiding

De aanwezigheid van associatie tussen nominale variabelen kan getoetst worden met Pearson's χ^2 -toets of met de likelihood-ratio toetsingsgrootheid G^2 . Naast het toetsen op significantie is het vaak wenselijk een maat te hebben voor de sterkte van een gevonden verband. Dit is met name het geval wanneer er veel waarnemingen zijn. Statistische significantie is dan geen garantie voor 'relevantie'. In dit rapport wordt een associatiemaat behandeld die nauw samenhangt met de G^2 -grootheid. Deze maat is asymmetrisch, dat wil zeggen geeft de associatie tussen variabelen weer, waarbij onderscheid gemaakt wordt tussen een afhankelijke en een onafhankelijke variabele. De associatiemaat is gebaseerd op het aan de informatietheorie ontleende begrip entropie. De entropie kan opgevat worden als een maat voor de variatie in een nominale variabele en de associatiemaat is te interpreteren als de proportie verklaarde variatie in die nominale variabele. De associatiemaat vertoont een zekere analogie met de correlatiecoëfficiënt. Als er meer dan één onafhankelijke variabele is, zijn er ook maten die analoog zijn aan de partiële en multiële correlatiecoëfficiënt. Een zeer uitgebreide en recente studie over deze associatiematen is het proefschrift van Lammerts van Bueren (1982).

In dit rapport wordt getoond hoe deze maat gebruikt kan worden om de associatie in een meerdimensionale tabel op zo eenvoudig mogelijke wijze te beschrijven. Allereerst beperken we ons tot die onafhankelijke variabelen die de meeste variatie in de afhankelijke variabele verklaren. De meerdimensionale kruistabel die gevormd wordt door deze onafhankelijke variabelen en de afhankelijke variabele kan vervolgens vaak verder gereduceerd worden door categorieën van de onafhankelijke variabelen samen te voegen. In het algemeen zal dit een verlies aan verklaarde variatie ten gevolge hebben. Maar als dit verlies slechts gering is, kan dit opwegen tegen de vereenvoudiging in de beschrijving van de data. Een ander type vereenvoudiging krijgen we door modellen voor de associatiestructuur te formuleren. In meerdimensionale tabellen is associatie mogelijk tussen combinaties van twee of meer variabelen. Door sommige van deze associaties gelijk aan nul te veronderstellen, zoals bij log-lineaire modellen, wordt de associatiestructuur vereenvoudigd. Ook deze vereenvoudiging zal in het algemeen een verlies aan

verklaarde variatie ten gevolge hebben. Dit verlies kan met de hier beschreven associatiemaat gemeten worden. De associatiemaat is dan op te vatten als een maat voor de fit van een log-lineair model.

De theorie wordt toegepast op gegevens uit het onderzoek Sociaal Milieu Voortgezet Onderwijs (SMVO). Dit onderzoek bevat gegevens van leerlingen die in de eerste klas van het voortgezet onderwijs zitten. De bedoeling is de keuze van voortgezet onderwijs te verklaren uit achtergrondvariabelen die te maken hebben met het sociaal milieu van de leerling en de variabelen prestatieniveau en advies van de onderwijzer.

2. Variatie en verklaarde variatie in nominale variabelen

2.1. Theorie

De entropie is een aan de informatietheorie ontleend begrip dat gebruikt kan worden om de variatie in een nominale discrete variabele te meten. De entropie van een discrete variabele X , $H(X)$, is gedefinieerd als

$$H(X) = - \sum_i p_i \ln p_i, \quad (1)$$

met $p_i \ln p_i = 0$ voor $p_i = 0$. Hierin is p_i de kans op een waarneming in categorie i van X . De entropie bereikt zijn maximum, $\ln I$, als alle categorieën de zelfde kans ($1/I$) hebben. De onzekerheid over de uitkomst van een waarneming is dan maximaal. $H(X)$ bereikt zijn minimumwaarde 0, als een van de categorieën kans 1 heeft en de andere categorieën dus kans 0 hebben. De onzekerheid over de uitkomst van een waarneming is dan minimaal; het is met zekerheid te zeggen in welke categorie de waarnemingen vallen. De entropie is een maat voor de onzekerheid over de uitkomst van een waarneming, of, anders gezegd, een maat voor de variatie in de kansverdeling van X .

De definitie (1) kan gegeneraliseerd worden naar multivariate verdelingen door de index i te laten lopen over alle cellen van een kruistabel die door meerdere variabelen gevormd wordt. Deze entropie van een multivariate verdeling kan ook genoteerd worden door meerdere indices te gebruiken. Zo geldt voor een bivariate verdeling

$$H(XY) = - \sum_i \sum_j p_{ij} \ln p_{ij}, \quad (2)$$

met p_{ij} de kans op een waarneming in categorie i van X en j van Y . Het minimum 0 wordt bereikt als een van de $p_{ij}=1$, het maximum is $\ln IJ$ en wordt bereikt als alle p_{ij} gelijk aan $1/IJ$ zijn (zie voor het bewijs, Theil (1972), blz. 8).

Ook de entropie van een conditionele verdeling is te berekenen. De entropie van Y gegeven een categorie i van X is

$$H_{X_i}(Y) = - \sum_j \frac{p_{ij}}{p_{i.}} \ln \frac{p_{ij}}{p_{i.}}, \quad (3)$$

met $p_{i.} = \sum_j p_{ij}$. Door het gewogen gemiddelde over i te nemen ontstaat

$$H_X(Y) = - \sum_i p_{i.} \sum_j \frac{p_{ij}}{p_{i.}} \ln \frac{p_{ij}}{p_{i.}} = - \sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{p_{i.}}. \quad (4)$$

$H_X(Y)$ is de gemiddelde conditionele entropie van Y gegeven X . Deze conditionele entropie kan uitgedrukt worden in de $H(XY)$ en $H(X)$:

$$\begin{aligned} H_X(Y) &= - \sum_i \sum_j p_{ij} \ln p_{ij} + \sum_i \sum_j p_{ij} \ln p_{i.} = \\ &= - \sum_i \sum_j p_{ij} \ln p_{ij} + \sum_i p_{i.} \ln p_{i.} = H(XY) - H(X). \end{aligned} \quad (5)$$

Vergelijkingen (4) en (5) zijn eenvoudig te generaliseren naar meerdere X en Y variabelen. Wanneer er m Y variabelen zijn, dan kunnen deze m variabelen worden opgevat als één variabele met als aantal categorieën het product van de aantallen categorieën van de m variabelen. Evenzo kunnen n X variabelen ook als één 'product variabele' worden opgevat. De conditionele entropie van de variabelen $Y_1, \dots, Y_m$ gegeven de variabelen $X_1, \dots, X_n$ is dus

$$H_{X_1, \dots, X_n}(Y_1, \dots, Y_m) = H(X_1, \dots, X_n, Y_1, \dots, Y_m) - H(X_1, \dots, X_n). \quad (6)$$

De entropie van de bivariate verdeling is ook uit te drukken in de beide univariate entropiën plus een restterm:

$$\begin{aligned}
 H(XY) &= \sum_i \sum_j p_{ij} \ln \frac{1}{p_{ij}} = \\
 &= \sum_i \sum_j p_{ij} \ln \frac{1}{p_{i.}} + \sum_i \sum_j p_{ij} \ln \frac{1}{p_{.j}} - \sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{p_{i.} p_{.j}} = \\
 &= H(X) + H(Y) - \sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{p_{i.} p_{.j}} .
 \end{aligned} \tag{7}$$

De restterm (zonder het teken) wordt in de informatietheorie de informatie-overdracht ('transmitted information') tussen X en Y genoemd. De informatie-overdracht wordt aangegeven door het symbool $I(XY)$. De schatter voor $I(XY)$, $\hat{I}(XY)$, is gerelateerd aan de G^2 -groottheid:

$$\hat{I}(XY) = \sum_i \sum_j \hat{p}_{ij} \ln \frac{\hat{p}_{ij}}{\hat{p}_{i.} \hat{p}_{.j}} = G^2/2n \tag{8}$$

met n het aantal waarnemingen. De informatie-overdracht en G^2 zijn op meer algemene wijze aan elkaar gerelateerd. In paragraaf 4 wordt hier nader op ingegaan. Als X en Y onafhankelijk zijn is $I(XY)=0$. $I(XY)$ is op te vatten als een maat voor de afhankelijkheid van X en Y. Uit (1), (4) en (7) volgt dat voor $I(XY)$ geldt

$$I(XY) = H(Y) - H_X(Y) . \tag{9}$$

Omdat $I(XY) \geq 0$ betekent dit dat

$$H_X(Y) \leq H(Y) . \tag{10}$$

Dus door te conditioneren op X en daarna te middelen over X neemt de entropie van Y af, behalve als X en Y onafhankelijk zijn omdat dan $H_X(Y)=H(Y)$.

De informatie-overdracht tussen X en Y kan ook conditioneel op een derde variabele, Z, berekend worden. Hierbij maken we gebruik van de regel dat in de informatietheorie iedere vergelijking van kracht blijft als aan alle termen hetzelfde subscript wordt toegevoegd. Deze regel staat bekend als de regel van McGill en Garner (zie Ashby, 1965). Met gebruikmaking van deze regel volgt uit (9)

$$I_Z(XY) = H_Z(Y) - H_{XZ}(Y) . \quad (11)$$

Omdat $I_Z(XY)$ evenals $I(XY)$ altijd groter of gelijk is aan nul geldt

$$H_{XZ}(Y) \leq H_Z(Y) . \quad (12)$$

Dus het toevoegen van een tweede conditioneringsvariabele zal in het algemeen de onzekerheid in Y verder reduceren. Het toevoegen heeft geen reductie in onzekerheid tot gevolg als $I_Z(XY)=0$, dus als X en Y onafhankelijk zijn gegeven Z.

Met behulp van de entropie en de conditionele entropie is een asymmetrische associatiemaat te definiëren. Hierbij wordt gebruik gemaakt van het PRE (Proportional Reduction in Error) principe. Een PRE grootheid heeft de volgende vorm:

$$PRE = \frac{\text{verklaarde variatie in de afhankelijke variabele}}{\text{totale variatie in de afhankelijke variabele}} .$$

Als de variatie in Y gemeten wordt door $H(Y)$ is $H(Y)-H_X(Y)$ op te vatten als de door X verklaarde variatie in Y. De proportie verklaarde variatie in Y wordt dan

$$PRE_{Y|X} = \frac{H(Y) - H_X(Y)}{H(Y)} = \frac{I(XY)}{H(Y)} . \quad (13)$$

Omdat de entropie een maat is voor de onzekerheid over de uitkomst van een waarneming wordt deze PRE grootheid ook wel PRU (proportional reduction in uncertainty) genoemd (Teachman, 1980). Het minimum van deze maat is nul en wordt bereikt als X en Y onafhankelijk zijn omdat dan geldt $H_X(Y)=H(Y)$. Het maximum wordt bereikt bij maximale associatie van X en Y. De waarde van dit

maximum hangt echter af van het aantal categorieën van X en van Y . In figuur 1a is de situatie weergegeven die zich voordoet bij maximale associatie (bij gegeven marginalen) in een vierkante tabel. Iedere rij en iedere kolom bevat slechts een kans (aangegeven door a t/m c) ongelijk aan nul. In dit geval geldt $H(XY)=H(X)=H(Y)$. Hieruit volgt (met behulp van (5)) dat $H_X(Y)$ gelijk aan nul is. De $PRU_{Y|X}$ is dus 1. Als Y meer (niet lege) categorieën heeft dan X is het niet meer mogelijk dat er slechts een kans ongelijk aan nul voorkomt in iedere rij. Deze situatie is geschetst in figuur 1b. Hier geldt $H(XY) \neq H(X)$, en dus $H_X(Y) \neq 0$. Omdat de conditionele entropie nooit negatief is geldt $H_X(Y) > 0$. Dit betekent dat $PRU_{X|Y}$ kleiner dan 1 is. De situatie dat de afhankelijke variabele minder categorieën heeft dan de onafhankelijke variabele is ook te illustreren met figuur 1b door X als afhankelijke en Y als onafhankelijke variabele op te vatten. Voor figuur 1b geldt $H(XY)=H(Y)$, dus $H_Y(X)=0$. De $PRU_{Y|X}$ is dus 1. De conclusie is dat een PRU de maximale waarde 1 slechts kan bereiken als de afhankelijke variabele evenveel of minder categorieën heeft dan de onafhankelijke variabele. Heeft de afhankelijke variabele meer categorieën dan is de PRU kleiner dan 1.

Figuur 1a. Maximale associatie
in een vierkante tabel

X	Y		
	1	2	3
1		b	
2	a		
3			c

Figuur 1b. Maximale associatie
in een rechthoekige tabel

X	Y			
	1	2	3	4
1		e		
2	d			g
3			f	

De PRU is te generaliseren naar meer afhankelijke en meer onafhankelijke variabelen. De algemene vorm is

$$PRU_{Y_1, \dots, Y_m | X_1, \dots, X_n} = \frac{H(Y_1, \dots, Y_m) - H_{X_1, \dots, X_n}(Y_1, \dots, Y_m)}{H(Y_1, \dots, Y_m)} \quad (14)$$

Analoog aan de partiële correlatie is ook een partiële PRU maat te definiëren. Deze maat geeft aan welke proportie variatie in Y nog door X verklaard wordt gegeven Z. Dus

$$PRU_{Y.Z|X} = \frac{H_Z(Y) - H_{XZ}(Y)}{H(Y)} \quad (15)$$

Voor deze partiële PRU geldt

$$PRU_{Y.Z|X} = \frac{H(Y) - H_{XZ}(Y)}{H(Y)} - \frac{H(Y) - H_Z(Y)}{H(Y)} = PRU_{Y|XZ} - PRU_{Y|Z} \quad (16)$$

De PRU maten zijn asymmetrisch. Interpretatie is alleen zinvol als er onderscheid gemaakt wordt tussen afhankelijke en onafhankelijke variabelen. Abrahamse en Lammerts van Bueren (1979) stellen voor om symmetrische associatie te meten met

$$PRU_{XY} = \frac{2I(XY)}{H(X) + H(Y)} \quad (17)$$

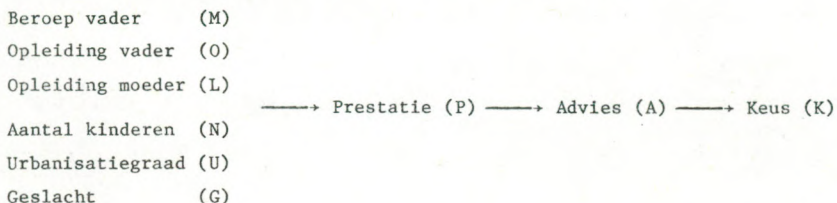
De teller van PRU_{XY} is de som van de tellers van $PRU_{Y|X}$ en $PRU_{X|Y}$ en de noemer is de som van de noemers van $PRU_{Y|X}$ en $PRU_{X|Y}$. Als X en Y onafhankelijk zijn worden $H_Y(X)$ en $H_X(Y)$ gelijk aan respectievelijk $H(X)$ en $H(Y)$, waardoor PRU_{YX} nul wordt. Het maximum 1 wordt bereikt als X en Y elkaar volledig bepalen, dus als $H_Y(Y)=H_X(X)=0$. Dit maximum wordt dus alleen bereikt in een vierkante tabel.

2.2. Toepassing

Door het CBS wordt onderzoek gedaan naar de invloed die een aantal achtergrondvariabelen uitoefenen op de schoolloopbaan van leerlingen die de lagere school net verlaten hebben. Voor dit onderzoek zijn een aantal variabelen gemeten bij een steekproef van ongeveer 37 000 leerlingen die in 1977 in de eerste klas van het voortgezet onderwijs zaten (zie CBS, 1979). De theorie die in dit rapport behandeld wordt zal gebruikt worden om de samenhang tussen een aantal achtergrondvariabelen en de variabele schoolkeus te analyseren. In de analyse worden de volgende variabelen opgenomen: school-

keus (K), het advies van de onderwijzer met betrekking tot de te kiezen vorm van voortgezet onderwijs (A), het prestatieniveau van de leerling gemeten door een Cito-toets (P), het beroep van de vader (M), het hoogst bereikte opleidingsniveau van de vader (O), het hoogst bereikte opleidingsniveau van de moeder (L), het aantal kinderen in het gezin (N), de urbanisatiegraad van de plaats waar de school zich bevindt (U) en het geslacht van de leerling (G). De categorie indeling van deze variabelen en de bijbehorende aantallen staan in bijlage A.

Bij de analyse wordt een bepaalde causale ordening van deze variabelen verondersteld. Deze is weergegeven in het volgende diagram:



Aangenomen wordt dat de achtergrondvariabelen (M, O, L, N, U en G) invloed kunnen uitoefenen op het prestatieniveau, dat de achtergrondvariabelen en het prestatieniveau invloed kunnen uitoefenen op het advies en dat de schoolkeuze beïnvloed kan worden door alle overige variabelen. Geen van deze relaties geldt in omgekeerde richting. Deze veronderstelling sluit niet uit dat de achtergrondvariabelen invloed uitoefenen op elkaar, noch dat er interacties optreden. Zo kan de opleiding van de vader verband houden met de opleiding van de moeder en kan de invloed van de opleiding van de vader op de schoolkeuze anders zijn voor jongens en meisjes. Alle mogelijke relaties tussen de variabelen kunnen slechts bepaald worden in de kruistabel die gevormd wordt door alle variabelen. De omvang van deze tabel maakt analyse echter praktisch onmogelijk. Daarom zal de invloed van ieder van de achtergrondvariabelen op Prestatie, Advies en Keus apart bepaald worden. Deze beperking houdt in dat relaties tussen de achtergrondvariabelen en interacties die meer dan een achtergrondvariabele bevatten niet onderzocht worden.

Het effect van een achtergrondvariabele X op het prestatieniveau kan bepaald worden met behulp van $PRU_{P|X}$. Het effect van X op het advies wordt conditioneel op het prestatieniveau bepaald, dus met behulp van $PRU_{A.P|X}$. Het effect van X op de schoolkeus wordt conditioneel op het prestatieniveau en het advies bepaald, dus met behulp van $PRU_{K.AP|X}$. De schattingen voor deze PRU's en partiële PRU's staan in tabel 1.

Tabel 1. Effecten van de achtergrondvariabelen op P, A en K

X		$PRU_{P X}$	$PRU_{A.P X}$	$PRU_{K.AP X}$
Opleiding vader	(O)	0,0346	0,0245	0,0144
Beroep vader	(M)	0,0248	0,0179	0,0104
Opleiding moeder	(L)	0,0230	0,0149	0,0103
Urbanisatiegraad	(U)	0,0058	0,0079	0,0082
Aantal kinderen	(N)	0,0018	0,0030	0,0036
Geslacht	(G)	0,0006	0,0008	0,0026

Het is duidelijk dat de belangrijkste effecten op Prestatie, Advies en Keus uitgaan van de variabelen O, M en L. Deze variabelen kunnen alle drie geïnterpreteerd worden als indicatoren voor het sociaal milieu waaruit de leerling afkomstig is. Hoewel in het meeste onderzoek op dit gebied het beroep van de vader als indicator voor het sociaal milieu genomen wordt blijkt dat het effect van de opleiding van de vader groter is. De drie milieuvariabelen hebben het grootste effect op het prestatieniveau, een minder grote invloed op het advies en de minste invloed op de schoolkeus. Voor de overige variabelen geldt juist het omgekeerde; het effect op de schoolkeus is het grootste en het effect op het prestatieniveau het kleinste. Opvallend is verder de zeer geringe invloed van de variabele geslacht, vooral op Prestatie en op Advies.

Met behulp van de schattingen $PRU_{A|P}=0,2517$ en $PRU_{K|PA}=0,4849$, is voor iedere achtergrondvariabele de totale verklaarde variatie in Advies en Keus te berekenen volgens respectievelijk

$$PRU_{A|PX} = PRU_{A|P} + PRU_{A.P|X}$$

en

$$\text{PRU}_{K|PAX} = \text{PRU}_{K|PA} + \text{PRU}_{K.PA|X}.$$

Zo is bijvoorbeeld de maximale verklaarde variatie in Advies $\text{PRU}_{A|P0}=0,2762$ en de maximale verklaarde variatie in Keus $\text{PRU}_{K|PA0}=0,4993$.

3. Het samenvoegen van categorieën

3.1. Theorie

Naast het vaststellen van de effecten van de achtergrondvariabelen kan met de PRU-maat ook het effect van het samenvoegen van categorieën bestudeerd worden. Het samennemen van categorieën zal in het algemeen namelijk een verlies aan verklaarde variatie, dus een reductie in de PRU ten gevolge hebben. In deze paragraaf zal deze reductie in informatie berekend worden. Uitgangspunt is een onafhankelijke variabele X en een afhankelijke variabele Y. De categorieën van X worden verdeeld in elkaar niet overlappende verzamelingen van categorieën v ($v=1, \dots, V$). Dit maakt het mogelijk om het effect van het samenvoegen van alle categorieën binnen een of meer verzamelingen v te tonen.

De Y-marginalen van de subtabel, gevormd door de categorieën i $\in$ v van X en j van Y zijn:

$$p_i^v = \sum_{j \in v} p_{ij}. \quad (18)$$

De kans op een waarnemingen in deze subtabel is $p_{\cdot}^v$. $I(xy)$ is nu te schrijven als

$$I(XY) = \sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{p_{i \cdot} p_{\cdot j}} = \sum_i \sum_j p_{ij} \ln \frac{p_{ij} p_j^v p_{\cdot}^v}{p_{i \cdot} p_{\cdot j} p_j^v p_{\cdot}^v} =$$

$$\begin{aligned}
&= \sum_i \sum_j p_{ij}^v \ln \frac{p_{ij}^v / p_{i.}^v}{\frac{p_{j.}^v}{p_{..}^v} \cdot \frac{p_{i.}^v}{p_{..}^v}} + \sum_i \sum_j p_{ij}^v \ln \frac{p_{j.}^v}{p_{..}^v} = \\
&= \sum_{v=1}^V p_{..}^v \sum_{i \in v} \sum_j \frac{p_{ij}^v}{p_{..}^v} \ln \frac{p_{ij}^v / p_{i.}^v}{\frac{p_{j.}^v}{p_{..}^v} \cdot \frac{p_{i.}^v}{p_{..}^v}} + \sum_{v=1}^V \sum_j p_{j.}^v \ln \frac{p_{j.}^v}{p_{..}^v} = \\
&= \sum_{v=1}^V p_{..}^v I^v(XY) + I(X^*Y), \tag{19}
\end{aligned}$$

Waarin X^* een variabele is met categorieën $v=1, \dots, V$, die wordt gevormd door X te sommeren over iedere i binnen v . $I(XY)$ is de informatie-overdracht in de oorspronkelijke tabel en $I(X^*Y)$ de informatie-overdracht in de (grovere) tabel, die ontstaat door de rijen binnen iedere verzameling v samen te voegen. $I^v(XY)$ is de informatie-overdracht in een subtabel van de oorspronkelijke tabel gevormd door de categorieën $i \in v$ van X en j van Y . Het is eenvoudig in te zien dat als een verzameling v van categorieën van X slechts één categorie bevat, de corresponderende term van $I^v(XY)$ gelijk is aan nul. Als nu de categorieën van één van de verzamelingen v worden samengenomen, en we berekenen op de nieuwe tabel een nieuwe $I(XY)$, zeg $I'(XY)$, dan verschilt deze $I'(XY)$ alleen van $I(XY)$ in de term $p_{..}^v I^v(XY)$, voor die v waarvan de categorieën zijn samengenomen. Voor $I'(XY)$ wordt deze term nul omdat de betreffende subtabel tot een rij is gereduceerd.

Dus als alle categorieën binnen een verzameling v samengenomen worden, vermindert $I(XY)$ met $p_{..}^v I^v(XY)$. Als X en Y onafhankelijk zijn binnen de subtabel verandert $I(XY)$ niet. Omdat $\text{PRU}_{Y|X} = I(XY)/H(Y)$ neemt de $\text{PRU}_{Y|X}$ door het samenvoegen van de categorieën uit v af met $p_{..}^v I^v(XY)/H(Y)$. Als er categorieën van de afhankelijke variabele y worden samengenomen is het niet te zeggen hoe de $\text{PRU}_{Y|X}$ verandert omdat dan zowel $I(XY)$ als $H(Y)$ afnemen.

Omdat $\hat{I}(XY)$ gelijk is aan $G^2/2N$ geldt volgens (19)

$$G^2 = \sum_{v=1}^V G_v^2 + G_{*}^2, \quad (20)$$

met G_v^2 de G^2 -waarde binnen een subtabel en G_{*}^2 de G^2 -waarde van de $X^{*} \times Y$ -tabel. Met G_v^2 kan de onafhankelijkheid binnen een subtabel worden getoetst. Volgens (20) is deze G_v^2 -waarde te berekenen door het verschil te nemen van de G^2 -waarden voor de oorspronkelijke tabel en de, door het samenvoegen van de categoriën binnen v , gereduceerde tabel.

De formules uit deze paragraaf kunnen net als bij (6) en (14) eenvoudig worden gegeneraliseerd naar meer dan één x -variabele en meer dan één y -variabele.

3.2. Toepassing

Voor twee variabelen, de variabelen opleidingsniveau van de vader en prestatieniveau van de leerling, zijn alternatieve categorie indelingen onderzocht. Opleidingsniveau van de vader (0) heeft de volgende categorieën: lagere school (1), lager beroepsonderwijs (2), voortgezet lager onderwijs (3), middelbaar beroepsonderwijs (4), middelbaar algemeen onderwijs (5), hoger beroepsonderwijs (6) en academisch onderwijs (7). Deze indeling geeft niet alleen het hoogst bereikte onderwijsniveau aan, maar maakt ook een onderscheid in beroeps- en algemeen onderwijs. Een nieuwe categorie indeling van deze variabele die uitsluitend het niveau weerspiegelt en geen onderscheid meer maakt tussen beroeps- en algemeen onderwijs kan verkregen worden door de categorieën (2) en (3) en de categorieën (4) en (5) samen te voegen. In tabel 2 staat aangegeven hoe groot de reductie in verklaarde variatie van de variabele schoolkeus is die dit samenvoegen ten gevolge heeft.

Tabel 2. Reductie in $PRU_{K|OPA}$ door samenvoegen van categorieën van opleiding vader (0).

Categorie indeling 0	$PRU_{K OPA}$	Reductie	G^2	df*
1, 2, 3, 4, 5, 6, 7	0,4993	0	33994	372
1, 2+3, 4, 5, 6, 7	0,4980	0,0013	33904	321
1, 2, 3, 4+5, 6, 7	0,4986	0,0007	33945	324
1, 2+3, 4+5, 6, 7	0,4973	0,0021	33855	267

* Omdat er nulcellen voorkwamen in de marginale tabel OPA is er een correctie op het aantal vrijheidsgraden toegepast (zie bijlage B).

De reductie in de PRU door naar 5 categorieën te gaan is slechts gering. De ingedikte variabele Opleiding vader blijft de belangrijkste milieuvariabele (vergelijk tabel 1). De reductie's in regel 2 en 3 van tabel 2 tellen op tot de reductie in de onderste regel (zie formule (19)). De afname in de G^2 -waarde door geen associatie te veronderstellen in de subtabellen Keus×Opleiding vader categorie (2) en (3) en Keus × Opleiding vader categorie (4) en (5) is 139 bij 105 vrijheidsgraden (zie formule (20)). Dit is net significant op het 1% niveau. Omdat we hier echter meer waarde hechten aan de PRU-maat dan aan statistische significantie geven we toch de voorkeur aan de gereduceerde tabel om de associatie te beschrijven.

Een andere reductie in de omvang van de tabel is mogelijk door van de variabele Prestatieniveau de laagste twee en de hoogste twee categorieën samen te nemen. Er ontstaat dan een variabele Prestatieniveau met drie categorieën die ieder ongeveer evenveel waarnemingen bevatten. In tabel 3 staan de gevolgen voor de PRU van deze hercategorisering.

De afname van de PRU door het prestatieniveau in drie categorieën in te delen in plaats van vijf is ruim drie keer zo groot als de afname ten gevolge van het samenvoegen van categorieën van de variabele Opleiding vader. De PRU daalt nu onder het niveau dat in de tabellen MPAK, LPAK en UPAK bereikt werd (vergelijk blz. 9). Met name de associatie in de subtabel Keus×Prestatieniveau categorie (4) en (5) blijkt van belang te zijn. De G^2 -waarde voor het verschil tussen de indeling in 5 categorieën en de indeling in 3 cate-

gorieën is 434, bij 123 vrijheidsgraden is de overschreidingskans kleiner dan 0,0001. Een reductie van de tabel door het sommeren van categorieën voor Prestatieniveau lijkt dus minder wenselijk.

Tabel 3. Reductie in de $PRU_{K|APO}$ door het samenvoegen van categorieën van Prestatieniveau

Categorie indeling P	$PRU_{K APO}$	Reductie	G^2	df*
1, 2, 3, 4, 5	0,4993	0	33994	372
1+2, 3, 4, 5	0,4975	0,0018	33870	324
1, 2, 3, 4+5	0,4948	0,0046	33684	300
1+2, 3, 4+5	0,4929	0,0064	33560	249

* De aantallen vrijheidsgraden zijn gecorrigeerd (zie bijlage B).

4. De aanpassing en selectie van modellen voor associatie

4.1. De PRU als passingsmaat voor een associatiemodel

Voor geselecteerde variabelen met een bepaalde categorie-indeling kan de associatie verder geanalyseerd worden. Stel dat er een driedimensionale tabel geselecteerd is met variabelen X, Y en Z, waarbij Y de afhankelijke variabele is. Voor deze tabel zijn de volgende 5 hypothesen (modellen) van belang (de modellen zijn genoteerd in termen van gefitte marginalen zoals bij log-lineaire analyse gebruikelijk is):

Model	Marginalen
H_1	XZ Y
H_2	XZ XY
H_3	XZ ZY
H_4	XZ XY ZY
H_5	XZY

De eerste hypothese veronderstelt geen associatie tussen afhankelijke en onafhankelijke variabelen. De hypothesen H_2 en H_3 veronderstellen partiële associatie, respectievelijk tussen X en Y en tussen Z en Y . Hypothese H_4 veronderstelt partiële associatie tussen X en Y en tussen Z en Y . Hypothese H_5 tenslotte veronderstelt alle partiële bivariate associaties en de trivariate associatie. Om de fit van een model te beoordelen kan de $PRU_{Y|XZ}$ berekend worden onder de veronderstelling dat het model geldt. Als deze PRU veel lager is dan de in de tabel maximaal haalbare PRU , dan mist het model belangrijke associaties.

De $PRU_{Y|XZ}$ onder een model m is

$$PRU_{Y|XZ}^m = \frac{H^m(Y) - H_{XZ}^m(Y)}{H^m(Y)} = \frac{H^m(Y) - H^m(XZY) + H^m(XZ)}{H^m(Y)}, \quad (21)$$

waarin entropieën en conditionele entropieën die zijn aangeduid met het superscript m , berekend worden uit de onder model m geschatte kansverdelingen. Omdat onder modellen H_1 t/m H_5 de univariate marginale verdeling van Y gefixeerd is, geldt ook $H^m(Y) = H(Y)$. Omdat bovendien de bivariate marginale verdeling van X en Z gefixeerd is geldt ook $H^m(XZ) = H(XZ)$. Formule (21) is daarom te schrijven als

$$PRU_{Y|XZ}^m = \frac{H(Y) - H^m(XZY) + H(XZ)}{H(Y)}. \quad (22)$$

$H(Y)$ en $H(XZ)$ kunnen eenvoudig uit de data geschat worden. Voor het schatten van $H^m(XZY)$ is de trivariate kansverdeling van X , Y en Z onder model m nodig. De maximum likelihood (ML) schattingen voor alle mogelijke modellen m zijn bekend (zie Birch, 1963). In sommige gevallen zijn de schattingen direct te berekenen uit de marginale kansen, in andere gevallen zijn hiervoor iteratieve algoritmes nodig zoals die geïmplementeerd zijn in standaard programmatuur voor log-lineaire analyse (in dit geval alleen voor model H_4).

De PRU -maat voor de fit van een model vertoont een zekere gelijkenis met de gebruikelijke toetsingsgrootte, G^2 , voor log-lineaire modellen. De G_m^2 -waarde is minus twee maal de log-likelihood ratio van model m ten opzichte van het verzadigde model (H_5):

$$G_m^2 = -2n \sum_{ijk} [(\hat{p}_{ijk}^m \ln \hat{p}_{ijk}^m) - (\hat{p}_{ijk} \ln \hat{p}_{ijk})], \quad (23)$$

waarin de kansen $\hat{p}_{ijk}^m$ de ML schattingen zijn voor de kansen onder model (m) en $\hat{p}_{ijk}$ de ML schattingen zijn voor de kansen onder het verzadigde model. Omdat ieder model (m) genest is in het verzadigde model geldt hier de volgende gelijkheid voor geneste modellen (zie Bishop et al. (1975), p. 126):

$$\sum_{ijk} \hat{p}_{ijk} \ln \hat{p}_{ijk}^m = \sum_{ijk} \hat{p}_{ijk}^m \ln \hat{p}_{ijk}^m. \quad (24)$$

Dus G_m^2 is te schrijven als

$$\begin{aligned} G_m^2 &= -2n[(\sum_{ijk} \hat{p}_{ijk}^m \ln \hat{p}_{ijk}^m) - (\sum_{ijk} \hat{p}_{ijk} \ln \hat{p}_{ijk})] = \\ &= 2n[H^m(XYZ) - H(XYZ)]. \end{aligned} \quad (25)$$

Dus de G^2 -waarde is op een factor $2n$ na gelijk aan de toename van de entropie als gevolg van bepaalde restricties (gedefinieerd door het model) op de celkansen. De entropie neemt toe door de restricties omdat de kansen meer aan elkaar gelijk worden, dus de onzekerheid wordt groter. Door substitutie van (25) in (22) krijgen we

$$PRU_{Y|XZ}^m = \frac{H(Y) - H_{XZ}(Y)}{H(Y)} - \frac{G_m^2/2n}{H(Y)}. \quad (26)$$

Dus door een model m te specificeren voor de associatie in de tabel, vermindert de PRU met $(G_m^2/2n)/H(Y)$. Voor het verzadigde model (H_5) is G^2 gelijk aan nul, dus $PRU_{Y|XZ}^m = PRU_{Y|XZ}$.

Een andere maat voor de fit van een associatiemodel is de door Goodman (1970) geïntroduceerde maat R^2 :

$$R^2 = [G^2(H^0) - G^2(H^*)]/G^2(H^0). \quad (27)$$

Deze maat meet de proportionele vermindering van de G^2 van model H^* ten opzichte van de G^2 van een basis hypothese H^0 . Voor de modellen H_1 t/m H_5 kiezen we als basishypothese het model H_1 . De R^2 -waarde voor de andere modellen geeft dan de reductie in G^2 aan ten gevolge van het toevoegen van bepaalde associaties tussen de onafhankelijke variabelen en de afhankelijke variabele. Als we G^2 opvatten als een maat voor de variatie in de tabel, dan is de R^2 -waarde van een model m de onder model m verklaarde variatie in de tabel. Het verschil tussen de R^2 en de PRU is dus dat de PRU een maat is voor de verklaarde variatie in de afhankelijke variabele en de R^2 een maat is voor de verklaarde variatie in de tabel. Dit betekent dat wanneer de onafhankelijke variabelen weinig samenhangen met de afhankelijke variabele de PRU klein zal blijven ook al breidt men het model uit tot een verzadigd model. De R^2 zal echter altijd 1 worden voor het verzadigde model; alle variatie in de tabel is dan verklaard. Hoewel Goodman de analogie van de R^2 met de multiële correlatiecoëfficiënt benadrukt lijkt de PRU door bovenstaande interpretatie meer analoog aan de multiële correlatiecoëfficiënt dan de R^2 . De overeenkomst tussen de R^2 en de PRU is hun nauwe relatie met de G^2 . Aan de formules voor de R^2 en de PRU onder een model is te zien dat wanneer een model een grotere G^2 -waarde heeft dan een ander model, het eerste model dus een slechtere fit vertoont, dat eerste model een lagere PRU en een lagere R^2 -waarde heeft dan het tweede model. De R^2 , PRU en G^2 -waarde zullen de modellen dus op eenzelfde manier ordenen naar fit.

4.2. Toepassing

Voor de variabelen Opleiding vader, Prestatie, Advies en Keus is een causale structuur verondersteld (zie paragraaf 2). Deze structuur houdt in dat Prestatie, Advies en Keus als afhankelijke variabelen opgevat kunnen worden in respectievelijk de tabellen OP, OPA en OPAK. De verklaarde variatie onder een model voor de associatiestructuur zal nu in ieder van deze tabellen onderzocht worden.

Tabel OP is een triviaal geval. Het enige niet verzadigde model is het model dat O en P onafhankelijk veronderstelt. De resultaten met Prestatieniveau als afhankelijke variabele staan in tabel 4.

Tabel 4. Resultaten met Prestatieniveau als afhankelijke variabele

Model	Marginalen		G^2	df	P	$PRU_{P O}$	R^2
H_1	O	P	2584	16	$<0,001$	0	0
H_2	OP		0			0,034	1

De G^2 -waarde van het onafhankelijkheidsmodel is erg groot vergeleken bij het aantal vrijheidsgraden ($P<0,001$). De verklaarde variatie in de tabel en de verklaarde variatie in de variabele Prestatie zijn uiteraard nul onder dit model. Als O en P niet onafhankelijk verondersteld worden, wordt de variatie in de tabel volledig verklaard, de variatie in Prestatieniveau wordt echter slechts voor een klein deel door Opleiding vader verklaard.

De resultaten voor tabel OPA, waarbij Advies als afhankelijke variabele is opgevat, staan in tabel 5. Uit deze resultaten blijkt dat het model H_6 , het model zonder de driefactor interactie van O, P en A, bijna evenveel variatie in Advies verklaard als het verzadigde model. Dit model verklaard ook bijna alle variatie in de tabel. Het model zonder driefactor interactie, H_6 , verdient ons inziens de voorkeur hoewel het significant van het verzadigde model verschilt.

Tabel 5. Resultaten met Advies als afhankelijke variabele

Model	Marginalen			G^2	df	P	$PRU_{A OP}$	R^2
H_3	OP	A		18904	72	$<0,001$	0	0
H_4	OP	OA		14979	60	$<0,001$	0,057	0,208
H_5	OP	PA		1603	60	$<0,001$	0,252	0,915
H_6	OP	OA	PA	74	48	0,009	0,274	0,996
H_7	OPA			0	0		0,275	1

De modellen voor de associatie tussen de variabele Keus en de drie onafhankelijke variabelen O, P en A staan in tabel 6. Selectie van model H_{15} wordt gesuggereerd door de $\text{PRU}_{K|OPA}$ en de R^2 ; de toename in deze grootheden door driefactor interacties toe te voegen is niet groot. Vooral de associaties tussen Prestatie en Advies, en Keus en Advies blijken belangrijk te zijn.

Tabel 6. Resultaten met Keus als afhankelijke variabele

Model	Marginalen	G^2	df *	P	$\text{PRU}_{K OPA}$	R^2
H_8	OPA K	33855	267	<0,001	0	0
H_9	OPA OK	29455	255	<0,001	0,065	0,120
H_{10}	OPA PK	17483	255	<0,001	0,241	0,484
H_{11}	OPA AK	2484	258	<0,001	0,459	0,924
H_{12}	OPA OK PK	15513	243	<0,001	0,269	0,542
H_{13}	OPA OK AK	1894	246	<0,001	0,469	0,944
H_{14}	OPA PK AK	924	246	<0,001	0,484	0,973
H_{15}	OPA OK PK AK	310	234	0,001	0,493	0,991
H_{16}	OPA OPK AK	259	176	<0,001	0,494	0,992
H_{17}	OPA OAK PK	232	182	0,005	0,494	0,993
H_{18}	OPA PAK OK	227	174	0,003	0,494	0,993
H_{19}	OPA PAK OAK	149	122	0,020	0,495	0,996
H_{20}	OPA OPK OAK PAK	99	62	0,001	0,496	0,997
H_{21}	OPAK	0	0		0,497	1,000

* De vrijheidsgraden zijn gecorrigeerd voor nullen in de gefitte marginalen (zie bijlage B).

5. Een recursief stelsel vergelijkingen

De modellen H_2 , H_6 en H_{15} voor de drie tabellen OP, OPA en OPAK, waarbij respectievelijk P, A en K onafhankelijke variabelen waren, zijn op te vatten als een stelsel simultane recursieve modellen die samen de vierdimensionale tabel beschrijven. Dit stelsel simultane modellen noemen we H_{22} . De G^2 -waarde voor model H_{22} is de som van de G^2 -waarden van de drie afzonderlijke modellen (zie Goodman, 1973, Fienberg, 1977). In tabel 7 zijn de G^2 -waarden van de drie afzonderlijke modellen en het simultane model weergegeven.

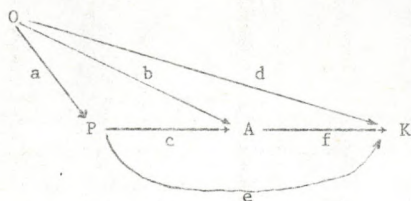
Tabel 7. Samenstelling van de G^2 -waarde voor model H_{22}

Model	G^2	df	P
H_2	0	0	
H_6	74	48	0,009
H_{15}	310	234	0,001
	— +	— +	
H_{22}	384	282	<0,001

We zien dat model H_{22} significant verschilt van het verzadigde model voor de vierdimensionale tabel. Niettemin verklaart dit model 3,4% van de variatie in Prestatie, 27,4% van de variatie in Advies en 49,3% van de variatie in Keus, terwijl maximaal (onder het verzadigde model) respectievelijk 3,4%, 27,5% en 49,7% verklaard kan worden. Gezien dit geringe verlies in verklaarde variatie en het relatief geringe aantal vrije parameters (112 parameters voor model H_{22} tegen 399 voor het verzadigde model) beschouwen we H_{22} als een adequate en spaarzame beschrijving van de data.

De verklaarde variatie in P, A, en K onder de modellen H_2 , H_6 en H_{15} (of, equivalent hieraan, model H_{22}) kan worden uitgesplitst naar de verschillende verklarende variabelen. In figuur 2 staan de mogelijke effecten aangegeven.

Figuur 2. Mogelijke causale relaties tussen O, P, A en K



De effecten op een afhankelijke variabele worden conditioneel op elkaar bepaald. Zo is b het effect van O op A, gegeven het effect van P op A en c het effect van P op A, gegeven het effect van O op A. Deze effecten zijn te kwantificeren door partiële PRU's. Effect b, de door O verklaarde variatie in A gegeven P, is gelijk aan $PRU_{A.P|O}$ en effect c is $PRU_{A.O|P}$. Deze twee partiële PRU's zijn te berekenen uit de $PRU_{A|OP}$ waarden die in tabel 5 staan. $PRU_{A.P|O}$ is, onder de veronderstelling dat het geselecteerde model H_6 geldt, het verschil tussen $PRU_{A|OP}$ onder model H_6 en $PRU_{A|OP}$ onder model H_5 . Op analoge wijze zijn de andere effecten te berekenen. De resultaten hiervan staan in tabel 8.

Tabel 8. Schattingen voor de effecten a t/m f

Effect	Geschatte waarde
a = $PRU_{P O}$	0,035
b = $PRU_{A.P O}$	0,022
c = $PRU_{A.O P}$	0,217
d = $PRU_{K.PA O}$	0,009
e = $PRU_{K.OA P}$	0,023
f = $PRU_{K.OP A}$	0,223

Uit tabel 8 blijkt dat het effect van Advies op Keus (f) ongeveer even groot is als het effect van Prestatie op Advies (c). Deze twee effecten zijn aanzienlijk groter dan de overige effecten.

6. Conclusies

Wanneer we bij nominale variabele onderscheid maken tussen onafhankelijke en afhankelijke variabelen kan de verklaarde variatie in de afhankelijke variabele gemeten worden met de PRU. Deze maat is op te vatten als een analogon van de proportie verklaarde variantie, R^2 , bij continue variabelen. De PRU heeft een nauwe relatie met de G^2 , waardoor deze associatiemaat toegepast kan worden in die gevallen waarin de G^2 als toetsingsgrootte gehanteerd wordt. Met behulp van de PRU kan het verlies aan verklaarde varia-

tie vastgesteld worden dat ontstaat door het samenvoegen van categorieën van de onafhankelijke variabelen of door restricties op te leggen aan de associatiestructuur zoals die geformuleerd worden door log-lineaire modellen. Door Goodman (1970) is de proportie verklaarde G^2 als analogon voor de R^2 voorgesteld. Goodmans maat heeft echter het nadeel dat deze altijd de maximumwaarde 1 bereikt voor een model zonder restricties. De PRU bereikt zijn maximum waarde 1 alleen als de afhankelijke variabele volledig verklaard wordt. Analooq aan de partiële correlatiecoëfficiënt is er ook een partiële PRU te definiëren. Met de partiële PRU is het effect van iedere onafhankelijk variabele conditioneel op de andere onafhankelijke variabelen te bepalen. Deze effecten zijn ook te berekenen onder de restricties die een model aan de associatiestructuur oplegt. In dat geval is een partiële PRU een samenvattende maat voor de grootte van een effect uit een log-lineair model.

De PRU is gebruikt om de effecten van een aantal achtergrondvariabelen op de schoolkeuze vast te stellen. Opleiding vader, Prestatieniveau en Advies van de onderwijzer verklaren samen 49,9% van de variatie in Schoolkeus. Door categorieën van Opleiding vader samen te voegen en alle driefactor en vierfactor interacties gelijk aan nul te stellen daalt de verklaarde variatie tot 49,3%. Het aantal te schatten parameters wordt door deze restricties teruggebracht van 559 tot 88. De partiële effecten van Advies op Keus en Prestatie op Advies bleken de grootste effecten te zijn. Opleiding vader had het grootste effect op het prestatieniveau, een minder groot partiël effect op het advies en het minste partiële effect op de schoolkeuze. De opleiding van de vader beïnvloedt de schoolkeuze dus vooral via het prestatieniveau en het advies.

Bijlage A. Categorie indelingen en steekproefaantallen

Opleiding	Vader (O)	Moeder (L)
1. Lagere school	8 703	12 947
2. Lager beroepsonderwijs	1 435	2 923
3. Voortgezet onderwijs	6 553	6 406
4. Middelbaar beroepsonderwijs	4 879	991
5. Middelbaar algemeen onderwijs	643	568
6. Hogere beroepsopleiding	2 980	1 287
7. Academisch onderwijs	985	161

Beroep vader (M)

1. Arbeider	8 504
2. Zelfstandige zonder Personeel	2 097
3. Zelfstandige met Personeel	1 578
4. Lagere employe	3 329
5. Middelbare employe	5 341
6. Hogere employe + vrije beroepen	3 140
7. Overige	5 301

Onderwijselement	Advies (A)	Keus (K)
1. Lager beroepsonderwijs	6 533	7 605
2. MAVO	10 586	10 443
3. HAVO	5 502	3 342
4. VWO	3 575	4 806

Geslacht (G)

1. Man	12 704
2. Vrouw	13 492

Urbanisatiegraad (U)

1. Plattelandsgemeenten	3 726
2. Verstedelijkte plattelandsgemeenten (< 5 000 inw.)	2 450
3. Verstedelijkte plattelandsgemeenten (5 000 - 30 000 inw.)	4 327
4. Verstedelijkte plattelandsgemeenten (specifieke forensengemeenten)	4 236
5. Gemeenten met een stedelijk karakter (2 000 - 10 000 inw.)	777
6. Gemeenten met een stedelijk karakter (10 000 - 30 000 inw.)	2 737
7. Gemeenten met een stedelijk karakter (30 000 - 50 000 inw.)	1 852
8. Gemeenten met een stedelijk karakter (50 000 - 100 000 inw.)	3 016
9. Gemeenten met een stedelijk karakter (100 000 en meer inw.)	6 153

Aantal kinderen (N)

1. Een	1 975
2. Twee	8 823
3. Drie	7 418
4. Vier of meer	6 052

Prestatieniveau (P)

1. Laag	2 002
2.	6 503
3.	8 766
4.	7 058
5. Hoog	1 867

Bijlage B. Marginale nullen

In een aantal marginale frequentietabellen kwamen nullen voor. Dit heeft tot gevolg dat bij ieder model dat een van deze marginalen bevat, al die geschatte celfrequenties die optellen tot een marginale nulcel, als nul geschat zullen worden. In zulke gevallen moet het aantal vrijheidsgraden aangepast worden (zie Bishop et al. (1975), p. 115). De aantallen vrijheidsgraden in tabel 2 en tabel 3 zijn verminderd als gevolg van het optreden van nullen in de marginale tabel OPA. De aantallen vrijheidsgraden in tabel 6 zijn verminderd omdat er nullen voorkwamen in de marginale tabellen OPA, OPK, OAK en PAK. In tabel 9 zijn de aantallen marginale nullen en de vermindering in aantallen vrijheidsgraden die dit ten gevolge had aangegeven.

Tabel 9. Vermindering in aantallen vrijheidsgraden als gevolg van marginale nullen

marginale tabel	aantal nullen	vermindering in df
OPA (O=1,2,3,4,5,6,7)	15	45
OPA (O=1,2+3,4,5,6,7)	13	39
OPA (O=1,2,3,4+5,6,7)	12	36
OPA (O=1,2+3,4+5,6,7)	10	30
OPA (P=1,2,3,4,5)	15	45
OPA (P=1+2,3,4,5)	4	12
OPA (P=1,2,3,4+5)	11	33
OPA (P=1+2,3,4+5)	0	0
OPK	4	12
OAK	4	16
PAK	6	24

Literatuur

- Abrahamse, A.P.J. en W.M. Lammerts van Bueren, 1979, Looking at multiway tables. Report 7931/S (Econometrisch Instituut, Erasmus Universiteit, Rotterdam).
- Ashby, W.R., 1965, Measuring the internal informational exchange in a system. *Cybernetica* 9, blz. 5-22.
- Birch, M.W., 1963, Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society B* 25, blz. 220-233.
- Bishop, Y.M.M., S.E. Fienberg en P.W. Holland, 1975, Discrete multivariate analysis, theory and practice (MIT Press, Cambridge).
- CBS (Centraal Bureau voor de Statistiek), 1979, Schoolloopbaan en herkomst van leerlingen bij het voortgezet onderwijs 1977/'78; Mededelingen no. 7699 (CBS, Voorburg).
- Fienberg, S.E., 1977, The analysis of cross-classified categorical data (MIT Press, Cambridge).
- Goodman, L.A., 1970, The multivariate analysis of qualitative data: interactions among multiple classifications. *Journal of the American Statistical Association* 65, blz. 226-256.
- Goodman, L.A., 1973, The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika* 60, blz. 179-192.
- Lammerts van Bueren, W.M., 1982, Measuring association in nominal data. Proefschrift (Erasmus Universiteit, Rotterdam).
- Teachman, J.D., 1980, Analysis of population diversity, measures of qualitative variation. *Sociological Methods and Research* 8, blz. 341-362.
- Theil, H., 1972, Statistical decomposition analysis (North-Holland, Amsterdam).