

Multivariate analysemethoden voor discrete variabelen, een vervolg:

Samenhang binnen is binnen?

W.M. Lammerts van Bueren*)

Samenvatting

Deze op uitnodiging van de redactie van KM geschreven notitie vormt een aanvulling op het artikel "Multivariate analysemethoden voor discrete variabelen" van Israëls e.a. (1981). De aanvulling betreft de door Israëls c.s. voorgestelde opsplitsing van samenhang in "tussen" en "binnen". Nadere beschouwing leert, dat "binnen" met twee maten gemeten kan worden.

1. Inleiding

In hun artikel "Multivariate analysemethoden voor discrete variabelen" bespreken Israëls e.a. (1981) een aantal technieken die geschikt zijn voor de analyse van meerdimensionale tabellen. Daarbij ligt, aldus de schrijvers, de nadruk op het "typeren", dus op de data-reductie, en speelt het "toetsen" op significantie van de opgespoorde effecten een ondergeschikte rol. Gegeven het enorme aantal waarnemingen waarover de schrijvers de beschikking hebben, ligt deze benaderingswijze voor de hand. In die omstandigheden immers zal, zoals de schrijvers het uitdrukken, "een bepaalde a priori restrictie vrijwel zeker worden verworpen". Of, om het met Buchanan (1974), p.642, te zeggen:

"... there appears to be a tendency to rely less on significance tests and more on the meaning of the data in context, especially when one is working with a large sample that will turn up many trivial but statistically significant differences."

Niettemin blijkt in het artikel het opsporen en meten van effecten zo nu en dan hand in hand te gaan met de toetsing ervan op significantie. Hoofdstuk 5, "Stapsgewijze selectiemethoden m.b.v. het G^2 -criterium", vormt daarvan een duidelijke illustratie. Enerzijds wijst het gebruik van G^2 , de likelihood

*) Vakgroep Statistiek, Economische Faculteit,
Erasmus Universiteit Rotterdam

ratio toetsingsgrootheid, op het confirmatieve (toetsende) karakter van de analyse, terwijl anderzijds de toepassing van de samenhangsmaatstaf ψ^2 het exploratieve (typerende) karakter onderstreept. Onderstaand commentaar richt zich op dit samenspel van G^2 en ψ^2 .

2. Meten van samenhang

Het is genoegzaam bekend dat het constateren van een bepaalde mate van samenhang tussen variabelen iets geheel anders is dan de vaststelling dat die samenhang al dan niet significant is. In een kleine steekproef moet de samenhang al erg sterk zijn, wil hij als significant uit de bus komen, terwijl in een grote steekproef een zwakke samenhang nog met vlag en wimpel de kwalificatie "significant" kan verdienen. Het significant-zijn van een samenhang zegt dus weinig over de sterkte van de betrokken relatie.

Juist in een situatie waarin ook een heel zwakke samenhang nog het stempel "significant" zal krijgen - en om die situatie gaat het in het artikel van Israëls c.s. - is het meten van samenhang onontbeerlijk. Het is dan ook alleszins begrijpelijk dat in hoofdstuk 5 naast het selectiecriterium G^2 ook nog een (asymmetrische) maatstaf voor samenhang wordt gehanteerd.

De bewuste maatstaf, ψ^2 , door Israëls c.s. "maatstaf voor voorspelbaarheid" genoemd, maar elders ook wel aangeduid als "uncertainty coefficient", zie bijvoorbeeld Nie e.a. (1975), p.226, of "covariability ratio", zie Abrahamse en Lammerts van Bueren (1980), behoort tot de categorie der zogeheten PRE-maatstaven, waarin PRE staat voor "proportional-reduction-in-error", zie ook van den Ende (1973), par. 6.5.3, en Hays (1980), p.751.

Hoewel de maatstaf opgevat kan worden als een geschaalde versie van de aan Pearson's χ^2 verwante toetsingsgrootheid G^2 , onderscheidt hij zich nadrukkelijk van geschaalde versies van χ^2 . Deze laatste, zoals ϕ^2 , zie Yule en Kendall (1973), p.53, en V^2 , zie Cramér (1966), p.443, staan in het algemeen een interpretatie in termen van proportionele reductie niet toe, zie bijvoorbeeld Costner (1965).

Voorbijgaand aan de details geldt voor $J(\underline{x}, \underline{y})$ uit definitie (5.1), in symbolen van Israëls c.s., dat

$$J(\underline{x}, \underline{y}) = H(\underline{y}) - H_{\underline{x}}(\underline{y})$$

waarbij de marginale grootheid $H(\underline{y})$ gedefinieerd is als $\sum_{j=1}^N p_{+j} \ln \frac{1}{p_{+j}}$ en de

partiële grootheid $H_{\underline{x}}(\underline{y})$ als $\sum_{i=1}^M p_{i+} \sum_{j=1}^N p_{j|i} \ln \frac{1}{p_{j|i}}$.

Omdat $H(\underline{y})$ op te vatten is als een maatstaf voor de marginale spreiding van $\underline{y}$ en $H_{\underline{x}}(\underline{y})$ als een maatstaf voor de (gemiddelde) spreiding van $\underline{y}$, gegeven de verschillende waarden van $\underline{x}$, dus de partiële spreiding van $\underline{y}$, is de maatstaf

$$\psi^2 = \frac{J(\underline{x}, \underline{y})}{H(\underline{y})} = \frac{H(\underline{y}) - H_{\underline{x}}(\underline{y})}{H(\underline{y})}$$

inderdaad te interpreteren als de (gemiddelde) proportionele reductie in spreiding van $\underline{y}$, tengevolge van het constant houden van $\underline{x}$.

Behalve de verwantschap van ψ^2 met G^2 (additiviteit!) is het juist deze interpretatiemogelijkheid die ψ^2 tot zo'n aantrekkelijke maatstaf voor samenhang maakt. Met name de getalswaarden die ψ^2 aanneemt krijgen hierdoor een intuïtief aansprekende betekenis.

In dit opzicht dringt een vergelijking met bijvoorbeeld het kwadraat van de correlatiecoëfficiënt, of beter nog, de correlatieratio, zich op. Deze geven immers ook de proportionele reductie in spreiding weer, de eerste onder de veronderstelling van een lineair verband.

3. Meten van samenhang "tussen" en "binnen"

In het licht van bovenstaande interpretatie van ψ^2 en zijn verwantschap met samenhangsmaatstaven voor kwantitatieve variabelen, is uitdrukking (5.9) een nadere beschouwing waard. Deze uitdrukking,

$$\Phi_{\text{Totaal}}^2 = \Phi_{\text{Tussen}}^2 + \Phi_{\text{Binnen}}^2$$

is verkregen door de G^2 -termen uit optelregel (5.8) te delen door $2 n \hat{H}(\underline{y})$.

Men kan zich nu twee dingen afvragen:

- 1) In hoeverre laat elk van de drie grootheden die in uitdrukking (5.9) voorkomen een PRE-interpretatie toe?
- 2) In hoeverre stemt de opsplitsing van Φ_{Totaal}^2 overeen met bijvoorbeeld de opsplitsing van de gekwadrateerde multipele correlatiecoëfficiënt in het geval van één te verklaren variabele, $\underline{y}$, en twee verklarende variabelen, $\underline{x}_1$, en $\underline{x}_2$?

Bij de beantwoording van deze vragen zal blijken dat Φ_{Totaal}^2 in uitdrukking

(5.9) wordt opgesplitst op een wijze die weliswaar numeriek volledig correct is, maar wellicht toch minder voor de hand liggend dan men bij oppervlakkige kennisneming geneigd is te denken. Het gaat daarbij met name om de definitie van de grootheid $\hat{\Phi}_{\text{Binnen}}^2$.

Onderstaand eenvoudig voorbeeld moge het specifieke karakter van deze grootheid illustreren (ter wille van een vlotte verificatie is een extreem voorbeeld gekozen):

Voorbeeld

Een 6×3 frequentietabel wordt ontbonden in twee 3×3 deeltabellen; de zes rijen worden dus in twee groepen ingedeeld.

Tabel (a)				Tabel (b_1)				Tabel (b_2)				Tabel (c)			
y				y				y				y			
10	0	0	10	10	0	0	10	0	0	10	10	10	20	30	60
0	20	0	20	0	20	0	20	0	20	0	20	0	20	10	60
0	0	30	30	0	0	30	30	0	0	30	30	0	0	40	120
0	0	10	10	10	20	30	60	0	0	10	10	40	40	40	120
0	20	0	20					0	20	0	20				
30	0	0	30					30	0	0	30				
40	40	40	120					30	20	10	60				

Tabel (a) is de oorspronkelijke tabel; op basis hiervan kan $\hat{\Phi}_{\text{Totaal}}^2$ worden berekend. De tabellen (b_1) en (b_2) stellen de twee groepen voor. Tabel (c) tenslotte bestaat uit de kolomtotalen van de twee tabellen (b) en levert

$\hat{\Phi}_{\text{Tussen}}^2$ op.

Zonder rekenen is gemakkelijk in te zien dat $\hat{\Phi}_{\text{Totaal}}^2 = 1$ en dat $0 < \hat{\Phi}_{\text{Tussen}}^2 < 1$. Immers, in Tabel (a) is de spreiding in y per rij geheel verdwenen, dus de proportionele reductie in spreiding is 1, terwijl Tabel (c) weliswaar reductie in spreiding laat zien, maar geen volledige (om precies te zijn: $\hat{\Phi}_{\text{Tussen}}^2 = .079$). Volgens uitdrukking (5.9) volgt hieruit dat $0 < \hat{\Phi}_{\text{Binnen}}^2 < 1$ ($\hat{\Phi}_{\text{Binnen}}^2 = .921$).

Met deze laatste constatering wordt onmiddellijk duidelijk wat hierboven met "het specifieke karakter" van $\hat{\Psi}_{\text{Binnen}}^2$ bedoeld werd: $\hat{\Psi}_{\text{Binnen}}^2$ dient niet opgevat te worden als een maatstaf voor de samenhang die zich (gemiddeld) binnen de groepen manifesteert. Immers, terwijl zowel in Tabel (b_1) als Tabel (b_2) de samenhang maximaal is, neemt $\hat{\Psi}_{\text{Binnen}}^2$ een waarde aan die kleiner is dan 1. Deze paradox laat zich overigens gemakkelijk verklaren. De grootheid $\hat{\Psi}_{\text{Binnen}}^2$ meet weliswaar een (gemiddelde) proportionele reductie in spreiding, maar relateert daarbij de (gemiddelde) absolute reductie aan de marginale spreiding van y in Tabel (a) in plaats van aan de (gemiddelde) marginale spreiding van y in de tabellen (b). De grootheid $\hat{\Psi}_{\text{Binnen}}^2$ is immers verkregen door deling van G_{Binnen}^2 door $2 n \hat{H}(y)$ en niet door $\sum_k 2 n_k \hat{H}_k(y)$, zoals blijkt uit de overgang van opsplitsing (5.8) naar (5.9) in Israëls e.a. (1981).

Hoewel er rekenkundig niets op tegen is de grootheid $\hat{\Psi}_{\text{Binnen}}^2$ te definiëren als de (gemiddelde) reductie in spreiding die zich binnen de groepen manifesteert, relatief ten opzichte van de oorspronkelijke spreiding, wordt het PRE-karakter van ψ^2 door deze wijze van definiëren wel aangetast. De getalswaarde die $\hat{\Psi}_{\text{Binnen}}^2$ aanneemt, staat niet langer voor de (gemiddelde) samenhang *binnen* de tabellen die aan de orde zijn.

Wellicht dat de volgende enigszins afwijkende aanpak daarom in gevallen als deze toch de voorkeur verdient boven de door Israëls c.s. voorgestelde schaling van de G^2 -termen uit optelregel (5.8).

Hierbij zij aangetekend dat er geen reden is de maatstaven $\hat{\Psi}_{\text{Totaal}}^2$ en $\hat{\Psi}_{\text{Tussen}}^2$ aan te passen. Zij zijn beide verkregen door deling van de overeenkomstige G^2 -termen door $2 n \hat{H}(y)$ en zijn dus echte PRE-maatstaven, waarbij de reductie in spreiding gerelateerd wordt aan de bijbehorende marginale spreiding. Om bij het bovenstaande voorbeeld te blijven: de drie kolomtotalen van respectievelijk Tabel (a) en Tabel (c) komen overeen met de marginale verdeling van y .

4. Meten van samenhang "binnen", een alternatief

Terwille van enige notationele eenvoud definiëren we

$$\mu_k = 2 n_k \hat{H}_k(y) \quad \text{en} \quad \mu = 2 n \hat{H}(y)$$

$$\text{zodat } \frac{\sum_k \mu_k}{\mu} = 1 - \hat{\Psi}_{\text{Tussen}}^2.$$

Uitdrukking (5.9) kan nu als volgt herschreven worden:

$$\begin{aligned}
 \hat{\Psi}_{\text{Totaal}}^2 &= \hat{\Psi}_{\text{Tussen}}^2 + \frac{\sum_k \hat{\Psi}_{\text{Binnen } k}^2}{\mu} \\
 &= \hat{\Psi}_{\text{Tussen}}^2 + \sum_k \left[\frac{\left(\sum_k \mu_k \right)}{\mu} \cdot \frac{\mu_k}{\left(\sum_k \mu_k \right)} \cdot \frac{\hat{\Psi}_{\text{Binnen } k}^2}{\mu_k} \right] \\
 &= \hat{\Psi}_{\text{Tussen}}^2 + \frac{\sum_k \mu_k}{\mu} \cdot \frac{\sum_k \mu_k \cdot \hat{\Psi}_{\text{Binnen } k}^2}{\sum_k \mu_k} \\
 &= \hat{\Psi}_{\text{Tussen}}^2 + (1 - \hat{\Psi}_{\text{Tussen}}^2) \cdot \hat{\Psi}_{\text{Binnen}}^2
 \end{aligned}$$

Hiermee is de (partiële) maatstaf $\hat{\Psi}_{\text{Binnen}}^2$ gedefinieerd als een gewogen gemiddelde van de (conditionele) maatstaven $\hat{\Psi}_{\text{Binnen } k}^2$, welke laatste de samenhang binnen de afzonderlijke groepen meten. In het voorgaande voorbeeld neemt $\hat{\Psi}_{\text{Binnen}}^2$ dan ook de waarde 1 aan.

Uit bovenstaande afleiding blijkt overigens ook dat $\hat{\Psi}_{\text{Binnen}}^2$ overeenkomt met

$$\text{het al eerder gesuggereerde quotiënt } \frac{\hat{\Psi}_{\text{Binnen}}^2}{\sum_k 2 \frac{n_k}{n} \hat{H}_k(y)}.$$

Het zal duidelijk zijn dat $\hat{\Psi}_{\text{Binnen}}^2$ als een, zij het partiële, PRE-maatstaf in de gebruikelijke zin kan worden opgevat. De getalswaarde die de maatstaf aanneemt heeft betrekking op de (gemiddelde) samenhang die binnen de groepen wordt aangetroffen.

Het ondubbelzinnig karakter van ψ^2 , opgevat als maatstaf voor samenhang in tabellen, lijkt door deze gemodificeerde aanpak dan ook beter gewaarborgd. Om die reden wordt in Abrahamse en Lammerts van Bueren (1980), p.10 en 11, aan deze benadering de voorkeur gegeven.

Hoewel Israël's c.s. geen gewag maken van de partiële maatstaf $\hat{\Psi}_{\text{Binnen}}^2$, wordt het gebruik van de conditionele maatstaf $\hat{\Psi}_{\text{Binnen } k}^2$ wel door hen gepropageerd. In Israël's en van Driel (1980), p.15, bijvoorbeeld, wordt $\hat{\Psi}_{\text{Binnen}}^2$ gedefinieerd als $\sum_k \lambda_k \hat{\Psi}_{\text{Binnen } k}^2$, waarbij λ_k staat voor $\frac{\mu_k}{\mu}$ en $\hat{\Psi}_{\text{Binnen } k}^2$ overeenkomt met $\hat{\Psi}_{\text{Binnen } k}^2$. Ook in Figuur 5.1 van het artikel dat nu aan de orde is, worden de verschillende $\hat{\Psi}_{\text{Binnen } k}^2$ afzonderlijk berekend. Zo valt bijvoorbeeld bij de eerste uitsplitsing in Figuur 5.1 af te lezen dat $\hat{\Psi}_{\text{Totaal}}^2 = .193$, $\hat{\Psi}_{\text{Tussen}}^2 = .103$, $\hat{\Psi}_{\text{Binnen } L_1}^2 = .079$ en $\hat{\Psi}_{\text{Binnen } L_2, L_3, L_4}^2 = .105$. Het gewogen gemiddelde

van de laatste twee volgt dan uit

$$\frac{\frac{7495}{.079} \times .079 + \frac{45605}{.105} \times .105}{\frac{7495}{.079} + \frac{45605}{.105}} = .100 = \Phi_{\text{Binnen}}^{*2}$$

erwijl dit resultaat uiteraard ook verkregen kan worden met behulp van de vergelijking

$$.193 = .103 + (1 - .103) \Phi_{\text{Binnen}}^{*2}$$

n het licht van hun definitie van Φ_{Binnen}^2 zou overigens verdedigbaar geweest zijn als Israëls en van Driel $\Phi_{\text{Binnen } k}^2$ hadden gedefinieerd als $\frac{n}{n_k} \cdot \frac{G_{\text{Binnen } k}^2}{\mu}$ n plaats van als $\frac{G_{\text{Binnen } k}^2}{\mu_k}$ ($= \Phi_{\text{Binnen } k}^{*2}$). Immers, indien het gemiddelde van de conditionele reducties in spreiding gerelateerd wordt aan de (oorspronkelijke) marginale spreiding van y , zie de definitie van Φ_{Binnen}^2 , waarom dan niet ook de conditionele reducties daaraan gerelateerd?

5. "Part" versus "partial"

Tot nu toe is steeds de beantwoording van de vraag betreffende het PRE-karakter van de gehanteerde maatstaven aan de orde geweest. Daarbij is gebleken dat Φ_{Totaal}^2 en Φ_{Tussen}^2 een "gewone" PRE-interpretatie toelaten, maar dat Φ_{Binnen}^2 als maatstaf voor binnen-samenhang in dit opzicht minder voor de hand ligt dan $\Phi_{\text{Binnen}}^{*2}$. Tegenover de opsplitsing

$$(1) \Phi_{\text{Totaal}}^2 = \Phi_{\text{Tussen}}^2 + \Phi_{\text{Binnen}}^2$$

zou dan ook de uitdrukking

$$(2) \Phi_{\text{Totaal}}^2 = \Phi_{\text{Tussen}}^2 + (1 - \Phi_{\text{Tussen}}^2) \Phi_{\text{Binnen}}^{*2}$$

gebruikt kunnen worden.

In hoeverre nu stemmen deze opsplitsingen overeen met gebruikelijke opsplitsingen van de gekwadrateerde multipele correlatiecoëfficiënt? Het antwoord op deze vraag verduidelijkt wellicht nog meer dan het voorgaande het verschillende karakter van Φ_{Binnen}^2 en $\Phi_{\text{Binnen}}^{*2}$.

Onder verwijzing naar Nie e.a. (1975), p.333, waar een even korte als heldere uiteenzetting te vinden is, zij hier vermeld dat, in het geval van één te

verklaren variabele, Y, en twee verklarende variabelen, X_2 en X_1 , geldt dat

$$(3) R^2_{y.12} = R^2_{y.2} + r^2_{y(1.2)} \quad , \text{ zie hun uitdrukking (28),}$$

en dat

$$(4) R^2_{y.12} = R^2_{y.2} + (1 - R^2_{y.2})r^2_{y1.2} \quad , \text{ zie (29).}$$

Hierin staat $R_{y.12}$ voor de multipele correlatiecoëfficiënt en $R_{y.2}$ voor de enkelvoudige correlatiecoëfficiënt van Y en X_2 . De opsplitsingen (3) en (4) zijn volstrekt analoog aan respectievelijk (1) en (2). Wat de interpretatie van de derde term in uitdrukking (3) betreft, $r^2_{y(1.2)}$, stellen Nie c.s.:

"one simple interpretation of part (also called semipartial) correlation $r_{y(1.2)}$ is that it is a simple correlation between the original Y and the residual of independent variable X_1 from which the effects of independent variable X_2 are taken out. (...) In other words, for the part correlation the effect of X_2 is taken out *only* from the independent variable X_1 . (...) The square of the part correlation (...) is (...) the *absolute* increment of R^2 due to the addition of X_1 to the equation already containing X_2 ."

En ten aanzien van de grootheid $r^2_{y1.2}$ in uitdrukking (4):

"In contrast, the partial correlation is a simple correlation between two residuals, the residual of Y and the residual of X_1 from both of which the effects of X_2 have taken out. (...) For the partial correlation the effect of X_2 is first taken out of *both* Y and X_1 , and then the resulting two residuals (...) are correlated with each other. (...) The square of the partial correlation between Y and X_1 is the *proportional* increment in the explained variation due to X_1 , expressed as a proportion of the variation unexplained by X_2 . That is, it represents a *proportional* reduction of *unexplained* variation."

(alle cursiveringen van Nie c.s.)

Met deze citaten lijkt ook het verschil in karakter tussen $\hat{\sigma}^2_{\text{Binnen}}$ en $\hat{\sigma}^{*2}_{\text{Binnen}}$ thans voldoende uit de doeken gedaan.

6. Conclusie

In deze notitie is de door Israëls e.a. (1981) voorgestelde opsplitsing van

ψ^2_{Totaal} in ψ^2_{Tussen} en ψ^2_{Binnen} ter discussie gesteld, waarbij als alternatief voor ψ^2_{Binnen} de partiële maatstaf $\psi^{*2}_{\text{Binnen}}$ werd voorgesteld. In tegenstelling tot ψ^2_{Binnen} blijkt $\psi^{*2}_{\text{Binnen}}$ een zuivere PRE-maatstaf te zijn. De keuze tussen beide grootheden is analoog aan die tussen het meten van "part correlation" en "partial correlation" in het geval van kwantitatieve variabelen.

Literatuur

- Abrahamse, A.P.J., en W.M. Lammerts van Bueren (1980), Nominal variables: dependence and predictive power, Report 8017/S, Econometrisch Instituut, Erasmus Universiteit Rotterdam.
- Buchanan, W. (1974), Nominal and ordinal bivariate statistics: the practitioner's view, American Journal of Political Science, 18, 625-646.
- Costner, H.L. (1965), Criteria for measures of association, American Sociological Review, 30, 341-353.
- Cramér, H. (1966), Mathematical methods of statistics, 11^e druk, Princeton University Press, Princeton.
- Ende van den, H.W. (1973), Beschrijvende statistiek, 2^e druk, Agon Elsevier, Amsterdam/Brussel.
- Hays, W.L. (1980), Statistics for the social sciences, 2^e ed., herdruk, Holt, Rinehart & Winston, London.
- Israëls, A.Z. e.a. (1981), Multivariate analysemethoden voor discrete variabelen, Kwantitatieve Methoden, 2, 87-149.
- Israëls, A.Z. en J. van Driel (1980), The use of the chi-square statistic in the analysis of multivariate tables, interne C.B.S.-nota nr. 9449-80-M1.
- Nie, N.H. e.a. (1975), Statistical package for the social sciences, 2^e ed., McGraw-Hill, New York.
- Yule, G.U. en M.A. Kendall (1973), An introduction to the theory of statistics, 14^e ed., 6^e druk, Charles Griffin & Co., London.