

HET METEN VAN ASSOCIATIE
BIJ VRAGEN MET MEERVOUDIGE ANTWOORDCATEGORIEËN

J.Th. Geilenkirchen en C. Ouwerkerk^{*}

Samenvatting.

In dit artikel wordt eerst een beknopte inleiding gegeven in de associatie-maatstaf ζ , welke de mate van samenhang meet tussen twee of meer nominale variabelen.

In het geval van vragen met meervoudige antwoordcategorieën ofwel multiple response vragen worden deze gehercodeerd volgens de methode van Goodman en Kruskal.

Vervolgens wordt deze maatstaf toegepast op twee voorbeelden.

Het eerste voorbeeld is overgenomen uit een artikel van Tanaka en Tarumi. Het tweede voorbeeld betreft een beperkt aantal gegevens afkomstig van een nog lopend onderzoek.

^{*} Erasmus Universiteit, Postbus 1738, 3000 DR Rotterdam.
tel. 010-4081519/4081318.

1. Inleiding.

In het sociaal-economisch onderzoek wordt bij het verkrijgen van informatie dikwijls gebruik gemaakt van vragenlijsten. Een vaak voorkomend fenomeen hierbij is het verschijnsel dat er bij een vraag meer antwoordmogelijkheden aangegeven kunnen worden. In deze situatie spreekt men van vragen met meervoudige antwoordcategorieën, in het engels multiple response vragen genoemd. Dit type vraag treedt bijv. op bij het stellen van de vraag welk of welke ochtendbladen men leest, waarbij gekozen kan worden uit de volgende antwoordmogelijkheden:

het Algemeen Dagblad	de Telegraaf	de Volkskrant
Ander ochtendblad	Geen.	

Voor een uitgebreide beschrijving van dit soort vragen verwijzen wij naar een aantal handboeken op market research gebied zoals Kinnear & Taylor (1987), Tull & Hawkins (1984) en Weiers (1984).

In het volgende zullen wij de engelse benaming 'multiple response' gebruiken, omdat deze meer gangbaar is dan 'vragen met meervoudige antwoordcategorieën'.

Voor het geschikt maken van een multiple response vraag voor computerverwerking bestaan meerdere codeermogelijkheden. Een overzicht van deze mogelijkheden wordt gegeven door Nie (1983). Voor de verwerking met behulp van de computer van multiple response vragen bieden de procedures Multiple Response en Tables in SPSSX en Tabulate in SAS de mogelijkheid om deze vragen te verwerken tot rechte tellingen en kruistabellen, waarbij als beschrijvende maatstaven percentages gegeven worden. Voor een vergelijking van de mogelijkheden welke SPSSX en SAS bieden zie Stemerding (1987).

In dit artikel kiezen wij voor de methode van coderen waarbij elke antwoordmogelijkheid van de multiple response vraag beschouwd wordt als één variabele. Combinaties van deze afzonderlijke variabelen worden samengestelde variabelen genoemd.

Indien men geïnteresseerd is in de mate van samenhang van twee variabelen, waarbij één of zelfs beide variabelen samengesteld zijn, laten deze programma's en de meeste statistiekboeken het afweten.

In Goodman & Kruskal (1979) wordt een methode gegeven hoe te handelen in het geval men associatie wenst te meten tussen twee samengestelde variabelen X en Y. Voor het meten van de samenhang tussen de samengestelde variabele X bestaande uit twee variabelen met de uitkomsten A of a en B of b enerzijds en de samengestelde variabele Y bestaande uit twee variabelen C of c resp. D of d anderzijds kan de volgende kruistabel worden opgesteld:

Y\X	AB	Ab	aB	ab
CD				
Cd				
cD				
cd				

Met behulp van deze tabel kan vervolgens de associatie gemeten worden tussen de 'rij-variabele' en de 'kolom-variabele'.

Tanaka & Tarumi (1985) beschrijven in hun artikel over samengestelde variabelen een methode om in dit soort situaties associatie te meten welke gebaseerd is op een kwantificering van de verschillende uitkomsten. Bij deze methode worden aan de uitkomsten numerieke waarden toegekend, zodanig dat de canonische correlatie tussen de betreffende variabelen maximaal is. Deze methode heeft als belangrijkste nadeel dat de uitkomsten moeilijk interpreteerbaar zijn.

In dit artikel zullen wij in par. 2 een andere maatstaf behandelen om samenhang te meten, in par. 3 wordt deze maatstaf toegepast op gegevens van Tanaka & Tarumi, vervolgens wordt in par. 4 een tweede voorbeeld behandeld en tenslotte worden de resultaten van beide methoden vergeleken.

2. De associatiemaat ζ .

De maatstaf ζ is een maat om de samenhang te meten in een tabel van twee of meer variabelen. Zij kan gebruikt worden om de associatie in de totale tabel, de associatie in een marginale tabel van minder variabelen en de

partiële associatie tussen twee variabelen te berekenen. In SPSSX en SAS kan bij het maken van een tabel ook ζ berekend worden. In deze pakketten wordt ζ de Uncertainty Coefficient genoemd. Bij een tabel van meer dan twee dimensies berekenen zowel SPSSX als SAS ζ apart voor elke subtabel.

De associatiemaat ζ is gebaseerd op het al langer bekende begrip entropie of variabiliteit, zie Theil (1967) en Khinchin (1957).

Gegeven is een variabele X met k klassen, gecodeerd 1 t/m k . De frekwenties van de uitkomsten zijn f_i en het totaal aantal waarnemingen is n . De relatieve frekwentie van klasse i is gelijk aan $p_i = f_i/n$.

De variabiliteit van X , $H(X)$, is vervolgens gedefiniëerd als:

$$H(X) = - \sum_{i=1}^k p_i \log p_i$$

waarbij $\log$ de natuurlijke logaritme is en $0 \log 0 = 0$.

H bereikt zijn minimum, 0, indien alle waarnemingen in één klasse liggen, er bestaat dus volledige overeenstemming onder alle respondenten. H bereikt zijn maximum, $\log k$, indien de waarnemingen evenredig over alle klassen gespreid zijn. H meet dus de mate van spreiding binnen een tabel.

Voor een kruistabel van twee variabelen, X en Y met resp. k_1 en k_2 klassen en relatieve frekwenties p_{ij} , is $H(X, Y)$ gedefiniëerd als:

$$H(X, Y) = - \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} p_{ij} \log p_{ij}$$

Ook nu heeft H weer een minimum van 0, dat bereikt wordt als alle waarnemingen in één cel liggen, en een maximum van $\log k_1 k_2$ (de $\log$ van het totaal aantal cellen van de kruistabel) als de waarnemingen geheel evenredig over alle cellen verspreid zijn.

Bij twee variabelen kunnen de volgende H 's onderscheiden worden:

- voor de totale tabel: $H(X, Y)$

- voor elke marginale tabel (de rij- en kolom-totalen): $H(X)$ en $H(Y)$:

$$H(Y) = - \sum_{j=1}^{k_2} p_{+j} \log p_{+j} \quad \text{waarbij } p_{+j} = \sum_{i=1}^{k_1} p_{ij}$$

- voor een subtabel van Y binnen de klasse $X = x$: de conditionele variabiliteit $H(Y \mid X = x)$:

$$H(Y \mid X = x) = - \sum_{j=1}^{k_2} \frac{p_{xj}}{p_{x+}} \log \frac{p_{xj}}{p_{x+}} \quad \text{waarbij } p_{x+} = \sum_{j=1}^{k_2} p_{xj}$$

- de partiële variabiliteit van Y, dit is het gewogen gemiddelde van de verschillende $H(Y \mid X = i)$ met als wegingsfactoren de relatieve frekwenties p_{i+} : notatie $H_X(Y)$:

$$H_X(Y) = \sum_{i=1}^{k_1} p_{i+} H(Y \mid X = i)$$

welke ook te schrijven is als: $H_X(Y) = H(X, Y) - H(X)$

Deze partiële variabiliteit van Y meet de verwachte mate van onzekerheid omtrent Y welke resteert als de uitkomst van X bekend zou zijn.

Het verschil tussen de marginale en de partiële variabiliteit van Y, $H(Y) - H_X(Y)$, geeft aan in hoeverre de variabiliteit van Y naar verwachting zal verminderen als X bekend zou zijn.

De symmetrische associatiemaat $\zeta(X, Y)$ is gedefiniëerd als:

$$\zeta(X, Y) = \frac{H(X) + H(Y) - H(X, Y)}{(H(X) + H(Y))/2}$$

Als $\zeta(X, Y) = 0$, dan zijn X en Y onderling onafhankelijk.

Als $\zeta(X, Y) = 1$, dan is er sprake van volledige samenhang.

In de overige gevallen zal $\zeta(X, Y)$ tussen 0 en 1 liggen.

Bij deze symmetrische versie van ζ wordt enkel de mate van samenhang tussen twee variabelen gemeten zonder dat er sprake is van te verklaren en verklarende variabelen.

Indien er sprake is van een te verklaren variabele Y en een verklarende variabele X (dit wordt aangegeven door middel van een dubbele punt achter de te verklaren variabele) dient men de asymmetrische versie van ζ te hanteren.

Deze asymmetrische associatiemaat $\zeta(Y: X)$ is gedefiniëerd als:

$$\zeta(Y: X) = \frac{H(Y) - H_X(Y)}{H(Y)}$$

De rol van ζ voor nominale variabelen is vergelijkbaar met die welke de verschillende correlatie-coëfficiënten spelen voor variabelen gemeten in een interval- of ratioschaal. Bij de interpretatie van de hoogte van ζ enerzijds en de correlatiecoëfficiënt anderzijds is er wel een duidelijk verschil tussen beiden. Wordt bij economisch onderzoek een correlatiecoëfficiënt van minstens 0,50 interessant, m.b.t. de hoogte van ζ blijkt uit ervaring dat een uitkomst van 0,25 al hoog genoemd mag worden.

In dit artikel beperken wij ons tot twee dimensies. Voor meer dan twee variabelen zijn de verschillende versies van H analoog gedefiniëerd, zie Lammerts van Bueren (1982). Voor de formules van de standaardfouten van ζ verwijzen wij naar Abrahamse & Geilenkirchen (1987).

3. Het voorbeeld van Tanaka & Tarumi.

In hun artikel geven beide auteurs een voorbeeld waarbij zij hun methode toepassen op vier vragen uit een vragenlijst ingevuld door 379 studenten. De vragen die in dit verband relevant zijn, luiden:

A. What were the main factors by which you chose the faculty and/or department ? Select at most two of the following categories:

1. Aptitude or interest
2. Scholastic attainments
3. Advantage in finding employment
4. As a second choice

1 = Selected 0 = Not Selected

B. Are you contented with your choice of faculty/department ?

1 = Yes 0 = No

Hierbij is A een multiple response vraag. De volledige tabel van A en B met alle voorkomende combinaties van A luidt als volgt:

A B	1	2	3	4	1 2	1 3	1 4	2 3	2 4	3 4	Tot.
1	69	4	14	7	60	87	13	21	14	12	301
0	5	3	3	15	9	10	11	6	8	8	78
Tot.	74	7	17	22	69	97	24	27	22	30	379

Tanaka & Tarumi komen na veel rekenwerk tot de conclusie dat er een zwakke correlatie bestaat tussen A en B. Omdat zij hierbij geen causale richting veronderstellen, is deze analyse van de samenhang vergelijkbaar met de symmetrische versie van ζ .

Wij kunnen op de gegevens in deze tabel een ζ -analyse toepassen.

Voor het uitvoeren van de ζ -analyses hebben wij gebruik gemaakt van het computerprogramma ZETA van Geilenkirchen (1988).

In eerste instantie hebben wij zowel de symmetrische als de asymmetrische ζ 's berekend voor de subtabellen bestaande uit telkens een van de vier antwoordmogelijkheden van A (A1 t/m A4) en B. De uitkomsten van deze analyse zijn:

	$\zeta(A_i : B)$	$\zeta(B : A_i)$	$\zeta(A_i, B)$
A1 :	0,057	0,069	0,063
A2 :	0,000	0,000	0,000
A3 :	0,005	0,007	0,006
A4 :	0,111	0,119	0,115

Al de gevonden samenhangen zijn erg zwak, enkel die tussen A4 en B is redelijk hboog. Dit leidt tot de conclusie, welke ook Tanaka & Tarumi trekken, dat er een redelijk sterk verband bestaat tussen 'tweede keuze' en 'ontevreden met de keuze'.

Bovenstaande analyses geven echter de marginale ζ 's tussen B en A1 t/m A4 afzonderlijk. De variabele A is zodoende niet als één samengestelde variabele behandeld. Voor de toepassing van een ζ -analyse levert het bestaan van multiple response vragen echter geen enkele beperking. Waar in de gegeven formules gesproken wordt over 'de variabelen X en Y', kunnen overal volgens de door Goodman & Kruskal voorgestelde methode X en/of Y vervangen worden door samengestelde variabelen X en Y, elk bestaande uit

groepen variabelen. De volgorde en de nummering van deze cellen is hierbij irrelevant omdat H en ζ onafhankelijk zijn van de gebruikte klasse-indeling. In een formule kan dit opstellen van de samengestelde variabele A eenvoudig gebeuren door gebruik te maken van de binaire codering:

$$A = A_1 + 2 * A_2 + 4 * A_3 + 8 * A_4$$

Van de 16 mogelijke klassen van deze samengestelde vraag A zijn er bij de betreffende gegevens maar 10 cellen gevuld, omdat de studenten maximaal twee antwoorden mochten opgeven.

De ζ -analyse tussen A en B resulteerde in de volgende ζ 's met tussen haakjes vermeld de daarbij behorende standaardfouten:

$$\begin{aligned}\zeta(A: B) &: 0,039 & (0,010) \\ \zeta(B: A) &: 0,157 & (0,039) \\ \zeta(A, B) &: 0,063 & (0,016)\end{aligned}$$

Tanaka & Tarumi komen tot een gekwadrateerde canonieke correlatiecoëfficiënt van 0,1666 met een standaardfout van 0,0583.

Bij onze analyse laat $\zeta(B: A)$ de hoogste waarde zien. Deze geeft aan dat door de kennis van A de onzekerheid omtrent de uitkomst van B met ongeveer 16% vermindert.

Daarentegen is $\zeta(A: B)$ veel lager: de onzekerheid omtrent A was hoog: $H(A) = 2,040$ en kennis van B zal hier niet veel aan veranderen.

Tenslotte ligt $\zeta(A, B)$ tussen beiden in.

4. Een tweede voorbeeld.

Een tweede voorbeeld is gebaseerd op gegevens uit een onderzoek op medisch terrein waarbij 260 patienten een enquête met meer dan 200 vragen moesten beantwoorden. In een aantal vragen moesten zij aangeven in welke mate zij last hadden van pijn in verschillende lichaamsdelen; hierbij moest gekozen worden uit vier antwoorden:

1 = geen, 2 = een beetje, 3 = tamelijk veel en 4 = zeer veel.

Voor dit artikel hebben wij 2 blokken van elk 3 vragen gekozen om de mogelijkheden van ζ -analyse bij samengestelde variabelen te illustreren.

Het eerste blok betreft de mate van last in nek, schouders en armen/handen. Het tweede blok omvat de onderrug, heup/liës/bovenbeen en knieën/onderbenen/voeten.

Wij hebben elk blok als een samengestelde variabele behandeld. Hierbij is gebruik gemaakt van de wijze van hercoderen zoals paragraaf 3 beschreven. Omdat bij elke vraag vier antwoorden mogelijk zijn, worden de vermenigvuldigingsfactoren 1, 4 en 16 gebruikt. Dit resulteert in twee variabelen met elk 64 klassen.¹⁾

De associatie tussen deze twee blokken van elk drie vragen is nu eenvoudig te berekenen en geeft een $\zeta = 0,426$ met een standaardfout van 0,018 te zien, hetgeen duidt op een zeer sterke samenhang. Omdat de samenhang slechts tussen twee (samengestelde) variabelen onderzocht is, zijn de marginale en partiële ζ 's niet van toepassing.

De constatering van deze zeer sterke samenhang toont duidelijk aan dat het de moeite waard is om verder onderzoek te doen naar de relatie tussen deze variabelen, dit om een meer gedetailleerd inzicht te verkrijgen in de mate en de richting van de samenhang tussen de verschillende categorieën. Dit onderzoek valt echter buiten het kader van dit artikel.

5. Conclusie.

Met de summier weergegeven theorie en de twee voorbeelden hopen wij duidelijk te hebben gemaakt dat de associatiemaat ζ goed bruikbaar is voor het meten en analyseren van samenhang tussen nominale variabelen. Multiple response vragen vormen voor de toepassing van ζ geen probleem. De maatstaf ζ is inzichtelijker en eenvoudiger dan het gebruik van canonieke correlatie, welke Tanaka & Tarumi hebben toegepast.

Technisch gezien kunnen multiple response vragen echter nog niet rechtstreeks verwerkt worden met behulp van de op dit moment beschikbare software. De door ons voorgestelde eenvoudige hercodering van multiple response vragen tot een samengestelde variabele lost deze belemmering elegant op. Deze hercodering biedt daarbij ook het voordeel dat de analyse

¹⁾ De volledige tabel met alle frekwenties is vanwege de omvang niet opgenomen; desgewenst kan zij bij de auteurs opgevraagd worden.

nu niet uitsluitend met het door ons gebruikte computerprogramma kan gebeuren maar eveneens met de procedures Freq in SAS en Crosstabs in SPSSX, zij het dat deze programma's enkel de marginale en conditionele associaties en niet de totale en partiële associaties berekenen.

Literatuur.

- Abrahamse, A.P.J. & J.Th. Geilenkirchen (1987), The asymptotic distribution and small-sample behaviour of the covariability coefficient Zeta, Report 8723/A, Econometric Institute, Erasmus University, Rotterdam.
- Geilenkirchen, J.Th. (1988), The computerprogram ZETA, (unpublished).
- Goodman, L.A. & W.H. Kruskal (1979), Measures of association for cross classifications, Springer Verlag, New York.
- Khinchin, A.I. (1957), Mathematical foundations of information theory, Dover Publications, Inc., New York.
- Kinnear, T.C. & J.R. Taylor (1987), Marketing Research, an applied approach, 3th ed., MacGrawHill Book Co., Singapore.
- Lammerts van Bueren, W.M. (1982), Measuring association in nominal data, Ph.D.Thesis, Rotterdam.
- Nie, N.H. (1983), SPSSX Users guide, McGraw Hill Book Co, New York.
- Sas Institute Inc (1985), Sas/Stat Guide for personal computers, Version 6 Edition, Cary NC.
- Stemerding, G.J. (1987), Complexe tabellen een vergelijking van SPSS-X Tables en SAS Tabulate, in: SSS87, Symposium Statistische Software, Groningen.
- Tanaka, Y. & T. Tarumi (1985), Statistical methods for multiple response items and their applications, in: Statistical Theory and Data Analysis, Proceedings of the Pacific Area Statistical Conference, Amsterdam, North-Holland.
- Theil, H. (1967), Economics and information theory, North-Holland Publishing Company, Amsterdam.
- Tull, D.S. & D.I. Hawkins (1984), Marketing research measurements and methods, 3th ed., MacMillan, New York.
- Weiers, R.M. (1984), Marketing research, Prentice Hall, Englewood Cliffs/New Jersey.