KM 2(1981)2
pag 150-155

# Multivariate analysemethoden voor discrete variabelen, een vervolg: Kijken naar grafieken en standaardisatie.

P.I.M. Schmitz [1]

## Samenvatting.

Een vierdimensionale tabel met CBS-gegevens omtrent inkomen in relatie tot leeftijd, scholing en bedrijfstak werd door Israëls e.a. (1981) bij wijze van experiment met behulp van zeven verschillende technieken geanalyseerd. In deze notitie wordt een aanvullende beschouwing gegeven over de eerste van deze technieken: "Kijken naar de gegevens, percenteren, permuteren en standaardiseren". Het betreft een eenvoudige grafische presentatie van de gegevens en enkele opmerkingen over standaardisatie.

## 1. Inleiding.

Een interessant onderzoek betreft het analyseren door Israëls e.a. (1981) van een vierdimensionale tabel met gegevens omtrent inkomen, leeftijd, scholing en bedrijfstak (zie elders in dit nummer van KM). Zeven verschillende technieken worden gebruikt, waarvan de eerste samengevat neerkomt op "kijken, percenteren, permuteren en standaardiseren". Uiteraard kan alleen het standaardiseren een techniek worden genoemd in de modelmatige betekenis van het woord. Ik meen dat de gepresenteerde standaardisatietechniek enkele nadere opmerkingen verdient, zeker daar het in feite de eenvoudigste van de zeven methoden is en het daarom niet onaantrekkelijk is deze te gebruiken. De voorwaarden voor het mogen toepassen van deze techniek worden in het kort geschetst.
Een tweetal beperkingen bij de hierna volgendebeschouwingen moeten echter worden genoemd. Er is alleen een analyse verricht van het zogenoemde "onthulde" bestand van Israëls e.a. (zie de tabel in de bijlage van Israëls e.a.). Verder zal alleen aandacht worden besteed aan indirecte standaardisatie. Voor directe standaardisatie gelden analoge voorwaarden,

---

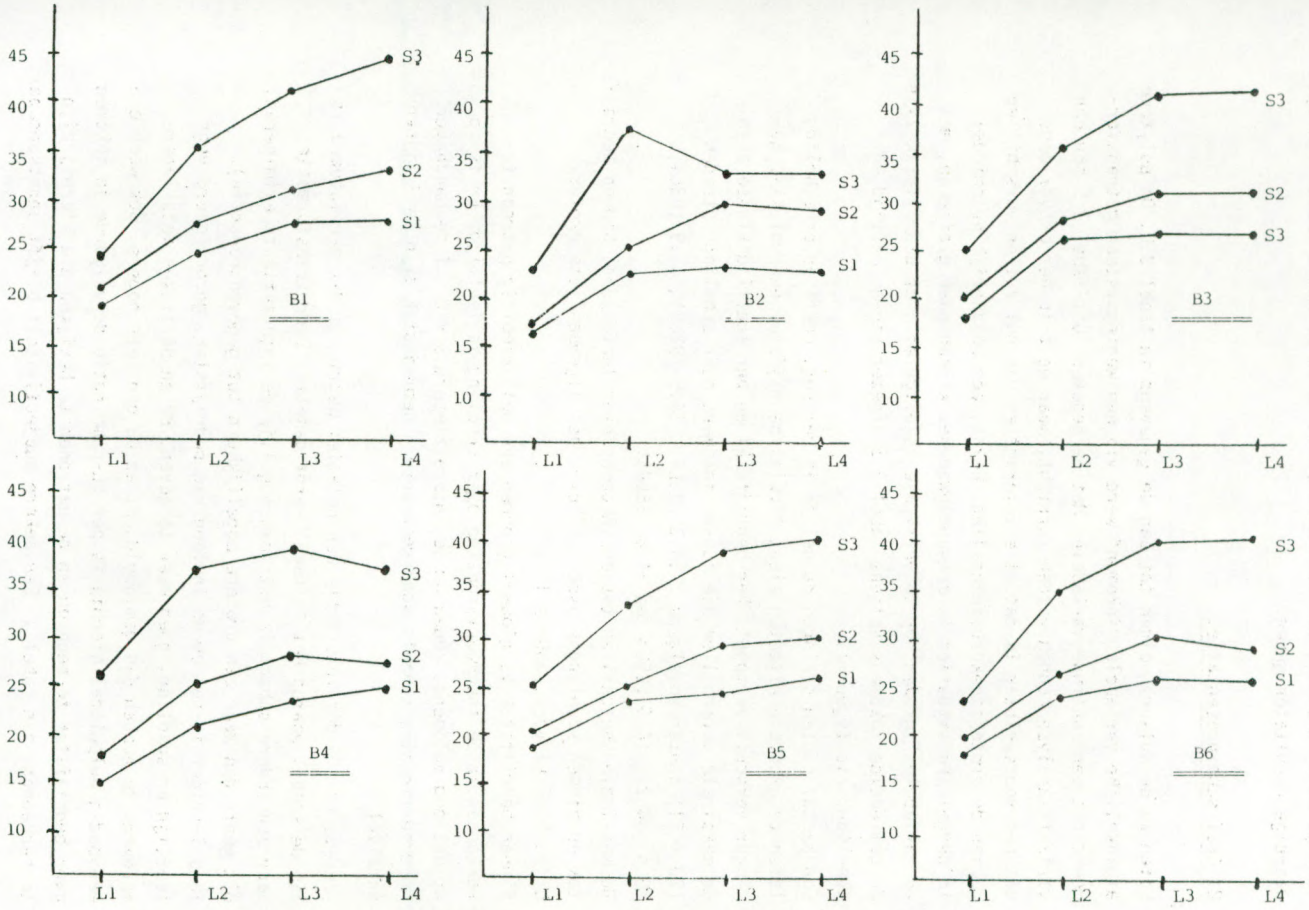
[1] Instituut voor Biostatistica,
Erasmus Universiteit Rotterdam

terwijl beide methoden van standaardiseren in de praktijk meestal gelijkwaardige resultaten geven.

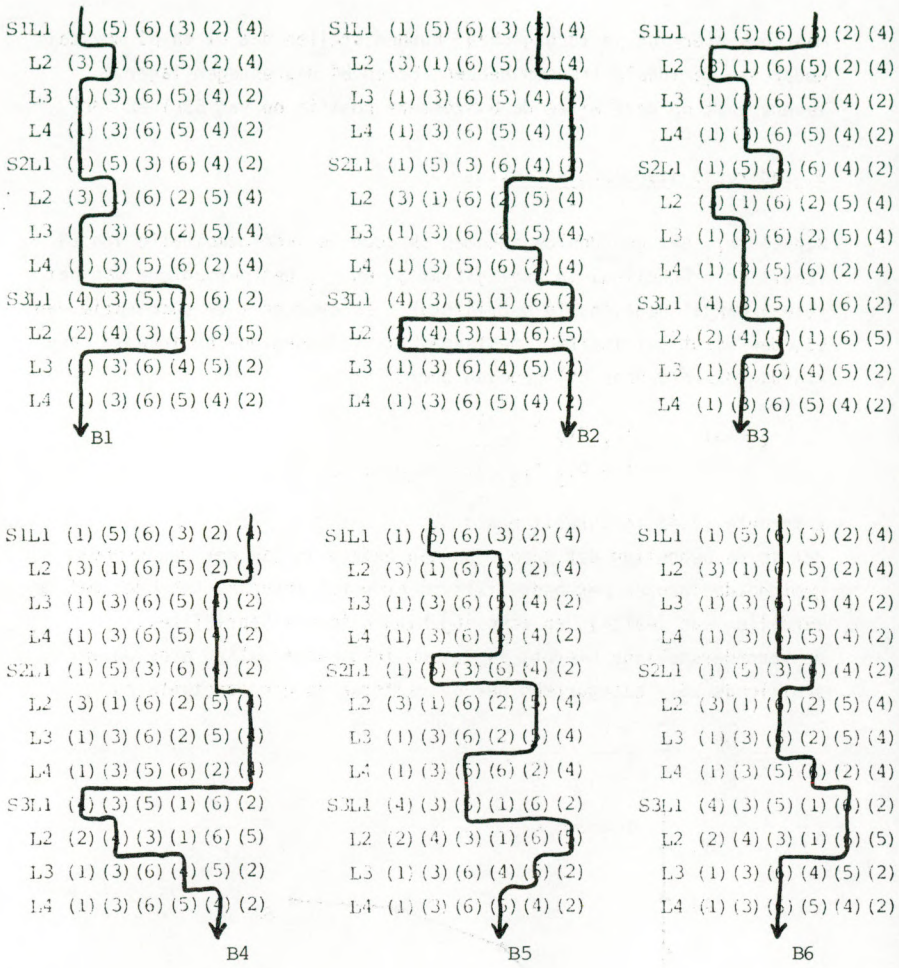
2. Grafische presentatie.

Laten we de auteurs volgen tot aan de gegevens in tabel 3.5. De polytome afhankelijke variabele "inkomen" werd via een wegingsysteem getransformeerd tot een continue variabele. Met de gegevens uit tabel 3.5 zou een variantieanalyse kunnen worden verricht, maar we zijn het erover eens dat het voorspelbaar is dat alle interakties van nul zullen verschillen gezien de grote steekproefaantallen. Welnu, een belangrijk hulpmiddel in dergelijke situaties is eenvoudigweg het kijken naar grafieken, met de variabele "inkomen" op de verticale as uitgezet. Per bedrijfstak zijn de gemiddelde inkomens volgens tabel 3.5 afgezet tegen leeftijd en scholing: zie figuur 1.
Een aantal punten springen duidelijk in het oog. Leeftijd en scholing laten de grootste effecten zien, terwijl bedrijfstak een relatief geringer verschil vertoont. Per bedrijfstak en per leeftijdsklasse zijn de maksimale verschillen die tussen inkomens over scholing optreden (S3 - S1) achtereenvolgens : 16.3 (B1L4), 14.9 (B2L2), 14.9 (B3L4), 16.6 (B4L2), 14.6 (B5L3) en 14.6 (B6L4).
De maksimale verschillen tussen inkomens over leeftijden, binnen bedrijfstak en binnen scholingsklasse zijn over het algemeen iets groter, bijv. L4 - L1 voor S3B6 is 17.1.
Binnen bedrijfstak is globaal genomen een gelijksoortig patroon te herkennen in de inkomensverdeling over scholing en leeftijd. Er is zo op het oog voldoende reden om de interaktieterm S x L x I te handhaven op deze lineaire schaal: voor geen enkele bedrijfstak zijn de 3 lijnen parallel.

(Opgemerkt zij dat interaktie een relatief begrip is dat gerelateerd is aan de schaal waarop het "effect" wordt gemeten. Door transformatie naar een andere schaal is het soms mogelijk de interaktie te elimineren. Dit geeft dan weer een grotere mogelijkheid tot gegevensreductie).
Iets lastiger is het om de invloed van bedrijfstak, gecorrigeerd voor leeftijd en scholing, goed weer te geven. B2 en B4 lijken iets lagere inkomens te hebben dan gemiddeld, B1 en B3 een iets hogere. Ook weer een eenvoudig hulpmiddel hierbij is per SL-combinatie de volgorde in inkomen over bedrijfstak te bepalen en de patronen te bekijken die hierbij zijn te herkennen: zie tabel 1. Een zekere subjectiviteit blijft onontkoombaar,




Figuur 1. Gemiddelde inkomens (y-as) per Leeftijd (L), Scholing (S) en Bedrijfstak (B).




Tabel 1. Bedrijfstak en Inkomen.

Per categorie-combinatie Scholing (S) - Leeftijd (L) wordt de volgorde van inkomen (afnemend) van de 6 Bedrijfstakken (B) weergegeven. De 6 blokken zijn volkomen identiek, maar per blok wordt terwille van de overzichtelijkheid één bedrijfstak eruit gelicht.

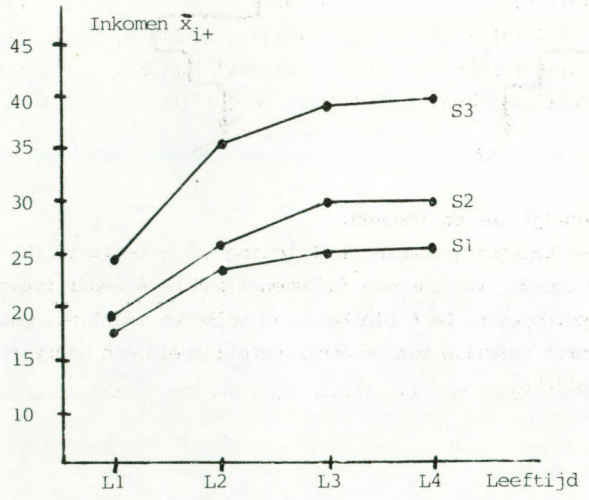
maar ook hier zou je toch globaal kunnen stellen dat B1 en B3 meestal hoger dan gemiddeld inkomen hebben, B2 en B4 daarentegen lager.
Tevens valt op deze wijze de bijzondere positie op van S3L1 en S3L2.

## 3. Indirecte standaardisatie.

Wanneer $x_{ij}$ het gemiddelde inkomen is voor de i-de combinatie van LS (Leeftijd x Scholing) in bedrijfstak j, en $q_{ij}$ het percentage van het totaal aantal gegevens in bedrijfstak j en over de i-de combinatie van LS, dan wordt het indirect gestandaardiseerde gemiddelde inkomen (gestandaardiseerd voor LS) gegeven door:

$$\bar{x}_j(is) = \frac{\left(\sum_i q_{ij}\, x_{ij}\right)\, \bar{x}_+}{\left(\sum_i q_{ij}\, \bar{x}_{i+}\right)}$$

( formule (3.4) in Israëls e.a.)
Het is de bedoeling dat deze maat een betere is dan het rekenkundig gemiddelde inkomen per bedrijfstak : er wordt rekening gehouden met de verdeling van leeftijd en scholing binnen iedere bedrijfstak.
De veronderstelling hierbij is dat de inkomensverdeling over de verschillende LS - categorieën per bedrijfstak op een constante na




Figuur 2. Gemiddelde inkomens $\bar{x}_{i+}$ per Scholing- en Leeftijdklasse.

gelijk is aan die van een standaardpopulatie, waarvoor meestal de sompopulatie (hier : som over bedrijfstakken) wordt genomen. De inkomensverdeling in deze sompopulatie is weergegeven in figuur 2.
Indirecte standaardisatie is alleen geoorloofd indien de inkomensverdeling over LS per bedrijfstak op een constante na dezelfde is als die in fig.2. Alle corresponderende LS-lijnen dienen evenwijdig te zijn (Gail, 1978). Zodra dit niet het geval is kan het gestandaardiseerde gemiddelde een slechtere maat betekenen dan het "ruwe" rekenkundige gemiddelde. Deze veronderstelling kan worden getoetst, maar ook hier zal dit ongetwijfeld tot een significant resultaat leiden i.v.m. de grote aantallen. Figuur 1 vormt weer een leidraad. Vooral bij de tweede bedrijfstak dient men op zijn hoede te zijn, verder lijkt de veronderstelling redelijk.
Bij een slotconclusie op basis van het indirect gestandaardiseerde gemiddelde kan tabel 3.6 (Israëls e.a.) dus goed worden gehanteerd, met uitzondering van B2. Merk op dat deze ook het grootste verschil met het "ruwe" rekenkundig gemiddelde vertoont.
De conclusie dat de volgorde in afnemend gemiddelde inkomen gegeven kan worden door ( B1,B3) , (B5,B6) , (B2,B4) blijft overigens ook op basis van de zes gestandaardiseerde waarden gehandhaafd.

## 4. Discussie.

De weging van de verschillende inkomensgroepen m.b.v. scores, zodat een continue afhankelijke variabele ontstaat, lijkt een goede procedure. Een voordeel van deze transformatie is de eenvoud en overzichtelijkheid waardoor een grafische benadering goed mogelijk is. Standaardisatie is alleen geoorloofd na grafische inspectie van de gegevens. Een acceptabele nadere kwantificering lijkt me residuenonderzoek ( zie Verbeek, dit nummer van KM). Een formele variantieanalyse leidt voor wat betreft het toetsen tot voorspelbare resultaten (alle interakties significant).
Een vergelijkbare aanpak is het dichotomiseren van de polytome variabele inkomen, bijv. I1 + I2 versus I3 + I4 + I5. Deze binaire afhankelijke variabele kan nu worden bestudeerd met behulp van een logitmodel waarbij eenvoudige grafische inspectie (logits versus L, S en B), standaardisatie en residuenanalyse eveneens mogelijk is.

Literatuur

1. A.Z. Israëls, J.G.Bethlehem,J.van Driel,M.E.Jansen,J.Pannekoek,S.J.M. de Ree en D.Sikkel: "Multivariate analysemethoden voor discrete variabelen", KM 2 (1981)
2. M.Gail : "The analysis of heterogeneity for indirect standardized mortality ratios",J.R.Statist.Soc.A.,141,part 2 (1978),p.224-234.