Multivariate Analyse voor de Sociale Wetenschappen: Wat Vooraf Komt

Ronan Van Rossem

Universiteit Gent

ACADEMIA
PRESS

© Uitgeverij Academia Press
Ampla House
Coupure Rechts 88
9000 Gent
België

www.academiapress.be

Uitgeverij Academia Press maakt deel uit van Lannoo Uitgeverij,
de boeken- en multimediadivisie van Uitgeverij Lannoo nv.

Van Rossem, Ronan
Multivariate Analyse voor de Sociale Wetenschappen: Wat Vooraf Komt
Gent, Academia Press, 2019, 172 p.

ISBN 9789401460231
D/2019/45/373
NUR 916

Eerste editie, 2007
Tweede editie, 2019

Inhoud

VOORWOORD

Multivariate analysetechnieken vormen het sluitstuk van de basis statistiekopleiding voor studenten in de sociale wetenschappen. Maar vooraleer men een dergelijke studie van multivariate technieken kan aanvatten, dient men toch enkele basis statistische technieken onder de knie te hebben. De multivariate analyse bouwt verder op technieken van bivariate analyse en statistische inferentie. In dit volume wordt een deel van de noodzakelijke voorkennis voor een studie van multivariate analysetechnieken herhaald.

Hierbij wordt er vanuit gegaan dat de lezer voldoende kennis heeft over de basisprincipes van de statistiek en de kansrekening, als ook van de matrixalgebra. Dit volume behandelt alleen enkele onderwerpen die rechtstreeks aansluiten bij de multivariate analyse, namelijk de principes van inferentiële statistiek, bivariate correlatie, bivariate regressie, en partiële correlatie. Het is ook niet uitsluitend een herhaling van eerder geziene stof maar af en toe wordt er een kleine uitbreiding toegevoegd die relevant is voor de studie van multivariate analysetechnieken (bv. kapitalisatie op toeval, maximum likelihoodschatting).

Sociale wetenschappers gebruiken statistische analyses overwegend om relaties tussen variabelen bloot te leggen gebruikmakend van steekproeven, en daaruit gevolgen trekken voor de populatie waaruit de steekproef getrokken was. Inferentiële (of inductieve) statistiek vormt de basis voor dergelijke veralgemeningen. Het toetsen van hypothesen over de effecten van variabelen is een centraal onderdeel van alle multivariate analysetechnieken. In het hoofdstuk over inferentiële statistiek worden de principes die aan de basis liggen van het hypothesetoetsen kort overlopen.

De correlatie, d.w.z., de relatie tussen twee variabelen, vormt de basis voor vele multivariate technieken, waaronder multipele regressie, factor analyse en padanalyse. In het tweede hoofdstuk komt de correlatieanalyse aan bod. Hierbij wordt vertrokken van de standaard productmomentcorrelatie tussen twee variabelen. Vervolgens komen aan bod hoe matrixalgebra kan gebruikt worden om de correlatie tussen meerdere variabelen te berekenen, alternatieve vormen van correlatie voor niet continue variabelen, en het probleem van de kapitalisatie op toeval bij meerdere hypothesetoetsen. Dit probleem wordt hier geïllustreerd aan de hand van correlatiematrices.

Van bivariate correlatie naar bivariate regressie is maar een kleine stap. In dit hoofdstuk wordt stap voor stap bivariate regressie

uitgelegd, waarbij ook aandacht besteed wordt aan verschillende schattingsmethoden. Twee methoden komen aan bod: de methode van de kleinste kwadraten en maximum likelihoodschatting. Verder wordt ook het toetsen van verschillen tussen regressiecoëfficiënten behandeld.

Het laatste hoofdstuk behandelt de partiële correlatie. Dit is in feite de meest eenvoudige multivariate techniek. Hierbij onderzoekt men wat er gebeurt men de correlatie tussen twee variabelen wanneer men controleert voor een derde variabele. De basisbegrippen die hierbij aan bod komen zijn dezelfde als deze die bij meer gesofisticeerde methoden gehanteerd worden.

Een achterliggend idee in dit volume is dat het niet voldoende is dat men weet op welke knopjes men moet drukken in statistische softwarepakketten, en dat men weet waar te kijken in de output geproduceerd door dergelijke pakketten, maar dat men ook dient te weten wat er zich onder de motorkap afspeelt. Er wordt in dit volume dat ook vrij veel aandacht besteed aan de wiskunde die aan de basis van deze technieken ligt, zonder dat we hierbij wensen te vervallen in een cursus wiskundige statistiek. Kennis en begrip van de onderliggende mechanismen is essentieel wanneer men zijn of haar beheersing van multivariate analysetechnieken verder wenst te voeren dan de standaard kookboekanalyses. Voor sociale wetenschappers is statistiek in de eerste plaats een toegepaste analysemethode. Daarom wordt er ook vrij veel aandacht besteed aan toepassingen van deze methoden, waarbij ingegaan wordt op hoe bepaalde analyses met behulp van SPSS (versie 12) kunnen uitgevoerd worden en hoe men de geproduceerde resultaten kan interpreteren.

Hoofdstuk 1: Principes van inferentiële statistiek

1.1. Inleiding

Statistische inferentie ligt aan de kern van de multivariate analyse. Een intuïtief begrijpen van de principes van statistische inferentie is dan ook noodzakelijk voor het adequaat toepassen en interpreteren van multivariate analysetechnieken. In dit hoofdstuk worden nogmaals in het kort de principes van inferentiële statistiek doorgenomen, van steekproevenverdelingen en de centrale limiet stelling, over type I en type II fouten tot en met de logica van hypothesetoetsen en de kapitalisatie op toeval.

De moderne sociologie maakt veelvuldig gebruik van statistische analysetechnieken om de relaties tussen variabelen te onderzoeken. Het zijn deze relaties tussen variabelen die centraal staan in het sociologisch onderzoek. Men is minder geïnteresseerd in de gemiddelde waarde van een variabele dan in hoe variabelen covariëren, of kennis van de waarde op één variabele ook iets zegt over de waarde op de andere variabele.

Neem nu bijvoorbeeld het recente artikel van Wejnert (2005) waarin de auteur de verspreiding van democratie sinds 1800 onderzoekt. Zij vraagt zich hierbij niet zozeer af wat het niveau van democratisering is in verschillende landen, maar wel hoe democratisering samenhangt met andere variabelen zoals ontwikkeling- en diffusie-indicatoren. In een ander recent artikel onderzoekt Sunmola (2005) bij vrachtwagenchauffeurs in Nigeria, een belangrijke vector voor de verspreiding van HIV/AIDS, hoe reproductief gezondheidsgedrag—seksueel risicogedrag, barrières voor condoomgebruik en percepties in verband met SOA-s en AIDS—covarieert met de sociodemografische kenmerken van de chauffeurs. De absolute prevalentie van dit gedrag en opvattingen zijn ook hier slechts van secundair belang, terwijl de relatie van dit reproductieve gezondheidsgedrag met de sociodemografische kenmerken van de chauffeurs centraal staat.

De aard van de relaties die men onderzoekt kunnen zowel associatief als causaal zijn. Bij associatieve relaties wordt gewoon gekeken naar de samenhang van twee variabelen maar worden er geen causaliteitsassumpties gemaakt. Bij causale relaties echter, staan—zoals de naam het al suggereert—de causaliteitsassumpties in de relatie centraal. Men gaat er van uit dat de waarde op één variabele aanleiding geeft tot een bepaalde waarde op een andere, of

dat de verandering in één variabele leidt tot een corresponderende verandering in een andere. Deze causaliteit wordt echter nooit bepaald door de statistische methoden die men gebruikt, maar is steeds deel van de theorie die men gebruikt om een bepaalde relatie te verklaren. Er zijn methodologisch wel enkele criteria waarop men bepaalde causaliteitsassumpties kan uitsluiten, zoals bv. dat een oorzaak temporeel steeds voorafgaand of gelijktijdig met het gevolg moet plaatsvinden, en dat veranderlijke kenmerken nooit onveranderlijke kenmerken kunnen veroorzaken. Bv. sekse (onveranderlijk kenmerk, temporeel prior) kan wel schoolresultaten (veranderlijk kenmerk, temporeel posterior) beïnvloeden, maar omgekeerd kan het niet.

Als sociologen hebben we niet de mogelijkheid onderzoek te voeren op een ganse populatie (in de statistische zin), maar moeten we ons steeds behelpen met steekproeven waarop we onze statistische analyses uitvoeren. Natuurlijk wil men zijn conclusies niet beperken tot deze ene steekproef, maar zou men het liefst de resultaten en de conclusies kunnen veralgemenen tot de populatie waaruit de steekproef getrokken is. Het is bij deze veralgemening dat de inferentiële statistiek haar intrede doet.

1.2. Steekproevenverdelingen

In de sociale wetenschappen gebruikt men doorgaans steekproeven om conclusies te trekken over de populatie. Gewoonlijk is het niet doenbaar om de ganse populatie bij het onderzoek te betrekken en moet men zich beperken tot een steekproef. Een probleem hierbij is dat naargelang de samenstelling van de steekproef de resultaten zullen verschillen. Zelfs indien de steekproef aselect getrokken werd uit de populatie en representatief is voor deze populatie—twee condities die in realiteit zelden vervuld zijn—zullen er nog steeds willekeurige fluctuaties optreden die maken dat twee steekproeven getrokken uit dezelfde populatie verschillende resultaten geven. Dit is wat men de steekproeffout (of 'sampling error') noemt. Als je bijvoorbeeld een steekproef van 1000 personen trekt om het stemgedrag bij de verkiezingen te voorspellen door middel van een exit poll, dan zal gewoon ten gevolge van de steekproeffout de resultaten bekomen door deze steekproef niet volledig overeenstemmen met de verkiezingsuitslag—zelfs niet indien de respondenten allen de waarheid vertellen. Elke steekproef, van welke omvang ook, getrokken uit een zelfde populatie zal steeds enigszins andere schattingen van de populatieparameters opleveren. Populatieparameters zijn o.a. het populatiegemiddelde en de populatievariantie, maar ook populatie regressiecoëfficiënten, correlatiecoëfficiënten.

Als men nu een oneindig aantal steekproeven trekt uit eenzelfde populatie, allen met dezelfde steekproefomvang, en men plot de waarden voor de populatieparameter die men op basis van elk van deze steekproeven geschat heeft, dan krijgt men de steekproevenverdeling van deze steekproef statistiek. De steekproevenverdeling is dus niets anders dan de theoretische verdeling van een geschatte parameter $\hat{\theta}$—een steekproef statistiek—geschat op basis van een steekproef van een bepaalde omvang uit een populatie met parameter θ. Deze verdeling blijft theoretisch omdat het praktisch niet mogelijk is om een oneindig aantal steekproeven te trekken. Kmenta definieert een steekproevenverdeling ('sampling distribution') als: "*a sampling distribution is a probability distribution of an estimator or of a test statistic*" (1971, p. 9, cursief in origineel).

Neem nu als voorbeeld dat je een populatie hebt die normaal verdeeld is met een gemiddelde μ van 100 en een standaardafwijking σ van 15, dus $N(100,15)$, bv. het IQ van een populatie, en je trekt daar steeds, met teruglegging, steekproeven uit met een omvang van 10 ($N = 10$), en je berekent de gemiddelden voor elk van deze steekproeven. Het steekproefgemiddelde $\bar{X}$ is hierbij de statistiek en het populatiegemiddelde μ de parameter die men wenst te schatten.

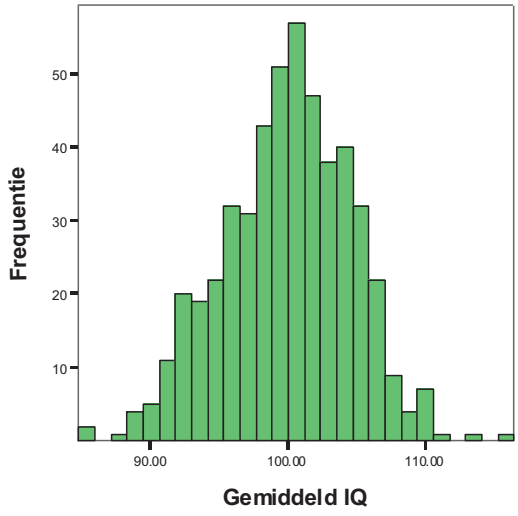
Tabel 1-1: Vijf aselecte steekproeven met omvang N = *10 getrokken uit een populatie* ~ N*(100,15)*

#	*Observatie*										$\bar{X}$
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	
1	129.1	108.3	73.8	91.7	124.9	89.5	98.4	94.7	122.3	89.5	102.2
2	94.0	116.8	74.7	118.5	98.1	90.1	107.5	117.7	89.3	101.3	100.8
3	103.5	77.8	86.6	73.6	109.3	102.6	109.4	104.4	79.4	118.9	96.5
4	82.4	98.6	107.1	111.5	91.0	69.5	85.7	97.0	111.3	86.6	94.1
5	90.8	103.0	69.9	94.5	107.2	86.3	82.5	73.3	79.7	106.1	89.3

Tabel 1-1 toont vijf dergelijke steekproeven getrokken uit een populatie ~ $N(100,15)$. Het steekproefgemiddelde $\bar{X}$ voor de eerste steekproef is 102.2, voor de tweede 100.8, voor de derde 96.5, voor de vierde 94.1, en voor de vijfde 89.3. Zoals men kan merken zit er heel wat spreiding op de verschillende steekproevengemiddelden. Bij deze vijf steekproeven varieert het van een hoogste waarde van 102.2 tot een laagste waarde van 89.3. Nochtans zijn deze vijf steekproeven allen getrokken uit eenzelfde steekproef, en zijn elk van deze statistieken (steekproefgemiddelden) een schatting van dezelfde populatieparameter $\mu = 100$.

Figuur 1-1 toont de verdeling van de statistieken, de steekproefgemiddelden, na vijfhonderd steekproeven van grootte 10 uit deze populatie getrokken te hebben. De steekproefgemiddelden variëren

van een minimum van 84.78 tot een maximum van 116.36 met een gemiddelde waarde voor de statistiek van 99.91. Deze laatste waarde is al vrij dicht bij de waarde van de populatieparameter maar is nog steeds niet exact gelijk. Geen van de steekproeven heeft een gemiddelde van exact 100. Zelfs indien men afrondt naar gehele getallen hebben maar 45 of 9% van de steekproeven een gemiddelde van 100, terwijl 144 (28.8%) steekproeven een gemiddelde hebben dat 5 of meer punten afwijkt van het populatiegemiddelde. Deze geobserveerde steekproevenverdeling heeft dus een gemiddelde van 99.91 en een standaardafwijking van 4.69, een standaardafwijking die substantieel kleiner is dan de populatiestandaardafwijking, terwijl het gemiddelde het populatiegemiddelde benadert.



Figuur 1-1: Geobserveerde steekproevenverdeling voor 500 steekproeven met N = 10 uit een populatie met verdeling N(100,15)

Dergelijke steekproevenverdelingen kunnen voor alle statistieken bepaald worden, zij het gemiddelden, regressiecoëfficiënten, associatiematen, en ook voor vergelijkingen van dergelijke statistieken. Het is op de kennis over de eigenschappen van de steekproevenverdelingen van de verschillende statistieken dat inferentiële statistiek gebaseerd is.

1.2.1. De centrale limiet stelling

Doordat men in de regel zelf geen steekproevenverdelingen kan samenstellen daar men slechts één steekproef ter beschikking heeft,