Van optimalisatie naar begrip en weer terug

Marieke E. Timmerman

Van optimalisatie naar begrip en weer terug

Rede uitgesproken bij de aanvaarding van het ambt van hoogleraar in de Statistische technieken voor analyse van multivariate gegevens uit gedragswetenschappelijk onderzoek aan de Rijksuniversiteit Groningen op dinsdag 21 juni 2016 door Marieke E. Timmerman

Leden van het College van Bestuur, zeer gewaardeerde aanwezigen,

Het woord 'Optimaliseren' is in de Dikke Van Dale omschreven als 'optimaal maken, in de meest gunstige omstandigheden of tot de gunstigste oplossing brengen'. Dat klinkt goed, wie wil dat nu niet? Ik wel, ik ben dol op optimaliseren. Wie mij een beetje kent, kent mijn voorliefde voor doelgerichtheid. Mijn favoriete aanpak om een klus te klaren bestaat uit twee stappen, en is daarmee een soort twee-trapsraket. In Stap 1 wordt het doel bepaald. Dat gaat weloverwogen, op basis van een gedegen voorbereiding en zo nodig na een goede discussie met alle betrokkenen. Is het doel eenmaal vastgesteld, dan gaan we over naar Stap 2. Die bestaat dan uit: Hup, aan de slag om het doel te bereiken. Terug naar Stap 1, oftewel de discussie over de doelstelling heropenen, dát mag van mij alleen als daar heel goede redenen voor zijn. Deze aanpak kun je zien als optimaliseren in de algemene betekenis van het woord.

Op mijn onderzoeksterrein, de statistische analyse van verzamelde gegevens, speelt 'optimaliseren' een grote rol. Dat is dan in een heel specifieke betekenis, namelijk die van de 'wiskundige optimalisatie'. Voor nu is het genoeg om te weten dat iedere statistische analyse van een bepaalde gegevensset gebruik maakt van een wiskundige optimalisatie. Het interessante is dat je voor dezelfde gegevensset verschillende optimalisaties kunt uitvoeren, met soms heel verschillende uitkomsten, en dus interpretaties. De vraag is dan natuurlijk: Welke moet ik kiezen? Welke optimalisatie moet ik nu serieus nemen? Ik wil deze materie graag met u bespreken, en laten zien dat uiteindelijk het woord 'Begrip' de centrale rol speelt. Daarbij wil ik een lans breken voor het slim inzetten van zogenoemde 'exploratieve analysemethoden', omdat juist die methoden heel belangrijk zijn voor het verwerven van nieuwe inzichten. Ik zal aangeven dat een verstandig gebruik van exploratieve analysemethoden betekent dat achtergrondkennis een expliciete rol speelt bij het modelleren, en laten zien hoe dit in praktijk kan. Ik hoop van harte dat exploratieve analyse uit de schaduw – voor sommigen uit het verdomhoekje – komt, en het de plek krijgt die het verdient.

Statistisch modelleren

Mijn werk gaat dus over het statistisch modelleren van empirische gegevens. Die gegevens zijn meestal uit psychologisch onderzoek. In goed onderzoek wordt het modelleren pas ingezet als de onderzoeker een duidelijk doel voor ogen heeft. Er zijn vele soorten psychologisch onderzoek, en de bijbehorende doelen zijn dan ook heel divers van aard. Maar ze hebben allemaal gemeen dat het uiteindelijk draait om inzicht te verwerven, om ervan te leren.

Om nu de rol van statistisch modelleren in psychologisch onderzoek te illustreren, geef ik u een voorbeeld. Psychologische tests worden tegenwoordig op veel plaatsen gebruikt, om eigenschappen van mensen te meten. Het bekendste voorbeeld is misschien wel de intelligentietest. Op basis van de antwoorden op een serie opgaven probeert men zicht te krijgen op het intelligentieniveau van een persoon. Er zijn ook instrumenten die bedoeld zijn om van een persoon aspecten van de persoonlijkheid of psychopathologie in kaart te brengen. Een voorbeeld is de zogenaamde 'Strengths and Difficulties Questionnaire', kortweg de SDQ [1,2]. Met de SDQ kan van een kind de ernst van psychosociale problemen gemeten worden. Vele Nederlandse ouders met een kind op de basisschool hebben de SDQ onder ogen gehad, met de vraag om de lijst in te vullen over hun kind. De SDQ is namelijk in de Nederlandse Jeugdgezondheidszorg het standaardinstrument om 5-jarigen te screenen op psychosociale problemen. Ook buiten Nederland is de SDQ een populair instrument. Om zo'n wijdverbreid gebruik te rechtvaardigen moet de SDQ natuurlijk wel goede meeteigenschappen hebben. Oftewel, de verkregen scores moeten informatief zijn over de mate waarin het kind psychosociale problemen heeft.

Bij het beoordelen van de kwaliteit van een psychologische test is een belangrijk hulpmiddel het gebruik van een statistisch model. Een onderzoek naar de kwaliteit van een test gaat dan ongeveer zo. De test wordt afgenomen onder een grote representatieve steekproef van de personen voor wie de test bedoeld is. De aldus verkregen scores op de vragen zijn de gegevens, ook wel data genoemd. Op basis van deze data wordt een statistisch model gemaakt, en wel zodanig dat het model helpt

bij het beantwoorden van de onderzoeksvraag. In het geval van een model voor de SDQ zou het model moeten laten zien welke vragen van de SDQ hetzelfde aspect meten, en ook hoe goed ze dat doen.

Zo'n statistisch model moet gebouwd worden. In de praktijk gaat men als volgt te werk. In stap 1 kiest men een bepaald type model uit, op basis van de verwachte eigenschappen van de data, de onderzoeksopzet en de onderzoeksvraag. Zo'n model heeft bepaalde vaste eigenschappen en bepaalde variabele eigenschappen, de zogenaamde vrije parameters. In stap 2 wordt de waarde van die variabele eigenschappen bepaald door het model op de data te passen, en dat gebeurt met behulp van wiskundige optimalisatie. Als het model de data goed past, dan kan het gehele model geïnterpreteerd worden en is de onderzoeker een flinke stap dichter bij het beantwoorden van zijn onderzoeksvraag.

Dat klinkt allemaal logisch en goed gefundeerd, en dat is het ook. In theorie werkt het uitstekend. In de praktijk is het niet altijd zo eenvoudig. U hoorde mij net zeggen: 'Als het model de data goed past, dan kan het gehele model geïnterpreteerd worden.' Een voorwaarde voor een zinvolle interpretatie is dus dat het model de data goed past. Nu past in de praktijk een model de geobserveerde data nooit perfect. Daar zijn twee oorzaken voor aan te wijzen.

Allereerst hebben we het welbekende fenomeen van de steekproeffluctuaties – het model geldt voor de populatie, en we moeten ons behelpen met een beperkte steekproef daaruit, die net wat andere eigenschappen zal hebben. Maar, zelfs als we de populatiegegevens zouden hebben, dan nog zal ons statistische model deze gegevens niet perfect beschrijven, oftewel er is modelfout. Dat komt simpelweg, omdat een model een vereenvoudigde weergave is van de werkelijkheid. Het model beschrijft dus de werkelijkheid, maar ís daarmee niet de realiteit [3,4]. Vergelijk het met een röntgenfoto van een kaak, waarop de botstructuur en de positie van de kiezen in het bot te zien is. Niet alles is er op te zien, maar die röntgenfoto levert wel *een* beschrijving van de werkelijkheid, en wel eentje die de tandarts kan helpen bij het opstellen van zijn behandelplan. Een statistisch model is een representatie van de werkelijkheid. De eis aan zo'n model is wat mij betreft, dat het helpt om

de werkelijkheid te begrijpen. In de regel is het begrip gebaat bij een simpel model dat goed past bij de data. De crux is dus om een model te vinden dat aan die eis voldoet. Modelselectie is daarmee één van de moeilijkste zaken bij statistisch modelleren. Om in analogie met Johan Cruijff te spreken: "Statistische modellen zijn simpel. Maar simpel statistisch modelleren blijkt vaak het moeilijkste wat er is.".

Laten we eens teruggaan naar het beoordelen van de kwaliteit van de SDQ. Zoals ik vertelde, wordt de SDQ gebruikt als screeningsinstrument voor de 5-jarigen in Nederland. Maar de SDQ wordt ook in een andere context gebruikt, namelijk voor kinderen met psychopathologie, of een vermoeden daarvan. In die context is de SDQ een hulpmiddel om tot een diagnose te komen. In beide gevallen, voor screening in de algemene populatie of diagnose in een klinische groep, wordt de ouders gevraagd om de SDQ in te vullen. Een goede vraag hierbij is of de SDQ in beide contexten wel op dezelfde manier meet. Het zou heel goed kunnen dat ouders van kinderen mét en kinderen zonder psychopathologie bepaalde vragen op een heel andere manier interpreteren, of er een heel verschillend referentiekader op na houden. Het is voorstelbaar dat dit zou gelden voor een stelling als ‘Heeft vaak driftbuien of woede-uitbarstingen’. Mijn oud-promovendus Iris Smits en ik hebben onderzocht in hoeverre de SDQ in beide contexten hetzelfde meet [5]. Dit deden we door op basis van ouderscores verzameld in beide contexten, een statistisch model te bouwen, en na te gaan of dat model gelijk kan zijn voor beide contexten. Als het model verschilt, heb je een indicatie dat de SDQ in verschillende contexten verschillend meet. Dat zou betekenen dat de resulterende scores niet vergelijkbaar zouden zijn tussen kinderen in verschillende contexten. Op basis van de statistische modellen zagen we geen indicatie van een verschillend soort meting. De SDQ is daarmee geslaagd voor dit examen, en de praktijkgebruiker van de SDQ zal dus opgelucht ademhalen.

Maar is die opluchting wel terecht? Ook als we rekening houden met steekproeffluctuaties, past het gebruikte model de data weliswaar redelijk, maar zeker niet perfect. Er is dus sprake van modelfout. De cruciale vraag is dan: Zit er belangrijke informatie verstopt in die

modelfout? Oftewel, zit in de modelfout nog iets dat, in dit geval, belangrijk is om de verschillen in contexten te karakteriseren?

Een aanpak om hier achter te komen is om een vrijer model toe te passen, om te zien of de modelfout verkleind kan worden. Dat kan door het gebruikte model handmatig een beetje aan te passen. Bij een simpel model en weinig modelfout kan dat prima. Maar bij complexere modellen en grotere modelfout kan het slim zijn een andere aanpak te gebruiken, namelijk meer exploratief.

Naar exploratief modelleren

Bij het voorbeeld van de SDQ spraken we over twee groepen, dat is nog overzichtelijk. Maar in andere gevallen zijn er veel meer groepen te onderscheiden. Een voorbeeld komt uit onderzoek naar emoties. In zulk soort emotie-onderzoek wordt aan personen gevraagd om regelmatig, zeg een aantal keren per dag gedurende een aantal dagen, een scoring te geven van een aantal emoties, zoals zij die op dat moment ervaren. De vraag is dan 'In hoeverre voelt u zich op dit moment...', met als emoties: nerveus, boos, angstig, blij, tevreden, ontspannen, etcetera. Een interessante vraag is in hoeverre bepaalde emoties in de regel op dezelfde momenten voorkomen, en welke emoties dat dan betreft. Het blijkt dat mensen hierin onderling behoorlijk verschillen. Sommigen ervaren altijd dezelfde soorten emoties gelijktijdig, wat er op neerkomt dat ze zich positiever dan wel negatiever voelen, en daarmee af. Anderen ervaren meer verschillende emoties gelijktijdig, en maken dus meer onderscheid in soorten emoties. Een beter onderscheid kunnen maken in verschillende soorten ervaren emoties wordt wel in verband gebracht met een betere mentale gezondheid[6]. De vraag is nu op welke manier personen onderscheid maken in soorten emoties, en hoe dit verschilt tussen de personen in ons onderzoek. Omdat er van te voren geen specifieke ideeën zijn over de emoties die samen ervaren worden, en al helemaal niet in hoeverre en op welke manier de verschillende personen hierin onderling verschillen, is een exploratieve aanpak nodig om dit in kaart te brengen.

Bij zo'n exploratieve aanpak wordt gebruikt gemaakt van een model met een bijbehorende optimalisatie. Dat is net als bij ons SDQ voorbeeld, maar in dit geval laat het model en de bijbehorende optimalisatie veel meer vrijheid toe, dat wil zeggen, past bij meer verschillende structuren in de gegevens. Aan de ene kant is die vrijheid heel mooi – het zorgt voor meer ruimte om speciale eigenschappen van de data in het model te vatten. Maar aan de andere kant heeft die vrijheid ook een nadeel – het kan er namelijk voor zorgen dat je een fenomeen aanziet voor een interessante speciale eigenschap, terwijl het niks anders is dan een toevalligheidje, iets dat je in welke dataset die je zou verzamelen dan ook, nooit weer zult zien. Daarnaast kan het ook gebeuren dat je door de bomen het bos niet meer zit. Kortom, het is een precaire balans tussen vrijheid en beperking. De truc is nu om ervoor te zorgen dat je díe beperkingen oplegt die de oplossing in de richting van een inhoudelijk interpreteerbaar en stabiel model sturen. U ziet dat er ook in exploratieve aanpakken beperkingen in het model worden opgelegd. Ook exploratief is dus maar relatief.

Wat heb je nu nodig voor zo'n exploratieve aanpak? Allereerst vraagt het dat je een arsenaal aan statistische modellen ter beschikking hebt, voor verschillende typen situaties. In de loop van de jaren heb ik onderzoek gedaan naar modellen voor de zogenaamde multi-set gegevens – gegevens die je kunt ordenen in meerdere sets, zoals meerdere personen. Dit ging in samenwerking met verschillende onderzoekers, onder wie, in chronologische startvolgorde, Henk Kiers, Age Smilde, Eva Ceulemans en Kim De Roover. Zo hebben we nieuwe modelvarianten voorgesteld, en laten zien hoe deze helpen om inzicht te krijgen in de data, in verschillende toepassingen.

Om nu te beoordelen in welke situaties de modellen mogelijk nuttig toepasbaar zijn, is het nodig de theoretische eigenschappen goed te begrijpen. Zo is het heel belangrijk te weten hoe de verschillende varianten zich tot elkaar verhouden. Dat is bijvoorbeeld te zien door de varianten in een raamwerk te plaatsen, oftewel ze te verdelen in verschillende families, met bijbehorende vertakkingen. Ik zal u de details besparen, maar neem van mij aan dat het leuke puzzels zijn.

Daarnaast is het belangrijk te weten hoe goed de modellen geschat kunnen worden, in allerlei soorten data. Een eerste stap is om zelf data te maken, te simuleren, onder makkelijker en moeilijker omstandigheden, en te kijken in hoeverre de modellen goed teruggevonden worden. Na een geslaagde test is het tijd om verder te kijken dan de eigen neus lang is, en te vergelijken met andere modelvarianten. Op basis van een vergelijkend warenonderzoek probeer je dan een goed beeld te krijgen van welke methode in welke situatie de beste weergave oplevert. Dat levert soms verrassende inzichten op. Een populaire methode komt lang niet altijd als beste uit de bus. Bijvoorbeeld, voor clusteranalyse wordt in empirische toepassingen vaak de zogenaamde K-means methode gebruikt. Dat is ook heel makkelijk te doen, omdat er een knopje voor is in het veel gebruikte statistische softwarepakket SPSS [7]. Uit vergelijkend warenonderzoek blijkt alleen dat de K-means methode alleen in nogal specifieke gevallen een goede clustering oplevert, en dat er betere alternatieven voor handen zijn die voor een veel bredere range aan gevallen een goede clustering geeft [8,9]. Aanbieders van statistische software zouden zich wel wat meer bewust mogen zijn van hun leidende rol in de keuze van statistische technieken door toepassers. Oftewel richt het pakket zo in dat de gebruiker zo min mogelijk in de verleiding gebracht wordt een minder geschikte techniek te gebruiken. Inzichten uit cognitieve ergonomie kunnen hierbij heel behulpzaam zijn.

Een ander voorbeeld: In bepaalde kringen is het gebruik van zogenaamde Sparse methoden enorm populair. Dat mag helpen om het voorspellend vermogen van regressiemodellen te verbeteren [10]. Maar voor goed beschrijvende multi-set modellen lijkt het beter om gebruik te maken van een doelgerichte rotatie [11]. Dit soort werk is belangrijk om inzicht te krijgen in de eigenschappen en prestaties van de methoden, in absolute en relatieve zin. In deze hoek is nog veel werk aan de winkel, al is het maar omdat modelbouwers uit verschillende tradities zich bezighouden met gelijkaardige modellen, die er op het oog heel verschillend uit kunnen zien.

Voor een zinvolle analyse van empirische gegevens hebben we aan kennis over de statistische methoden niet genoeg. Inhoudelijke

theorie en kennis is onontbeerlijk. Dit is ook precies waarom voor mij samenwerking met inhoudelijke onderzoekers zo belangrijk is. Door hen krijg ik inzicht in de beschikbare achtergrondkennis en de open vraagstukken, en daarmee richting voor zinvolle statistische modellen. Zo ging ik met Hannah Lennarz en Anna Lichtwarck-Asschoff op zoek naar typen emotiepatronen onder adolescenten [12], met Boele de Raad naar persoonlijkheidstrekken die universeel voorkomen in verschillende culturen [13], en met Edo Saccenti naar overeenkomsten tussen metabolietpatronen in de urine van rhesusapen en mensen [14]. Zulke patronen krijg je boven water met behulp van een exploratieve multi-set analyse. Voor mij als mede-ontwikkelaar van dit soort modellen is dat natuurlijk een mooi resultaat.

Hierbij past wel enige relativering. In heel veel toepassingen in de psychologie ligt de interesse helemaal niet in het opsporen van patronen. Mijn interesse is dan ook niet beperkt tot multi-set methoden. Ik werk graag samen met inhoudelijke onderzoekers die interesse hebben in het zorgvuldig toepassen van statistische methoden. Dergelijke samenwerkingen worden wel eens aangeduid als 'Statistische consultatie', maar ik houd niet zo van die term. Het roept bij mij het beeld op van een loket waar 'de statistische vraag' onder het raampje doorgeschoven wordt, en vervolgens 'het statistische antwoord' teruggeschoven. Zo'n loketaanpak werkt alleen als 'de statistische vraag' helemaal is uitgekristalliseerd. Op zo'n vraag is het antwoord meestal makkelijk op te zoeken, en daarmee het loket overbodig. Samenwerken verwijst direct naar een vruchtbare interactie, ieder vanuit de eigen expertise.

Een voorbeeld van een dergelijke samenwerking is de ontwikkeling van de Bayley-III-SNA (Special Needs Addition) [15-19] in samenwerking met Linda Visser, Selma Ruiter, Bieuwe van der Meulen en Wied Ruijssenaars. De Bayley-III-SNA is een versie van de veelgebruikte Bayley-III ontwikkelingstest, die speciaal aangepast is voor kinderen met een beperking, bijvoorbeeld in de motoriek en/of het gezichtsvermogen. Deze test is inmiddels uitgegeven en daarmee voor de klinische praktijk beschikbaar.

Een ander project waarin een psychologische test geëvalueerd wordt, gaat binnenkort van start. In samenwerking met Accare, TNO Leiden en de Kinderacademie Groningen gaan we na hoe goed de SDQ gebruikt kan worden voor screening op psychosociale problemen onder alle adolescenten. We denken dat de SDQ waarbij de adolescent over zichzelf moet rapporteren, wel eens ongeschikt zou kunnen zijn voor bepaalde groepen adolescenten. Daarom gaan we ook na welke goede, en praktisch haalbare alternatieven er zijn.

Dit soort projecten heeft gemeen dat de focus ligt op het beantwoorden van gedragswetenschappelijke onderzoeksvragen. Soms stuit ik hierbij op witte vlekken op de kaart van het statistisch modelleren. Een voorbeeld hiervan is op het gebied van testnormering. In praktijk worden meerdere methoden toegepast, maar het is onduidelijk welke methode het beste resultaat oplevert, dat wil zeggen de meest betrouwbare normen. Ook ontbreekt bijvoorbeeld een goede methode om tot een optimaal steekproefontwerp te komen. Promovendus Lieke Voncken verkent deze witte vlek, en heeft al een aantal goede, praktisch relevante inzichten verworven.

Bij inhoudelijk georiënteerd onderzoek is een gedegen expertise op het gebied van statistisch modelleren niet alleen heel nuttig om een goede onderzoeksopzet te kiezen – denk bijvoorbeeld aan een gefundeerde uitspraak over een minimale grootte van een steekproef. Het kan ook richting geven aan alternatieve denkwijzen die inhoudelijk van betekenis kunnen zijn. Om deze uitspraak wat concreter te maken, geef ik een voorbeeld.

Een klassieke manier van onderzoek bedrijven onder personen met een bepaalde klinische aandoening maakt gebruik van groepsvergelijking. Hierbij wordt een groep personen met de aandoening vergeleken met een groep zonder de aandoening, door de gemiddelde scores op één of meer relevante kenmerken te vergelijken. Nu blijkt maar al te vaak dat personen binnen de groepen behoorlijk van elkaar verschillen en wel zodanig dat er flinke overlap ontstaat tussen de groepen. Dan is het vergelijken van de gemiddelden van de twee groepen op de afzonderlijke kenmerken wel een heel simplistische samenvatting,

die echt te weinig informatie geeft. Het is dan zinnig om op een slimme manier op zoek te gaan naar onderscheidende subgroepen, waarbij informatie van meerdere kenmerken tegelijk wordt meegenomen. Dit vereist dus een exploratieve multivariate analyse. Op dit terrein valt nog het nodige te doen. Met Catharina Hartman werk ik momenteel aan zo'n model voor het beschrijven van typische patronen in de ontwikkeling onder adolescenten met en zonder een ADHD-diagnose.

U ziet dat zo'n samenwerking tussen inhoudelijke onderzoekers en statistische modelleerders vruchtbaar kan zijn. Wat mij betreft krijgt deze interactie ook meer gezicht in het onderwijs. Onderwijs in de statistiek behandelt traditioneel de statistische theorie en laat daarbij toepassingen zien ter illustratie. Het staat buiten kijf dat voor een goede interpretatie kennis van de statistische theorie essentieel is. Maar zinvol modelleren leer je pas als je ook eens vertrekt vanuit de toepassing, en de verschillende keuzemogelijkheden voor de statistiek leert zien, beoordelen en waarderen. Docenten spelen hierbij een cruciale rol, in het delen van inzichten, aanmoedigen tot vragen stellen en het stimuleren van een kritische reflectie.

Exploratief modelleren en begrip

Nu ik u een idee heb gegeven van het nut van exploratieve analysemethoden, wil ik graag even terug naar het woord 'Begrip'. In hoeverre dragen deze methoden nu echt bij aan het begrip? Het is essentieel dat de gekozen methode aansluit bij de achtergrondkennis – exploratieve methoden zijn weliswaar vrijer dan confirmatieve, maar ze gebruiken nog steeds een bepaald criterium. Dat betekent dat de analyseresultaten ook geïnterpreteerd moeten worden met het criterium in het achterhoofd – bepaalde aspecten van de data kunnen wel, en andere kunnen principieel niet boven tafel komen. Net als de röntgenfoto van de kaak wel kan prijsgeven hoe de positie van de kiezen is, maar niet of er een ontsteking in de wortel van een kies is. Voor een juiste en zinvolle interpretatie van een röntgenfoto moet een tandarts dus precies weten wat wel, en hoe dan, en wat niet op zo'n röntgenfoto afgebeeld kan zijn. Dit geldt precies zo voor de interpretatie van een statistische

analyse – je moet weten welke aspect van de data wel, en hoe dan, en vooral ook wat niet weergegeven kan worden.

Een zinvolle exploratieve analyse is in lijn met de processen en mechanismen die aan de data ten grondslag liggen. Om zo'n zinvolle analysemethode te kiezen helpt dus achtergrondkennis over die processen en mechanismen – maar vaak doen we juist onderzoek om die kennis op te kunnen doen. U ziet de spiraal vast voor u. De taak is in zekere zin vergelijkbaar met die van Baron van Münchhausen die zichzelf én zijn paard uit het moeras hielp, door zichzelf aan zijn haren omhoog te trekken. Bij dit proces kunnen exploratieve analysemethoden een goede bijdrage leveren. Dit geldt trouwens niet alleen voor exploratieve analysemethoden die gericht zijn op beschrijving van de data, maar ook voor exploratieve analysemethoden die zich richten op optimale voorspelling[20]. Bij gebrek aan voldoende theoretische kennis kunnen we dankbaar gebruik maken van een exploratieve analysemethode – en afhankelijk van het doel moet de beschrijving dan wel voorspelling de voorrang krijgen. Als ik wil weten of een bepaalde patiënt meer gebaat is bij therapie A dan bij therapie B dan heb ik een voorspellingsvraag. Een therapie-effect evaluatie studie zou zich dan ook vooral op voorspelling moeten richten, en niet zozeer op beschrijving.

Exploratief, maar hoe zit het dan met confirmatief?

Nu is het mogelijk dat het bij u ergens toch een beetje knaagt. Een exploratieve analysemethode, dat klinkt als gerelateerd aan exploratief onderzoek. En was exploratief onderzoek nu niet dubieus? Het klopt dat het voor velen verdacht is. Dat komt in mijn ogen vooral doordat de gehanteerde statistische analysestrategie niet optimaal is. Ik zal uitleggen waarom, en hoe het beter kan.

Volgens het boekje is confirmatief onderzoek gebaseerd op een theorie. Op basis van die theorie formuleert de onderzoeker een hypothese, bedenkt een slim onderzoeksdesign, en bepaalt welke statistische toets te zijner tijd uitgevoerd zal worden op de data, en tot welke conclusie dat dan zal leiden als het één dan wel het ander eruit zal

komen. Daarna verzamelt hij de gegevens, analyseert ze met behulp van de geselecteerde statistische toets, en rapporteert de resultaten.

Exploratief onderzoek is nodig als theorie niet voor handen is, en wordt gebruikt om een bepaald fenomeen beter te leren begrijpen, en juist de theorie te ontwikkelen. De aanpak voor exploratief onderzoek is door verschillende auteurs als volgt omschreven: De onderzoeker onthoudt zich van het opstellen van hypotheses, verzamelt data, en probeert daarna verschillende statistische analyses uit om 'de meest interessante en de veelbelovendste aspecten van de data' op te sporen [21-23]. In mijn ogen wordt met zo'n beschrijving het exploratief onderzoek vogelvrij verklaard – met alle ellende van dien. Onder het mom 'het is exploratief onderzoek' wordt dan opeens toegelaten dat men gaat schatzoeken in de data – ook wel dataharken [24] genoemd. Het aloude 'Zoekt en gij zult vinden' gaat ook hier op. Hét grote probleem is dat de gevonden schat maar al te vaak een eendagsvlieg blijkt te zijn. Dit grote probleem – en het feit dat dataharken ook gebruikt wordt vanuit minder edele motieven – is voor sommigen aanleiding geweest tot een zeer kritische houding ten opzichte van exploratief onderzoek en tot het gebruik van de term 'krakkemikkige statistiek' ('wonky stats') [23]. Deze onderzoekers pleitten – in navolging van anderen – voor het gebruik van preregistratie bij confirmatief onderzoek.

Bij preregistratie beschrijft de onderzoeker vóór de dataverzameling het studieprotocol – het steekproefontwerp, de te verzamelen data en het analyseplan. Dat is een uitstekende aanpak, want het leidt tot transparantie, voorkomt daarmee oneigenlijk gedrag, en het dwingt de onderzoeker tot heel goed voordenken. Preregistratie kan dus gezien worden als publiek voordenken – en voordenken helpt een goede toepassing van statistiek vooruit [25].

Die aanpak hoeft alleen beslist niet beperkt te worden tot confirmatief onderzoek – ook bij exploratief onderzoek kan het, en moet het. Er is geen exploratief onderzoek te vinden dat niet voortbouwt op eerdere observaties en bevindingen, en er is geen exploratief onderzoek dat gebruik maakt van een willekeurige dataset, met een willekeurige keuze aan analyse technieken. Ook als er geen specifieke hypotheses opgesteld kunnen worden, is er achtergrondkennis en die wordt, impliciet of expliciet, gebruikt. Met verstand exploratieve analyse toepassen betekent dus dat de onderzoeker voordenkt, en het resultaat van dit voordenken expliciet maakt. Preregistratie – in welke vorm dan ook – helpt daarbij, en is daarmee van harte aanbevolen.

De term voordenken danken we aan één van mijn wetenschappelijke inspirators. Ik prijs mij gelukkig dat ik heb kunnen profiteren van de rijke psychometrietraditie. In het algemeen ben ik dankbaar dat ik op de schouders van velen mag staan, in het verleden en in het heden. Ik spreek de hoop uit dat ik ook voor anderen een schouder kan betekenen.

Ik heb gezegd.

Referenties

1. Goodman R. The strengths and difficulties questionnaire: A research note. *Journal of child psychology and psychiatry*. 1997;38(5):581-586.

2. Van Widenfelt BM, Goedhart AW, Treffers PD, Goodman R. Dutch version of the strengths and difficulties questionnaire (SDQ). *Eur Child Adolesc Psychiatry*. 2003;12(6):281-289.

3. Box GE. Science and statistics. *Journal of the American Statistical Association*. 1976;71(356):791-799.

4. Hand DJ. Wonderful examples, but let's not close our eyes. *Statistical Science*. 2014;29(1):98-100.

5. Smits IAM, Theunissen MHC, Reijneveld SA, Nauta MH, Timmerman ME. Measurement invariance of the parent version of the strengths and difficulties questionnaire (SDQ) across community and clinical populations european journal of psychological assessment. *European Journal of Psychological Assessment*. In druk.

6. Lindquist KA, Barrett LF. Emotional complexity. In: Lewis M, Haviland-Jones JM, Barrett LF, eds. *Handbook of emotions.* 3rd. ed. Guilford Press; 2008:513-532.

7. IBM Corp. IBM SPSS statistics for Windows. 2013;Version 22.0.

8. McLachlan GJ, Peel D. *Finite mixture models.* New York: Wiley; 2000.

9. Timmerman ME, Ceulemans E, Roover K, Leeuwen K. Subspace K-means clustering. *Behavior Research Methods*. 2013:1-13.

10. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning.* Vol 1. 2nd, 10th printing ed. Springer series in statistics Springer, Berlin; 2013. (Verkregen op 09-06-2016 van http://statweb.stanford.edu/~tibs/ElemStatLearn/ printings/ESLII_print10.pdf).

11. Timmerman ME, Kiers HAL, Ceulemans E. Searching components with simple structure in simultaneous component analysis: Blockwise simplimax rotation. *Chemometrics and Intelligent Laboratory Systems*. 2016.

12. Lennarz HK, Lichtwarck-Aschoff A, Timmerman ME, Granic I. Emotion differentiation and its relation with emotional well-being in adolescents. In voorbereiding.

13. De Raad B, Barelds DPH, Timmerman ME, De Roover K, Mlacic B, Church AT. Towards a pan-cultural personality structure: Input from 11 psycholexical studies. *European Journal of Personality*. 2014;28(5):497-510.

14. Saccenti E, Tenori L, Verbruggen P, et al. Of monkeys and men: A metabolomic analysis of static and dynamic urinary metabolic phenotypes in two species. *PLoS ONE*. 2014;9(9):e106077.

15. Ruiter SAJ, Visser L, van der Meulen BF, Timmerman ME. *Bayley-III-NL Special Needs Addition (SNA).* Amsterdam: Pearson; 2014.

16. Visser L, Ruiter SAJ, van der Meulen BF, Ruijssenaars WAJJM, Timmerman ME. Validity and suitability of the Bayley-III low motor/vision version: A comparative study among young children with and without motor and/or visual impairments. *Res Dev Disabil*. 2013;34(11):3736-3745.

17. Visser L, Ruiter SAJ, Van der Meulen BF, Ruijssenaars WAJJM, Timmerman ME. Accommodating the Bayley-III for motor and/or visual impairment: A comparative pilot study. *Pediatric Physical Therapy*. 2014;26(1).

18. Visser L, Ruiter SAJ, Van der Meulen BF, Ruijssenaars WAJJM, Timmerman ME. Dynamic assessment with the Bayley-III among young children with developmental disabilities. *Journal of Cognitive Education and Psychology*. 2015:126-142.

19. Visser L, Ruiter SAJ, van der Meulen BF, Ruijssenaars WAJJM, Timmerman ME. Low verbal assessment with the Bayley-III. *Res Dev Disabil*. 2015;36(0):230-243.

20. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: Lessons from machine learning. 2016. (Verkregen op 09-06-2016 van http://jakewestfall.org/publications/Yarkoni_Westfall_choosing_prediction.pdf).

21. Sijtsma K. Playing with data—Or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*. 2015;1-15.

22. Groot A. The meaning of "significance" for different types of research [translated and annotated by E. Wagenmakers, D. Borsboom, J. Verhagen, R. Kievit, M. Bakker, A. Cramer, D. Matzke, G.J. Mellenbergh, and H.L.J. van der Maas]. *Acta Psychol*. 2014;148:188-194.

23. Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL, Kievit RA. An agenda for purely confirmatory research. *Perspect Psychol Sci*. 2012;7(6):632-638.

24. Kerr NL. HARKing: Hypothesizing after the results are known. *Pers Soc Psychol Rev*. 1998;2(3):196-217.

25. Molenaar IW. Vast versus variabel: De statistische splitsing. 2000.