J.J. Hox (1998). Er is nieuws onder de zon: nieuwe oplossingen voor oude problemen. *Kwantitatieve Methoden*, 19, 95-118

ER IS NIEUWS ONDER DE ZON

NIEUWE OPLOSSINGEN VOOR OUDE PROBLEMEN

Joop Hox

Samenvatting

In deze rede wordt met een aantal concrete voorbeelden aangegeven wat de methodenleer en statistiek de vakinhoudelijke onderzoekers in de sociale wetenschappen te bieden heeft. De voorbeelden lopen uiteen van het probleem van de ontbrekende gegevens, resampling en randomisatietechnieken, het bestrijden van non-respons, en het construeren van interviewvragen. De achterliggende gedachte daarbij is dat onderzoekstechnieken globaal ingedeeld kunnen worden als accepted good practice, current best methods, en state of the art. Beargumenteerd wordt dat vakinhoudelijke onderzoekers te vaak blijven steken in het automatisch toepassen van accepted good methods, terwijl current best methods zonder veel problemen binnen het bereik van vakinhoudelijke onderzoekers (zouden moeten) liggen.

Rede, uitgesproken bij de aanvaarding van het ambt van hoogleraar 'methoden en technieken van sociaal-wetenschappelijk onderzoek' aan de Universiteit Utrecht op donderdag 1 mei 1997. J.J. Hox, vakgroep Methodenleer en Statistiek, Faculteit Sociale Wetenschappen, Universiteit Utrecht. Heidelberglaan 2, 3584 CS Utrecht. <j.hox@fss.uu.nl>

1. Inleiding

Tussen de methodenleer en statistiek van sociaal-wetenschappelijk onderzoek[1] en de inhoudelijk georiënteerde vakgebieden binnen de sociale wetenschappen bestaat een wat ongemakkelijke verhouding. Enerzijds is het duidelijk dat methodenleer en statistiek er bij horen. In alle faculteiten binnen de sociale wetenschappen wordt in de eerste fase aan studenten onderwijs in methoden en technieken gegeven. Daarbij komt dat de landelijke onderzoekscholen, die de promotieopleiding organiseren, in hun algemene onderzoekersopleiding doorgaans een stevige methoden en technieken component kennen. Er is binnen de sociale wetenschappen zelfs een onderzoeksschool, de Inter-universitaire Onderzoeksschool voor Psychometrie en Sociometrie (IOPS), die zich geheel richt op het verder ontwikkelen van methoden en technieken van onderzoek. Het belang van methodenleer en statistiek voor het vakinhoudelijk georiënteerde onderzoek is kennelijk onomstreden.

Anderzijds zijn er ook tegenbewegingen waar te nemen. Bij bezuinigingen en de daaruit voortvloeiende reorganisaties binnen het wetenschappelijk onderwijs wordt het vakgebied methodenleer en statistiek niet ontzien. Landelijk staan verschillende vakgroepen op dit terrein op de nominatie om te worden opgeheven. Wat dat betreft kan de Universiteit Utrecht, waar binnen de sociale faculteit nog niet zo lang geleden tot een heroprichting van de vakgroep methodenleer en statistiek is overgegaan, zich als enigszins tegendraads beschouwen, misschien zelfs als een voorloper in een tegenbeweging. Maar toch niet zo heel erg tegendraads, want een groot enthousiasme voor een eerste fase opleiding methodenleer en statistiek, of voor deelname aan het eerder genoemde IOPS, is mij ook in Utrecht tot nu toe niet gebleken. Tenslotte, hoewel niemand mij dit ooit in mijn gezicht heeft medegedeeld, krijg ik soms wel eens de indruk dat methodologen worden beschouwd als abstracte luchtfietsers; interessante acrobatiek op de vierkante centimeter, maar wat moet je ermee.

In deze oratie zal ik proberen met mijn voeten dicht bij de grond te blijven, en met een aantal concrete voorbeelden aangeven wat de methodenleer en statistiek de vakinhoudelijke onderzoekers zo al te bieden heeft. Op de vraag, wat de relatie moet zijn van individuele onderzoekers met methoden en technieken, en de vraag of een sociaal-wetenschappelijke faculteit op dat terrein een voorziening in stand moet houden, inclusief eerste fase opleiding en promovendi, zal ik aan het eind van deze rede terugkomen.

[1] Met methoden en technieken bedoel ik zowel de statistiek als de algemene methoden zoals onderzoeksontwerp en dataverzameling. Evenzo betreft de term methodoloog ook statistici en psychometrici. Wanneer het specifiek een van deze onderwerpen betreft zal ik de meer beperkte term gebruiken.

2. Enige statistisch georiënteerde voorbeelden

Allereerst zal ik beginnen met een tweetal voorbeelden van recente ontwikkelingen in de statistiek, die voor zover ik het kan overzien aan de aandacht van de vakinhoudelijke onderzoekers grotendeels ontsnapt zijn.

2.1 Het probleem van de onvolledige gegevens

Het eerste voorbeeld gaat over het probleem van ontbrekende gegevens. De meeste sociaal-wetenschappelijke onderzoekers worden regelmatig geconfronteerd met het probleem van ontbrekende gegevens; gaten in de datamatrix. Ontbrekende gegevens kunnen ontstaan doordat respondenten of interviewers die vragenlijsten invullen dat soms wat slordig doen, en af en toe per ongeluk een vraag overslaan. Respondenten kunnen soms ook expliciet weigeren een vraag te beantwoorden, omdat zij de vraag te persoonlijk vinden en daarom van mening zijn dat hun antwoord de onderzoeker niets aangaat. Bij longitudinaal panelonderzoek, waarin een groep respondenten soms vele jaren wordt gevolgd en regelmatig wordt ondervraagd, kan het ook gebeuren dat een respondent er de brui aan geeft, en niet langer wil meedoen. Bij psychologisch en biomedisch onderzoek is de oorzaak eerder uitvallende apparatuur of computerstoringen, waardoor niet alle geplande waarnemingen verzameld kunnen worden. In alle gevallen is het resultaat dat de onderzoeker blijft zitten met onvolledige gegevens; bij sommige van de gestelde vragen is het antwoord bekend, maar bij andere vragen ontbreekt het antwoord.

Globaal zijn er drie strategieën te onderscheiden om met onvolledige gegevens om te gaan. Deze strategieën duid ik hier kort aan met: schrappen, inschatten, en modelleren. Het veel gebruikte computerprogramma SPSS, hetgeen oorspronkelijk stond voor Statistical Package for the Social Sciences, kent bij de meeste berekeningen twee vormen van schrappen, namelijk listwise deletion en pairwise deletion. Listwise deletion komt er op neer dat bij de analyse van een verzameling variabelen; een persoon die bij een van die variabelen een ontbrekende waarde heeft, geheel uit de analyse wordt weggelaten. Het zal duidelijk zijn dat dit soms tot een drastische reductie van de hoeveelheid beschikbare gegevens kan leiden. Bij multivariate analyses waarin veel variabelen betrokken zijn, kan het verlies gemakkelijk oplopen tot het wegvallen van bijna de helft van de verzamelde steekproef! Dit is duidelijk een ongewenste situatie. De andere optie, pairwise deletion, is minder verkwistend; de betreffende persoon wordt alleen weggelaten bij die berekeningen waarbij de variabele met de ontbrekende waarde betrokken is, bij alle andere variabelen wordt die persoon wel meegenomen. Een probleem bij pairwise deletion is wel, dat bij multivariate analyses van een aantal variabelen, de verschillende resultaten door dit selectief weglaten op steeds verschillende

groepen respondenten zijn uitgerekend. Dit kan soms tot vreemde uitkomsten leiden.

De tweede strategie om met onvolledige gegevens om te gaan, sinds kort eveneens beschikbaar in SPSS, is de gegevens volledig te maken door de ontbrekende gegevens in te schatten. Statistici noemen dit imputatie. Nadat de gegevens op een of andere manier volledig zijn gemaakt, kunnen de standaard statistische technieken gebruikt worden voor de analyse. Voor inschatten zijn allerlei fraaie methoden bedacht, maar in de praktijk worden doorgaans slechts eenvoudige systemen gebruikt. Een veel gebruikte methode is om bij een ontbrekende waarde eenvoudig het gemiddelde in te vullen van die variabele in de rest van de steekproef. Een betere manier is om bij het inschatten van de ontbrekende waarde gebruik te maken van de informatie die we wel hebben over die persoon. Weten we bijvoorbeeld de sekse, dan kunnen we in plaats van het gemiddelde van de gehele steekproef voor een ontbrekende waarde ook het gemiddelde invullen van de mannen of de vrouwen. In laatste instantie kunnen we proberen de ontbrekende waarden te voorspellen uit alle variabelen waarvan we wel gegevens hebben. Een groot nadeel van dit soort methoden is, dat het voorspellen van de ontbrekende waarden door andere informatie uit de steekproef, de in de gegevens aanwezige structuren versterkt. De hoeveelheid ruis in de gegevens wordt daardoor onderschat; de resultaten worden mooier dan zij in werkelijkheid zijn.

De derde strategie om met ontbrekende waarden om te gaan, die door de meeste statistici ook als de beste wordt beschouwd, is om het statistische model waarmee gerekend wordt zo aan te passen dat onvolledige gegevens gewoon kunnen worden meegenomen. Het analyseprogramma gaat er dan als het ware omheen. De betreffende statistische methode maakt gebruik van de methode van de maximale aannemelijkheid, ofwel Maximum Likelihood. Maximum likelihood methoden leiden tot parameterschattingen die de geobserveerde gegevens maximaal waarschijnlijk maken. In principe maakt het daarvoor niet uit of die geobserveerde gegevens volledig zijn of niet. Wanneer het wenselijk is om de gegevens volledig te maken, kan dezelfde methode gebruikt worden om de ontbrekende gegevens in te schatten. Voor multivariaat-normale gegevens is deze methode beschikbaar in het programma AMOS (zie Arbucle, 1996), voor categoriële gegevens in het programma *l*EM (vgl. Vermunt, 1996).

Het inbouwen van de onvolledige gegevens in het statistische model is evenmin als de eerder genoemde strategieën zonder problemen. Om te beginnen is het in de standaard statistische programma's zoals SPSS doorgaans niet beschikbaar. Verder geldt ook hier dat als de ontbrekende gegevens daadwerkelijk ingeschat worden, de aanwezige structuur in de gegevens vergroot wordt. De belangrijkste redenen waarom statistici dit de beste methode vinden, is dat de schattingen in veel gevallen nauwkeuriger zijn, en dat er sprake is van een expliciet statistisch model, zodat het mogelijk is wiskundig te analyseren wat er eigenlijk gebeurt. Wanneer deze meer geavanceerde methode gebruikt is, dan herkent U dat aan het statistisch jargon; er is dan sprake van Maximum Likelihood en het E-M algoritme. Ook de termen Missing Completely At Random en Missing At

Random vallen dan doorgaans. Wat dat betekent zal ik aan de hand van een voorbeeld uitleggen. Voor een statistische onderbouwing verwijs ik naar Little en Rubin (1987) en Schafer (1997).

Stel dat we een enquête hebben met daarin drie vragen: een vraag naar het inkomen, een vraag naar de opleiding, en een vraag naar het beroep. Op de vraag naar het inkomen krijgen we in 20% van de gevallen geen antwoord. De vraag is nu hoe deze ontbrekende waarden tot stand zijn gekomen. Het eenvoudigste is wanneer ze zuiver toevallig tot stand zijn gekomen; in de statistiek heet dat *missing completely at random* (MCAR). In dat geval zal vrijwel elke manier van inschatten goede resultaten geven. Het andere uiterste is wanneer het niet geven van een antwoord op de vraag naar het inkomen, afhangt van dat inkomen zelf. Bijvoorbeeld: met name personen met een hoog inkomen geven liever geen antwoord op deze vraag. In dat geval is het moeilijk tot goede statistische schattingen te komen. Tussen deze beide mogelijkheden in ligt de situatie dat het ontbreken van een antwoord op de inkomensvraag weliswaar niet toevallig is, maar niet afhangt van het inkomen zelf. Bijvoorbeeld: met name mensen met een hoge opleiding geven liever geen antwoord op de inkomensvraag. Deze laatste situatie wordt aangeduid met *missing at random* (MAR), en hier kan de statistiek goede schattingen produceren.

De tabel hieronder geeft de resultaten van schattingen van het gemiddelde inkomen in de situatie dat 20% van de antwoorden ontbreekt, en waarbij het ontbreken van dat antwoord samenhangt met de opleiding van de respondent. Achter elkaar worden gegeven de ware populatiewaarde, de schatting bij listwise deletion, de schatting bij pairwise deletion, en de directe schatting volgens de maximum likelihood methode (via Amos, Arbucle, 1996).

gemiddelde	populatie	listwise	pairwise	ML-methode
inkomen	4000	3700	3700	4040
beroep	5	4.4	5.0	5.0
opleiding	8	7.4	8.0	8.0

Het is duidelijk dat de maximum likelihood methode tot veel betere schattingen leidt, vooral voor de cruciale variabele inkomen. Hetzelfde geldt voor de correlaties tussen de variabelen; de maximum likelihood methode leidt daar ook tot veel betere schattingen. Het aardige van dit voorbeeld is, dat de meeste onderzoekers vermoedelijk denken dat zij door eenvoudig listwise deletion te gebruiken, weinig aannamen maken over de gegevens. Het tegendeel is waar; listwise en pairwise deletion veronderstellen dat de ontbrekende gegevens *completely at random* zijn weggevallen, en de maximum likelihood methode gaat uit van de zwakkere veronderstelling dat ze *at random* ontbreken. In vrijwel alle gevallen dat er enige structuur in het ontbreken van de gegevens zit, mogen we aannemen dat de maximum likelihood methode het beter doet dan het eenvoudig schrappen van personen met ontbrekende gegevens (zie bijvoorbeeld Arbucle, 1996, of Verleye,

1996). Wanneer het ontbreken van gegevens samenhangt met de (onbekende) waarde van de variabele die we willen meten, spreken we van *not missing at random* (NMAR) of *nonignorable nonrespons*. Deze situatie is wezenlijk ingewikkelder, omdat we dan het nonresponse mechanisme in het model moeten opnemen, wat doorgaans impliceert dat we over dat mechanisme aannamen moeten doen. Ik ga daar hier niet verder op in; een voorbeeld gevolgd door discussie is te vinden in Dingle (1994).

2.2 Computer intensieve statistiek

Een ander voorbeeld, nog steeds in de statistische hoek. Binnen de statistische analyse heeft zich op een voor inhoudelijke onderzoekers vrijwel ongemerkte, bijna stiekeme manier een nieuwe tak van statistiek, die ik hier aanduid met *computer intensieve statistiek*. Dat in de statistische analyse veel met computers wordt gewerkt weet U al, en als U het niet wist zal het U in ieder geval niet verbazen. Doordat computers steeds sneller en goedkoper zijn geworden, zijn statistici ze ook gaan gebruiken voor andere dingen dan het uitrekenen van statistische formules. Het statistisch jargon spreekt bij deze nieuwe toepassingen van bootstrappen, randomisatie- en permutatietesten (zie o.a. Efron & Tibshirani, 1993; Mooney & Duval, 1993; Noreen, 1989; een vroege toepassing is te vinden in Hox, 1986).

2.2.1 Bootstrappen

Wat dit allemaal inhoudt is het beste uit te leggen aan de hand van een voorbeeld. Stel dat U een vragenlijst hebt ontwikkeld, bijvoorbeeld tien vragen die samen een attitudeschaal vormen. Het is een goed gebruik om voor zo'n schaal in een onderzoek te bepalen hoe betrouwbaar deze meet. Dikwijls wordt daar een maat voor gebruikt die bekend staat als *coëfficiënt alfa*. Deze heeft een eenvoudige interpretatie: wanneer coëfficiënt alfa gelijk is aan één, dan meet U zonder enige meetfout. Is coëfficiënt alfa gelijk aan nul, dan meet U niets. Doorgaans wordt een waarde van 0.70 of hoger beschouwd als een indicatie dat U met voldoende betrouwbaarheid meet.

Nu hebt U in een vooronderzoek bij een steekproef van 50 personen voor Uw attitudeschaal een alfa van 0.75 berekend. Dat ziet er goed uit, het is in ieder geval groter dan de 0.70 die ik net noemde. Die 0.75 is echter alleen van toepassing op de toevallige steekproef van 50 respondenten van Uw vooronderzoek. Statistisch georiënteerde lieden zouden hiervoor graag een statistische toetsing uitvoeren, dat wil zeggen op formele wijze toetsen of het aannemelijk is dat in de populatie waar Uw steekproef vandaan komt de alfa groter is dan 0.70. Een gebruikelijke manier om zoiets te doen is om voor Uw steekproefresultaat van 0.75 een zogenaamd 95% betrouwbaarheidsinterval te

berekenen, dat wil zeggen het interval waarvan we met 95% zekerheid mogen aannemen dat de populatiewaarde ertussen valt.[1] Voor coëfficiënt alfa valt het vaststellen van een 95% betrouwbaarheidsinterval niet mee. Weliswaar is hiervoor een statistisch model ontwikkeld (Feldt, 1965), maar zoals bij alle statistische procedures worden daarin een aantal aannamen gedaan, en in dit geval zijn dat helaas aannamen die in de praktijk doorgaans onjuist zijn. Bovendien geldt het betreffende model pas bij grote aantallen vragen en bij grote steekproeven. Tien vragen en vijftig respondenten lijken daarvoor beide aan de lage kant.

Het 95% betrouwbaarheidsinterval van de coëfficiënt alfa kan echter ook worden vastgesteld door middel van een *bootstrap*. Voor de bootstrap gaan we helemaal terug naar de logica waarop statistische toetsen zijn gebaseerd. Die logica gaat er van uit dat als we meerdere malen een steekproef trekken uit een populatie, we uiteraard niet steeds precies dezelfde groep respondenten zullen krijgen, en daarom ook niet steeds dezelfde uitkomsten zullen vinden. De uitkomsten zullen van steekproef tot steekproef variëren, en we gebruiken een mathematisch statistisch model om te schatten hoe groot de variatie is die we mogen verwachten, om op basis daarvan betrouwbaarheidsintervallen en statistische toetsen uit te rekenen.

In plaats van een statistisch model te gebruiken, kunnen we natuurlijk ook eenvoudig duizend maal met teruglegging een steekproef trekken, de gewenste uitkomsten berekenen, en tenslotte bekijken hoe groot de variatie in uitkomsten is. In de praktijk is dat veel te bewerkelijk, maar in principe zou dat kunnen. De bootstrap methode doet iets dat hier op lijkt. In plaats van 1000 maal een nieuwe steekproef te trekken uit de gehele populatie, wordt in de computer 1000 maal een steekproef getrokken uit de reeds verzamelde gegevens. In ons voorbeeld van de attitudeschaal wordt uit de voorliggende steekproef van 50 respondenten 1000 maal een nieuwe steekproef getrokken, eveneens van 50 personen, en op die 1000 nieuwe steekproeven wordt dus 1000 maal een coëfficiënt alfa berekend. De variabiliteit van die 1000 uitkomsten wordt gezien als een indicatie van de grootte van de steekproefvariaties, en op basis hiervan wordt het 95% betrouwbaarheidsinterval berekend en worden eventuele statistische toetsen uitgevoerd. Net als indertijd de Baron van Münchhausen trekken we ons aan onze laarzen het statistische moeras uit, vandaar de naam: bootstrappen.

U vraagt zich misschien af, of we bij het trekken van steekproeven van 50 uit de geobserveerde gegevens van 50 respondenten niet steeds precies dezelfde uitkomst zullen krijgen. Neen, dat is niet zo, want we trekken die steekproeven met teruglegging. Dus, telkens als een respondent in de bootstrap steekproef getrokken is, wordt hij weer teruggelegd bij de oorspronkelijke gegevens. De bootstrap steekproeven verschillen dus van elkaar doordat in elke

[1] De interpretatie van een betrouwbaarheidsinterval is strikt genomen ingewikkelder en minder intuïtief dan hier omschreven; daar ga ik hier echter niet nader op in. Voor het voorbeeld is de precieze interpretatie niet belangrijk.

steekproef sommige respondenten ontbreken en andere weer meerdere malen voorkomen. Omdat bij de bootstrap steeds opnieuw steekproeven uit de oorspronkelijke steekproef worden getrokken heet deze strategie ook wel *resampling*.

Wanneer we in ons voorbeeld via een statistisch model het 95% betrouwbaarheidsinterval voor de gevonden coëfficiënt alfa bepalen, dan blijkt dit te lopen van 0.63 tot 0.84. Hoewel het grootste deel van dit betrouwbaarheidsinterval boven de 0.70 ligt, kunnen we er op basis van de steekproef niet voldoende zeker van zijn, dat de coëfficiënt alfa in de populatie inderdaad boven het gewenste minimum van 0.70 ligt. Echter, dit zijn de resultaten van een toets waarbij we weten dat aan de aannamen niet voldaan is. We kunnen ook uit onze gegevens 1000 bootstrap steekproeven trekken, 1000 maal coëfficiënt alfa berekenen, en bij de gevonden alfa's kijken welke de bovengrens en de ondergrens van de middelste 95% aangeven. Het resultaat is dan een 95% bootstrap betrouwbaarheidsinterval.

2.2.2 Randomisatie- en permutatietoetsen

Andere computerintensieve methoden zijn de zogenaamde randomisatie- en permutatietoetsen. Deze sluiten aan bij de gedachtengang in experimenteel onderzoek. Stel, U bent nog steeds bezig met Uw attitudeschaal van 10 vragen. Nu vraagt U zich af of het voor de betrouwbaarheid van deze schaal iets uitmaakt of U de vragen laat stellen door een interviewer in een persoonlijk interview, of dat U de respondenten een schriftelijke vragenlijst laat invullen. Volgens de beste principes van het experimentele onderzoek verzamelt U 50 respondenten, waarvan U er via loting 25 door een interviewer laat ondervragen, en de andere 25 een schriftelijke vragenlijst laat invullen. De interviewmethode levert een alfa op van 0.70, de vragenlijstmethode van 0.80. Mogen we nu concluderen dat de schriftelijke vragenlijst tot betrouwbaarder resultaten leidt?

Ook voor deze vraag is een statistische toets beschikbaar (Feldt, 1969; Hakstian & Whalen, 1976). Wanneer we deze toets toepassen, dan blijkt dat het verschil tussen de alfa's van beide groepen niet significant is; de overschrijdingskans is 0.35, hetgeen wil zeggen dat de kans dat dit soort verschillen op basis van toeval ontstaan 35% is. Die kans is vrij groot, en we accepteren de nulhypothese dat de wijze van vragenstellen niets uit maakt.

Ook voor deze toets geldt, dat we bij voorbaat weten dat aan de aannamen niet voldaan is. We kunnen de toetsing echter ook op een andere wijze uitvoeren, namelijk met een *randomisatie- of permutatietoets*. Bij een randomisatietoets gaan we terug naar de logica van het experimenteel onderzoek. De basisvorm van experimenteel onderzoek is, dat de proefgroep volgens een toevalsproces wordt verdeeld in een experimentele groep en een controlegroep. Vervolgens wordt het experiment uitgevoerd, en nagegaan of de experimentele groep verschilt van de controlegroep. Doordat de groepsindeling volgens toeval oftewel *at random* is geschied, mogen we aannemen dat

de beide groepen voor het experiment gelijk zijn. Wanneer we achteraf verschillen constateren, dan kunnen we die toeschrijven aan de experimentele manipulatie.

De rol van de statistische toets bij experimenteel onderzoek is, na te gaan of een eventueel gevonden verschil reëel is. Immers, wanneer we de proefpersonen at random over de experimentele en de controlegroep verdelen, zullen deze beide groepen doorgaans niet exact gelijk zijn. In ons voorbeeld; het is erg onwaarschijnlijk dat de coëfficiënt alfa in beide groepen in alle decimalen gelijk is. Kleine toevallig optredende verschillen zijn te verwachten. Ook hier zouden we de toetsing via een statistisch model kunnen omzeilen door het experiment een groot aantal malen te herhalen, met steeds weer andere toewijzingen van personen aan groepen. Om allerlei redenen is dat niet erg praktisch. Wat we wel kunnen doen is de personen waarvan we gegevens hebben, in de computer op alle mogelijke manieren in twee groepen verdelen, van beide groepen de coëfficiënt alfa bepalen, en daarvan tenslotte het verschil berekenen. De verdeling van al die verschillen tussen alfa's is een indicatie van de mate van variabiliteit in het verschil tussen twee alfa's die we op basis van toeval mogen verwachten. Wanneer het daadwerkelijk geobserveerde verschil, vergeleken met die referentieverdeling, nu uitzonderlijk groot is, dan concluderen wij dat het geobserveerde verschil niet op toeval berust maar reëel is. Analoog aan het 5% significantieniveau van de gebruikelijke statistische toets, kunnen wij afspreken dat we het gevonden verschil significant noemen als het bij de 5% grootste verschillen zit.

De zojuist beschreven toets heet een randomisatietoets, omdat deze toets het at random toewijzen van personen aan groepen in de computer nabootst. Het aantal manieren om 50 personen over twee groepen te verdelen is echter onhanteerbaar groot, het kan worden opgeschreven als een getal met 64 nullen (3E64). Gebruikelijk is om voor het construeren van de referentieverdeling hieruit een steekproef te trekken, en bijvoorbeeld 1000 random toewijzingen uit te voeren en door te rekenen. Dit heet dan een approximatieve randomisatietoets.

Bootstrap- en randomisatiemethoden bestaan al enige tijd. Ze worden met name toegepast wanneer langs de weg van de mathematische statistiek geen oplossing kan worden gevonden, of wanneer aan allerlei aannamen van de statistische toets niet voldaan is. De gedachte achter het laatste is in principe eenvoudig. Wanneer, bijvoorbeeld, de geobserveerde variabelen een zeer vreemde verdeling hebben, zodat aan de gebruikelijke aanname van een normale verdeling niet voldaan is, dan zal deze vreemde verdeling in iedere bootstrap steekproef ook opduiken. Het probleem is daarmee vanzelf verdisconteerd.[1]

[1] Overigens zijn bootstrap- en randomisatiemethoden ook gebonden aan aannamen, en daardoor geen panacee. Zo wordt de hier beschreven bootstrap methode wel aangemerkt als de naïeve of nonparametrische bootstrap. Wanneer de bootstrap methode gebruikt wordt voor het genereren van en nul-verdeling voor fit-indices zijn de resultaten doorgaans onbetrouwbaar (een voorbeeld hiervan wordt gegeven door Bollen & Stine, 1992). In zulke gevallen biedt de parametrische bootstrap uitkomst: de achtereenvolgende steekproeven worden dan niet

Het zal duidelijk zijn waarom hier gesproken wordt van computer intensieve methoden. Het berekenen van één enkele coëfficiënt alfa is betrekkelijk weinig werk, zeker op een steekproef van slechts 50 personen. Het berekenen van 1000 coëfficiënten alfa valt voor een computer ook nog wel mee. Wanneer het gaat om bootstrappen of randomisatietoetsen bij complexe multivariate modellen en grote steekproeven, dan is zelfs een moderne PC wel een tijdje bezig. Naarmate computers sneller en goedkoper worden, zullen dit soort methoden echter populairder worden. Voor veel toepassingen is nu nog een specialistisch computerprogramma nodig (bijvoorbeeld BOJA, Boomsma, 1991), maar de laatste versie van het standaard statistisch pakket SPSS bevat al enkele van dit soort methoden.

3. Enige voorbeelden uit de dataverzameling

Methodologische vernieuwingen zijn niet beperkt tot het terrein van de statistiek. Ook op andere terreinen, zoals dataverzameling en kwalitatief onderzoek, kan van nieuwe ontwikkelingen melding worden gemaakt. In deze rede beperk ik mij tot voorbeelden uit de sfeer van de dataverzameling.

3.1 Het probleem van de non-respons

Een groot probleem in sociaal-wetenschappelijk onderzoek is non-respons. Non-respons houdt in dat van sommige respondenten in de steekproef in het geheel geen gegevens verkregen kunnen worden, ofwel omdat zij niet bereikt worden, ofwel omdat zij expliciet weigeren aan het onderzoek mee te doen. Het probleem van non-respons is niet zozeer dat daardoor de steekproef te klein wordt; dat kan altijd opgevangen worden door bij het begin van het onderzoek de uitgezette steekproef wat groter te nemen. Het probleem is dat, wanneer er sprake is van non-respons, altijd de kans bestaat dat deze selectief is. Met andere woorden, wanneer de nonrespondenten niet geheel vergelijkbaar zijn met de respondenten, dan kunnen we eindigen met een vertekende steekproef. In feite speelt hier hetzelfde probleem dat ook speelt bij het analyseren van incomplete gegevens; wanneer respondenten ontbreken *completely at random* is er in feite geen probleem; zodra er sprake is van een selectieproces dat samenhangt met de te meten gegevens zijn correctie- of imputatieprocedures nodig (vgl. Lessler & Kalsbeek, 1992; Bethlehem & Kersten, 1986).

---

uit de data getrokken, maar met behulp van de geschatte modelparameters uit een theoretische statistische verdeling gegenereert. Voor een toepassing zie Van der Heijden, 't Hart & Dessens, 1997). Een nadeel hiervan is dat bootstrapmethoden vooral worden toegepast wanneer aan modelaannamen wordt getwijfeld. Bollen en Stine (1992) beschrijven een variant op de nonparametrische bootstrap die het probleem op andere wijze ondervangt.

Met non-respons is overigens in Nederland iets merkwaardigs aan de hand; de non-respons in Nederland is veel en veel groter dan in vergelijkbare westerse landen.[1] Wanneer statistische bureaus in de ons omringende landen het hebben over de 'Dutch disease' dan doelen zij daarmee op onze non-respons cijfers, en hopen dat zij niet besmet zullen raken. Bij wijze van voorbeeld, van één bepaald CBS onderzoek, het arbeidskrachtenonderzoek, is tussen 1983 en 1995 het responscijfer gedaald van 81% in 1983 naar 60% in 1995. Omringende landen als België en Engeland halen ondertussen voor vergelijkbaar onderzoek nog steeds 80% en in de Verenigde Staten haalt men nog steeds 95% respons (De Heer, 1996). De lage respons is daarbij niet specifiek voor het CBS; ook in het universitaire en het marktonderzoek is non-respons een groot en groeiend probleem. Zo heeft nog in 1994 de Sociaal-wetenschappelijke raad een verkennend onderzoek doen uitvoeren naar de non-respons bij universitair en para-universitair onderzoek (Kalfs & Kool, 1994). De voor (para-)universitair onderzoek 'normale' non-respons werd daarin geschat op 30 à 50 procent. Over de non-respons bij marktonderzoek is minder bekend, omdat responscijfers doorgaans tot gevoelige bedrijfsinformatie gerekend wordt, maar non-respons is ook daar een omvangrijk probleem.

Bij een non-respons van een dergelijke omvang is de mogelijkheid van een niet representatieve steekproef een reëel gevaar. In grote lijnen zijn er twee strategieën aan te wijzen om hiermee om te gaan: de non-respons bestrijden, of de vertekeningen in de verkregen steekproef corrigeren door weging.

3.1.1 Het bestrijden van non-respons

Voor het bestrijden van non-respons zijn in de loop van de tijd allerlei maatregelen voorgesteld. Voor een deel zijn deze maatregelen gebaseerd op informatie afkomstig van onderzoekers en veldmanagers die betrokken zijn bij het uitvoeren van enquêteonderzoek. Daarnaast is veel empirisch onderzoek gedaan waarin de effectiviteit van verschillende maatregelen vergeleken wordt. Veel van dit methodologische onderzoek lift als het ware mee men een inhoudelijk onderzoek; als onderdeel van het onderzoek of van het vooronderzoek worden verschillende responsverhogende maatregelen met elkaar vergeleken. Slechts weinig onderzoek is van meet af aan opgezet als een methodologisch experiment. Het resultaat is dat veel van deze onderzoeken slecht op elkaar aansluiten en arm zijn aan theorie. Toch zijn er wel resultaten geboekt. Doordat er betrekkelijk veel non-respons onderzoek is gedaan, is het mogelijk de uitkomsten van al deze onderzoeken bij elkaar te nemen en op hun beurt statistisch samen te vatten. Over non-respons onderzoek zijn inmiddels verscheidene van zulke meta analyses uitgevoerd. Een van de eerste is een overzicht van Heberlein en Baumgartner uit 1978. In hun overzichtsartikel wordt de respons bij 98 methodologische

[1] In Duitsland zijn de responsecijfers bijna net zo laag als in Nederland, zie Schell, 1997.

onderzoeken voorspeld uit kenmerken van de steekproef, de vragenlijst, en de gevolgde onderzoeksprocedures. Uit deze analyse is een formule af te leiden waarmee het responspercentage van een enquête van tevoren voorspeld kan worden. Een aantal kenmerken is voor de onderzoekers moeilijk te manipuleren. Bijvoorbeeld, marktonderzoek haalt systematisch een lagere respons dan onderzoek van (semi-)overheidsinstanties. Andere kenmerken zijn echter wel te manipuleren. Zo levert elke volgende contactpoging gemiddeld zeven procent extra respons op, en levert een kleine beloning gemiddeld zes procent extra respons op, en een grote twaalf procent. Het aardige is dat deze formule door De Leeuw en Hox in 1985 in Nederland is toegepast op bestaand onderzoek, met als resultaat een correlatie van 0.85 tussen de voorspelde en de behaalde respons (De Leeuw & Hox, 1985). Dit resultaat is later ook gerepliceerd door Van Rooy (1986).

Zoals gezegd is het meeste non-responsonderzoek nogal theorie-arm. Het nadeel daarvan is, dat het weinig aanknopingspunten biedt voor verbeteringen in de veldwerkprocedures of voor andere maatregelen die tot een grotere respons zouden kunnen leiden. Er is op dit moment ook geen allesomvattende theorie over het meewerken aan onderzoek. Wel zijn er een aantal theoretische modellen die onderzoeksdeelname vanuit een specifieke invalshoek benaderen. Ik noem er drie: sociaal-psychologische theorieën over overreding en ingaan op legitieme verzoeken, interviewer-respondent interacties, en rationele keuze-theorie.

Het *sociaal-psychologisch onderzoek naar overreding en toegeven* heeft een aantal resultaten opgeleverd, die zonder veel moeite vertaald kunnen worden naar de enquêtesituatie.

Het wel of niet ingaan op een verzoek, bijvoorbeeld het verzoek mee te werken aan een enquête, hangt samen met een groot aantal factoren. Een aantal van die factoren zijn inherent aan het verzoek, zoals de aantrekkelijkheid van het onderwerp van de enquête of de lengte van de vragenlijst. Daarnaast spelen een aantal meer algemene sociaal-psychologische principes een rol. Cialdini (1988, zie ook Groves, Cialdini & Couper, 1992) formuleert zes algemene heuristische principes die mensen toepassen bij de beslissing wel of niet in te gaan op een verzoek.

Het eerste principe is reciprociteit. Men zou dit ook het tit-for-tat principe kunnen noemen; mensen zijn geneigd om op een bepaald gedrag te reageren met gelijksoortig gedrag. Vertaald naar de enquêtesituatie betekent dit, dat mensen eerder geneigd zijn om mee te werken als het verzoek daartoe is voorafgegaan door een beloning. Het bijsluiten van een kleine beloning bij de vragenlijst werkt volgens dit principe.

Het tweede principe is consistentie. Mensen willen in eigen ogen graag consistent zijn. Hieruit kan worden afgeleid dat mensen bereid zullen zijn mee te werken aan een enquête, wanneer zij dat kunnen opvatten als consistent met hun algemene opvattingen. De consistentienorm is van toepassing bij longitudinale panelonderzoeken. Respondenten in een panel hebben bij het begin van het onderzoek toegezegd dat zij voor de hele periode mee zullen werken. Wanneer respondenten

dreigen af te haken bij het onderzoek, vertonen zij daarmee inconsistent gedrag. Pogingen om zulke respondenten alsnog bij het onderzoek te betrekken kunnen dus aansluiten bij het consistentieprincipe.

Het derde principe wordt gegeven door de relevante omgevingsnormen. Mensen weerspiegelen in hun gedrag de opvattingen en normen van hun omgeving. Dit principe verklaart waarom enquêtes waarin respondenten volgens een tweetrapsprocedure benaderd worden doorgaans een lagere respons hebben. Een tweetrapsprocedure wordt veel gebruikt in situaties waarin de privacy een belangrijke rol speelt, bijvoorbeeld omdat gebruik wordt gemaakt van adres-bestanden van derden. In zo'n geval is het gebruikelijk dat de beheerder van het adresbestand de doelgroep een brief stuurt met het verzoek om medewerking, met daarin een antwoordkaart waarop men kan aangeven of men wel of niet meewerkt. Hiermee wordt gecommuniceerd dat het niet meewerken aan het onderzoek een volstrekt normale zaak is. Dat de aangeschreven personen dit ook zo opvatten, is af te leiden uit het feit dat veel mensen wel degelijk de moeite nemen te reageren, maar dan wel met een weigering.

Het vierde principe is autoriteit. Mensen zijn eerder geneigd op een verzoek in te gaan, als dit verzoek afkomstig is van een gezaghebbende instantie, die het betreffende verzoek om gegronde redenen doet. In enquêtes weerspiegelt dit principe zich in de bevinding dat (semi-)overheidsinstanties doorgaans hogere responscijfers behalen dan marktonderzoekers.

Het vijfde principe is schaarsheid. Mensen gaan eerder in op verzoeken wanneer hun een zeldzame kans wordt geboden. Gegeven de grote hoeveelheid enquêteonderzoek die over Nederland wordt uitgestort is dit principe slechts beperkt van toepassing; alleen bij onderzoek van speciale groepen kan dit argument op geloofwaardige wijze worden gebruikt.

Het zesde principe is sympathie. Mensen gaan eerder in op verzoeken van personen of organisaties die men aardig vindt. Bij een enquête kan dit principe toegepast worden bij de presentatie van het onderzoek, maar ook door de interviewer. Interviewers die in kleding en optreden lijken op de respondent worden in het algemeen aardiger gevonden, en zullen daardoor een hogere respons bereiken.

Een meer specifieke invalshoek bij het onderzoeken van non-respons is het onderzoek naar de *interactie tussen interviewers en potentiële respondenten* op het moment dat het verzoek tot medewerking aan het onderzoek wordt gedaan. Bij deze interactie spelen de eerder genoemde zes principes natuurlijk een rol, maar ook meer algemene principes van sociale interactie, en daarmee ook de sociale vaardigheden waarover de interviewer beschikt. Een belangrijk gegeven is ook dat de beslissing van de respondent om wel of niet mee te werken aan het onderzoek doorgaans binnen één of twee minuten valt. Dat betekent dat er niet veel tijd is voor complexe afwegingen; het is waarschijnlijk dat de potentiële respondent beslist op een beperkt aantal gemakkelijk waar te nemen

aspecten, en daarbij geen ingewikkelde besliskundige regels volgt maar een simpele heuristiek. Onderzoek van interacties aan de deur (Morton-Williams, 1993) levert op dat de belangrijkste strategie voor interviewers is het inspelen op de potentiële respondent, en het voorkomen dat er definitief NEE wordt gezegd. Blijven praten, en vlot terugtrekken om het later nog eens te proberen is een betere strategie dan koste wat het kost door te zetten; een eenmaal geïncasseerde weigering alsnog omzetten in toestemming is buitengewoon moeilijk.

De derde theoretische invalshoek richt zich op het beslissingsproces bij de potentiële respondent. Een mogelijk model is het toepassen van *rationele keuze-theorie*. Rationele keuze-theorie gaat er van uit dat mensen een bepaald gedrag vertonen wanneer de opbrengst daarvan groter is dan de bijbehorende kosten. Een verwant model is de theorie van beredeneerd handelen, ook wel bekend als het Ajzen-Fishbein model. Dit model gaat er van uit dat het feitelijke handelen, in ons geval het wel of niet meewerken aan een enquête, geheel bepaald wordt door de intentie tot die handeling. Die intentie wordt op zijn beurt bepaald door de attitude van die persoon ten opzichte van de betreffende handeling, en opvattingen over sociale normen en de mate van controle die men over de betreffende handeling heeft. Door in te werken op de attitude of opvattingen is het in principe mogelijk de keuze voor bepaalde handelingen te beïnvloeden. De theorie van beredeneerd handelen is door Hox, De Leeuw en Vorst (1996) toegepast in een onderzoek naar respons bij postenquêtes. In dit onderzoek werd aan alle eerstejaars psychologiestudenten van de Universiteit van Amsterdam een vragenlijst afgenomen met vragen die betrekking hebben op centrale begrippen uit de theorie van beredeneerd handelen. Een half jaar later kregen alle studenten een postenquête toegezonden over studiegedrag en studievoortgang. Vervolgens is geprobeerd het wel of niet terugsturen van die enquête te voorspellen via het Ajzen-Fishbein model. De voorspelbaarheid van het responderen was helaas niet groot; de correlatie tussen de intentie en het responderen was slechts 0.24. Uit het model blijkt dat met name de specifieke kenmerken van de vragenlijst belangrijk zijn voor het responderen, en niet zozeer de globale attitude en opvattingen over enquêtes in het algemeen. Dat betekent, dat deze groep potentiële respondenten wel degelijk een beslissingsproces over het specifieke verzoek hebben doorlopen, en geen simpele heuristische regel hebben gevolgd. Bij een postenquête is dat uiteraard ook mogelijk, de beslissing om wel of niet mee te werken hoeft immers niet stante pede genomen te worden. Het zou aardig zijn, een dergelijk onderzoek te herhalen met een interview aan de deur of per telefoon.

Wat kunnen we uit het voorafgaande concluderen? Om te beginnen, dat de meest gebruikelijke manier om met non-respons om te gaan, namelijk bij de pakken blijven neerzitten, zeker niet de beste aanpak is. En vervolgens, dat er niet één enkele dramatische ingreep is waarmee de non-respons volkomen kan worden voorkomen. Er is eerder sprake van een losse verzameling tactieken

waarmee de respons kan worden verhoogd. Een goed theoretisch begrip van waardoor non-respons ontstaat is daarbij uiterst belangrijk. Wanneer 'zo maar' een aantal veelbelovende tactieken worden toegepast, dan is de kans groot dat ze elkaar tegenwerken. Het is belangrijk om na te gaan via welk mechanisme een bepaalde tactiek werkt, zodat de verschillende elementen in de benadering van de potentiële respondent elkaar versterken.

3.1.2 Wegen

Wanneer een steekproef door welke oorzaak dan ook niet representatief is voor de onderzochte populatie, dan is het mogelijk hiervoor te corrigeren door de steekproef te *wegen*. Bijvoorbeeld, wanneer wij een steekproef nemen uit de populatie van volwassen Nederlanders, dan weten wij dat we ongeveer evenveel mannen als vrouwen in onze steekproef zouden moeten hebben. Vinden we 40% mannen en 60% vrouwen, dan weten we dat onze steekproef wat betreft de variabele 'sekse' scheef is. Kennelijk hebben we een steekproefprocedure gevolgd waarin vrouwen een grotere kans hebben dan mannen om in de steekproef te worden opgenomen. Eén manier om dit probleem te ondervangen is vrouwen in de berekeningen een wat lager gewicht te geven dan mannen, waardoor de steekproef uiteindelijk statistisch wordt rechtgetrokken. Het berekenen van zulke gewichten is op zich niet ingewikkeld, er zijn relatief eenvoudige programma's voor beschikbaar (Bethlehem, 1996), en het analyseprogramma SPSS ondersteund de mogelijkheid. In ons voorbeeld zouden de mannen een weegfactor krijgen van 1.25, en de vrouwen 0.83. Wanneer we het gewogen percentage voor de variabele sekse berekenen, dan blijkt dit nu 0.5 te zijn. Door het wegen is de steekproef nu wel representatief geworden wat betreft de sekse, en we mogen hopen dat de steekproef ook meer representatief is geworden met betrekking tot andere variabelen, waarvoor we niet kunnen wegen omdat we de populatiegegevens niet kennen.

Ook met wegen is iets merkwaardigs aan de hand. Grote statistische bureaus, zoals in Nederland het CBS, gebruiken weging routinematig in vrijwel alle onderzoeken. Marktonderzoekers doen dat eveneens. Universitaire sociaal-wetenschappelijke onderzoekers doen dit, voor zover mijn blikveld reikt, vrijwel nooit. De standaard procedure bij universitair onderzoek is om het probleem van de niet-representatieve steekproef te negeren. Op zijn best wordt in een aparte paragraaf beschreven op welke onderdelen de steekproef afwijkt van de populatie, geconstateerd dat het allemaal wel meevalt, en vervolgens alsnog tot de orde van de dag overgegaan. Hier is sprake van een hoogst opmerkelijk verschil in de onderzoekspraktijk, waarvoor ook geen methodologische rationale gegeven kan worden, anders dan het bestaan van verschillende onderzoekstradities in verschillende organisaties.

Op zich is het niet corrigeren van de steekproef nog te verdedigen ook, al maak ik mij sterk dat de meeste sociale wetenschappers van deze verdediging niet op de hoogte zijn. Bij het CBS en

vergelijkbare instellingen is een van de belangrijkste doelstellingen het *beschrijven* van de populatie wat betreft een aantal kernvariabelen. Daarvoor is een goede steekproef uitermate belangrijk. In het universitaire onderzoek houden we ons meer bezig met het toetsen van theorieën, althans, zo zouden wij dat graag willen. Een theorie doet uitspraken over samenhangen tussen een stelsel van variabelen, en in de moderne onderzoekspraktijk wordt zoiets bij voorkeur getoetst door het opstellen en schatten van een statistisch model. Het geval wil, dat wanneer we een statistisch correct model schatten, dat dan de precieze samenstelling van de steekproef minder belangrijk is. De gedachte is dat door selectieve non-respons de gevonden gemiddelden wellicht vertekend zullen zijn, maar dat de modelparameters zoals regressiecoëfficiënten minder vertekend zullen zijn. Zolang we voornamelijk geïnteresseerd zijn in modelparameters lijkt er dan weinig aan de hand.

Het probleem bij deze redenering, die door Groves (1989) uitvoerig uiteengezet wordt, is dat de robuustheid van de model-resultaten alleen opgaat wanneer we een correct model schatten. Wanneer het model dat we schatten niet correct is, dan gaat het hele verhaal niet op. Zolang we niet op voorhand weten dat al onze modellen correct zijn (bijvoorbeeld bij correlaties het homoscedastische lineaire model), is het verstandig om de resultaten ook op basis van een gewogen steekproef te berekenen. Een opvallend verschil tussen de ongewogen en de gewogen resultaten kan dan opgevat worden als een indicatie van een specificatiefout in het model, met andere woorden: het model is dan onjuist of onvolledig.

3.2 Van theoretisch begrip naar interviewvraag

Het tweede voorbeeld uit de dataverzameling gaat over de route van theoretisch begrip naar interviewvraag. Hierbij doen zich allerlei methodologische keuzen voor, variërend van het verzinnen van de vraag op zich tot het beslissen of het antwoord op een vijf- of een zeven-punts schaal gemeten moet gaan worden. Voor de meeste van deze beslissingen geldt dat ze vallen onder de creatieve vrijheid van de onderzoeker. Wetenschapsfilosofen formuleren dit door te stellen dat zij vallen binnen de context van de theorievorming en niet de context van de bewijsvoering. In de context van de theorievorming zijn onderzoekers vrij in hun beslissingen, en hoeft zich van de raadgevingen van andere onderzoekers, inclusief methodologen en wetenschapsfilosofen, niets aan te trekken.

Dat neemt niet weg dat er nog steeds goede en minder goede beslissingen zijn. Bovendien is er door methodologen veel onderzoek gedaan naar het construeren en afnemen van vragenlijsten. Ik zal het proces van theoretisch begrip naar vraag in een vragenlijst hier nalopen, en de verschillende stappen van methodologische kanttekeningen voorzien.

Allereerst moet de vraag *bedacht* worden. Stel dat we iets willen meten, bijvoorbeeld

egoïsme, of pesten in de klas, en we besluiten dit te gaan meten met een vragenlijst. De eerste stap is het zorgvuldig uitwerken van het theoretisch begrip, en het verzinnen van vragen die daarvan een goede indicator zouden moeten zijn. Daar zijn geen vaste regels voor, maar er zijn wel degelijk strategieën aan te geven die ons helpen dit systematisch aan te pakken. Deze strategieën zijn globaal onder te verdelen in strategieën die beginnen bij de beoogde respondenten, en strategieën die beginnen bij het beoogde begrip. Bij strategieën die beginnen bij de respondenten moet U denken aan kwalitatieve onderzoeksmethoden, zoals een open interview, om te achterhalen wat de respondenten zelf denken bij bepaalde begrippen, en in welke termen zij daarover praten. Op basis van dit uitgangsmateriaal kunnen vervolgens enquêtevragen bedacht worden. Strategieën die beginnen bij het beoogde begrip beginnen doorgaans met het betreffende theoretische begrip op een systematische wijze uiteen te leggen in deelbegrippen of subdomeinen, waarna per deelbegrip vragen bedacht worden. Het aardige is nu dat dit soort strategieën niet alleen beschreven en aanbevolen kunnen worden (Hox, 1997), maar dat er verschillen lijken te zijn in de betrouwbaarheid en validiteit van de langs die weg geconstrueerde instrumenten. In een recente dissertatie aan de Universiteit van Amsterdam (Oosterveld, 1996) blijken twee theorie-georiënteerde strategieën (construct methode en de facet methode) het beter te doen dan twee data-georiënteerde strategieën (de prototypische en de externe methode). Of deze bevinding generaliseerbaar is, moet nog blijken, maar het intrigerende is dat er op dit terrein verschillende strategieën voorgesteld kunnen worden die voldoende duidelijk zijn dat ze in een vergelijkend experiment kunnen worden opgenomen, en dat er tussen zulke strategieën kennelijk systematische kwaliteitsverschillen kunnen bestaan.

De genoemde strategieën leiden niet direct tot een kant en klare vraag. Eerder is er sprake van een prototype van een vraag. Voor de uiteindelijke vraagformulering dienen de onderzoekers zich goed rekenschap te geven van de doelgroep aan wie de vraag gesteld gaat worden. Kinderen in groep zes van de basisschool moeten allicht anders benaderd worden dan professoren. Daarnaast moeten nog allerlei schijnbaar ondergeschikte beslissingen genomen worden, zoals de formulering: wordt het een vraag of een stelling? en het aantal antwoordmogelijkheden: even? oneven? met of zonder een categorie 'weet niet'? Voorts hebben we nog de keuze tussen een vragenlijst en een interview, al dan niet per telefoon, en de vraag of we de computer inschakelen bij het interview-proces.

Over al dit soort vragen kan eindeloos lang gedebatteerd worden. Veel van dit soort debatten zijn echter overbodig. Er is een enorme hoeveelheid methodologisch onderzoek gedaan naar het effect van variaties in de interviewprocedures en de vraagformulering op de betrouwbaarheid en validiteit van de verkregen antwoorden. De hoeveelheid onderzoek is omvangrijk genoeg, dat het zinvol is om meta analyse te gebruiken om de resultaten samen te vatten. Er loopt al meerdere jaren een groot internationaal samenwerkingsproject (vgl. Saris & Van Meurs, 1989) dat zich richt op de systematische evaluatie van meetinstrumenten, langs de weg van geplande meta-analyses van

experimentele studies. Op veel van de vragen waarop vragenlijstmakers hun tanden stukbijten, zijn de antwoorden in grote lijnen bekend.[1] Bijvoorbeeld, het aantal antwoordcategorieën. Er is een omvangrijke onderzoeksliteratuur op dit gebied, met als algemene uitkomst: hoe meer hoe beter, maar er is sprake van een soort wet van de verminderde meeropbrengst: meer dan zeven antwoordcategorieën levert naar verhouding weinig winst op. Een ander voorbeeld is het wel of niet opnemen van een neutrale middencategorie bij de antwoordalternatieven. Sommige onderzoekers doen dit bij voorkeur niet, met als argument dat de respondent zich op deze manier niet kan verschuilen, maar gedwongen wordt om een herkenbaar positief of negatief standpunt in te nemen. Andere onderzoekers vinden dit juist een nadeel; respondenten die écht een neutrale houding hebben, worden op deze manier gedwongen een positie in te nemen die zij in feite niet hebben. Methodologisch onderzoek naar de effecten van het wel of niet opnemen van een neutrale middencategorie levert wisselende uitkomsten op. De resultaten wijzen voornamelijk in een richting van: baat het niet, het schaadt ook niet. Met andere woorden, àls er effecten worden gevonden, gaan die vooral in een richting die wijst op voordelen van het opnemen van een middencategorie. De pragmatische oplossing is dus standaard een middencategorie te gebruiken, tenzij er heel duidelijke redenen zijn om dat niet te doen.

Ongeacht hoe de onderzoekers de vraag uiteindelijk hebben geformuleerd, voor een valide antwoord is het nodig dat een respondent de vraag verstaat en begrijpt op de manier die de onderzoekers bedoeld hebben. Dat is lang niet vanzelfsprekend. Reeds in de jaren zeventig deed Belson (Belson, 1981) onderzoek naar de vraag of respondenten bij het beantwoorden van een enquêtevraag wel in hun hoofd hebben wat de onderzoekers bedoelen. Zijn resultaten zijn bepaald schokkend; bij een analyse van 29 met opzet enigszins moeilijk gemaakte vragen, afgenomen aan 265 respondenten, gingen gemiddeld slechts 29% van de antwoordgevers uit van de door de onderzoekers bedoelde interpretatie. U moet dan denken aan vragen als 'Denkt U dat geweld op de televisie een negatieve invloed heeft op jongeren?' Bij doorvragen bleek dat de interpretatie van wat 'geweld' was nogal verschilde; bijvoorbeeld: horen de vechtpartijen in Donald Duck en andere tekenfilms er nu wel of niet bij? Ook het begrip 'jongere' was niet zo eenvoudig; er was een tachtigjarige respondent voor wie iedereen onder de veertig daar onder viel.

Het is al heel lang een goede gewoonte om een nieuwe vragenlijst in een vooronderzoek uit te testen. Aan het arsenaal van methoden om enquêtevragen aan de tand te voelen is sinds enige tijd een nieuwe verzameling methoden toegevoegd, ontleend aan de cognitieve psychologie. Gemeenschappelijk aan deze methoden, die bekend staan onder de verzamelnaam *cognitive*

[1] Veel relevante informatie over dit soort beslissingen is te vinden in de overzichtsartikelen van Alwin en Krosnick (1990), Andrews (1984), Krosnick & Fabrigar (1997), Rodgers, Andrews & Herzog (1992), en Scherpenzeel & Saris (1997).

laboratory methods, is dat ze uitgaan van een model voor het vraag-antwoordproces in enquêtes, en dat ze toegepast worden met het doel om betere vragen te ontwikkelen (Forsyth & Lessler, 1991). Het gehanteerde vraag-antwoordmodel is doorgaans vrij simpel: de respondent moet de vraag begrijpen, het antwoord uit het geheugen ophalen, en besluiten hoe het antwoord op de gestelde vraag uiteindelijk geformuleerd gaat worden. In elk van deze fasen kan iets mis gaan, en er zijn verschillende methoden beschikbaar om te onderzoeken wat er mis gaat en waarom. Voorbeelden van zulke methoden zijn kwalitatieve interviews, respondenten hardop laten denken terwijl ze de vragen beantwoorden, coderen van respondent- en interviewergedrag, meten van de tijd die nodig is om vragen te beantwoorden, en nog vele andere (Forsyth & Lessler, 1991, p397). Het soort problemen waar Belson (1981) op wijst, kunnen met cognitive laboratory methoden uitstekend aangepakt worden.

Cognitive laboratory methoden kunnen toegepast worden om nieuwe vragen te testen, maar ook om oude vragen te verbeteren. Wanneer bijvoorbeeld bekend is, dat een bepaalde vraag een groot percentage weigeringen oproept, en daardoor tot onvolledige gegevens leidt, dan kan geprobeerd worden met cognitieve methoden te achterhalen waarom de betreffende vraag zo problematisch is, en om een betere formulering te vinden die tot minder weigeringen leid. Met dit voorbeeld is de cirkel rond; we zijn immers weer terug bij de onvolledige gegevens waar ik het eerder ook al over had. En aangezien ook in de methodologie voorkomen beter is dan genezen, zal het duidelijk zijn dat op de lange duur het ontlokken van antwoorden aan respondenten mijn voorkeur heeft boven het inschatten ervan.

4. Methodologie en sociaal-wetenschappelijk onderzoek

Aan het begin van deze rede heb ik beloofd dat ik nog terug zou komen op de vraag wat de relatie moet zijn van individuele onderzoekers met methoden en technieken, en of een sociaal-wetenschappelijke faculteit op dat terrein een voorziening in stand zou moet houden. Het is tijd deze belofte in te lossen.

Voor de individuele, inhoudelijk georiënteerde onderzoeker, zou je kunnen redeneren dat die moet proberen naast de ontwikkelingen op het eigen vakgebied ook de methodologische en statistische ontwikkelingen bij te houden. Dat is een strategie die ik kan afraden. Aan de Erasmus Universiteit te Rotterdam is een centrum verbonden, Social Research Methodology, dat de sociaal-wetenschappelijke methodologische literatuur catalogiseert. Aan de CD-ROM waarop de betreffende literatuur te vinden is worden per jaar meer dan 2000 titels toegevoegd. Zelfs wanneer maar 10% hiervan de moeite waard is, dan nog zou men per jaar 200 publikaties moeten lezen om methodologisch bij te blijven. Dat is duidelijk niet te doen.

De strategie die ik inhoudelijk georiënteerde onderzoekers zou willen aanraden kan worden toegelicht aan de hand van drie sleuteltermen: *accepted good practice*, *current best methods*, en *state of the art*.

Geaccepteerde goede praktijken zijn die methoden en technieken welke al jaren in gebruik zijn, en waarvan methodologen en onderzoekers inmiddels de sterke en zwakke punten kennen. Dikwijls zijn het ook de methoden en technieken die de kern van het universitaire onderwijs vormen, en waarvoor goede software beschikbaar is. En, heel belangrijk: ze zijn ook bij tijdschrift-redacties en reviewers bekend en geaccepteerd. Door je primair van deze technieken te bedienen, kan je je als auteur eigenlijk geen buil vallen.

Current best methods zijn die methoden en technieken welke bij de methodologen goed bekend zijn, waarvan de methodologen zich inmiddels een redelijk beeld van de sterke en zwakke punten hebben gevormd, die dikwijls een verbetering zijn van de geaccepteerde standaardpraktijken, maar die doorgaans bij de inhoudelijke onderzoekers nog weinig worden toegepast. Het zijn het soort methoden waarbij inhoudelijk georiënteerde onderzoekers er verstandig aan doen om een methodoloog te consulteren, of samenwerking met een methodoloog te zoeken, al was het alleen omdat de standaard software deze methoden dikwijls nog niet ondersteunt.

State of the art methoden zijn die methoden waar de methodologen zelf nog volop mee in de weer zijn. In de methodologische tijdschriften wordt nog druk gepubliceerd over verbeteringen, variaties en alternatieven, en bij consultaties krijgt U over deze methoden van geen twee methodologen hetzelfde advies. Voor state of the art methoden geldt de klassieke zegswijze: *caveat emptor*, ofwel: de koper moet hier verschrikkelijk goed oppassen. In goed Hollands: garantie tot de deur, en misschien niet eens tot daar.

Mijns inziens zou in kwalitatief hoogwaardig universitair onderzoek zoveel mogelijk gebruik moeten worden gemaakt van current best methods. Deze term komt overigens niet uit de lucht vallen. Het is afkomstig uit de managementsfilosofie waarin gestreefd wordt naar een voortdurende bewaking en verbetering van de kwaliteit van het geleverde product. Dat gebeurt niet zozeer door het controleren van het eindproduct, maar door het controleren en verbeteren van het productieproces (Deming, 1982). Het sleutelwoord hierbij is Total Quality Management; het doel is de kwaliteit van alle deelprocessen te bewaken en te verbeteren. Current best methods spelen bij die kwaliteitsverbetering een belangrijke rol, omdat ze vastleggen en documenteren wat op dit moment de beste manier is om bepaalde doelen te bereiken.[1]

[1] Deze voorkeur voor current best methods wordt deels ingegeven door pragmatische overwegingen, zoals de beperkte tijd die inhoudelijk georienteerde onderzoekers hebben om zich te verdiepen in methodologische finesses, en het probleem, dat inhoudelijke tijdschriften nog wel eens huiverig zijn voor al te complexe methoden en analyses. Snijders (persoonlijke mededeling, 1997) heeft mij er op gewezen dat in een gegeven geval een state of the art methode absoluut de beste kan zijn, en dat juist de interactie met serieuze toepassingen de state of the art beproeft en verder brengt. Ik kan het hier geheel mee eens zijn, maar inhoudelijke

Een belangrijk aspect van current best methods is dat ze *current* zijn; ze omvatten de methoden die op het huidige moment de beste zijn. Daaruit spreekt de verwachting, dat ze op een gegeven moment vervangen zullen worden door nieuwe en betere methoden. Wat nu state of the art is, kan morgen current best method zijn, en ooit eindigen als een ouderwetse geaccepteerde standaardmethode. Een voorbeeld van zo'n beweging is gemakkelijk te geven. Het gebruik van multiniveau analyse was zo'n tien jaar geleden voorbehouden aan de experts, slechts enkele statistici en methodologen hielden zich hier mee bezig. Tegenwoordig vermoed ik dat een artikel dat een 'platte' analyse toepast op multiniveau-gegevens door een tijdschrift als het Tijdschrift voor Onderwijsresearch of Pedagogische Studiën niet meer geaccepteerd wordt. Het is de huidige beste methode, en hard op weg om standaardmethode te worden.

Wellicht dat sommige onderzoekers hun werk beleven als zo uniek en creatief dat het door het bewust selecteren van optimale methoden en technieken alleen maar in zijn mogelijkheden beknot kan worden. Voor die onderzoekers geldt het citaat 'In het rijk der geest is de methode als een kruk; de ware denker loopt vrij.' Een mooi citaat, maar voor wie het niet kent: het is afkomstig uit de kolderdetective *Bill Clifford* van Godfried Bomans.

Nadenken over de current best methods is daarbij niet alleen de taak van de inhoudelijke onderzoekers. Een belangrijk kenmerk van current best methods is dat ze *beschreven* zijn. Er moet gedocumenteerd zijn om welke methoden het gaat, en er moet gedetailleerd uitgelegd worden waarom en hoe die toegepast moeten worden. Vertaald naar de universitaire situatie betekent dit dat methodologen de moeite moeten nemen om wat zij als veel belovende methoden beschouwen op begrijpelijke wijze uit te leggen. Current best methods zouden ook meer dan nu het geval is aan bod moeten komen in met onderwijs in methodenleer en statistiek, bij voorkeur al in het eerste fase-onderwijs, maar zeker in de onderzoekersopleidingen. Wat dat betreft is het enigszins kortzichtig dat onderzoeksscholen het produceren van goed en vernieuwend onderwijsmateriaal niet als wetenschappelijke productie wensen te beschouwen. De theorieën en onderzoeksmethoden die jonge onderzoekers opdoen vormen het fundament voor de gehele verdere onderzoekersloopbaan. Wat dat betreft verwijs ik naar Max Planck, die gezegd heeft dat belangrijke vernieuwingen, waaronder wat mij betreft ook onderzoeksmethoden vallen, doorgaans niet algemene ingang vinden doordat gereputeerde onderzoekers er zich meester van maken. Integendeel, vernieuwing komt van de kant van de jonge onderzoekers, die van meet af aan met de nieuwe ideeën vertrouwd zijn geraakt (Mackay, 1991, p195).

---

onderzoekers dienen zich te realiseren dat ze zich potentieel problemen op de hals halen.

5. Tot slot

Zoals eerder gezegd, wat vandaag state of the art is, kan morgen current best method zijn, en overmorgen alweer ouderwets. Een grote sociale faculteit, die daartoe de middelen heeft, heeft daarom een groep onderzoekers nodig die zich bezig houden met het verder ontwikkelen van de state of the art op het gebied van methodenleer en statistiek. Soms kan dat wel eens de indruk wekken van luchtfietserij, maar sommige van de onderzochte methoden zullen zich ooit ontwikkelen tot standaardinstrumenten in de methodologische gereedschapkist. Het een en ander impliceert mijns inziens de noodzaak van een opleiding methodenleer en statistiek. Immers: waar denkt U dat jonge methodoloogjes vandaan komen? Anderzijds heeft de ontwikkeling van methodenleer en statistiek heeft er ook baat bij ingebed te zijn in een brede faculteit. Veel van de methodologische en statistische vernieuwingen zijn gestimuleerd doordat methodologen uit verschillende disciplines met elkaar in aanraking kwamen. Bovendien: wat in de ene discipline accepted good practice is, kan in een andere discipline volledig onbekend zijn, en de moeite waard om daar als current best method te propageren. De diversiteit aan praktische vragen die op je af komen is daarbij een stimulans om weer over fundamentele methodologische problemen na te denken. Ik ben daarom bijzonder erkentelijk, dat de faculteit der sociale wetenschappen van de Universiteit Utrecht mij in de gelegenheid heeft gesteld, mijn methodologische onderzoek en onderwijs in deze gevarieerde en stimulerende omgeving uit te voeren, temidden van een groep collega-methodologen die naast enige luchtfietserij niet bang zijn hun handen vuil te maken aan de problemen van hun inhoudelijk georiënteerde collega's. Ik verheug me in een plezierige samenwerking met mijn Utrechtse collega's.

Mijnheer de rector, dames en heren.

Eén belangrijk aspect van de wetenschap heb ik nog niet genoemd. Het beoefenen van de wetenschap verschaft niet alleen brood op de plank en enig maatschappelijk aanzien, het is boven alles leuk om te doen en spannend. Ik bedank mijn leermeesters die mij dit zo geleerd hebben en mijn collega's bij wie ik gelukkig dezelfde houding aantref. Onder die collega's reken ik overigens uitdrukkelijk de promovendi. Ik ga geen namen noemen, dat zouden er teveel worden. Tenslotte bedank ik mijn studenten voor de uitdaging die ze mij bieden om duidelijk te maken waarom onderzoek doen, inclusief het je meester maken van de benodigde methodenleer en statistiek, niet alleen moeilijk, maar ook leuk en spannend is.

Ik heb gezegd.

Referenties

Alwin, D.F. & Krosnick, J.A. (1990). The reliability of attitudinal survey measures: the role of question and respondent attributes. *Sociological Methods & Research*, 20, 139-181.

Andrews, F.M. (1984). Construct validity and error components of survey measures: a structural modeling approach. *Public Opinion Quarterly*, 48, 409-422.

Arbucle, J. (1996). Full information estimation in the presence of missing data. In G.A. Marcoulides & R.E. Schumacker (eds). *Advanced structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Belson, W.A. (1981). *The design and understanding of survey questions*. Aldershot: Gower.

Bethlehem, J. (1996). *Bascula for weighting sample survey data*. Voorburg/Heerlen: CBS.

Bethlehem, J.G. & Kersten, H.P.M. (1986). *Werken met non-respons*. Dissertatie, Universiteit van Amsterdam.

Bollen, K.A. & Stine, R.A. (1992). Bootstrapping goodness-of-fit measures in the context of structural equation models. *Sociological Methods & Research*, 21, 205-229.

Boomsma, A. (1991). *BOJA. A program for bootstrap and jackknife analysis*. Groningen: ProGamma.

Cialdini, R.B. (1988). *Influence: science and practice*. Glenview, IL: Scott & Foresman.

Deming, W.E. (1982). *Quality, productivity, and competetive posiition*. Cambridge, MA: MIT.

Dingle, P. (1994). Informative dropout in longitudinal data analysis. *Applied Statistics*, 43, 49-93.

Efron, B. & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Feldt, L.S. (1965). The approximate sampling distribution of Kuder-Richardson coefficient twenty. *Psychometrika*, 30, 357-370.

Feldt, L.S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*, 34, 363-373.

Forsyth, B.H. & Lessler, J.T. (1991). Cognitive laboratory methods: a taxonomy. In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz & S. Sudman (eds.) *Measurement errors in surveys*. New York: Wiley.

Groves, R.M. (1989). *Survey errors and survey costs*. New York: Wiley.

Groves, R.M., Cialdini, R.B. & Couper, M.P. (1992). Understanding the decision to participate in a survey. *Public Opinion Quarterly*, 56, 475-495.

Hakstian, A.R. & Whalen, T.E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika*, 41, 219-231.

Heberlein, T.A. & Baumgartner, R. (1978). Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature. *American Sociological Review*, 43, 447-462.

Heijden, P.G.M. van der, 't Hart, H. & Dessens, J. (1997). A parametric bootstrap procedure to perform statistical tests in a LCA of anti-social behavior. In J. Rost & R. Langeheine (eds.) *Applications of latent trait and latent class models in the social sciences*. Münster/New York: Waxmann.

Heer, W. de (1996). *International response trends: development and results of an international survey*. Paper, 4th International Social Science Methodology Conference, Esses, UK.

Hox, J.J. (1986). *Het gebruik van hulptheorieën bij operationalisering*. Dissertatie, Universiteit van Amsterdam.

Hox, J.J. (1997). From theoretical concept to survey question. In: L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz & D. Trewin (eds.) *Survey measurement and process quality*. New York: Wiley.

Hox, J.J., de Leeuw, E.D & Vorst, H. (1996). A reasoned action explanation for survey nonresponse. In S. Laaksonen (ed.). *International perspectives on nonresponse*. Helsinki: Statistics Finland.

Kalfs & Kool (1994). *Ervaringen met nonrespons*. Amsterdam: NIMMO.
Krosnick, J.A. & Fabrigar, L.R. (1997). Designing rating scales for efefctive measurement in surveys. In: L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz & D. Trewin (eds.) *Survey measurement and process quality*. New York: Wiley.
Leeuw, E.D. de & Hox, J.J. (1985). Recente oontwikkelingen bij schriftelijk onderzoek per post. In: *Jaarboek van de Nederlandse Vereniging van Marktonderzoekers, 1985*. Haarlem: De Vrieseborgh.
Lessler, J.T. & Kalsbeek, W.D. (1992). *Nonsampling error in surveys*. New York: Wiley.
Little, R.J.A. & Rubin, D.B. (1987) *Statistical analysis with missing data*. New York: Wiley.
Mooney, C.Z. & Duval, R.D. (1993). *Bootstrapping. A nonparametric approach to statistical inference*. Newbury Park, CA: Sage.
MacKay (1991). A dictionary of scientific quotations. Bristol: Adam Hilger.
Morton-Williams, J. (1993). *Interviewer approaches*. Aldershot, UK: Dartmouth.
Noreen, E.W. (1989). *Computer-intensive methods for testing hypotheses: An introduction*. New York: Wiley.
Oosterveld, P. (1996). *Questionnaire design methods*. Nijmegen: Berkhout.
Rodgers, W.L., Andrews, F.M. & Herzog, A.R. (1992). Quality of survey measures: a structural modeling approach. *Journal of Official Statistics*, 8, 251-275.
Saris, W.E. & Van Meurs, A. (eds) (1989). *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies*. Amsterdam: North-Holland.
Schafer, J. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
Schell, (1997). *Nonresponse in Bevölkerungsumfragen*. Opladen: Leske und Budrich.
Scherpenzeel, A.C. & Saris, W.E. (1997). The validity and reliability of survey questions: a meta-analysis of MTMM studies. *Sociological Methods & Research*, 25, 341-383.
Van Rooy, C. (1986). Responsvoorspellingen: toverformules of realisme? In: *Jaarboek van de Nederlandse Vereniging van Marktonderzoekers, 1986*. Haarlem: De Vrieseborgh.
Verleye, G. (1996). *Missing at random data problems in attitude measurement using maximum likelihood structural equation modelling*. Dissertatie, Vrije Universiteit Brussel.
Vermunt, J. (1996). *Log-linear event history analysis: a general approach with missing data, unobserved heterogeneity, and latent variables*. Tilburg: Tilburg University Press.