HOE TE LEVEN MET STATISTIEK

Rede, uitgesproken bij de aanvaarding
van het ambt van gewoon hoogleraar
in 'de statistiek ten behoeve van de sociale
wetenschappen',
aan de Universiteit Utrecht
op vrijdag 17 juni 1994

door

Peter G.M. van der Heijden

uni 1994

ıctie-afdeling Faculteit Sociale Wetenschappen

ISOR Universiteit

Het is nu veertig jaar geleden dat het boek "How to lie with statistics" van Darrell Huff[1] verscheen. De titel van het boek is waarschijnlijk geinspireerd door een uitspraak van de staatsman Benjamin Disraeli, die ooit beweerde: "there are three kinds of lies: lies, damned lies, and statistics".[2] Het boek wees het grote publiek op het verkeerde gebruik van statistiek in, bijvoorbeeld, krantenartikelen, tijdschriftartikelen en advertenties. Er wordt bijvoorbeeld op gewezen dat steekproeven representatief behoren te zijn en dit vaak niet zijn; dat het duidelijk moet zijn of men bij het woord gemiddelde het rekenkundig gemiddelde, de modus of de mediaan bedoelt; en bijvoorbeeld dat bij rapporteren van gegevens uit heel kleine steekproeven gecheckt hoort te zijn of de resultaten significant zijn. Velen onder u kunnen zich allerlei fouten of onjuiste voorstellingen van zaken indenken, en hebben die zelf ook wel eens gezien.

Op sommige punten is dit boek heel duidelijk verouderd. Door de activiteiten van de Consumentenbond worden steeds minder vaak misleidende voorstellingen in advertenties gegeven. Ook wordt er veel vaker dan 40 jaar geleden op significantie getoetst. Ik zal hier later nog op terugkomen. Op andere punten zou het boek nog steeds actueel kunnen zijn: zo heb ik mijn collega Engbersen wel eens horen klagen over CBS-statistieken op het gebied van armoede, waar de volgens hem door selectieve nonrespons echt armen niet in meegenomen zouden zijn.[3] Het is aardig te weten dat dit onderwerp reeds als eerste voorbeeld wordt behandeld in het boek van Huff .

De engelse term "statistics" heeft een dubbele betekenis: enerzijds betekent het "het vakgebied statistiek", en anderzijds verwijst het naar het meervoud van een statistiek, een getal dat volgens een bepaald voorschrift is berekend en een bepaalde kansverdeling volgt. De Nederlandse vertaler van het boek van Huff, Mr. H.A.M. van der Heyden, geen familie overigens, vond dit een probleem bij de vertaling van de titel van Huff´s boek: hij koos uiteindelijk voor de titel "Gebruik en misbruik van de statistiek".[4] Zo zijn we ook die scherpe term "liegen" uit de titel kwijt. Ik houd niet zo van het gebruik van dit soort grote woorden met ethische implicaties, en

sluit me daarom in deze oratie graag bij mijn naamgenoot's vertaling aan. Onder misbruik van statistiek versta ik dan zowel bewust als onbewust verkeerd gebruik van de statistiek. Ik wil in deze oratie aandacht schenken aan enkele voorbeelden van nieuwere vormen van gebruik en misbruik van statistiek. Ook wil ik enkele gedachten aan u voorleggen hoe men hiermee zou kunnen leven. Ik begin met het formuleren van mijn visie op de hedendaagse wetenschapsbeoefening in de sociale wetenschappen. Daarna bespreek ik enkele voorbeelden van mijn onderzoek, die de belofte in zich dragen een oplossing te bieden voor bestaand misbruik van statistiek.

Het gebruik van statistiek in sociaal wetenschappelijk onderzoek houdt mij al lange tijd bezig, zowel in gedachten als in tijdbesteding. Ik heb als student-assistent nog meegedraaid in inhoudelijk sociaal-wetenschappelijk onderzoek, en verrichtte daar de data-analyses. Later kwam ik vooral met inhoudelijk onderzoek in aanraking in het kader van de statistische consultatie. Wat mij al die jaren steeds is opgevallen is dat zo veel wetenschappelijk onderzoek niet volgens de regels verloopt zoals dat volgens de gangbare handboeken in de methodologie zou moeten. De regels schrijven voor dat wij op een rationele en rechtlijnige manier van de oorspronkelijke vraagstelling tot de rapportage van de onderzoeksgegevens komen. Zo onderscheidt bijvoorbeeld de Groot in zijn beroemde standaardwerk getiteld "Methodologie" de volgende fasen van de empirische cyclus[5]: observatie (het verzamelen en groeperen van empirisch feitenmateriaal), inductie (het formuleren van hypothesen), deductie (afleiding van speciale consequenties uit de hypothesen in de vorm van toetsbare voorspellingen), toetsing (van de hypothesen) en evaluatie (van de uitslag van de toetsing, in termen van de hypothese, of de theorie, in verband met mogelijke nieuwe, aansluitende onderzoekingen).

In welke zin wijkt de huidige onderzoekspraktijk daar dan van af? In de statistische consultatie merk ik dat de onderzoekscyclus in de praktijk eerder chaotisch dan rationeel lijkt. Ook zijn er dikwijls geen duidelijke fasen in het onderzoek te ontdekken, of ontbreken er zelfs fasen. Ik neem meestal voor het eerst kennis van een onderzoeksproject nadat de data reeds zijn verzameld. Het komt dan zeer vaak voor dat de onderzoeker en ik in gezamelijk overleg besluiten de

oorspronkelijke vraagstelling een specifiekere invulling te geven, of zelfs te veranderen. Een reden hiervoor kan zijn dat er met de data meer beantwoord kan worden dan de onderzoeker besefte, omdat het statistisch instrumentarium uitgebreider is dan de onderzoeker wist bij het formuleren van de vraagstelling. Soms kan de oorspronkelijke vraagstelling niet beantwoord worden omdat het statistisch instrumentarium niet beschikbaar is, en ook niet eenvoudig te ontwikkelen is. Soms komen onderzoekers pas langs voor consultatie nadat zij zelf geruime tijd bezig zijn geweest met het analyseren van hun data, en dan door de bomen het bos niet meer zien. Soms echter heeft men al analyserend de oorspronkelijke vraagstelling beter leren doorgronden. In deze gevallen heeft de oorspronkelijke vraagstelling meestal te weinig houvast geboden om rechtlijnig te werk te gaan, omdat de vraagstelling bijvoorbeeld nogal exploratief van aard was. Meestal zijn enorme pakken computeroutput wel te reduceren tot enkele analyses, die antwoord lijken te geven op de kern van de onderzoeksvraagstelling. En dan komt de fase van het onderzoek waarover ik mij het meest verwonder: de rapportage. In de fase van de rapportage wordt de relatieve chaos weer gladgestreken tot een rationeel onderzoeksproces. De onderzoeksresultaten worden het eerst beschreven, dan wordt de vraagstelling zo beschreven dat de onderzoeksresultaten hierop exact aansluiten. Men doet alsof men zich geheel volgens het boekje heeft gedragen. Ik wil hier benadrukken dat er natuurlijk chaotisch handelen is dat het gevolg is van de ondeskundigheid van een onderzoeker. Echter, ik wil ook benadrukken dat deskundige onderzoekers evenzeer grote moeite hebben zich aan de empirische cyclus te houden.

Een eerste vraag die men zich zou kunnen stellen is: hoe vaak gebeurt dit nu? Ik heb dit nooit bijgehouden, maar mijn indruk is dat dit geldt voor de bulk van het sociaal-wetenschappelijke onderzoek dat mij in de consultatiefase onder ogen komt. Dit onderzoek is zeker niet representatief voor al het sociaal wetenschappelijke onderzoek dat wordt uitgevoerd. Ik zie bovendien vooral veel survey-onderzoekers en mijn ervaring is dat in survey-onderzoek de vraagstellingen een stuk minder duidelijk geformuleerd zijn dan in, bijvoorbeeld, experimenteel onderzoek. Kwalitatieve onderzoekers zie ik helemaal niet, maar door gesprekken die ik van tijd

tot tijd met sommigen van hen heb, heb ik de indruk dat ook voor hen geldt dat gedurende het onderzoek de vraagstelling wordt aangepast aan de opgedane ervaringen. Al met al denk ik dat wij hier praten over een probleem dat niet verwaarloosbaar is.

Een volgende vraag is dan, hoe erg is dit eigenlijk? Hoe erg is het indien onderzoekers chaotisch te werk gaan, en zich niet houden aan de strikte regels van de empirische-cyclus methodologie, om vervolgens in tijdschriften het onderzoeks-proces veel rationeler voor te stellen dan het daadwerkelijk is? Eerst de vraag naar het chaotisch te werk gaan. Laat ik beginnen met te stellen dat het mij niet wenselijk lijkt te pogen chaotisch handelen uit te bannen. Veel interessante vondsten zijn bij toeval gedaan, en niet omdat men zo rationeel te werk ging. Het lijkt me dus ook niet per definitie wenselijk in alle gevallen te streven het boekje te volgen. Met andere woorden, het strikt volgen van de empirische cyclus ontneemt ons vele mogelijkheden interessante en belangwekkende resultaten te vinden. Daarnaast heb ik zeer grote twijfels of het uberhaupt mogelijk is door een betere opleiding van onderzoekers en goede begeleiding chaotisch werk uit te bannen.

Hoe erg is het dan dat in tijdschriften het onderzoeksproces geregeld veel rationeler wordt voorgesteld dan het daadwerkelijk is? Dit is eigenlijk een vreemde gang van zaken, waar ook door anderen van tijd tot tijd, en in verschillende bewoordingen, op is gewezen. Ik vind dit voor de wetenschap niet zo erg, maar het is wel vervelend dat weinigen weten dat wetenschap zo chaotisch verloopt. Om met dit laatste te beginnen: beginnend onderzoekers voelen zich hierdoor ten onrechte ongelukkig, schuldig, of dom, of een slecht mens doordat zij hun chaotische werk gladstrijken in de rapportage van hun onderzoek. Het is de dwang van tijdschrift-redacties die er voor zorgt dat vrijwel niemand het zich kan permiteren niet te conformeren.

Voor het wetenschappelijk bedrijf is het verder niet zo erg. Ik zal dit toelichten aan de hand van een eis die ook vaak in methodologie-handboeken wordt gesteld: de eis van repliceerbaarheid van wetenschappelijke resultaten. Waar het mij om gaat is of de in het artikel gepresenteerde analyseresultaten repliceerbaar zijn. Het

antwoord hierop is eenvoudig: indien de data een goede weerspiegeling vormen van de werkelijkheid, dan is dit geen probleem. Immers, indien er opnieuw data worden verzameld, en men repliceert nu zonder omwegen direct de analyses die in het artikel zijn gerapporteerd, dan zal men vaak weer hetzelfde vinden indien de werkelijkheid niet veranderd is. Het enige probleem is dus steekproeffluctuatie. Dat wil zeggen, door het toeval kunnen de gegevens in de eerste steekproef er wat anders uitzien dan de gegevens in de tweede steekproef.

Indien men onderzoek doet volgens het boekje, beschermt men zich tegen steekproeffluctuatie door bijvoorbeeld de type 1 fout op 5 % te stellen, dat wil zeggen men zal slechts in 5 % van de gevallen de nulhypothese ten onrechte verwerpen. Echter, doordat men dezelfde data gebruikt om een grote hoeveelheid toetsen uit te voeren, waarbij de uitkomsten van sommige toetsen vaak weer een aanleiding zijn voor het uitvoeren van andere toetsen, is de uiteindelijke bescherming tegen steekproeffluctuatie niet te bepalen. Indien er zeer veel analyses zijn gedaan, dan kunnen wij de in het artikel gerapporteerde overschrijdingskansen niet serieus nemen. Men spreekt hier wel van kanskapitalisatie. Kanskapitalisatie speelt een grotere rol bij overschrijdingskansen die dicht bij het verwerpingsgebied, zeg maar .05, liggen. Hierdoor speelt voor grote steekproeven kanskapitalisatie meestal een geringe rol, omdat in grote steekproeven de nul-hypothese meestal altijd verworpen wordt met overschrijdingskansen die veel kleiner zijn dan .05, maar voor kleine steekproeven wordt de interpretatie van de gerapporteerde overschrijdingskansen problematisch.

Dit probleem van kanskapitalisatie is door statistici wel onderkend. Wat in de literatuur wordt aangeraden is om kruisvalidatie toe te passen. Bij kruisvalidatie houdt men voordat de data-analyses aanvangen een deel van de data, bijvoorbeeld de helft van alle personen, apart. Men gaat dan exploratief door de andere helft van de data, en komt daar uiteindelijk op een bepaald model uit. Dit model wordt dan vervolgens getoetst in de nog ongebruikte eerste helft van de steekproef. Zo doe je dus als het ware je eigen replicatieonderzoek. Deze aanpak is bruikbaar als de oorspronkelijke steekproef groot genoeg is. Bij een kleine steekproef is deze aanpak

niet haalbaar. Omdat onderzoekers vaak de grootste moeite hebben om tot een aanvaardbare steekproefgrootte te komen voor de oorspronkelijke steekproef, lijkt deze aanpak me al met al niet vaak bruikbaar.

Op de vraag of het erg is indien veel onderzoekers chaotisch te werk gaan geef ik dus als antwoord: niet zo erg, want het enige gevaar dat je loopt is dat ten gevolge van kanskapitalisatie niet repliceerbare zaken worden gevonden. Ik noem dit geringschattend het enige gevaar, omdat er wel meer oorzaken zijn die leiden tot het trekken van verkeerde conclusies in onderzoek, zoals nonresponse in steekproeven en slechte datakwaliteit. Ik sluit niet apriori uit dat het probleem van kanskapitalisatie in het niet valt bij deze twee andere problemen. De tijd ontbreekt me om deze twee problemen hier verder te bespreken. Mijn betoog onderstreept nog eens het belang van daadwerkelijk replicatieonderzoek.

Is er een koninklijke weg voor een statisticus om met deze problematiek in de statistische consultatie om te gaan? Hier wordt verschillend over gedacht. Laat ik eerst mijn eigen mening geven. Ik gaf al eerder aan dat veel belangwekkende zaken door het toeval zijn gevonden. Daarnaast hecht ik, net als anderen, aan daadwerkelijke replicatie van onderzoek. Dit impliceert volgens mij dat de statisticus de inhoudelijk onderzoekers ondersteunt in wat zij willen. Hij moet met hen overleggen of de door hen gestelde vraag de juiste is, aangeven welke methode voor data-analyse het meest aangewezen is om te gebruiken, maar ook aangeven welke methoden een antwoord geven op een andere voor het onderzoek relevante vraag. Onderzoekers vinden het vaak moeilijk van te voren te kiezen voor een bepaalde analysestrategie. Ik raad dan vrijwel altijd aan alles maar te doen. Men krijgt zo een beeld van de robuustheid van de resultaten onder de verschillende analysetechnieken.[6] Ik ben geneigd de onderzoeker voor te stellen dit dan te rapporteren in een noot van het uiteindelijke artikel, indien dit de publicatie van het artikel niet in gevaar brengt. Ik moet toegeven dat enig opportunisme in deze me niet vreemd is.

Ik concludeer dus dat ik geneigd ben om de chaotische sociaal-wetenschappelijke onderzoekspraktijk te accepteren zoals deze is. In mijn visie zouden methodologische handboeken een realistischer beeld moeten geven van deze praktijk. Verschillende

van mijn collega's denken hier anders over. Ik heb hier niet de tijd uitgebreid op hun meningen in te gaan, en kan slechts hopen dat de paar zinnen die ik aan een aantal van hen wijdt geen karikatuur maken van hun standpunt.

Dit jaar is de twaalfde editie van de Groot's Methodologie uitgebracht.[7] Don Mellenbergh, hoogleraar aan de Universiteit van Amsterdam, heeft een voorwoord geschreven waarin hij vol lof is over dit werk. Hij stelt dat het werk niet gedateerd is, en stelt dat "voor gevorderd studenten, assistenten en onderzoekers in opleiding het boek nog altijd het beste uitgangspunt voor verdere methodologische scholing is". In Amsterdam lijkt de empirische cyclus dus nog in full swing. Daarnaast lijkt Ivo Molenaar, hoogleraar aan de RijksUniversiteit Groningen, nog steeds optimistisch over de mogelijkheid om onderzoekers beter op te leiden, om zo te zorgen dat zij zich meer conform de empirische cyclus gedragen. Hij heeft een reeks artikellen geschreven, die onder andere zijn verschenen in inhoudelijke sociaal-wetenschappelijke tijdschriften zoals Mens en Maatschappij en Psychologie en Maatschappij.[8] Hij stelt onderzoekers voor meer voor te denken in plaats van na te denken, waarmee hij bedoelt dat onderzoekers beter zouden moeten nadenken voor het uitvoeren van hun onderzoek, om later, bijvoorbeeld in de data-analyse fase, geen onverwachte problemen tegen te komen. Daarentegen lijkt mijn promotor Jan de Leeuw, nu hoogleraar aan de UCLA, blijkens een artikel in Statistica Neerlandica[9] bijzonder negatief over het realisme van het gehele model dat achter statistische toetsing zit. Ook hij benadrukt het belang van daadwerkelijke replicatie van onderzoek om zo problemen van de gangbare methodologische voorschriften te ondervangen. Ik heb de indruk dat mijn visie hier het dichtst bij ligt.

Is hier nu sprake van misbruik van statistiek? Mijn antwoord is: ja, in de zin van foutief gebruik. Men rapporteert foutieve overschrijdingskansen. Deze overschrijdingskansen geven wetenschappelijke resultaten een status die zij niet altijd verdienen. Dit is een moderne vorm van misbruik van statistiek die Huff in 1954 nog niet had kunnen voorzien. Door de beschikbaarheid van snelle computers en gebruikersvriendelijkere software neemt de neiging om chaotisch in plaats van doordacht data te analyseren toe. Hierdoor zijn de gerapporteerde overschrijdings-

kansen niet juist, en deze zijn ook onmogelijk exact te bepalen. Zo is aan de roep van Huff om alleen resultaten te rapporteren die op significantie zijn getoetst tegemoet gekomen op een manier die hij niet kon bevroeden. Echter, door het belang dat algemeen gehecht wordt aan daadwerkelijke replicatie meen ik dat er met dit misbruik van de statistiek best is te leven.

Onderzoek

Na deze bespiegeling over de statistische consultatie zou ik u iets willen vertellen over onderzoek dat uitgevoerd wordt door de vakgroep Methodenleer en Statistiek. Wij werken momenteel een onderzoeksprogramma uit. Zoals het zich nu laat aanzien zullen wij onze onderzoeksactiviteiten concentreren op vier thema's, namelijk de analyse van categorische data, de analyse van longitudinale data, schaaltechnieken en dataverzameling. Ik zal u nu iets vertellen over een onderzoeksproject dat deel uitmaakt van de projecten categorische data en dataverzameling, namelijk het schatten van de omvang van moeilijk te bereiken populaties.

Bij de het schatten van de omvang van moeilijk te bereiken populaties kunt u denken aan vragen als "hoeveel autokrakers zijn er actief in Den Bosch", "hoeveel illegalen zijn er werkzaam in het Westland", "hoeveel wordt er gefraudeerd in de ww", of "hoeveel bezitters van illegale vuurwapens zijn er in Nederland". Deze vragen liggen in de strafrechtelijke sfeer, maar de methoden waarnaar ik onderzoek wil doen zijn ook bruikbaar voor andere onderwerpen waar men liever niet over praat, zoals sexueel gedrag van risicogroepen, en geweld binnen het gezin.

Ik heb er voor gekozen om u iets over dit onderwerp te vertellen, omdat het onderzoek dient te leiden tot een verbetering van schattingen, die nu geregeld niet meer zijn dan een slag in de lucht. Dergelijke schattingen suggereren bij velen een grotere exactheid dan waargemaakt kan worden. Het lijkt er vaak op dat maatschappelijke groeperingen met overdreven schattingen aankomen om zo meer geld van de overheid los te weken. Dergelijke schattingen zijn vaak niet op systematisch onderzoek gebaseerd, maar gaan geregeld wel een eigen leven leiden. Dit is een voorbeeld van misbruik van statistiek. Ons onderzoek zou dergelijke schattingen

moeten beteren.

Wij ontwikkelen statistische methoden die bruikbaar zijn om uitspraken te doen over zaken waarover geen directe gegevens beschikbaar zijn. Ik ga u twee deelonderzoeken beschrijven.[10] Ik heb aan deze onderzoeken gewerkt met Ger van Gils van het bureau Beleidsonderzoek en -Advies, met Filip Smit en met Renee van Rongen, en geregeld maken wij gebruik van adviezen van andere leden van de vakgroep.

Vangst-hervangst[11] methoden

Op het probleem van de autokrakers in Den Bosch stuitte ik via een krantenartikel. Op 15 februari 1990 stond in de Volkskrant een klein stukje waarin melding werd gemaakt van een experiment in Den Bosch om autokrakers op te sporen. De recherche maakte gebruik van een lokauto, dat wil zeggen een bij autokrakers gewilde auto. Een dergelijk voertuig was voorzien van geluidsapparatuur en een speciaal stil alarm. De locatie van de lokauto is bij alle surveillancewagens en de meldkamer bekend, en daardoor is snel ingrijpen mogelijk als de auto wordt opengebroken. Niet alleen werden veel autokrakers bij de lokauto aangehouden, maar bovendien werden verschillende autokrakers twee maal bij deze lokauto aangehouden, en een autokraker zelfs drie maal.

Dit zette me aan het denken: indien je weet hoeveel autokrakers een keer zijn aangehouden, hoeveel twee keer, hoeveel drie keer, zou je dan niet op een of andere wijze het aantal autokrakers kunnen berekenen dat actief is maar nul keer is aangehouden? De vraag stellen is hem beantwoorden: dat kan inderdaad, als je tenminste bereid bent om een aantal zaken te veronderstellen. Een nogal simpele veronderstelling is dat elke autokraker over een bepaalde periode een constante kans heeft om opgepakt te worden. In een erg korte periode zal zo'n kraker wel of niet worden opgepakt, maar over een langere periode zal een kraker nul, 1 of meerdere keren worden opgepakt. Het aantal keren dat een kraker wordt opgepakt volgt dan een zogenaamde Poissonverdeling. Indien men bereid is aan te nemen dat alle autokrakers ieder dezelfde Poissonverdeling volgen, dan is het aantal autokrakers dat

nul keer is gepakt eenvoudig te schatten.

We krijgen zo natuurlijk een getal dat aangeeft hoeveel autokrakers er zijn, maar dit getal is erg afhankelijk van de veronderstellingen die wij hebben gemaakt. Deugen deze veronderstellingen niet, dan is het mogelijk dat het gevonden getal erg afwijkt van de werke lijkheid. Na wat zoeken in de literatuur bleek dat er vooral in de biologie veel aandacht aan vergelijkbare problemen was besteed.[12] Het probleem is daar, bijvoorbeeld, hoeveel vissen zitten er in het meer, hoeveel apen in het bos, e.d..

Het uitdagende van het onderzoeksgebied is om veronderstellingen te doen die zo nauw mogelijk aansluiten bij de werkelijkheid, deze veronderstellingen op te nemen in het model, dat dan vervolgens een schatting geeft van de nul-frekwentie. Een model dient dan om een zo adekwaat mogelijke beschrijving van de werkelijkheid te geven. Wij werken niet volgens de empirische cyclus, maar volgens de zogenaamde modelmethodologie.[13] Hierbij is men op zoek naar een model dat aan bepaalde doelstellingen voldoet. Uitgangspunt is dat een model nooit waar is, maar slechts een versimpeling van de werkelijkheid. Men werkt dan ook met wegwerp-modellen, in de zin dat, indien men een model vindt dat nog beter aan onze doelstellingen voldoet, het eerder gevonden model in de prullenbak belandt.

Samenwerking met mensen uit het veld is in dit opzicht van belang om te komen tot een adekwate beschrijving van de werkelijkheid. Bijvoorbeeld, om terug te komen op de autokrakers uit Den Bosch, voor de schatting van het aantal autokrakers zou men eigenlijk moeten weten hoe autokrakers zich gedragen. Blijven alle autokrakers werkzaam in Den Bosch gedurende de gehele periode van het tellen, of zijn er ook autokrakers waarvoor dit niet geldt, bijvoorbeeld omdat zij van stad veranderen, gearresteerd worden, of stoppen? Blijft de kans om gepakt te worden over de gehele periode constant, of wordt de kans om gepakt te worden, nadat men eenmaal is gepakt, kleiner, omdat men leert uit handen van de politie te blijven? Is de kans om gepakt te worden voor elke autokraker gelijk, of zijn er verschillen tussen de kansen voor verschillende autokrakers; en indien dit laatste het geval is, hoe kunnen we deze verschillen dan in ons model meenemen?

Wij hebben een simpele versie van deze modellen recent toegepast op gegevens die verzameld zijn door Willem de Haan, werkzaam bij het Willem Pompe Instituut voor Strafrechtswetenschappen.[14] Hij heeft onderzoek gedaan naar diefstal met geweld (meestal tasjesroof) in Amsterdam Zuidoost. Experts maakten schattingen van het aantal tasjesrovers ingedeeld in een incidentele groep, een terugkerende groep en een notoire groep. Om deze schatting te onderzoeken deelde Willem de Haan de tussen 1 januari 1991 en 1 januari 1994 aangehouden tasjesrovers in op basis van hun antecedenten en aanwijzingen van de politie. Het aantal incidentele tasjesrovers werd door de experts geschat op 2000, en door het model tussen de 944 en 2381; het aantal terugkerende tasjesrovers op 50 door de experts, en door het model tussen de 141 en de 338; het aantal notoire tasjesrovers op 50 door de experts, en door het model tussen de 84 en de 176. De experts zaten er aantoonbaar naast, omdat het aantal gepakte terugkerende en notoire daders hun schatting van het totale aantal reeds overtrof. Bijvoorbeeld, er waren 65 notoire tasjesrovers 1 of meer keren gepakt, en hier moet dan nog de schatting bij worden opgeteld van het aantal notoire tasjesrovers dat nooit is gepakt; de experts hadden het totale aantal, zoals gezegd, op 50 geschat. Verder is de accuraatheid van de getallen die het model oplevert niet te checken. We kennen de werkelijke omvang van de onderscheiden groepen tasjesrovers immers niet.

Een ander voorbeeld van onderzoek, dat wij momenteel verrichten, is een onderzoek naar het aantal overtredingen van de vuurwapenwet in Nederland. Dit onderzoek vindt plaats in opdracht van het Ministerie van Binnenlandse Zaken, dat zich voor het probleem gesteld ziet hoe de regionale afdelingen van de politie te financieren op basis van werklast. Een belangrijk deel van de werklast van de politie is te meten met behulp van zogenaamde slachtofferenquetes. Dit zijn grote aselecte steekproeven waarin onder andere gevraagd wordt of men het slachtoffer is geworden van een misdrijf of een overtreding. Er zijn echter ook overtredingen en misdrijven zonder slachtoffers, en ook deze leveren werkdruk op. Overtredingen van de vuurwapenwet zijn hier een voorbeeld van. De vraag is hoe de omvang van dit probleem te meten.

Wij zijn als volgt te werk gegaan. Wij maken gebruik van het zogenaamde Herkenningssysteem van de politie, waarin alle aanhoudingen van personen worden geregisteerd die leiden tot een ten laste legging. Hierin zoeken wij per politieregio op hoeveel personen een keer zijn aangehouden, hoeveel twee keer, hoeveel drie keer, enzovoort. Op basis hiervan schatten wij met een bepaald statistisch model per regio het aantal overtreders dat nul keer is aangehouden.

Het gehele onderzoeksgebied is erg opwindend. Ik denk dat dit onderzoek kan leiden tot een stap vooruit op het gebruik van schattingen die louter gebaseerd zijn op oordelen van experts. Hierbij worden inzichten van experts gebruikt om tot een bepaald model te komen, dat een schatting oplevert. Onze indruk is dat de schattingen die voortgebracht worden met deze benadering leiden tot een aanzienlijke verbetering van eindschattingen.

Randomized Response

Ik wil u nu iets vertellen over een tweede onderzoeksproject dat de omvang van populaties probeert te schatten. In zojuist beschreven onderzoeksgebied bestuderen we gedrag van mensen. Soms is het gemakkelijker om mensen naar hun gedrag te vragen. We bestuderen dan gerapporteerd gedrag. Echter, mensen zijn soms niet geneigd hun gedrag te rapporteren, omdat dit gedrag bijvoorbeeld niet sociaal wenselijk is. De systematische meetfout die men in dergelijk onderzoek oploopt is te bepalen in zogenaamde validatiestudies. Hierin is reeds bekend welk gedrag mensen vertonen, en wordt hun vervolgens gevraagd of zij dit gedrag vertonen. De meetfouten kunnen zeer aanzienlijk zijn. Ik zal u enkele voorbeelden geven.

Als eerste voorbeeld noem ik een onderzoek van Marianne Junger.[15] Hieruit bleek dat slechts 38 % van autochtone jongeren die in het laatste jaar contacten met de politie hebben gehad, dit toegeeft. Niet meer dan 9 % van Turkse jongeren geeft hun contacten toe. Hierbij zijn gegevens van de politie vergeleken met de gegevens die jongeren zelf verstrekken. Dit leverde dus een onderrapportage op van maar liefst 62 en 91 %. Een tweede voorbeeld is een onderzoek naar uitkeringsfraude in de WW, uitgevoerd in opdracht van de Federatie van Bedrijfsverenigingen. Hierin

gaf 43 % van de respondenten een reeds geconstateerde fraude niet toe. Een derde voorbeeld betreft een telefonische enquete naar het betalen van motorrijtuigenbelasting in een onderzoek van Bert Berghuis en Max Kommer.[16] Van de personen waarvan bekend was dat zij ten onrechte geen belastingen betaald hadden, gaf slechts tussen de 8 en 13 % dit toe. Bovengenoemde voorbeelden hebben betrekking op situaties dat de respondenten kunnen vermoeden dat hun antwoorden consequenties kunnen hebben voor henzelf. Het is bekend dat dergelijke problemen ook belangrijk zijn bij allerlei ander onderzoek waar sociale wenselijkheid een rol speelt, zoals bijvoorbeeld onderzoek naar abortus, AIDS en extreem rechts. Ik hoop met bovengenoemde voorbeelden te hebben aangegeven dat, in geval van gevoelige onderwerpen, de mogelijkheden beperkt zijn om met vragenlijsten en interviews valide antwoorden te krijgen.

Momenteel werken wij aan de verdere ontwikkeling van een methode die mensen meer veiligheid zou moeten bieden om te antwoorden op gevoelige vragen. Deze methode heet randomized response. Deze techniek werd in 1965 voor het eerst voorgesteld door Warner[17], en is vervolgens in veel verschillende versies uitgewerkt. De eerste versie was als volgt: Stel men doet een onderzoek naar belastingfraude. De geinterviewde heeft twee stellingen voor zich liggen, namelijk "ik fraudeer" en "ik fraudeer niet". Men laat een geinterviewde nu met een dobbelsteen werpen. Indien er een 1, 2, 3 of 4 bovenkomt dan antwoord deze persoon op de eerste stelling, "ik pleeg fraude", met "dat klopt" of "dat klopt niet". Komt een 5 of 6 boven dan antwoordt hij of zij op de tweede stelling, "ik pleeg geen fraude", met "dat klopt" of "dat klopt niet". De term "randomized" in randomized response slaat dus op het feit dat het door kans bepaald wordt op welke stelling de persoon antwoordt. In ons voorbeeld is de kans op antwoord op de stelling "ik pleeg fraude" 2/3. De geinterviewde gooit de dobbelsteen, en laat de uitkomst niet zien aan de interviewer. Stel nu dat een geinterviewde zegt "dat klopt". De interviewer kan dan niet vaststellen op welke stelling deze persoon "dat klopt" heeft geantwoord: het is niet te achterhalen of deze persoon heeft gefraudeerd of niet. Daarom kan deze persoon antwoord geven op de stelling die door de dobbelsteen aan hem voorligt

zonder zich bloot te geven. Echter, doordat de kans op de stelling "ik pleeg fraude" 2/3 is, en de kans op de stelling "ik pleeg geen fraude" 1/3, heeft een fraudeur zo een kans van 2/3 om "dat klopt" te zeggen, terwijl een niet-fraudeur een kans van 1/3 heeft om "dat klopt" te zeggen. Voor een enkele respondent is dus aan de hand van het antwoord niet vast te stellen of hij fraudeert of niet, maar na wat eenvoudig rekenwerk kan men wel een schatting krijgen van het aantal respondenten dat heeft gefraudeerd.

De zojuist genoemde procedure levert dus een schatting van het aantal fraudeurs op, maar deze schatting heeft een veel groter betrouwbaarheidsinterval dan je zou vinden bij een directe vraag naar fraude in een vragenlijst zonder deze randomized response-procedure. Dat komt omdat wij een extra meetfout hebben geintroduceerd. Een directe vraag naar fraude zou echter waarschijnlijk weer een hoge onderrapportage geven, en daarmee ben je nog verder van huis. Maar het is goed te constateren dat je dus niet iets voor niets krijgt.

Ik zei u al dat er verschillende varianten van randomized response zijn.[18] Er zijn andere randomized response procedures die tot een kleiner betrouwbaarheidsinterval voor de proportie fraudeurs leidt. Een ervan werkt als volgt: we leggen twee stellingen voor, bijvoorbeeld "ik pleeg fraude" en "mijn moeder is jarig in oktober, november of december". Met de dobbelsteen bepaalt de geinterviewde vervolgens of hij de eerste of de tweede stelling zal beantwoorden. D.w.z. bij een 1, 2, 3 of 4 beantwoordt men de stelling "ik pleeg fraude", en bij een 5 of 6 beantwoordt men de stelling van de verjaardag van de moeder. Indien we aannemen dat de kans dat iemands moeder jarig is in oktober, november en december 1/4 is, dan kan op eenvoudige wijze voor de gehele groep de kans op fraude worden geschat.

Men moet bij de keuze voor de alternatieve vraag de respondent voldoende veiligheid bieden. Er is hier een spanningsveld voor de onderzoeker, want minder veiligheid voor de respondent levert een kleiner betrouwbaarheidsinterval voor de onderzoeker op, maar wellicht ook minder eerlijke antwoorden. Men biedt bijvoorbeeld te weinig veiligheid door als de alternatieve vraag te nemen: "is uw moeder jarig op 25 december ?". Deze situatie geldt toevallig voor vijf mensen in

deze zaal, waarvan er vier familieleden van mij zijn, en ikzelf de vijfde ben. Mocht u niet weten wanneer onze moeder jarig is, dan zal een "dat klopt"-antwoord van ons toch snel uitgelegd worden als het toegeven van fraude, omdat u de kans dat onze moeder op 25 december jarig is natuurlijk veel kleiner inschat dan de kans dat wij frauderen.

We kunnen ons de vraag stellen: werkt randomized response altijd? De randomized response-procedure heeft wisselende successen gekend. Een bekend succes is het voorkomen van abortus in Canada in 1975. In gewone mondeling interviews gaven 1150 vrouwen toe een of meer abortussen te hebben gehad. In schriftelijke interviews, die wat anoniemer zijn, waren dit er 3060, dus bijna 3 maal zo veel. Bij randomized response was het aantal 12320, vier maal zoveel als bij schriftelijke interviews, en 10 maal zoveel als bij de mondelinge interviews. Bij onzorgvuldige instructies kan er echter ook wel een lacherige sfeer bij geinterviewden ontstaan. In de VS speelt dan ook het probleem dat respondenten wel eens met een dobbelsteen weigeren te gooien, of kaarten weigeren te trekken, omdat zij niet willen gokken. Randomized response-onderzoek dient dan ook zorgvuldig te worden opgezet, waarbij veel tijd wordt besteed aan het uitleggen van het doel van de procedure. Een voorstudie waarin de gehele procedure wordt uitgeprobeerd is absoluut noodzakelijk.

Men zou kunnen denken dat de procedure zo moeilijk is dat hij alleen werkt bij hoger opgeleiden. Dat is niet juist. Er zijn voorbeelden van vruchtbare toepassing van randomized response bij lager opgeleiden. Ook het omgekeerde geldt niet: de procedure kan ook mislopen bij hoger opgeleiden. Een maand geleden vertelde een collega van de University of Maryland, Mitchel Dayton, mij dat hij een schriftelijke randomized response studie had ondernomen onder de hoogleraren van zijn universiteit.[19] Het onderwerp was academisch bedrog. De studie ging over zaken als het toeeigenen van ideeen van anderen, bijklussen zonder de verdiensten van dit bijklussen op te geven of af te dragen, en dergelijke. Hij kreeg een groot aantal vragenlijsten terug met scheldwoorden en obsceniteiten in de kantlijn. Een hoogleraar vroeg hem in een brief hoe hij zo stom had kunnen zijn het onderzoek

op deze wijze op te zetten: nu kon hij de vragenlijsten toch net zo goed in de prullenbak gooien?

Voorzover wij kunnen nagaan is in Nederland randomized response slechts eenmaal toegepast, namelijk in een klein proefonderzoek naar zwart werk van Kazimier en van Eck van het CBS.[20] Bij dit onderzoek werd de randomized response-procedure zonder omwegen direct aan respondenten voorgelegd. Respondenten vonden het onderzoek "vreemd", "wel grappig" of "onbegrijpelijk". Een groot aantal personen gaf te kennen dat de techiek misschien wel aardig was, maar in hun geval niet zinvol, omdat ze toch niets te verbergen hadden. Zij weigerden vervolgens medewerking. Het proefonderzoek was hierdoor mislukt, volgens ons doordat respondenten de indruk hadden dat toestemmen aan meedoen met randomized response aangaf dat men wat te verbergen had.

Al met al is het onderzoeksgebied naar de bruikbaarheid van de randomized response-procedure nogal rommelig. Er is nauwelijks enig systematisch onderzoek naar de condities waaronder randomized response wel of niet werkt. Wij willen dit gaan uitzoeken met een meta-analyse. In een meta-analyse worden de onderzoeksresultaten uit verschillende onderzoeken op een statistische wijze gecombineerd. Men kan zo onderzoeken welke aspecten van randomized response behulpzaam zijn bij het laten slagen van de procedure.

Naast het onderzoek naar de bruikbaarheid van de randomized response procedure willen wij gaan werken aan de verdere ontwikkeling van data-analytische technieken voor randomized response-data. In deze technieken dienen de gerandomiseerde antwoorden gerelateerd te worden aan verklarende variabelen, of aan responsvariabelen. Enkele basale technieken zijn reeds uit de literatuur bekend.[21] De totale data-analytische gereedschapskist voor randomized response-data is evenwel leger dan de gereedschapskist voor gewone data.

Specifiek bruikbaarheidsonderzoek zal zeer waarschijnlijk binnenkort aanvangen met een randomized response-studie voor het Ministerie van Sociale Zaken, naar uitkeringsfraude. Dit onderzoek voeren wij uit samen met BOA. Ons voornemen is in een vooronderzoek twee randomized response-technieken middels een experiment

te vergelijken. De eerste techniek heet "geforceerd antwoord". Wij laten in deze conditie ontvangers van een uitkering met twee dobbelstenen gooien. Indien de som van de twee gelijk is aan 2,3 of 4, dient de respondent "ja" de antwoorden op de vraag of hij of zij fraudeert. Indien de som gelijk is aan 11 of 12, dient hij of zij "nee" te antwoorden. Indien de som ligt tussen 5 en 10, vragen wij om eerlijk antwoord te geven op de vraag. Wij kiezen voor deze procedure omdat uit psychologisch onderzoek bekend is dat respondenten de kans op een waarde tussen 5 en 10 onderschatten, en zich dus veiliger voelen dan zij in werkelijkheid zijn. De andere conditie probeert te voorkomen dat respondenten "ja" moeten antwoorden op een vraag naar fraude. Voor de respondent liggen twee spellen kaarten. In de linker stapel is 3/4 van de kaarten rood, en in de rechter is 1/4 van de kaarten rood. Wij vragen de respondent om uit elke stapel een kaart te trekken, en ons niet te laten zien welke kleur de kaarten hebben. Indien men heeft gefraudeert dient men de kleur te noemen van de kaart uit de linker stapel, en indien men niet heeft gefraudeert de kaart uit de rechter stapel. Zo zal onder fraudeurs vaker rood worden gezegd dan onder niet-fraudeurs. De proportie fraudeurs is met deze methode vrij eenvoudig te schatten.

Samenvattend, ik denk dat de statistiek hier een aantrekkelijke methode te bieden heeft om te komen tot betere schattingen. Deze methode kan er voor zorgen dat in geval van gevoelige onderwerpen nog steeds gewerkt kan worden met vragenlijsten en gestandaardiseerde mondelinge interviews. Eenvoudige statistische principes zorgen voor een grotere geldigheid van uitspraken gebaseerd op steekproefgegevens, en daarom spreek ik ook in dit geval van statistiek waar zeker mee te leven valt.

Tot slot

Meneer de Rector, Dames en Heren.

De titel van deze oratie is "hoe te leven met statistiek". Ik heb me willen afzetten tegen het negatieve beeld dat over statistiek bestaat, en dat tot velen van u is gekomen door de titel van Huff's boek "how to lie with statistics". In het eerste deel van mijn oratie heb ik aangegeven dat delen van de problematiek van Huff nog

steeds actueel zijn. In het tweede deel heb ik enkele delen van mijn onderzoek voor het voetlicht willen brengen, die aan moeten geven dat er met statistiek best te leven valt, omdat wij met statistiek dingen kunnen die zonder statistiek niet mogelijk zijn.

Aan het eind van mijn oratie gekomen wil ik graag enkele woorden van dank uitspreken.

Geachte leden van het College van Bestuur en het Faculteitsbestuur. Ik wil u bedanken voor het blijkens de benoeming in mij gestelde vertrouwen. Meer in het bijzonder wil ik de decaan, de heer Adriaansens, oud-directeur mevrouw Gloudi, de waarnemend directeur de heer Dirksen en het bestuurslid de heer Raaymakers bedanken voor de energie die zij hebben gestoken in het normaliseren van de organisatievorm van de methodengroep, door zich hard te maken voor de oprichting van een vakgroep Methodenleer en Statistiek. Ik zal mijn best doen U niet teleur te stellen in de verwachtingen die u hebt inzake de rol van de vakgroep voor de faculteit.

Geachte leden van de faculteit. Ik vind het bijzonder aantrekkelijk om werkzaam te zijn ten behoeve van de gehele Faculteit Sociale Wetenschappen. De bruikbaarheid van de meeste modellen en technieken overschrijdt de discipline-grenzen, en daarom lijkt een facultaire vakgroep voor het vakgebied de meest wenselijke constructie. Dit zorgt echter ook voor een grote verscheidenheid aan consultatievragen. Dit ervaar ik als een uitdaging. Ik hoop dat deze oratie er toe bijdraagt dat u mij voor consultatie benadert.

Studenten. Velen van u vinden statistiek een moeilijk vak. Op ons rust de taak u dit vakgebied begrijpelijk uit te leggen. Wij zullen ons hier tot het uiterste toe inspannen.

Graag neem ik ook de gelegenheid te baat om allen te bedanken die aan mijn academische vorming hebben bijgedragen. Zonder anderen tekort te doen wil ik in het bijzonder vier personen uit mijn Leidse tijd bedanken, namelijk John van de Geer, Jan de Leeuw, Ab Mooijaart en Leo van der Kamp. John, jij hebt me in het verleden op het spoor van correspondentie-analyse gezet. Dit is een heel vruchtbaar

spoor gebleken. Jan, jij hebt me laten zien hoe mooi ons vakgebied is. Zonder jouw inspiratie had ik hier nooit gestaan. Het was een genoegen om met jou te kunnen werken. Ab, gedurende onze samenwerking heb ik van jou veel geleerd; ik hoop dat wij onze samenwerking nog lange tijd kunnen voortzetten. En Leo, ik heb altijd veel waardering gehad voor de ruimte die jij me hebt geboden toen ik in jouw vakgroep werkte. Je hebt me op beslissende momenten in mijn carriere waardevolle adviezen gegeven. Zo herinner ik me bijvoorbeeld dat je me er in 1988 van hebt weerhouden de automatisering in te gaan. Bedankt daarvoor.

Leden van de vakgroep Methodenleer en Statistiek. Ik wil hier graag jullie bedanken voor de prettige sfeer die wij met elkaar delen. Ik besef maar al te goed dat dit klimaat voor een belangrijk deel te danken is aan onze collega Harm 't Hart, die de methodengroep door de voor haar moeilijke bezuinigingsjaren heeft heengeloodst, zonder dat de groep onherstelbare averij heeft opgelopen. Dit is een prestatie die niet te overschatten is.

En tenslotte wil ik jou bedanken, Marianne. Ik zou je tekort doen door te proberen mijn dank in woorden te expliciteren.

Ik heb gezegd.

1. Gepubliceerd in 1954 te New York door Norton.

2. Leo van der kamp maakte mij er op opmerkzaam dat het citaat aan Disraeli wordt toegeschreven door Mark Twain, in Autobiography I, 246. Of Disraeli het werkelijk heeft gezegd is de vraag.

3. Zie ook: G. Engbersen (1990). Publieke bijstandsgeheimen: het ontstaan van een onderklasse in Nederland. Leiden: Stenfert Kroese, p. 244-245.

4. Gepubliceerd in 1960 te Utrecht door Het Spectrum (Prisma nr. 539). De vertaler heeft overigens ook alle echte Amerikaanse voorbeelden uit het boek vervangen door fictieve Nederlandse om "de zaak wat huiselijker te maken" (p. 7).

5. A.D. de Groot (1961) Methodologie. Grondslagen van onderzoek en denken in de gedragswetenschappen. Den Haag: Mouton. In 1994 als twaalfde druk heruitgegeven door van Gorcum, Assen, met een voorwoord door Don Mellenbergh.

6. Zie ook Gifi (1990, p. 36-39).

7. Zie noot 5.

8. Bijvoorbeeld in 1977: Ik wordt ziek van de statistiek, of: er van weten zonder er naar te handelen. Mens en maatschappij, 52, 58-71; in 1988: Formal Statistics and informal data analysis, or why laziness should be discouraged. Stastica Neerlandica, 42, 83-90; in 1990: Statistiek smaakt beter in een denksandwich. Psychologie en maatschappij, 53, 357-366.

9. Models and techniques. Statistica Neerlandica, 1988, 42, 91-98; vergelijk ook Gifi (1990), Nonlinear multivariate analysis, New York: Wiley, p. 19-64; en de Series Editor's Introduction van Jan de leeuw in J.P. van de Geer (1993), Multivariate analysis of categorical data: Theory en Multivariate analysis of categorical data: Applications, beiden uitgegeven door Sage, Newbury park.

10. Voor een overzicht van andere methoden verwijs ik naar van der Heyden, Smit en van Gils (1993). Schattingen van het aantal slachtofferloze delicten. Politica Nova 3, uitgegeven door het Ministerie van Binnenlandse Zaken en naar Smit, van der Heyden en van Gils, 1994, Enkele weinig gebruikte methoden om de omvang van criminaliteit te schatten. Tijdschrift voor Criminologie (in druk).

11. In het engels: capture-recapture.

12. Belangrijke overzichten uit de biologie zijn van G.A.F. Seber, in 1983: The estimation of animal abundance and related parameters (2nd ed.) London: Griffin; in 1986: A review of estimating animal abundance. Biometrics, 42, 267-292; in 1992: A review of estimating animal abundance II. In: International Statistical Review, 60, 129-166. In de criminologie zijn twee relevante artikelen in 1990 verschenen in de Journal of Quantitative Criminology, namelijk van M.F. Collins and R.M. Wilson, Automobile theft: estimating the size of the criminal population (p. 395-409) en van D.K. Rossmo and R. Routledge: Estimating the size of criminal populations (p. 293-314).

13. Zie bijvoorbeeld Meerling (1989), Methoden en Technieken van psychologisch onderzoek. Deel I. Model, observatie en beslissing. (vierde druk). Meppel: Boom.

14. Een onderzoeksverslag, getiteld Berovers in de Bijlmer is in april uitgegeven door het Willem Pompe Instituut. De hieronderstaande resultaten zijn hierin verwerkt.

15. p. 22 van Marianne Junger (1990), Delinquency and Ethnicity. Deventer: Kluwer.

16. A.C. Berghuis en M.M. Kommer (1982). Effecten van voorlichting en controle: een experiment bij de motorrijtuigenbelasting. Den Haag: Ministerie van Justitie (W.O.D.C.).

17. S.L. Warner (1965). Randomized response: a survey techniquefor eliminating evasive answer bias. Journal of the American Statistical Association, 60, 63-69).

18. Voor een inleidende overzichten, zie J.A. Fox and P.E. Tracy (1986). Randomized response: a method for sensitive surveys. Newbury Park: Sage, and A. Chaudhuri and R. Mukerjee (1988). Randomized response. Theory and Techniques. New York: Dekker.

19. De onderstaande resultaten zijn niet gepubliceerd.

20. Zie bijlage van B. Kazemier, R. van Eck en C.C. Koopmans. Economische aspecten van belasting en premieontduiking en misbruik van sociale voorzieningen. C.B.S.

21. Zie de referenties in noot 18.