H. Elffers & P.J. van Koppen (2002) Methoden van de rechtspsychologie. In: P.J. van Koppen, D.J. Hessing, H. Merckelbach & H.F.M. Crombag (red.), *Het recht van binnen: Psychologie van het recht* (pp. 1005-1030). Deventer: Kluwer

53

Methoden van de rechtspsychologie

Henk Elffers
Peter J. van Koppen

In de voorgaande hoofdstukken zijn rechtspsychologische inzichten gepresenteerd die voor een belangrijk deel berusten op empirisch onderzoek. In dit hoofdstuk geven wij in kort bestek een inleiding in dat soort onderzoek. Het hoofdstuk bestaat uit twee delen. In het eerste deel bespreken wij de manier waarop in de rechtspsychologie empirisch onderzoek wordt uitgevoerd. Dit gebeurt aan de hand van een voorbeeld dat als rode draad door het eerste deel van dit hoofdstuk loopt, te weten onderzoek naar egoïsme en regelovertredend gedrag. Steeds als begrippen of technieken worden geïntroduceerd, wordt getoond hoe die begrippen en technieken eruit komen te zien in dit concrete onderzoek. Het voorbeeld is gekozen op grond van onze eigen onderzoeksbelangstelling en -ervaring, zonder dat het exact gelijk is aan feitelijk door ons uitgevoerd werk.

Egoïsme en regelovertredend gedrag

Ter introductie vertellen wij hoe het onderzoek naar egoïsme en regelovertredend gedrag tot stand is gekomen. In ons onderzoek werd in eerste instantie veel aandacht besteed aan de determinanten van belastingontduiking: welke factoren bepalen of iemand overgaat tot het ontduiken van de belastingen en welke factoren vormen nu juist een rem op die neiging. Het bleek daarbij vruchtbaar om na te gaan hoe belastingbetalers een afweging maken tussen hun eigen financiële belang en het algemeen belang, vertegenwoordigd door de staat. Betrapte belastingontduikers hadden vaak een sterk op zichzelf en hun eigen belang gerichte instelling, terwijl degenen die zich aan de belastingregels hielden vaker een op de gemeenschapsbelangen gerichte instelling vertoonden. Voortborduurend op dat thema was de vraag: is het in het algemeen zo dat regelovertreders sterker een egoïstische inslag hebben dan zij die de regels volgen?[1]

1. Wie alles over dit onderzoek wil weten leze Weigel, Hessing & Elffers (1999), Elffers & Hessing (1999), Verkuyten, Rood-Pijpers, Elffers & Hessing (1994), van Giels, Hessing & Elffers (1992), Hessing, Elffers, Robben & Webley (1992), Webley, Robben, Elffers & Hessing (1991) en Elffers (1991).

Vraagstelling

De vraagstelling van een onderzoek hangt samen met het *waarom*, het doel van het onderzoek. Als men wil proberen regelovertreding tegen te gaan, doet men ander onderzoek dan wanneer men wil begrijpen hoe egoïsme tot stand komt. In het eerste geval onderzoekt men bijvoorbeeld of regelovertreding kan worden voorkomen door mensen minder egoïstisch te maken. In het tweede geval gaat het bijvoorbeeld om de vraag of egoïsme door geslaagde regelovertreding wordt bevorderd.

In elk empirisch onderzoek doet men er goed aan nauwkeurig te omschrijven welke vraag men wil beantwoorden. Daarbij zijn zeer algemeen geformuleerde vragen gevaarlijk, omdat ze veelal zo ruim zijn dat ze een industrieel potentieel aan onderzoeksactiviteit zouden vereisen. Bescheidenheid bij het formuleren van het doel van een onderzoek is raadzaam. Een vraag formuleren die binnen afzienbare tijd onderzoekbaar is, helpt daarbij.

> *Voorbeeld:* De algemene vraag of egoïstische mensen meer geneigd zijn tot regelovertreding dan op de gemeenschap gerichte mensen, geeft een goed kader voor een onderzoek dat een leven lang kan duren. Een concrete, daarvan afgeleide vraag kan zijn: rijden egoïsten vaker door rood? Maar ook: leidt egoïsme vaker tot kindermishandeling? Of tot belastingontduiking? Ook zou men zich kunnen voorstellen te onderzoeken of egoïsme op een aantal weinig (of juist sterk) gerelateerde terreinen tot overtredingsgeneigdheid leidt. In alle gevallen hoort daar een ander onderzoek bij.

Het is belangrijk de concrete onderzoeksvraag niet zodanig in te perken dat men achteraf onmiddellijk moet toegeven dat het wel een heel erg klein onderdeel van de algemene vraagstelling dekt. Wie denkt dat het de moeite niet waard is, beginne niet aan een dergelijk werk. Verderop in dit hoofdstuk (bij 'interpretatie') zullen we nog nagaan hoezeer deze kwestie terugkomt.

Het is aanbevelenswaardig ruim de tijd te nemen om na te denken over de onderzoeksvraag. In het voorbeeld hadden we het over 'geneigdheid' tot overtreden. Dat is een mooie frase, maar bedoelen we dat ook? Bedoelen we niet eigenlijk 'feitelijk overtreden'? Dat is nogal wat anders. De vraag: 'willen egoïsten vaker door rood rijden,' is immers van een geheel andere orde dan de vraag 'rijden egoïsten vaker door rood.' Het is raadzaam om, zodra men een onderzoeksvraag heeft geformuleerd, die formulering te bespreken met zoveel mogelijk buitenstaanders. Veelal doet men zijn voordeel met hun commentaar: wat zij zich bij die vraag voorstellen en al of niet interessant vinden, is vaak een *eye-opener*, en dwingt de onderzoeker tot precisie en tot het nemen van weloverwogen besluiten.

Een vraag moet niet alleen onderzoekbaar, maar ook interessant zijn. De psychologie wordt nogal eens geplaagd door onderzoeksresultaten die met kracht een open deur intrappen. Het heeft weinig zin om een vraagstelling te formuleren waarvan verwacht kan worden dat die alleen leidt tot resultaten die iedereen toch al weet. Voor het oplossen van dit probleem stelde Hofstee zijn weddenschapsmodel voor: begin uitsluitend aan een onderzoek met een vraagstelling als

je iemand anders kan vinden die bereid is een weddenschap aan te gaan over de uitkomsten van het onderzoek.[2]

Operationalisatie

Als er eenmaal een onderzoeksvraag is vastgesteld, komt het opzetten van het eigenlijke onderzoek. Om concreet te worden zal men de in de onderzoeksvraag opgevoerde onderwerpen benaderbaar – dat wil zeggen: onderzoekbaar – moeten maken. Wat bedoelen we met egoïsme en hoe stellen we precies vast welke mensen egoïst zijn en welke niet? Wie telt als roodrijder, wie niet? Het vaststellen van de concrete meetprocedure voor een gegeven concept, heet operationaliseren. Operationaliseren is geen gemakkelijke taak. Veel discussie tussen wetenschappers gaat over de gebruikte meetinstrumenten, waarbij mode en voorkeuren vaak een belangrijke rol spelen, maar ook min of meer standaardprocedures in omloop zijn om na te gaan of een bepaalde operationalisatie voldoende kwaliteit heeft. We benadrukken dat er geen vaste procedure is om tot een operationalisatie van een concept te komen. Neem het voorbeeld van egoesme. Hoe stelt men vast of iemand als egoïst geldt? Laten we eens twee denkbare procedures bekijken.

Als eerste voorbeeld nemen we de operationalisatie door middel van deelname aan een laboratoriumexperiment.[3] De proefpersoon zit in een groep van tien mensen die allemaal 10 euro van de proefleider hebben gekregen. Elk van hen wordt gevraagd om, ongezien voor de anderen, al of niet een euro in de pot te doen. Als er minstens 7 euro in de pot terechtkomt, krijgt iedereen twee euro uitbetaald, ongeacht of hij iets heeft bijgedragen in de pot. Zit er minder in de pot, dan krijgt niemand iets en is degenen die iets in de pot heeft gestopt zijn inzet kwijt. Egoïstisch gedrag wordt gedefinieerd als het *niet* inzetten van een euro. Dit spelletje wordt tien maal herhaald. De uiteindelijke egoïsmescore – de experimentele egoïsmemaat – is het aantal maal dat men *geen* euro in de pot heeft gestopt, een score die van nul (helemaal niet egoïstisch) tot 10 (volkomen egoïstisch) kan uiteenlopen. Dit soort experimentele operationalisaties – zogeheten sociale dilemma-operationalisaties, waarvan vele varianten in omloop zijn[4] – wordt vaak gekozen omdat men met dit 'echte gedrag' dicht bij de werkelijkheid komt.

Een radicaal andere manier van operationaliseren krijgen we als wij het egoësmeconcept gaan meten met een Egoïsmeschaal, samengesteld uit antwoorden op een aantal schriftelijke vragen.[5] De respondent krijgt dan pakweg een twaalftal stellingen voorgelegd van het type 'De beste manier om met mensen om te gaan is hen te vertellen wat ze graag willen horen', 'Mensen moeten alleen die wetten naleven die hun redelijk lijken' en 'Het is moeilijk om vooruit te komen als je het al te nauw neemt.' De respondent wordt verzocht te reageren op deze stellingen door te kiezen uit de antwoordmogelijkheden 'helemaal oneens',

2. Zie W.K.B. Hofstee (1980). Ook op andere punten is dit boek lezenswaardig voordat men aan een empirisch onderzoek begint.
3. Met laboratorium wordt hier bedoeld: een kunstmatige situatie die door de experimenterende psycholoog is gecreëerd.
4. Liebrand & Van Lange (1989).
5. Dit is de manier die door Weigel, Hessing & Elffers (1999) is voorgesteld.

'oneens', 'eens noch oneens', 'eens', 'helemaal eens'. Deze antwoorden worden gescoord als respectievelijk 1, 2, 3, 4 en 5. De totale egoïsmescore is de som van de antwoordscores op alle twaalf stellingen. Deze loopt dus uiteen van 12 (volstrekt niet egoïstisch) tot 60 (buitengewoon egoïstisch).

Duidelijk is dat beide manieren om egoïsme te operationaliseren, de experimentele egoïsmemaat en de Egoïsmeschaal, sterk verschillen. Voor beide manieren van operationaliseren is wat te zeggen en op beide is ook kritiek mogelijk. Het dient daarom een doel om na te gaan wat van een goede operationalisatie wordt verwacht. Daarbij worden vaak twee aspecten onderscheiden: *betrouwbaarheid* en *geldigheid*.

Betrouwbaarheid

Van belang is dat een voorgestelde meetprocedure een aantal kwaliteiten heeft. Allereerst behoort de procedure niet gevoelig te zijn voor irrelevante verschillen in de meetomstandigheden. Of iemand als egoïst wordt geklasseerd met behulp van de experimentele egoïsmemaat moet niet teveel afhankelijk zijn van of de vorige deelnemers er aardig uitzien, chinees hebben gegeten of Engels spreken, om maar iets te noemen. Men noemt een procedure die (relatief) ongevoelig is voor irrelevante invloeden betrouwbaar (*reliable*). Merk op dat verschillen van mening denkbaar zijn over de vraag of bepaalde kenmerken irrelevant zijn of niet. Veelal is men het erover eens dat herhaalde afname van de meting bij dezelfde onderzoekseenheid bij verschillende gelegenheden hetzelfde moet opleveren (herhalingsbetrouwbaarheid of *test-retest reliability*). Een instrument dat een respondent 's ochtends als egoïst, maar 's avonds als altruïst kenschetst, is onbetrouwbaar. Ook is het in het algemeen problematisch als de ene interviewer met een interviewschema een bepaalde persoon wel als egoïst ziet, maar een andere interviewer juist niet (inter-beoordelaarsbetrouwbaarheid of *interjudge reliability*).

Geldigheid of validiteit

Het tweede kwaliteitscriterium is of het instrument wèl gevoelig is voor relevante verschillen. Een meetinstrument voor geneigdheid tot delinquent gedrag dat geen verschil ziet tussen veroordeelden en een groep rechters is waardeloos. Meetinstrumenten die gevoelig zijn voor relevante verschillen worden valide of geldig genoemd. Merk op dat het niet eenvoudig is om vast te stellen óf een instrument valide is. Men moet dan immers uit anderen bronnen weten dat er sprake is van relevante verschillen, en bij gebrek aan een instrument is dat nu juist onbekend. De Egoïsmeschaal die hierboven werd geïntroduceerd, moet egoïsten een hogere score geven dan niet-egoïsten, maar zonder die schaal kunnen we de mensen nu juist niet indelen. Als indirecte oplossing voor dit probleem wordt vaak een reeks verwante meetinstrumenten met elkaar vergeleken, waarbij als eis wordt gesteld dat ze onderling voldoende verband vertonen. Zijn de mensen die zich in een experimentele situatie egoïstisch gedragen ook degenen die door de Egoïsmeschaal als zodanig worden herkend? Zo ja, dan zit het wel goed (men spreekt van concurrente validiteit). Als dat niet het geval is, wat is er dan aan de hand? En hoeveel verschil van mening tussen beide instrumenten vinden we nog acceptabel?

Dit soort vragen leidt vaak tot de noodzaak om op een nog onontgonnen terrein eerst onderzoek te doen om een instrument te ontwikkelen, voordat in het hoofdonderzoek van het aldus ontwikkelde instrument gebruik kan worden gemaakt. De gemakzuchtige tendentie om te denken dat een haastig in elkaar geflanste vragenlijst wel goed genoeg is, leidt niet zelden tot instrumenten van treurig lage geldigheid, en vaak zonder dat de onderzoeker het zich zelfs maar realiseert.

Overigens is geldigheid van een instrument tot op zekere hoogte een relatief begrip: wat wil je met het instrument? Als het erom gaat vast te stellen of leerlingen voldoende Frans beheersen om aan een conversatieles deel te nemen, is een eenvoudige luistertoets voldoende (en daarmee geldig voor het betreffende doel), maar als het erom gaat de betreffende leerlingen een advies te geven om al of niet Frans als eindexamenvak te nemen, liggen de eisen heel wat hoger: voor dat doel is zo'n eenvoudige toets wellicht niet valide. In het voorbeeld van regelovertreding zien we iets vergelijkbaars: het is heel moeilijk vast te stellen of iemand de belasting ontduikt, dus een valide en algemeen bruikbaar ontduikingsinstrument is heel moeilijk te construeren. Het is gemakkelijker om de virtuoze belastingontduikers te herkennen, dus voor een studie waarin het er alleen om gaat een aantal overduidelijke gevallen te selecteren (bijvoorbeeld voor een nader interview) is een instrument dat valide genoeg is wel voorstelbaar.

Intermezzo: Betrouwbaarheid in het recht

Het zal de juridische lezer al zijn opgevallen dat wat psychologen onder betrouwbaarheid verstaan iets geheel anders is dan wat men in juridische kringen daaronder verstaat.[6] Als daar een getuige betrouwbaar wordt genoemd, bedoelt men dat zijn verklaring overeenkomt met hetgeen in werkelijkheid is gebeurd. Psychologen zouden hier spreken over validiteit. Het juridisch equivalent voor het psychologische begrip 'betrouwbaarheid' is 'consistentie'. Een getuige die bij opeenvolgende verklaringen steeds in grote lijnen hetzelfde verhaal vertelt, zouden psychologen betrouwbaar noemen.

Deze uiteenzetting maakt niet alleen duidelijk dat validiteit en betrouwbaarheid – in psychologische zin – twee verschillende begrippen zijn, maar ook een zekere relatie hebben. Een betrouwbare, consistente getuige vertelt steeds hetzelfde verhaal, maar kan liegen dat ie barst. Als een meetinstrument betrouwbaar is, hoeft dat derhalve niet te impliceren dat het ook een valide meetinstrument is. Een onbetrouwbaar meetinstrument is echter altijd invalide. Om weer de vergelijking met het recht te trekken: een getuige die bij de rechter-commissaris een ander verhaal ophangt dan hij tegen de politie vertelde en ter terechtzitting weer met een geheel andere versie komt, is niet alleen inconsistent en onbetrouwbaar, maar zijn verklaringen zijn ook niet valide. Welke versie zou men immers moeten geloven om de waarheid vast te stellen?

Kwantitatieve en kwalitatieve methoden

Bij psychologische onderzoek wordt vaak een onderscheid gemaakt tussen kwantitatieve en kwalitatieve methoden. Dit zijn containerbegrippen waarvan de betekenis niet altijd even duidelijk is. Veelal wordt er een verschillende stijl van

6. Zie daarover ook hoofdstuk 21.

operationalisatie mee aangeduid. Met kwantitatieve methoden worden meestal instrumenten bedoeld die sterk zijn gestandaardiseerd (in de zin dat vooraf bekend is welke de verzameling van mogelijke uitkomsten bij afname van het instrument is), onder gelijke omstandigheden bij veel onderzoekseenheden kunnen worden afgenomen en waarbij gepoogd wordt de verkregen resultaten voor cijfermatige analyse geschikt te maken. Voorbeelden zijn bijvoorbeeld vragenlijstonderzoekingen, zoals hierboven met de Egoïsmeschaal.

Onder kwalitatieve methoden worden 'andere dan kwantitatieve methoden' verstaan. Het gaat om methoden waarvan de uitkomst niet gestandaardiseerd is in bovenstaande zin en bij verschillende onderzoekseenheden steeds op een andere manier kunnen worden afgenomen. De verwerking van de onderzoeksresultaten gebeurt meestal niet met statistische analyses, maar op verbaal vlak. Voorbeelden zijn onder andere: ongestructureerde interviews en gevalsbeschrijvingen (*case*-studies of casus).

Er is een neiging om kwantitatief onderzoek met betrouwbaar maar niet valide en de kwalitatieve benadering met geldig maar onbetrouwbaar te identificeren. In zijn algemeenheid is dat een grove simplificatie. Toch is de achterliggende redenering wel aansprekend: kwalitatief onderzoek, zo luidt die redenering, maakt meer gebruik van wat in de concrete situatie van het onderzoek als belangrijk naar voren komt zonder zich, zoals kwantitatief onderzoek, aan gestandaardiseerde schema's te houden. Daardoor kan het 'dichter op de huid' van het onderzochte komen en onderkennen wat voor de onderzochte persoon belangrijk is. Dit zou daardoor tot meer geldige, want met relevantie samenhangende, antwoorden kunnen leiden. Anderzijds leidt het verwaarlozen van standaardisering tot beïnvloeding door wellicht niet-relevante omstandigheden, met als gevolg teruglopende betrouwbaarheid.

Sommige onderzoekers zweren bij kwantitatieve, andere bij kwalitatieve methoden. Vaak is dat een kwestie van smaak en de techniek waarmee men zich vertrouwd voelt. Ons inziens hebben beide methoden hun sterke en zwakke kanten, die in concrete gevallen wel of niet naar voren kunnen komen, terwijl bovendien verschillende methoden elkaar dikwijls goed kunnen aanvullen.

Afhankelijke en onafhankelijke variabelen

Een geoperationaliseerd begrip wordt vaak een variabele genoemd, dat wil zeggen een kenmerk dat kan variëren. Vaak berust onderzoek, impliciet of expliciet, op een onderscheid tussen onafhankelijke (verklarende) variabelen en afhankelijke (te verklaren) variabelen. Dit onderscheid zit niet in de aard van de variabelen, maar in de rol die ze binnen de vraagstelling van het onderzoek krijgen toebedeeld. Als we willen onderzoeken of egoïstische mensen vaker door rood rijden, dan is de afhankelijke variabele het al of niet door rood rijden, of de frequentie van het door een rood stoplicht rijden, en de onafhankelijke variabele is in dit geval de egoïsmescore. Maar als de vraag is of regelovertreding in het verkeer tot verhoogd egoïsme leidt, dan is egoïsme (of preciezer geformuleerd: het verschil in egoïsme voor en na regelovertreding) de afhankelijke en al of niet door rood rijden de onafhankelijke variabele. Vaak is er sprake van meerdere onafhankelijke en meerdere afhankelijke variabelen (men spreekt dan van een multivariate vraagstelling).

Als voorbeeld: leiden egoïsme en gebrekkige opleiding tot meer door rood rijden en tot meer fout parkeren? De eerste twee zijn hier de onafhankelijke variabelen, wat allerminst wil inhouden dat ze onderling onafhankelijk zouden moeten zijn, de laatste twee de afhankelijke. De termen verklarende en te verklaren variabelen geven het onderscheid wellicht wat duidelijker weer dan het begrippenpaar 'onafhankelijk'-'afhankelijk'.

Meetschaal van variabelen

Sommige geoperationaliseerde begrippen kunnen slechts een beperkt aantal waarden aannemen. Wanneer we afstuderende juristen indelen in staatsrechtjuristen, strafrechtspecialisten en privatisten, dan hebben we een voorbeeld van een zogeheten nominale schaal: er is sprake van een aantal klassen en er is geen impliciete rangorde van die klassen. De klassennamen zijn een soort etiketten en je kunt niet zeggen dat iemand in de ene klasse meer of minder van enig kenmerk heeft dan in de andere klasse. Als we om redenen van afkortingsgemak cijfers als etiket gebruiken (1=staatsrecht; 2=strafrecht; 3=privaatrecht), dan loopt men het risico door de conventionele betekenis van cijfers in de val te lopen: 3 staat, in de rekenkunde, verder af van 1 dan 2 van 1, maar dat zegt niet dat privatisten meer gemeen hebben met strafrechtspecialisten dan met staatsrechtjuristen. Er zou dan ook niks tegen zijn om een willekeurig andere etikettering te kiezen (zoals 1=straf, 2=staats, 3=privaat). Als we de 'neutrale' etikettering bedoelen, dan spreken we van een nominale schaal.

Als we op grond van hun gedrag in een psychologisch laboratorium mensen classificeren als egoïst, individualist of altruïst, dan bedoelen we (of kunnen bedoelen: ook dit is een keuze) wel een ordening aan te brengen: altruïsten staan verder af van egoïsten dan van individualisten. Bij numerieke etikettering (1=egoïst; 2=individualist; 3=altruïst) willen we de volgorde-eigenschap van de cijfers juist wel erven; we spreken dan van een ordinale schaal. Toch kunnen we deze ordinale etiketten vervolgens niet gebruiken om mee te rekenen: hoezeer ook 3 driemaal zoveel is als 1, de bewering dat een altruïst (score 3) driemaal zo weinig egoïstisch is als een egoïst (score 1) is onzin.

Bij etikettering waarbij we wel gebruik willen maken van een deel van de 'gewone' numerieke eigenschappen van cijferlabels spreken we van een intervalschaal. Nemen we als voorbeeld de Egoïsmeschaal. Deze is samengesteld uit een aantal vragen waarmee men het in meerdere of mindere mate eens kan zijn. De score op de schaal loopt uiteen van 12 (volstrekt niet egoïstisch tot 60 (uitgesproken egoïstisch). Allereerst is deze schaal zeker ordinaal van karakter: hogere scores betekenen dat de zo geklasseerde respondent als egoïstischer wordt beschouwd. Maar er is meer aan de hand: als we zien dat Jan 50 scoort en Piet 40, Marie 35 en Mieke 45, dan bedoelen we daarmee dat we Jan even veel egoïstischer achten dan Piet, als Mieke ten opzichte van Marie.[7] We spreken van een intervalschaal, omdat aan gelijke verschillen (=intervallen) tussen scores gelijke betekenis wordt gehecht.

De absolute schaal is de etikettering die het meest betrekking heeft op de normale eigenschappen van getallen. Als we noteren hoe vaak iemand bij twintig gelegenheden door rood rijdt, dan loopt het antwoord van 0 tot 20; de schaalsco-

7. Nagaan of en onder welke condities een meetinstrument voldoet aan de eisen van een intervalschaal heet schaalanalyse en vormt onderdeel van een instrumentontwikkelingsonderzoek.

res hebben de eigenschappen van een intervalschaal, maar je mag er ook mee vermenigvuldigen en delen: wie 10 keer door rood rijdt, is een twee keer zo intense roodrijder als iemand die 5 keer door rood rijdt. Dat men zo kan rekenen met schaalscores op een absolute schaal is het gevolg van het bestaan van een natuurlijk nulpunt op de schaal. Bij de Egoïsmeschaal is dat niet het geval was: we zullen iemand die de hoogste score van 60 haalt niet tweemaal zo egoïstisch willen noemen als iemand die de score 30 haalt.

Met het karakter van een schaal hangt samen welke statistische procedures er op van toepassing kunnen zijn. We komen daar in de betreffende paragrafen op terug.

Onderzoeksontwerp en gegevensverzameling

Nu we na de operationalisatiefase weten wat we willen meten, volgt de vraag: bij wie gaan we wanneer en hoe vaak die metingen verzamelen? Deze vraag staat bekend als de vraag naar het onderzoeksontwerp of -design.

Observationele en experimentele opzetten

Er valt in ieder geval onderscheid te maken tussen observationele en experimentele opzet. Bij een experimentele opzet manipuleert de onderzoeker de omstandigheden (condities) waaronder de meting wordt verricht en kijkt of in de ene set omstandigheden iets anders wordt waargenomen dan in de andere set. Hij interpreteert de aangetroffen verschillen vervolgens als het gevolg van de verschillende condities.

> *Voorbeeld:* In een experiment bedoeld om te kijken onder welke omstandigheden egoïsten parkeren op parkeerplaatsen voor invaliden, manipuleert de onderzoeker de omstandigheden. Op een parkeerplaats bij een supermarkt zorgt hij de ene dag dat alle parkeerplekken tot 100 meter van de winkel bezet zijn, op één voor invaliden gereserveerde plaats na. De andere dag zorgt hij voor ruime parkeergelegenheid op 50 meter, naast de nog altijd vrije invalidenplaats. Hij observeert parkeergedrag van aankomende automobilisten, en neemt bij hen – na het parkeren – een Egoïsmeschaal af. Hij wil zo nagaan of egoïsme vooral in schaarstesituaties tot regelovertreding aanleiding geeft.

Tegen experimentele opzetten voert men vaak als bezwaar aan dat ze moeilijk te generaliseren zijn naar 'werkelijke' gedrag, omdat de ingrepen in de situatie wellicht juist niet-natuurlijk gedrag uitlokken.

Bij de observationele opzet neemt de onderzoeker waar 'wat zich voordoet,' zonder de omstandigheden naar zijn hand te zetten. Het is dan vaak moeilijker om een causale interpretatie aan gevonden verschillen te geven.

> *Voorbeeld:* Wanneer op bovengenoemde parkeerplaats geen experimentele ingrepen worden gedaan, dat wil zeggen dat er geen omstandigheden – zoals het verplaatsen van parkeerplaatsen voor invaliden – worden gemanipuleerd door de onderzoeker, wordt er slechts parkeergedrag en egoïsme gemeten. Stel nu dat egoïsme samenhangt met wangedrag, dan is het niet

zonder meer mogelijk het verband tussen egoïsme en regelovertreding te leggen, omdat er misschien concurrente verklaringen zijn. Wellicht komen egoïsten juist op drukke tijden, als er weinig plaatsen zijn, en de condities dus eerder tot wangedrag noden.

Controlegroep

Een ander onderscheid bij proefontwerpen is dat tussen die met en zonder controlegroep. Wie wil onderzoeken of een voorlichtingscursus over gebruik van alcohol helpt om dronken rijders op het rechte pad te houden, kan proberen na te gaan hoeveel procent van de deelnemers aan zo'n cursus binnen een jaar niet meer recidiveert. Als dat percentage 75 procent blijkt te zijn, kan hij de ontwerper van de cursus feliciteren. Of niet? Als blijkt dat onder dronken rijders die de cursus niet volgden óók 75 procent niet recidiveert, is er natuurlijk geen sprake van succes. Op het veronachtzamen van dit feit berust veel slecht onderzoek. Door een controlegroep in het onderzoeksontwerp in te bouwen kan vaak worden voorkomen dat men zich te gemakkelijk tot ongefundeerde conclusies laat verleiden. Is een controlegroep altijd nodig? Nee, soms is genoegzaam bekend hoe de zaak ligt in 'normale' condities en kan men zich de moeite besparen. Toch is een controlegroep vaak wel zo veilig.

Voorbeeld: Wij willen onderzoeken of een TBS-behandeling van verkrachters ertoe leidt dat zij in de toekomst, na hun ontslag, minder zullen verkrachten. En inderdaad: TBS-ers verkrachten minder dan voor hun behandeling. Maar weten wij nu zeker dat het door de behandeling komt? Het is best mogelijk dat het feit dat zij opgepakt zijn en veroordeeld op zichzelf genomen al ertoe leidt dat zij minder zullen verkrachten. Of misschien kan de afname wel toegeschreven worden aan het simpele feit dat na ontslag de oud-TBS-ers door hun leeftijd seksueel minder actief zijn geworden en daardoor minder gaan verkrachten. De onderzoeksvraag is daarom alleen te beantwoorden als de experimentele groep wordt vergeleken met een controlegroep die ongeveer even oud is wel is veroordeeld voor verkrachting, maar geen TBS-behandeling heeft gekregen.

Observatie

Een ander onderdeel van het onderzoeksontwerp is de specificatie van de gegevens- of dataverzameling. Natuurlijk is dit onderdeel van het ontwerp sterk verweven met de operationalisatie van de betrokken begrippen, maar we behandelen het hier apart.

Er is een aantal manieren om het verzamelen van gegevens te onderscheiden. De meest voor de hand liggende wijze is observatie: ter plekke zijn (dat wil zeggen: dáár waar het te onderzoeken gedrag plaatsvindt) en tot je laten doordringen wat er gebeurt.[8] De wijze van vastleggen kan verschillen: aantekeningen maken (al dan niet in een vastliggend observatieschema), bandopnamen maken (en die later analyseren), video-opnamen maken (idem), *real time* classificeren (voorbeeld: tijdens een teamsport voor een zekere deelnemer op computertoets

8. Zoals gebeurde in het onderzoek door Van de Bunt onder officieren van justitie. Zie Van de Bunt (1985).

'e' drukken als hij egoïstisch gedrag vertoont; de computer registreert hoelang die toets wordt ingedrukt). Kenmerkend voor observatie is dat de observator probeert er te zijn als het te onderzoeken gedrag plaats vindt en het dan waar te nemen, zonder zelf een rol in het gebeuren op zich te nemen.

Voorbeeld: Egoïsme en regelovertreding. Een typerende observatiestudie in dit verband is die naar het rijden door een rood stoplicht. Op een bepaald kruispunt werd telkenmale geobserveerd of de eerst aankomende auto stopte of doorreed nadat het licht op rood was gesprongen. Natuurlijk was het zaak daarbij een nauwkeurig observatievoorschrift op te stellen: er moesten minstens 2 seconden zijn verlopen na op rood springen, voor een observatie meetelde (om twijfelgevallen te voorkomen), en behalve het door-rood-rijden-gedrag als zodanig werd ook geobserveerd of de bestuurder passagiers bij zich had, of het een man of een vrouw was, en of de auto herkenbaar was als een bedrijfswagen. De rest van de in dit onderzoek relevante vragen werd vergaard via een naderhand toegestuurde Egoïsmevragenlijst.

Participerende observatie

Veel mensen zijn geneigd de term 'observatie' welhaast zonder nadenken van het epitheton 'participerende' te voorzien. Kennelijk bestaat er een diepgewortelde gedachte dat participerende observatie mooier, beter en indrukwekkender is dan zo maar observatie. Participerende observatie is echter heel wat anders dan observatie zoals boven omschreven. Van participerende observatie is sprake wanneer de onderzoeker deel uitmaakt van de sociale omgeving waarin het te onderzoeken gedrag plaatsvindt, hij aan die omgeving deelneemt en niettemin probeert waar te nemen en verslag te doen van wat er plaatsvindt.

Voorbeeld: Milieudelicten op zee zijn zeer moeilijk waar te nemen. De onderzoeker besluit daarom als matroos aan te monsteren op een olietanker om te onderzoeken wanneer en onder welke omstandigheden men op zee olie en ander afval loost.

Participerende observatie vindt soms plaats zonder dat de omgeving weet dat de onderzoeker onderzoekt (Walraff-stijl),[9] maar vaak ook terwijl de betrokkenen wel weten dat er (ook) sprake is van een onderzoeksdoel. Het hoeft geen betoog dat deze vorm van observeren erg arbeidsintensief is en meer dan enig andere vorm van dataverzameling gevoelig is voor beoordelaarsonbetrouwbaarheid.

Observatie leidt vrijwel altijd tot de noodzaak het beschouwde gedrag te klasseren in categorieën, die al of niet tevoren vastgesteld zijn of tijdens of na de observatie worden gedefinieerd. Veelal is het noodzakelijk de beoordelaarsbetrouwbaarheid van deze classificatieprocedure nauwkeurig na te gaan. Wanneer het geobserveerde gedrag is vastgelegd – op geluidsband of video – is het mogelijk meerdere beoordelaars onafhankelijk van elkaar de klassering te laten uitvoeren, verschillen te analyseren en te laten bespreken.

9. Günter Walraff is een Duitse journalist die bekendheid verwierf doordat hij vermomd als Turkse arbeider in de Duitse industrie werkte en zijn ervaringen te boek stelde. Zie Walraff (1985).

Interview

Een tweede veel gebruikte methode is het vraaggesprek of interview. Er dienen zich onmiddellijk enige varianten aan. Allereerst is er het voorgestructureerde interview. Dit interview is in zijn extreme vorm eigenlijk een soort voorgelezen vragenlijst, waarbij echter de geïnterviewde niet gebonden is aan een vast antwoordformaat (zoals meestal bij de vragenlijst). Een tweede variant is het halfgestructureerde interview, dat door de interviewer wordt gevoerd aan de hand van een lijstje van onderwerpen, die alle de revue moeten passeren, maar waarbij het aan de interactie tussen bevraagde en ondervrager wordt overgelaten wat er precies aan de orde komt en hoe uitgebreid dat wordt behandeld. De interviewer heeft vooral als taak de aandacht van de geïnterviewde bij het centrale onderwerp te houden ('Ja, maar ik wilde het eigenlijk vooral over door rood licht rijden hebben ...'). Een, door zijn onderwerp vrij bijzondere, vorm van het interview is het levensgeschiedenisinterview (*event-history*-interview). Daarbij probeert de onderzoeker zijn gesprekspartner te verleiden tot het opbiechten van voor het onderzoek belangrijke gebeurtenissen in zijn leven.

> *Voorbeeld:* Het is weinig zinvol om te trachten aan de hand van geregistreerde misdrijven van criminelen hun criminele levenswandel te reconstrueren, omdat slechts een klein deel van de misdrijven wordt opgelost. Wil men toch uitspraken doen over de misdrijven die mensen in hun leven pleegden, dan is men aangewezen op zo'n levensgeschiedenisinterview.

Net zoals de term 'observatie' te pas en te onpas wordt toegerust met het adjectief 'participerend,' zo wordt de term 'interview' vaak opgezadeld met de uitbreiding 'diepte,' maar lang niet altijd terecht. Een diepte-interview of vrije attitude-interview is een arbeidsintensieve vorm van interviewen waarbij de interviewer eigenlijk alleen het onderwerp ter sprake brengt, en de geïnterviewde stimuleert om zo veel mogelijk – al dan niet associatief – over het onderwerp te vertellen. Wat de geïnterviewde ook te berde brengt, alles wordt in dank aanvaard.

De drie interviewvormen stellen andere eisen aan de voorbereiding, maar ook aan het codeerwerk achteraf. Het vrije attitude-interview brengt bergen werk met zich mee om al het gezegde onder te brengen in zinvolle categorieën, geconstrueerd op grond van het door de geïnterviewden naar voren gebrachte. Beoordelaarsbetrouwbaarheid vormt een cruciaal punt. Daarentegen brengt het voorgestructureerde interview meer werk vooraf mee: opstellen van de vragen en uittesten of iedere geïnterviewde in die vragen zijn ideeën en gevoelens wel kwijt kan. Bij deze vorm is met name de validiteit van de verzamelde gegevens een punt van zorg.

De diverse vormen van interviews geven goed weer hoezeer het onderscheid kwalitatief-kwantitatief betrekkelijk is: algemeen worden diepte-interviews tot de kwalitatieve methoden gerekend, voorgestructureerde interviews heten een kwantitatieve methode ten zijn en over halfgestructureerde interviews lopen de meningen uiteen.

Vragenlijsten

Een derde veel gebruikte methode is de vragenlijst (*survey*). De meest op het voorgestructureerde interview gelijkende vorm is de mondelinge vragenlijst,

waarbij echter niet alleen de vragen, maar ook de mogelijke antwoorden van te voren zijn geformuleerd. Soms wordt de respondent ook die antwoordmogelijkheden voorgelezen en moet hij zelf kiezen. In andere gevallen brengt de interviewer zelf het gegeven antwoord onder in de vastgestelde categorieën. Een speciale vorm is het telefonisch interview.

De schriftelijke vragenlijst is een variant waarbij de ondervraagde zelf de vraag leest en een antwoord invult, vaak in voorgecodeerde vorm (aankruisen). Veelal worden schriftelijke vragenlijsten als postenquête uitgevoerd. Vragenlijsten hebben in het algemeen niet zo'n goede naam en krijgen veelal het verwijt door hun voorgestructureerdheid de respondent in een niet passend keurslijf te dwingen. Anderzijds maakt de standaardisering van vragenlijsten een diepgaande analyse van hun eventuele tekortkomingen mogelijk. Een mengvorm is de open vraag. Waar de meeste vragenlijsten gesloten vragen hanteren ('Heeft u de laatste week wel eens door rood gereden? ❑ nee, nooit; ❑ 1 maal ❑ 2 maal; ❑ meer dan twee maal; ❑ ik weet het niet meer'), hanteren andere vragenlijsten een gestandaardiseerde, maar open vorm. Daarbij moet de respondent zijn antwoord in eigen woorden weergeven ('Waarom bent u de laatste week wel door rood gereden? Omdat, … (vul in)'). Eigenlijk zijn open vragen een niet altijd geslaagde poging om het interview in de schriftelijke vragenlijst te incorporeren. De bijbehorende coderingsproblematiek wordt daarmee namelijk eveneens geïmporteerd.

Een speciale vorm van de schriftelijke vragenlijst is het zogeheten telepanel. Het Nederlands Instituut voor de Publieke Opinie en het Marktonderzoek, NIPO, heeft een groep van ruim duizend Nederlandse huishoudens uitgerust met een PC met modem waarop wekelijks een vragenlijst wordt geplaatst, die de leden van het huishouden op de PC beantwoorden, waarna de antwoorden per modem naar het NIPO worden doorgeseind.

Scenario's

Een bijzondere vorm van meting vormen de scenario's. Daarbij worden de ondervraagden geconfronteerd met een korte schets van een levensechte situatie. Vervolgens wordt hen gevraagd: 'Wat zou u doen in deze situatie?' Scenario's zijn eigenlijk een soort semi-experimenten. Door bepaalde onderdelen in de scenario's te wijzigen, kan men proberen het effect van de in dat onderdeel vervatte factoren te benaderen. Zo kan men zich voorstellen dat de boven aangehaalde parkeerstudie in plaats van als levensecht veldexperiment (kostbaar, moeizaam, tijdrovend) als scenariostudie (snel, goedkoop, gemakkelijk, maar wellicht ook minder 'levensecht' en daarmee misschien minder valide) wordt uitgevoerd.

Voorbeelden: (1) Een onderzoek naar attitudes over *date-rape* wordt als scenario-studie uitgevoerd: terwijl de meeste vragen (onder andere die naar egoesme) met een standaardvragenlijstmethode werden gesteld, werden de vragen naar *date-rape*-gedrag met scenario's vormgegeven. Het volgende fragment is een illustratie:

Robert en Carla gaan al enige maanden regelmatig met elkaar uit; ook vanavond hebben ze een disco bezocht; ze zijn allebei nogal aangeschoten, en als ze naar huis gaan gaat Robert nog even bij Carla mee naar binnen, en ze drinken nog wat; op de bank gezeten knuffelen ze wat, maar als Robert verder wil gaan, zegt Carla geen zin in vrijen te hebben; Robert houdt ech-

ter aan en ondanks Carla's protest drukt hij haar na enige tijd op de bank neer en heeft seks met haar. De vraag (aan mannen) is dan: hoe waarschijnlijk is het dat jij in een dergelijke situatie je als Robert zou gedragen? Antwoordcategorieën variëren van 'uitgesloten' tot 'zeker'.

(2) Ook de studie van Ten Kate en Van Koppen naar informatiebehoefte bij rechters maakte gebruik van scenario's: er werd een civielrechtelijke casus geschetst, en de, zich in de rol van rechter verplaatsende, respondent kon de onderzoekers vragen om nadere informatie, die – geanticipeerd door de onderzoekers – veelal ook aanwezig was.[10]

Simulatie

Een verder doorgevoerde vorm van de scenariostudie is de simulatie. De ondervraagde wordt in een zo levensecht mogelijke situatie gebracht, en zijn gedrag wordt genoteerd. Een simulatieomgeving kan fysiek zijn, maar vaak wordt de computer voor dat doel gebruikt. Dit onderwerp is vrij specialistisch, en we verwijzen naar het onderzoek van Webley en collega's.[11]

Voorbeeld: Belastingontduiking werd bestudeerd in een simulatie van een kleine supermarkt.[12] De ondervraagde werd geacht zich als ondernemer in een supermarkt te gedragen, een supermarkt die op een PC werd gesimuleerd. De ondervraagde kon prijzen bepalen, adverteren, personeel in dienst nemen, investeren en zodoende een omzet maken. Vervolgens kon hij – en daar ging het allemaal om – zijn belastingformulier invullen. Omdat alle cijfers op de PC waren gegenereerd kon ondubbelzinnig worden nagegaan of er sprake was van ontduiking binnen de simulatie. In hoeverre dat ook iets zegt over 'echt' ontduikingsgedrag wordt in de aangehaalde literatuur nauwkeurig besproken.

Dossieronderzoek

Zeker in het juridische veld is een veel voorkomende vorm van onderzoek het dossieronderzoek. Eigenlijk is dossieronderzoek een vorm van observatie, maar dan niet van dat wat feitelijk gebeurt, maar van de schriftelijke weergave ervan. Dat betekent dat de onderzoeker een indirecte vorm van waarneming toepast: eerst geeft de dossieropsteller (politieman, griffier of notulist) een impressie – vaak gereglementeerd door vormvoorschriften – van wat er is geschied. Daarna probeert de onderzoeker aan de hand van dit verslag te achterhalen wat voor zijn onderzoek relevant is. Ten opzichte van directe waarneming is dossieronderzoek zowel behept met voordelen als met nadelen. Het feit dat er een – niet door onderzoeksmotieven geïnspireerde – tussenpersoon actief is, maakt het verband met het gebeurde zwakker: de notulisten hebben met het maken van het betreffende dossier veelal een ander doel dan het leven van de onderzoeker te veraangenamen. Anderzijds vormen dossiers vaak een niet door gewone observatie haalbare compilatie van diverse processen, zoals: wat de politie rapporteert, wat

10. Ten Kate & Van Koppen (1984).
11. Webley, Robben, Elffers & Hessing (1991).
12. Robben (1991) en Webley, Robben, Elffers & Hessing (1991).

de officier van justitie zegt, hoe de rechter-commissaris ertegenaan kijkt, wat de verdediging te berde brengt, wat de rechtbank vroeg, wat het vonnis was, etc. In zo'n geval is het haast ondenkbaar dat men in een vroeg genoeg stadium aanwezig zou kunnen zijn om alle fasen actueel te observeren.

> *Voorbeeld:* Bij onderzoek naar belastingontduiking werden door een speciaal aangezochte groep van drie ervaren belastingambtenaren belastingdossiers nog eens doorgelicht op de noodzaak tot het aanbrengen van een correctie en of er in hun ogen sprake was van opzettelijke ontduiking.[13]

Een bijzondere vorm van dossieronderzoek is inhoudsanalyse van documenten. Daarbij wordt gepoogd beschikbare documenten (en dat kan van alles zijn, zoals historische actes, egodocumenten, krantenartikelen en dagboeken) door te nemen om na te gaan of er een beeld over een bepaald onderwerp uit oprijst.

Fysieke metingen

Tenslotte noemen we de fysieke metingen. Soms is het mogelijk om bepaalde te onderzoeken onderwerpen vast te stellen op grond van fysieke metingen. Voorbeelden in de verkeerscontext zijn: snelheidsmetingen en bloedalcoholgehalten, maar ook kan men denken aan identificatiemiddelen zoals vingerafdrukken en DNA-patronen.

> *Voorbeeld:* Bij onderzoek naar egoïsme en dronken rijden kan het laatste worden vastgesteld met behulp van de bloedproef of ademtest.

Steekproeven

In veel onderzoek gaat het om het vaststellen van de waarde van variabelen zoals die voorkomen in een grote groep personen waarop het onderzoek betrekking heeft. Deze groep wordt de onderzoekspopulatie genoemd, en kan uit zeer diverse individuen (eenheden van analyse) bestaan. Vaak gaat het om mensen (voorbeeld: alle sociale advocaten), combinaties van mensen (voorbeeld: alle echtparen, bij een onderzoek naar huwelijksvermogensrecht) of gebeurtenissen (voorbeeld: bankovervallen).

Omdat populaties doorgaans erg omvangrijk zijn, is het meestal niet mogelijk bij alle eenheden in de populatie de waarde van de onderzoeksvariabelen vast te stellen. Dan richt het onderzoek zich op het nauwkeurig bekijken van een deel van de populatie, een deel dat dan steekproef wordt genoemd. Soms bepaalt de onderzoeker, vaak aan de hand van een toevalsmechanisme, welke eenheden uit de populatie in de steekproef terechtkomen.

> *Voorbeeld:* uit de ledenlijst van de broederschap van notarissen wordt elke naam op een briefje geschreven, en in een grote doos gestopt; na goed roeren neemt men zonder te kijken een greep van 100 briefjes, en de betreffende notarissen worden aangeschreven voor een interview.

13. Elffers (1991).

Ook wordt vaak door de inrichting van het onderzoek slechts een bepaald deel van de populatie bekeken (voorbeeld: bij een stoplicht wordt gedurende 3 dagen van 7-12 uur elke automobilist die als eerste bij het op rood springende licht aankomt in het onderzoek opgenomen). Soms is de populatie niet eens zo duidelijk, omdat eigenlijk alleen de steekproef zelf welomschreven is. In het laatste voorbeeld kan je je afvragen wat eigenlijk de populatie is: alle automobilisten, alle automobilisten die wel eens (welke periode?) bij het betreffende stoplicht komen of alle automobilisten die op de betreffende drie dagen in hun auto hebben gereden?

Aselecte steekproef

De pretentie van veel onderzoek is dat wat wordt aangetroffen bij de steekproef een goede afspiegeling is van hetgeen zou worden gevonden als men de hele populatie doorlicht. Het gevaar dat men loopt is dat de steekproef door de wijze waarop zij is samengesteld een vertekening geeft ten opzichte van de populatie. De best hanteerbare remedie tegen die kwaal is de steekproef op basis van het toeval samen te stellen (toevalssteekproef, aselekte steekproef of *random sample*). Kenmerk van een toevalssteekproef is dat elk element van de populatie door de trekkingsprocedure gelijke kans heeft om in de steekproef terecht te komen. In dat geval is het met behulp van de wiskundige statistiek mogelijk om te kwantificeren hoe groot het verschil tussen steekproefresultaat en onbekend populatieresultaat met grote waarschijnlijkheid ten hoogste is. Op deze techniek komen we hieronder terug.

Bij het bovenaangehaalde voorbeeld van een onderzoek onder Nederlandse notarissen is sprake van een aselecte steekproef. Vaak maakt men in de praktijk gebruik van quasi-toevalsmechanismen: in plaats van de omslachtige procedure met de lootjes met notarisnamen pakt men de namenlijst, begint op een willekeurig punt en neemt elke tiende naam tot men de steekproef gevuld heeft. Strikt genomen is dat geen aselecte steekproef, maar men behandelt zulke procedures alsof dat wel het geval is (systematische steekproef). Het is vaak echter niet mogelijk om een echte aselecte steekproef te trekken, of als het al mogelijk is, dan is het te omslachtig en duur. In het stoplichtenvoorbeeld is geen sprake van een echte aselecte steekproef en het is ook niet duidelijk hoe men die zou kunnen trekken, allereerst omdat de populatie niet welomschreven is. Maar zelfs als dat wel zo zou zijn (populatie: alle Nederlanders met een rijbewijs), dan is de waargenomen reeks automobilisten niet in redelijkheid met een toevalssteekproef te vergelijken: Nederlanders uit woongebieden ver van het betreffende kruispunt hebben een veel kleinere kans te worden waargenomen, mensen die niet 's ochtends rijden hebben geen kans, mensen die veel rijden hebben een grotere kans, etc. Toch wordt veelal in onderzoek zonder veel omhaal aangenomen dat het onderzoeksmateriaal representatief is voor een veel grotere populatie. Terwijl een onderzoek eigenlijk tot stand is gekomen door een groep studenten die op een bepaalde dag in Rotterdam het tweedejaars college Rechtspsychologie bijwoonden een vragenlijst te laten invullen, worden de resultaten soms moeiteloos gepresenteerd als staande voor alle Rotterdamse tweedejaars rechtenstudenten, alle Rotterdamse rechtenstudenten, alle Rotterdamse studenten, alle Nederlandse studenten, alle Nederlanders, alle mensen. Dat kan terecht zijn, maar het is goed in te zien dat zo'n generalisatie eigenlijk berust op het vertrouwen van de onderzoeker dat er tussen de betrokken groepen geen verschil

van betekenis zal bestaan. Of die aanname klopt, staat niet op voorhand vast en het is verstandig daar bij concreet onderzoek over na te denken. Is het reëel om het materiaal als een toevalssteekproef uit een populatie te beschouwen? Welke vertekening heeft eventueel plaatsgevonden? Wie onderzoek doet naar de motieven van bankovervallers door er 20 in de gevangenis te bezoeken, kan niet volhouden dat zijn steekproef alle bankovervallers representeert: degenen die niet gepakt of veroordeeld zijn, komen niet voor, en de langgestraften (ernstigere gevallen?) hebben een grotere kans te worden geïnterviewd dan de kortgestraften. Men spreekt dan van vertekening of *bias*.

Een beruchte vorm van bias treedt bijvoorbeeld op als bij telefonisch interviewen alleen gebruik wordt gemaakt van het telefoonboek. Men mist dan mensen zonder telefoon en mensen met geheime nummers. In juridisch onderzoek is een groot gevaar van vertekening gelegen in het zich baseren op (een steekproef uit) gepubliceerde arresten, als men als populatie 'alle arresten' voor ogen heeft, omdat immers de reden van publicatie veelal is dat er iets bijzonders met de zaak aan de hand was.

Non-respons

Een belangrijke bron van vertekening kan de zogeheten non-respons zijn. Aangezien onderzoek veelal de medewerking of toestemming van betrokkenen vergt, is het onvermijdelijk dat niet iedereen die voor onderzoek wordt benaderd daaraan ook feitelijk meedoet. Vaak moet men vrezen dat de resterende steekproef vertekend is, omdat de beslissing al of niet mee te werken, samenhangt met het onderwerp van onderzoek.

> *Voorbeeld:* Egoïsme en regelovertreding. Medewerking aan een onderzoek zal door egoïsten wellicht vaker geweigerd worden en regelovertreders willen misschien minder graag aan de tand gevoeld worden dan zij die zich aan de regels hielden.

Als cijfers over de hele populatie bekend zijn, is het soms mogelijk om na te gaan of de steekproef op een aantal relevante kenmerken vergelijkbaar is met de populatie.

> *Voorbeeld:* In onderzoek onder belastingplichtigen over hun kijk op de belastingdienst is sprake van veel non-respons (meer dan 50%). Hangt dit samen met de behandeling die betrokkenen van de dienst hebben ondervonden? Er werd in de steekproef aangetroffen dat een derde van de ondervraagden de voorafgaande twee jaar een correctie van de belastingdienst op hun aangifte IB had ontvangen. Uit cijfers van de dienst is bekend dat in de populatie van particuliere IB-plichtigen dat percentage 36 is. Er is dus nauwelijks verschil, en op dit punt is er geen sprake van vertekening. Dat neemt niet weg dat er op andere punten wel sprake kan zijn van vertekening.

Error

In onderzoek hoopt men dat een meetinstrument identieke resultaten geeft als het twee keer bij dezelfde persoon wordt toegepast en de relevante omstandighe-

den zich niet gewijzigd hebben (herhalingsbetrouwbaarheid). Dat is echter zelden exact het geval. Eenzelfde respondent zal op eenzelfde reeks vragen niet altijd exact gelijk antwoorden, doordat zijn stemming wisselt, hij net andere ervaringen heeft meegemaakt, de ene interviewer de andere niet is en wat niet al. Men kijkt tegen dit treurige feit veelal aan door de gegeven antwoorden te beschouwen als een steekproef uit de hypothetische populatie van alle denkbare replicaties van het afnemen van het instrument. Het (onbekende) verschil tussen de feitelijke resultaten en de 'echte' hypothetische stand van zaken in de populatie van alle afnamen wordt veelal aangeduid met de term *error*. Die term geeft goed weer wat men ervan hoopt: dat het om kleine niet belangrijke fouten gaat, die door irrelevante omstandigheden worden veroorzaakt. Door een betrouwbaar instrument te gebruiken, of zonodig door meerdere afnames te doen, kan de error klein gehouden worden.

Samenvattend: steekproefresultaten kunnen afwijken van een populatieresultaat door drie bronnen:
- Het feit dat het om een aselecte steekproef gaat (statistisch onder controle te krijgen);
- Vertekening, doordat de steekproef geen echte aselecte steekproef is;
- Error, omdat de feitelijke resultaten afwijken van de 'echte' (ruisvrije) resultaten.

Kwantitatieve data-analyse: beschrijvende statistiek

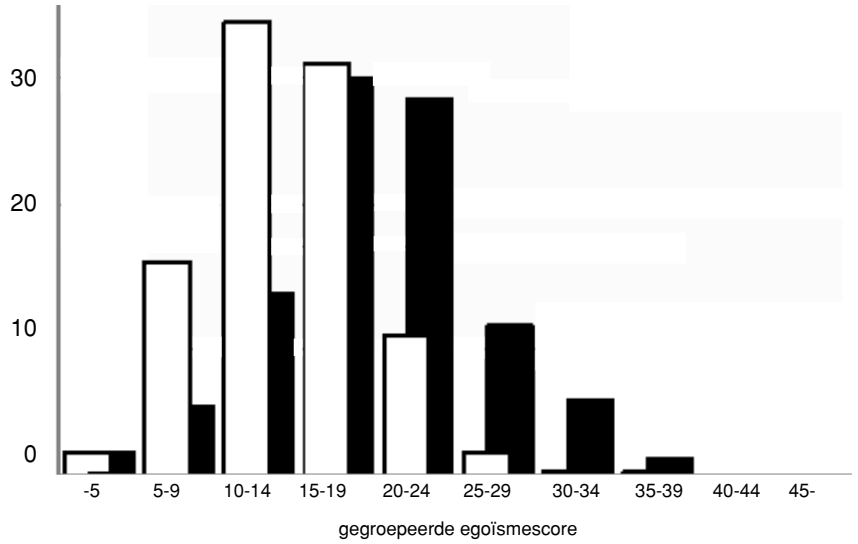
In deze paragraaf wordt een overzicht gegeven van veel gebruikte statistische methoden om de resultaten van kwantitatieve onderzoeksmethoden te presenteren. Men spreekt ook wel van beschrijvende statistiek. In de volgende paragraaf zal het verband tussen steekproef en populatie aan de orde komen. Dan gaat het om de toetsende of inferentiële statistiek.

De ruwste vorm van presenteren van gegevens is het blootweg opsommen van alle verkregen resultaten, maar meestal is dat te onoverzichtelijk. Beschrijvende statistiek biedt een reeks standaardprocedures om een samenvatting van deze complete reeks van gegevens te verschaffen. Samenvatten leidt tot het benadrukken van bepaalde eigenschappen van de gegevens (data), hetgeen inzicht geeft in die eigenschappen, maar anderzijds andere eigenschappen verdoezelt. De keuze van wat er gepresenteerd wordt is dan ook belangrijk. Tot de belangrijkste middelen behoren enerzijds grafische presentaties ('plaatjes'), anderzijds samenvattende grootheden, zoals bijvoorbeeld gemiddelden. In deze tekst zal niet worden ingegaan op hoe men precies die grootheden kan berekenen – dat gebeurt meestentijds trouwens toch door een computerprogramma – maar op de erin vervatte gedachte en de interpretatie ervan. Wij slaan ook de meest voor de hand liggende maten over, omdat die in ieder eenvoudig statistiekboek te vinden zijn. Wij beperken de bespreking tot maten die gaan over de samenhang van variabelen.

Onderling verband tussen variabelen

Eén van de hoofddoelen van statistische presentatie is vaak het weergeven van het verband tussen twee waargenomen variabelen (die aan dezelfde onderzoekseenheden zijn waargenomen). Men spreekt dan van bivariate waarnemingen.

Een voorbeeld is het verband tussen verklarende en te verklaren variabelen. Als bepaalde waarden van de ene variabele bij veel analyse-eenheden in combinatie voorkomen met bepaalde waarden van de andere variabele, spreekt men van verband. Figuur 1 geeft een voorbeeld van zo'n verband: hoge waarden op de Egoïsmeschaal komen relatief vaak voor in combinatie met de waarde 'USA' op de variabele land van herkomst, en lage egoïsmewaarden met de waarde 'Nederland' op die variabele.



Figuur 1: Voorbeeld vergelijking Egoïsmeschaal in Nederland en de Verenigde Staten. (witte balken Nederland; zwarte balken Verenigde Staten)

Bij nominale en ordinale schaaltypen is de kruistabel de meest gebruikte manier om verband te laten zien. Tabel 1 zet de variabelen afstudeerrichting (nominaal) af tegen geslacht. De vermelde percentages zijn rijpercentages, dat wil zeggen dat ze per rij optellen tot 100 procent. We lezen dus af dat 23 procent van de mannen staatsrechtelijk afstudeert, etc.

Aan de hand van de absolute aantallen is moeilijk te zien of er sprake is van verband tussen geslacht en afstudeerrichting, maar aan de hand van de rij-percentages is gemakkelijk te zien dat de verdeling van de afstudeerrichtingen voor de beide geslachten niet veel van elkaar afwijkt: er is nauwelijks verband. Wanneer we onder de indruk zijn van verschillen tussen twee groepen (verdelingen), is een kwestie van smaak. In veel gevallen is een verschil van meer dan 10 procent tussen de rijpercentages in dezelfde kolom zeker de moeite van het vermelden waard, maar het is heel goed denkbaar dat bij sommige data veel kleinere verschillen al buitengewoon interessant kunnen zijn. In het bijzondere geval dat beide variabelen in een kruistabel slechts twee waarden hebben (dichotome variabelen) noemt men de kruistabel veelal een 2-bij-2-tabel.

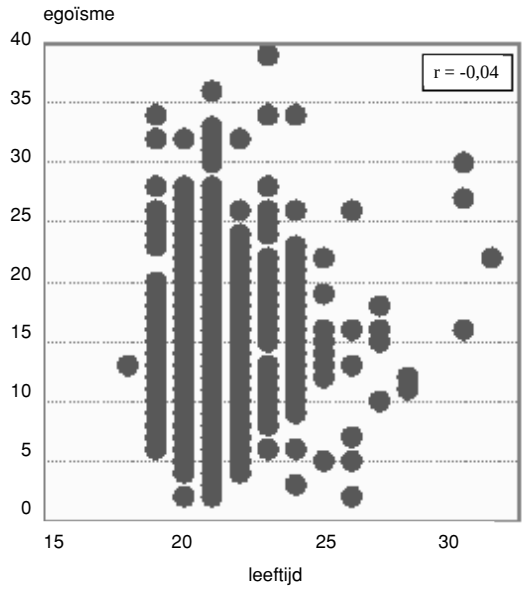
Tabel 1: Kruistabel (fictieve cijfers; aantallen en rij-percentages).

geslacht	afstudeerrichting			totaal (kolom %)
	privaatrecht	staatsrecht	strafrecht	
mannen	38 51%	17 23%	20 26%	75 (61%)
vrouwen	22 46%	12 25%	14 29%	48 (39%)
totaal	60 48%	29 24%	34 28%	123

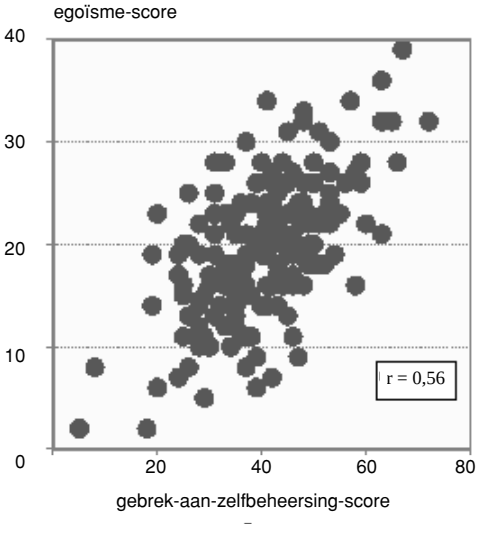
Wanneer de ene variabele nominaal of ordinaal is (land van herkomst) en de andere een intervalschaal heeft (egoïsme), dan is het mogelijk om ook een kruistabel te maken, mits de intervalvariabele in niet teveel klassen wordt samengevat. Ook is het mogelijk om een grafiek als figuur 1 te maken, maar het meest gebruikt is wel het vergelijken van de gemiddelde in de subgroepen bepaald door de nominale variabele: Amerikanen hebben g=19.3 en Nederlanders g=14.0. Is dat een verschil om opgewonden over te raken? Dat is een inhoudelijke, en geen statistische vraag, maar we willen wel een handige vuistregel voorstellen: we vinden een verschil in gemiddelden tussen twee groepen interessant als het groter is dan een derde van de standaarddeviatie.[14] (Als de standaarddeviaties sterk verschillen is dat op zich interessant, en kan men in de vuistregel uitgaan van het gemiddelde van de twee standaarddeviaties). In het voorbeeld is de (gemiddelde) standaarddeviatie 6, zodat het aangetroffen verschil van 5.3 zeker opvallend is. Merk op dat als de nominale/ordinale variabele veel categorieën heeft, grafieken als figuur 1 (waarin voor elke categorie een eigen verdeling van de intervalvariabele-score wordt getekend) aan overzichtelijkheid inboeten. Het bestuderen van de gemiddelden komt dan meer in aanmerking.

Wanneer er sprake is van twee intervalvariabelen (of ook wel bij twee ordinale variabelen) kan men, na opdeling van de scores in groepen, ook weer gebruikmaken van kruistabellen, maar men kan ook, en dan zonder de scores te hoeven groeperen, een puntenwolk (scattergram) tekenen. Dat is een plaatje waarin men alle analyse-eenheden door een punt representeert, met als coördinaten op de X-as de waarde op de ene variabele, en op de Y-as de waarde op de tweede variabele. Figuren 2 en 3 laten twee puntenwolken zien. De eerste, waar leeftijd is uitgezet tegen de egoïsmescore in een onderzoek onder Rotterdamse en Amerikaanse studenten, is een voorbeeld van een gebrek aan verband: het is niet het geval dat de jongere studenten over het algemeen hogere of juist lagere scores op de Egoïsmeschaal hebben.

14. De standaarddeviatie is een maat voor de spreiding van een verdeling.



Figuur 2: Puntenwolk, geen verband.



Figuur 3: Puntenwolk met positief verband.

Figuur 3 laat het verband bij de Amerikaanse studenten zien tussen de Egoïsmeschaal en de Gebrek-aan-Zelfbeheersingschaal (GZB-schaal). Gebrek aan zelfcontrole en impulsiviteit wordt in criminologisch onderzoek vaak als mogelijke verklaarder voor overtredingsgedrag genoemd.[15] Hier zien we een positief verband: over het geheel genomen gaan lage waarden op de Egoïsmeschaal samen

15. Zie bijvoorbeeld M.R. Gottfredson & Hirschi (1990).

met lage waarden op de GZB-schaal (dat wil dus zeggen: sterke zelfbeheersing), en hoge met hoge. Toch is het verband niet zo sterk dat bij één waarde op de GZB-schaal ook maar één waarde op de Egoïsmeschaal voorkomt. Integendeel: bij een bepaalde GZB-score komen nog heel wat verschillende egoïsmescores voor. Naarmate de puntenwolk meer op een dunne sigaar lijkt die min of meer diagonaal loopt, is er sprake van een sterker verband: bij een score op de X-as horen maar weinig verschillende scores op de Y-as. Bij een dikke sigaar is het verband zwakker, en ontbreken van associatie ziet men, zoals in figuur 2 terug in de vorm van een amorfe puntenwolk.

We spreken in figuur 3 van een positief verband, omdat hogere scores op de ene schaal samengaan met hogere scores op de andere schaal. Het kan ook voorkomen dat de puntenwolk van linksboven naar rechtsbeneden verloopt: hoge scores op de ene schaal zijn dan geassocieerd met lage scores op de andere, en vice versa. Men spreekt dan van een negatief (of omgekeerd) verband.

Associatiematen

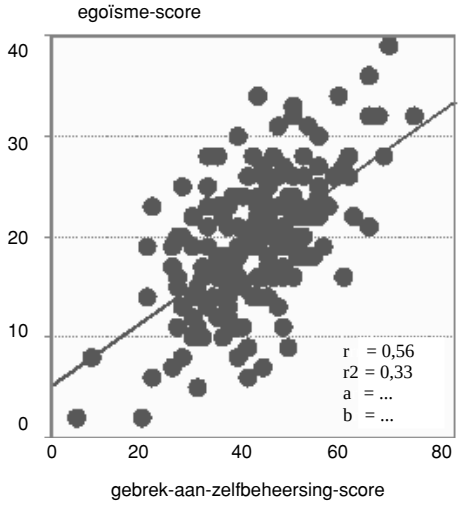
Met associatiematen wordt geprobeerd een puntenwolk samen te vatten. De meest gebruikte associatiemaat is de correlatiecoëfficiënt, die voluit Pearson's productmoment-correlatiecoëfficiënt heet, maar veelal wordt afgekort tot correlatie, vaak met r of ρ (rho, de Griekse r) aangeduid. Wij zullen voorbijgaan aan de wijze waarop deze coëfficiënt wordt berekend, en volstaan met te vermelden dat de coëfficiënt varieert tussen -1 en 1. Gebrek aan associatie weerspiegelt zich in een r van 0, perfecte positieve correlatie (alle punten van de puntenwolk liggen dan op één stijgende lijn van 45 graden) in r=1, perfecte negatieve correlatie in r=-1, positieve samenhang leidt tot een positieve r, negatieve samenhang tot een $r < 0$. Over wanneer een correlatiecoëfficiënt als groot beschouwd mag worden verschillen de meningen, en het hangt natuurlijk ook weer van het probleem af waarvoor de correlatie wordt bestudeerd. Wij houden de maatstaf aan als weergegeven in tabel 2.

Tabel 2: Beoordeling van correlatiecoëfficiënten (r).

Ondergrens	Bovengrens	Waardering
	≤ -0.50	sterk negatief verband
- 0.50 <	≤ -0.30	matig negatief verband
- 0.30 <	≤ -0.15	zwak negatief verband
- 0.15 ≤	< +0.15	nagenoeg geen verband
+ 0.15 ≤	< +0.30	zwak positief verband
+ 0.30 ≤	< +0.50	matig positief verband
+ 0.50		sterk positief verband

Ook bij variabelen die op ordinaal niveau zijn gemeten kan men een (ordinale) associatiemaat berekenen, waarbij veelal de zogeheten rangcorrelatiecoëfficiënt Kendalls ~ (tau, de Griekse t) wordt gebruikt. De berekening van ~ en r is verschillend, maar de gedachte achter de maten is analoog. Waarden van ~ zijn vaak wat lager dan overeenkomstige waarden van r. Tabel 2 kan men voor ~ gebruiken als men de grenswaarden met 2/3 vermenigvuldigt.

Ook voor het verband tussen nominale variabelen in een kruistabel wordt wel een associatiecoëfficiënt opgeven als uitdrukking van de sterkte van het verband. Een veelgebruikte maat voor dit soort variabelen is Cramers *V*. Omdat er bij nominale variabelen geen sprake is van een richting van een verband (er is immers geen sprake van 'hoger' of 'lager' scoren), wordt Cramers *V* uitgedrukt op een schaal van 0 (geen verband) tot 1 (volledige afhankelijkheid). Over welke waarden van *V* als uitdrukking geven aan een sterk, matig, of zwak verband zijn, verrassend genoeg, geen duidelijke maatstaven in omloop. Wij opteren voor $V <$ 0.05 : nagenoeg geen verband, $0.05 \leq V <.10$ zwak verband, $.10 \leq V < .175$ matig verband, $V \geq .175$ sterk verband.



Figuur 4: Regressielijn, gebrek aan zelfbeheersing–egoïsme.

Regressieanalyse

Wanneer er sprake is van een verklarende (onafhankelijke) variabele *V* en een te verklaren (afhankelijke) variabele *W*, die in een puntenwolk een duidelijke correlatie laten zien, gaat men vaak over tot het trekken van een regressielijn in de puntenwolk (figuur 4). Een regressielijn is te zien als een korte beschrijving van het verband dat tussen beide variabelen bestaat: het is de as van de 'sigaar' die een puntenwolk met duidelijke samenhang vormt. Bekijk eens alle eenheden in de puntenwolk die een waarde van de verklarende variabele *V* hebben die (ongeveer) gelijk is aan V_0 (zie in figuur 4 bijvoorbeeld bij V_0=60). Men ziet dat de waarden die deze eenheden op de te verklaren variabele *W* hebben, gecentreerd liggen rond de waarde W_0 op de regressielijn (Bij V_0=60 hoort een W_0 van 26.1; de *W*-waarden rond die lijn bij V_0=60 lopen uiteen van 16 tot 36, en dat is toch een veel kleiner bereik dan de totale waardenrange van *W*, die van 2 tot 39 loopt). Omdat een lijn in het (*V*,*W*)-vlak geschreven kan worden als $W = \alpha + \beta * V$ kan men, als men eenmaal de zogeheten regressiecoëfficiënten α en β kent, de waarde *W* bij elke *V* voorspellen. In figuur 4 is $\alpha = 5.11$ en $\beta = 0.35$. We kunnen de regressiecoëfficiënt â aldus interpreteren: gemiddeld gesproken geldt in de groep Amerikaanse studenten, dat wie een punt hoger scoort op de GZB=schaal, 0.35 punt hoger scoort op de Egoïsmeschaal. Natuurlijk is die voorspelling niet

voor elke eenheid exact correct: er zijn eenheden met waarden van *W* boven de lijn, en ook eenheden met waarden van *W* onder de lijn, maar toch is de voorspelling beter dan als we *V* niet zouden weten (dan kunnen we voor *W* immers niet beter voorspellen dan het gemiddelde van alle *W*). De voorspelling is beter naarmate de puntenwolk meer op een dunne sigaar lijkt, dat wil zeggen als de correlatiecoëfficiënt *r* dichter bij 1 (of bij -1) ligt. Eigenlijk is *r* dus een goede uitdrukking van de sterkte van de regressievoorspelling, maar men geeft meestal in plaats van *r* liever r^2 op (daarmee is men meteen het minteken bij negatieve samenhang kwijt). De waarde ervan wordt meestal determinatiecoëfficiënt of verklaarde variantie genoemd. Deze laatste benaming komt voort uit bovenstaande voorstelling in termen van voorspelling: rond de stippellijn (waar *V*=60 is) is een maat voor de voorspellingsfout van de waarde *W* als we *V* weten juist de standaarddeviatie s_V, terwijl als we *W* niet weten de voorspellingsfout weergegeven wordt door de standaarddeviatie van alle *W*-scores, *s*. Nu blijkt het zo te zijn dat $r^2=(s^2-s_V^2)/s^2$. Dat wil zeggen de totale variantie minus de resterende variantie (als fractie van de totale variantie), ofwel dat gedeelte van de aanvankelijke voorspellingsonzekerheid dat door de regressielijn is weg-'verklaard.'

Kwantitatieve data-analyse: Inferentiële statistiek

Steekproef als benadering van populatie

Eerder hebben we aangeroerd dat empirisch onderzoek veelal gebruik maakt van steekproeven in plaats van het te omvangrijke, zo niet onmogelijke onderzoek van een complete populatie. Natuurlijk brengt dat een risico met zich mee. De steekproef is immers niet identiek aan de populatie, dus kan men zich vergissen als men afgaat op de steekproefresultaten en conclusies trekt die in de populatie helemaal niet gelden. Door gebruik te maken van toevalssteekproeven (aselecte steekproeven) dekt men zich enigermate in tegen dit risico. Met behulp van de waarschijnlijkheidsrekening kan worden afgeleid hoe groot de kans is dat (onder zekere voorwaarden) het steekproefresultaat in het geval van aselecte steekproeven 'gemiddeld' gelijk is aan het populatieresultaat.

Toegegeven, deze formulering is nogal cryptisch. Laten we eens een voorbeeld nemen. Stel wij nemen een aselecte steekproef *S* van 20 studenten. Wij berekenen het gemiddelde *W* van hun 20 egoïsmescores om daarmee een indruk te krijgen van de gemiddelde egoïsmescore ì in de populatie van 'alle studenten'. Wij kunnen erop vertrouwen dat deze operatie 'gemiddeld' goed zit, in de betekenis dat als we alle mogelijke steekproeven $S_1, S_2, \ldots S_N$ zouden bekijken (*N* is duizelingwekkend groot, maar daar gaat 't nu niet om), en telkenmale de gemiddelde egoïsmescore $m_1, m_2, \ldots, m_N$ zouden uitrekenen, dan is het zo dat het 'gemiddelde van alle gemiddelden' , dus $(m_1+m_2+\ldots+m_N)/N$ juist exact gelijk is aan het populatiegemiddelde ì van de egoïsmescores van alle studenten. In die zin is steekproefonderzoek 'gemiddeld goed,' ofwel het geeft, zoals het ook wel wordt uitgedrukt, *in the long run* de juiste resultaten. We merken evenwel op dat dit in zekere zin maar een schrale troost is, want wat heb je eraan of een procedure *in the long run* bevredigend is, als je helemaal niet te maken hebt met een *long run,* maar met slechts één steekproef? Het steekproefresultaat van juist deze steekproef kan immers toevalligerwijs nogal misleidend zijn (zonder dat je het weet).

We merken allereerst op dat de kans op een misleidend resultaat groot is als de spreiding onder alle denkbare steekproefuitkomsten $m_1, m_2, \ldots, m_N$ groot is: dan zijn er immers veel onderling sterk verschillende *W*'s, ofwel er zijn nogal wat steekproeven met véél te kleine *W* ten opzichte van het populatiegemiddelde ì en ook heel wat met een veel te grote *W*.

Nu is de spreiding van de steekproefgemiddelden afhankelijk van twee grootheden: (a) de steekproefomvang en (b) de spreiding van de scores zelf in de populatie. Grote steekproeven leiden met grote kans tot minder misleidende resultaten omdat toevallig voorkomen van elementen met erg kleine score veelal weer wordt gecompenseerd door elementen met grote score. En in populaties waar een kleine spreiding is, mag men nauwkeuriger resultaten verwachten – alles lijkt daar immers nogal op elkaar – dan in populaties met grote spreiding. Als we iets weten over de steekproefomvang (en die is bekend) en over de spreiding in de populatie (die is niet bekend, maar op grond van de spreiding in de steekproef hebben we daar weer wel een indruk van), dan is het mogelijk na te gaan hoe groot binnen bepaalde zekerheidsmarges maximaal het verschil is tussen steekproef- en populatieresultaat. Als bijvoorbeeld de gemiddelde egoïsmescore van alle studenten 10 is en de spreiding is klein, dan is het erg onwaarschijnlijk dat een aselecte steekproef van 20 studenten ooit een gemiddelde egoesmescore van 30 oplevert. We kunnen dit ook zo uitdrukken: de hypothese dat in de populatie de gemiddelde egoïsmescore 10 zou zijn, is niet goed rijmbaar met een geobserveerde gemiddelde score van 30.

Statistisch toetsen

De gedachte dat een grote discrepantie tussen steekproef en populatie onwaarschijnlijk is, is geformaliseerd in de toetsingstheorie. Als we aannemen dat in de populatie een zekere stand van zaken geldt en we treffen een daarmee sterk contrasterend steekproefresultaat aan dan is het van tweeën één:

òf (1) je hebt toevalligerwijs een onwaarschijnlijke (maar niet onmogelijke) a-typische steekproef te pakken
òf (2) het is niet waar dat de steekproef uit de bedoelde populatie komt.

Als het eerste alternatief maar onwaarschijnlijk genoeg is, dan kiest men al gauw voor de tweede mogelijkheid. Het is de conventie in statistisch onderzoek om als grens een waarschijnlijkheid van 1 op de 20 keer te nemen of 1 op de 100 keer. Is de eerste mogelijkheid onwaarschijnlijker dan 1 op de 20 keer in de reeks van alle denkbare steekproeven (respectievelijk 1 op de 100 keer), dan spreekt men, conventioneel, af dat men dan liever geloof hecht aan het tweede alternatief. Men zegt dan dat men 'bij significantieniveau 1 op 20 (resp. 1 op 100)' tot de tweede conclusie besluit. Afgekort: 'de steekproef wijkt significant af van de veronderstelde (hypothetische) waarde die men voor de populatie had verondersteld, bij een significantieniveau van 1/20 (of: 0.05, of 5%), resp. 1/100 (.01, 1%).' Het is goed zich te realiseren dat als de populatiewaarde wèl geldt, men toch 1 op de 20 keer (1 op de 100 keer) tot deze, alsdan onjuiste, conclusie komt. Merk navenant op dat als een steekproefresultaat in de buurt van het gehypothetiseerde populatieresultaat wordt geobserveerd, dat niet zegt dat die hypothese over het populatieresultaat juist was: allerlei niet te veel van dat populatieresultaat verschillende resultaten zijn óók niet onwaarschijnlijk.

Inferentiële statistiek is die tak van de wiskundige statistiek waarin redeneringen als de bovenstaande op wat rigoureuzere manier worden gekwantificeerd en voor veel statistische problemen is een toets ontwikkeld die aansluit bij veel voorkomende vragen. Het in de populatie veronderstelde resultaat wordt meestal aangeduid met de nulhypothese en bekeken wordt of die nulhypothese rijmbaar is met de geobserveerde waarden of, met een bepaald significantieniveau, moet worden verworpen. Verwerpt de toets de nulhypothese, dan stelt men het omgekeerde ervan, de alternatieve hypothese, te hebben aangetoond. Statistische leerboeken staan vol met toetsen, en statistische programma's berekenen de toetsuitslagen. Wij zullen hier niet op de berekening van zulke toetsen ingaan. Van belang is dat men de toetsuitslag kan interpreteren.

Een statistische toets is in het algemeen een formule die op grond van de steekproefwaarnemingen een waarde (de toetsingsgrootheid) berekent en een rekenvoorschrift waarmee zich laat bepalen hoe groot de kans is dat mogelijkheid (1) zich voordoet. Soms is dat rekenvoorschrift een expliciete formule, soms een voorschrift om de toetsingsgrootheid in een tabel op te zoeken. Tegenwoordig is dit rekenvoorschrift bijna altijd in een computerprogramma ingebouwd en wij hoeven ons niet te bekommeren om de finesses, als we de toetsuitslag (wel of niet verwerpen van de nulhypothese) maar kunnen begrijpen.

Helaas zijn er nogal wat termen in gebruik om uitslagen weer te geven. De toetsuitslag wordt soms eenvoudig weergegeven met de mededeling: de toets verwerpt de nulhypothese en vaak staat daar dan bij: bij significantieniveau α (met α zoals boven aangegeven conventioneel veelal op 1% of 5% gesteld). Soms wordt in plaats van de term significantieniveau ook de term 'onbetrouwbaarheid' of 'onbetrouwbaarheidsdrempel' gebruikt. Vaak ook geeft men de overschrijdingskans, (soms '*p*-waarde' genoemd). Dat is eigenlijk de kans dat mogelijkheid (1) in bovenstaand dilemma zich voordoet, dus de kans (als de nulhypothese in werkelijkheid wel waar is) op een steekproefuitslag die net zo onwaarschijnlijk is als de feitelijk gevondene. Als die overschrijdingskans kleiner is dan de gekozen onbetrouwbaarheidsdrempel, wordt de nulhypothese verworpen. Soms zegt men: de toetsuitslag is significant (wederom met vermelding van de onbetrouwbaarheidsdrempel α).

Is significant ook interessant?

Stel eens dat we na willen gaan of mannen en vrouwen even egoïstisch zijn. De nulhypothese is dat er geen verschil is en we doen een studie waarbij we een steekproef van driehonderd mannen vergelijken met een steekproef van driehonderd vrouwen. De passende toets verwerpt de nulhypothese. Is het nu tijd om opgewonden een artikel te schrijven over het geslachtsverschil en egoïsme? Dat is niet zonder meer het geval. Laten we even nagaan wat een significante toetsuitslag betekent: het betekent waarschijnlijk dat mannen en vrouwen ook écht verschillen in hun mate van egoïsme. Dat verschil kan zeer subtiel zijn. Kortom: significantie zegt iets over de zekerheid waarmee we aantonen dat er verschil bestaat, maar niets over de grootte en dus over de relevantie van het verschil. Of een verschil relevant is, hangt af van de onderzoekscontext. In sommig onderzoek zijn minieme verschillen al relevant, in andere zijn zelfs flinke verschillen nauwelijks om van op te kijken. Het is in wezen geen statistische vraag, maar een

inhoudelijke. Dat impliceert dat we de controle of aangetoonde verschillen ook relevant zijn los van de statistiek moeten uitvoeren.

Het is dus goed steeds na te gaan of een aangetroffen verschil èn significant (statistische vraag), èn relevant (inhoudelijke vraag) is (zie tabel 3).

Tabel 3: Classificatie van verschillen.

	Significant	Niet significant
Relevant	verschil is zeker, en interessant van omvang	verschil is niet zeker
Niet relevant	verschil is zeker, maar te klein om interessant te zijn	

Alhoewel relevantie uiteindelijk een inhoudelijke vraag is, is het toch vaak wel mogelijk om een statistische indicatie van de relevantie van verschillen te geven. Vaak is het van belang de gevonden verschillen – bij intervalvariabelen – te vergelijken met de standaarddeviatie van de scores in kwestie. Immers, omdat we weten dat 95 procent van de scores binnen de grenzen van ± 2 * de standaarddeviatie ligt, geeft een verschil dat groot is ten opzichte van de standaarddeviatie aan dat de verdelingen flink verschoven zijn. Cohen heeft een aantal maatstaven voorgesteld dat als vuistregel voor relevantie kan worden aangemerkt.[16] Hier noemen we alleen Cohens maatstaf voor wanneer we kijken naar het verschil in gemiddelden bij twee populaties, zoals hierboven bijvoorbeeld voor de egoïsmescores van Amerikaanse en Rotterdamse studenten werd gedaan. Cohen noemt een verschil 'groot' als het groter is dan 0.8 keer de standaarddeviatie van de scores, 'redelijk' als het groter dan 0.5 keer de standaarddeviatie is, en 'klein' als het minstens 0.2 maal de standaarddeviatie is. Onder die waarde vindt Cohen het verschil niet de moeite waard. Tabel 2 geeft maatstaven voor de relevantie van correlatiecoëfficiënten. Voor andere toepassingen verwijzen we naar Cohens boek.

Aanbevolen literatuur

Een elementaire introductie in onderzoeksmethodiek geven Baarda & Goede (1990). Wat diepgaander is Segers (1977). J. Hagan, Gills & Brownfield (1996) is een fraai leerboek voor empirisch onderzoek in een criminologische context. Voor onderzoeksopzetten kunnen Kuypers (1989) en Verschuren (1988) worden genoemd.

Voor data-analyse en statistische toetsen noemen we Moore & McCabe (1994), van Peet, van den Wittenboer & Hox (1995) en van Peet, van den Wittenboer & Hox (1997)..

16. J. Cohen (1977a).