Ik word ziek van de statistiek, of: er van weten zonder er naar te handelen*

W. Molenaar

1. Inleiding

De opmars van de statistiek in de sociale wetenschappen bezorgt een statisticus, behalve dagelijks brood, ook wel eens koude rillingen. Mede omdat de voorlichting vaak meer de rekentechniek dan het juiste gebruik accentueert, zijn er nogal wat uiterst onoordeelkundige toepassingen te signaleren.

Wie daarop zijn kritiek afvuurt, in de hoop dat dit een zinvolle toepassing bevordert, loopt het gevaar, argumenten te leveren aan de 'backlash' van moedelozen en wereldvreemden, die het statistisch hulpmiddel als bedreigend ervaren of het voor maatschappijkritisch onderzoek overbodig achten: statistiek leidt tot fouten, dus weg met de statistiek!

Elk technisch hulpmiddel dat op grote schaal beschikbaar wordt gesteld, zal in de eerste fase te pas, maar ook nogal eens te onpas worden gebruikt. Men is geneigd om het nieuwe speelgoed overal op los te laten, en te menen dat het al onze problemen kan oplossen. Zo'n stadium is te onderscheiden bij de auto, bij de computer, en ik verwacht het binnenkort bij de systeemtheorie en video-apparatuur.

Door een fout van de opdrachtgever of de programmeur wordt de rijksstudietoelage te laat uitgekeerd of klopt het bedrag niet (men spreekt dan van een 'fout van de computer'). Doordat de auto voor iedere verplaatsing wordt gebruikt, verstoppen wij tijdens het spitsuur onze steden en veranderen de weilanden er om heen in zeeën van asfalt en beton. Volgt daaruit dat we de computer en de auto moeten afschaffen? Nee, we moeten leren om ze spaarzaam en oordeelkundig te gebruiken. We schaffen ook geen

* Een eerdere versie verscheen als Heymans-Bulletin 74-HB-165-EX.
De auteur is verbonden aan de Rijks Universiteit Groningen.

bijlen af, omdat die wel eens gebruikt zijn om iemand het hoofd in te slaan: door een preek of door opsluiting in de gevangenis proberen we de gebruiker duidelijk te maken dat hij een misbruiker was.

Ik houd niet van opsluiten, maar wel van tegen u preken. Mijn preek zal een aantal statistiek-misverstanden en -misbruiken op een rijtje zetten. Ongetwijfeld heeft u er wel eens van gehoord, maar misschien weet u toch niet steeds dienovereenkomstig te handelen. In traditionele statistiekboekjes wordt er ook weinig over gezegd — een loffelijke uitzondering wordt gevormd door Van den Ende en Verhoef (1973) — en nu de computerpakketten de statistische berekeningen binnen ieders bereik hebben gebracht, is er des te meer reden om een aantal bekende waarschuwingen tegen misbruik te herhalen. Als u wel een voorbeeldig gebruiker van de statistische methoden bent, kunt u zich aan de komende pagina's zacht spiegelen.

Vooraf past een waarschuwing. Waarschijnlijkheidsrekening en mathematische statistiek zijn disciplines waarin op volstrekt logische wijze stellingen en axioma's worden bewezen. Maar in het grensgebied van statistiek en methodologie maakt wiskundige zekerheid plaats voor doordachte, maar onvermijdelijke subjectieve overtuiging. Ik heb wel eens het idee dat menig collega-statisticus dat grensgebied om die reden mijdt en veilig in zijn wiskunde-rijk blijft.

De onjuistheid van bepaalde opvattingen over statistiek kan ik wel beargumenteren, maar niet strikt bewijzen, en over sommige hier aangesneden problemen bestaat tussen de experts een diepgaand verschil van inzicht. Dat is lastig, omdat u van een hulpwetenschap duidelijke aanwijzingen verwacht en geen methodologische twistgesprekken; ik selecteer dus, zo nodig. Het beknopte overzicht pretendeert niet volledig te zijn, en evenmin origineel.

2. *De twee overspanningen*

> *'That is to say, he must have some understanding of both the theoretical background of statistics and the substance of his science even if he requires little or no knowledge of the mathematics of statistics'.*
> (Novick en Jackson, 1974, blz. 13)

De beschrijvende statistiek poogt een gegeven verzameling van data zo compact en duidelijk mogelijk te beschrijven. De inductieve statistiek (inferential statistics) mikt echter op conclusies die een wijdere geldigheid

hebben en probeert vanuit een steekproef (waarin waarnemingen worden gedaan of waarmee een experiment wordt uitgevoerd) te generaliseren naar een populatie. Het heeft zin, daarbij onderscheid te maken tussen de populatie waaruit de steekproef in feite is getrokken (sampling population) en de populatie waarover men uitspraken zou willen doen (target population). Wanneer beide populaties identiek zijn, kan de inductieve statistiek de mate van betrouwbaarheid berekenen van uitspraken over de populatie, gebaseerd op de steekproef, althans wanneer die steekproef volgens zekere eisen is getrokken. Maar meestal vallen de genoemde populaties niet samen en wordt een wijdere generalisatie beoogd, zoals blijkt uit het volgende citaat van Cornfield en Tukey (1956).

The two spans of the bridge of inference. In almost any practical situation where analytical statistics is applied, the inference from the observations to the real conclusion has two parts, only the first of which is statistical. A genetic experiment on Drosophila will usually involve flies of a certain species. The statistically based conclusions cannot extend beyond this race, yet the geneticist will usually, and often wisely, extend the conclusion to (a) the whole species, (b) all Drosophila, or (c) a larger group of insects. This wider extension may be implicit or explicit, but it is almost always present. If we take the simile of the bridge crossing a river by way of an island, there is a statistical span from the near bank to the island, and a subject-matter span from the island to the far bank. Both are important. By modifying the observation program and the corresponding analysis of the data, the island may be moved nearer to or farther from the distant bank, and the statistical span may be made stronger or weaker. In doing this it is easy to forget the second span, which usually can only be strengthened by improving the science or art on which it depends. Yet a balanced understanding of, and choice among, the statistical possibilities requires constant attention to the second span. It may often be worth while to move the island nearer to the distant bank, at the cost of weakening the statistical span — particularly when the subject-matter is weak.

Conclusie: het is aan de auteur, eventueel aan de lezer, om op vak-inhoudelijke gronden aannemelijk te maken dat zijn conclusies wijdere geldigheid hebben dan voor de populatie waaruit de steekproef werd getrokken.

Generalisaties naar de tijd, naar de plaats en naar de groep van proefpersonen zijn in doorsnee in de sociale wetenschappen moeilijker dan bijvoorbeeld in de medische wetenschap, waar de chirurg van het R.K. Ziekenhuis in Heerlen in 1977 wel direct toepasbaar zal achten wat zijn collega in het Algemeen Ziekenhuis in Winschoten in 1974 omtrent de operaties van enkelfracturen had gepubliceerd.

3. De aard van de steekproef

Meer dan negentig procent van de publikaties in statistische vakbladen

gaat uit van enkelvoudig aselecte steekproeven met teruglegging, maar meer dan negentig procent van de steekproeven waarop de resultaten van die publikaties worden toegepast, heeft die eigenschappen niet. Sommige afwijkingen zijn niet storend: in een populatie van honderdduizend mensen maakt het weinig uit, of een steekproef van honderd mèt dan wel zònder teruglegging was getrokken, en wanneer twee van de honderdduizend niet geregistreerd waren en dus niet getrokken konden worden, zal dat ook niet veel uitmaken. Vooral bij survey-onderzoek werkt men echter vaak met systematische steekproeven, cluster-steekproeven of tweetrapssteekproeven. Schattingen voor gemiddelden en fracties kan men dan vaak met behulp van speciale formules nog wel maken, dikwijls zelfs nauwkeuriger dan bij aselect trekken. Correlatie, chi-kwadraat, regressie of verdelingsvrij toetsen wordt een hachelijke zaak. Evenals bij non-respons kan men de onderzoeker de last toeschuiven om aannemelijk te maken dat de feitelijk verkregen steekproef kan worden beschouwd als een enkelvoudig aselecte, maar dat is weinig bevredigend.

Bij de keuze tussen een grote cluster-steekproef en een (even dure) kleine, strikt aselecte, zou het nuttig zijn als de statistici meer aanwijzingen konden geven over de mate van verstoring die door het cluster-karakter wordt veroorzaakt. Hoewel er op dit terrein enige vordering wordt geboekt, is het vrijwel onmogelijk om algemene regels te geven — een euvel dat robuustheidsonderzoek nu eenmaal vaak plaagt. Een zorgvuldige methodologische verantwoording is bijvoorbeeld nagestreefd door Gadourek (1963) en Knol (1976).

4. Beslissing of gevolgtrekking

Beslissen in onzekere situaties is het onderwerp van de statistische beslissingstheorie, waar men, uitgaande van verliesfuncties, gedefinieerd voor elke combinatie van ware toestand en beslissing, een optimale strategie ontwikkeld heeft om op grond van onvolledige informatie een keuze te maken. Gestelde vragen zijn bijvoorbeeld: Is machine A nauwkeuriger dan machine B? Heeft de patiënt ziekte C of ziekte D? Zal de geteste persoon de studie volbrengen of niet? Geven de waarnemingen meer steun aan theorie E of aan theorie F?

Gemeenschappelijk aan deze vraagstellingen is een soort principiële gelijkwaardigheid van beide alternatieven, al laat de theorie toe dat het ene op voorhand waarschijnlijker wordt geacht (Bayesiaans standpunt of dat de verliezen bij beide foute beslissingen ongelijk zijn. Nodig is, dat de

alternatieven duidelijk van elkaar gescheiden zijn in hun consequentie voor de waargenomen grootheid. De vaststelling van de verliesfuncties en eventueel de a priori verdelingen is bepaald niet eenvoudig. Maar wie deze methoden als 'te subjectief' terzijde schuift, dient te bedenken dat in de keuze van model en proefopzet bij schijnbaar meer objectieve methoden impliciet oordelen op dit terrein verborgen kan zijn.

Heel anders is het gesteld met het 'statistisch bewijs' van een onderzoekshypothese. Hier wordt het tegendeel van die onderzoekshypothese als nulhypothese geformuleerd en vervolgens wil men aantonen dat de waarnemingen zeer weinig compatibel zijn met die nulhypothese. Lukt dit, dan wordt de nulhypothese verworpen en is het 'bewijs' geleverd (uiteraard behoudens de beroemde kans alfa op ten onrechte verwerpen). Lukt het niet, d.w.z. zijn de waarnemingen niet flagrant in strijd met de nulhypothese, dan kan deze niet verworpen worden, maar daarom is zij nog niet bewezen. Ze is alleen maar op de proef gesteld en heeft die proef doorstaan. Evenals bij Popper kan men dus wel falsifiëren, maar niet bewijzen, en dat is precies de reden dat men het tegendeel van de onderzoekshypothese als nulhypothese kiest. De onderzoeker zegt als het ware tegen een lid van het denkbeeldig wetenschappelijk forum: 'Wilt u beweren dat mijn onderzoekshypothese fout is? Laten we eens even aannemen dat u gelijk hebt' (nulhypothese). 'Dan zijn de door mij verkregen waarnemingen toch wel erg onwaarschijnlijk' (overschrijdingskans). Wanneer er niet 'onwaarschijnlijk', maar 'onmogelijk' stond, zouden we het een bewijs uit het ongerijmde noemen.

Een zinnige theorie zal meestal de richting van verband of verschil kunnen specificeren, zodat er eenzijdig getoetst kan worden. Van meervoudige nulhypothesen hebben we daarbij geen last, als we het 'moeilijkste' enkelvoudige deel toetsen en de rest a fortiori verwerpen. Voorbeeld: als delinquenten in neuroticisme verschillen van anderen, zullen ze wel neurotischer zijn. Nulhypothese: delinquenten even neurotisch of minder, zijn de waargenomen scores niet compatibel met 'even neurotisch', omdat het steekproefgemiddelde bij delinquenten duidelijk hoger is, dan zijn ze nog veel minder compatibel met 'minder neurotisch', zodat de gehele nulhypothese kan worden verworpen.

Er zijn onderzoekers die het neuroticisme van delinquenten met een tweezijdige toets onderzoeken, ook al verwachten zij op grond van een theorie of op grond van hun ervaring dat delinquenten neurotischer zijn. Zij zeggen desgevraagd dat zij beducht zijn voor een mogelijk verschil in de andere richting, dat misschien ook wel ooit gevonden is of zou kunnen worden. Wie om die reden tweezijdig toetst, neemt volgens mij zijn

eigen inzicht niet erg serieus. Om in het onverwachte geval nog iets te kunnen aantonen, offert hij namelijk een stuk onderscheidingsvermogen op, en verkleint dus zijn kansen om een afwijking in de verwachte richting te kunnen aantonen. Hoeveel die kansen dalen, hangt uiteraard af van het verschil tussen de populaties: als dat zeer groot is, leidt zowel de eenzijdige als de tweezijdige toets tot verwerping; als het miniem is, vermoedelijk geen van beide. Maar wie tevoren zijn steekproefomvang zo kiest dat hij bij een matig groot verschil een redelijk onderscheidingsvermogen verwacht, moet niet verbaasd zijn als hij bij de tweezijdige toets bijvoorbeeld twintig of dertig procent meer waarnemingen nodig heeft om aan die eis te voldoen.

Er is nog een bijkomend probleem, van meer formele aard. Wie bij tweezijdig toetsen van de nulhypothese 'correlatie nul' deze verwerpt omdat de steekproefcorrelatie duidelijk positief is, mag strikt genomen alleen constateren dat de waarnemingen niet compatibel zijn met de nulhypothese. Het is gebruikelijk, daaraan de conclusie te verbinden dat de populatiecorrelatie positief is. Dit houdt in dat men, als de echte populatiewaarde zwak negatief is, niet alleen een risico loopt om de nulhypothese ten onrechte niet te verwerpen (fout van de tweede soort), maar ook nog een risico van een falikant verkeerde conclusie: zou er, ondanks de zwak negatieve populatiewaarde, een vrij duidelijke positieve steekproefwaarde zijn ontstaan, dan verklaren we het statistisch bewezen dat er een positief verband is, geheel ten onrechte. In de meeste toepassingen is de kans op zo'n falikant verkeerde conclusie gelukkig verwaarloosbaar klein voor alle alternatieven die op enige afstand van de waarde volgens de nulhypothese liggen.

De statistische beslissingstheorie laat een uitwerking van het geschetste dilemma toe. Er zijn drie natuurlijke toestanden: correlatie positief, correlatie nul en correlatie negatief in de populatie. Er zijn ook drie beslissingen:

– 'significant positief', als steekproefwaarde $r \geqslant r_1$;
– 'significant negatief', als steekproefwaarde $r \leqslant -r_2$;
– 'nulhypothese niet verworpen', als $-r_2 < r < r_1$.

Bij iedere combinatie van natuurlijke toestand en beslissing kan men een verlieswaarde stellen, die bijvoorbeeld nul is als de juiste beslissing wordt genomen en heel groot als de falikant verkeerde uitspraak wordt gedaan. Het zou te ver voeren, hier de Bayesiaanse of minimax-strategie voor dit probleem af te leiden, maar wel willen wij opmerken dat men door een asymmetrische keuze van de verlieswaarden kan bereiken dat de ontdekkingskans van positieve (volgens de theorie verwachte) afwijkingen wordt

vergroot zonder dat afwijkingen in de andere richting geheel worden ver
waarloosd: in het correlatievoorbeeld zou men dan r_1 kleiner dan
kiezen.

Soms zegt de onderzoekshypothese dat er geen verschil of geen verband
bestaat. Het tegendeel als nulhypothese kiezen, brengt ons in moeilijk
heden, omdat elke waarnemingsreeks compatibel is met 'wel verschil' (da
verschil zou nl. extreem klein kunnen zijn). Maar wie de onderzoeks
hypothese zelf als nulhypothese neemt, wordt gedwongen om haar òf te
verwerpen òf genoegen te nemen met de conclusie dat de waarnemingen
haar niet flagrant tegenspreken. Volgens mij is het dilemma mede door
slordig taalgebruik veroorzaakt. In een populatie zal er maar zelden
absoluut geen verschil of verband zijn, maar dat is niet erg: we bedoelen
immers 'geen verschil of verband van belang'. De kunst is nu, de onder-
zoekshypothese door operationalisering van dat 'belang' te formaliseren,
bijv. als 'hoogstens drie eenheden verschil' of 'hoogstens een correlatie
van 0,08', waarna we nu wel een statistisch bewijs kunnen leveren (al
vereist dat vrij veel waarnemingen). Zijn er twee theorieën, waarvan de
ene wel verschil voorspelt en de andere niet, dan kunnen we na dezelfde
herformulering ook de statistische beslissingstheorie gebruiken.

5. Hoe significant relevant is, en hoe irrelevant significant

Er bestaat een mythe dat significante resultaten belangrijk, waar en publi-
cabel zijn, terwijl niet-getoetste of niet-significante resultaten als mislukt,
onjuist en rijp voor de prullenmand worden beschouwd. Van alle misvat-
tingen over statistische toetsing is dit misschien wel de rampzaligste. Veel
onderzoeksituaties lenen zich in het geheel niet voor inductieve statistiek,
bijv. omdat er geen duidelijke populatie is of omdat de te onderzoeken
begrippen zich slecht laten operationaliseren. Maar ook wanneer er wel
reden tot toetsen is, zou het uitermate jammer zijn, de niet-significante
uitslagen niet te publiceren. Niet alleen vormen ze als 'schip op het strand'
een baken voor verder onderzoek, maar ook leidt deze publikatiepolitiek
tot een veel grotere fractie onjuiste gevolgtrekkingen dan de 'officieel' op-
gegeven alfa. Immers, wie voortdurend onjuiste onderzoekshypothesen
formuleert, bijv. omtrent te verwachten verschillen tussen twee groepen
die in feite niet verschillen, mag verwachten, na gemiddeld negentien keer
een niet-significant onderzoek naar de oud-papierhandel te hebben ge-
bracht, de twintigste keer een significant resultaat aan een tijdschrift te
kunnen aanbieden.

Even ellendig is de 'sleepnet-procedure': wie tevoren geen hypothesen stelt, maar achteraf gaat zoeken naar bijv. paren van variabelen die op vijf-procent-niveau significant correleren, zal in een 20-bij-20-correlatiematrix allicht iets van zijn gading vinden. Als aanwijzing voor verder onderzoek (statistische detectie, Hemelrijk, 1958) is dat prachtig, maar totdat het onderzoek is gerepliceerd, is er helemaal niets statistisch bewezen. En met replicatie-onderzoek is het net als met het weer: iedereen praat er over, maar niemand doet er iets aan.

De jacht op significantie leidt tot rijke buit bij het gebruik van zeer omvangrijke steekproeven. Bij het ondervragen van 1600 volwassen Nederlanders vindt men dat alcoholgebruik significant ($P < 0{,}05$) correleert met druggebruik ($r = -0{,}059$), en concludeert: 'drinkers are definitely underrepresented among the actual drug users.' Bij 1236 produktiewerkers leidt een r van 0,14 tussen autonomie en intrinsieke werkmotivatie ($P < 0{,}01$ tot: 'Produktiewerkers met een relatief grote autonomie hebben een grotere gemiddelde intrinsieke werkmotivatie dan hun collega's met een betrekkelijk geringe vrijheid in het werk.' Beide conclusies zijn formeel juist: ik weet nagenoeg zeker dat de correlatie niet nul is. Maar ik weet even zeker dat zij maar erg weinig van nul verschilt: pogingen om de ene variabele uit de andere te voorspellen, zullen heel weinig succes hebben. De fractie 'verklaarde variantie' is r^2 en dat is minder dan vier promille resp. twee procent. Wel is het van belang dat een samenhang in de tegengestelde richting blijkbaar in strijd is met de waarnemingen. En wat te denken van de (fictieve) psycholoog die na het testen van twintigduizend zesdeklassers zou vinden dat het gemiddeld verbaal IQ van de jongens significant (alfa = 0,01) hoger is dan dat van de meisjes, en wel resp. 99,8 en 100,2? Ik vrees dat hij er de actualiteitenrubrieken mee zou halen, maar vergeleken met de spreiding in IQ binnen de groepen lijkt het me nog geen reden om de coëducatie weer af te schaffen of alle meisjes remedial teaching toe te dienen. De absurditeit van zo'n onderzoek wordt duidelijk, als we bedenken dat de gemiddelden in de populaties ongetwijfeld in de zoveelste decimaal wel enig verschil zullen vertonen: wie hun gelijkheid tweezijdig toetst, is van succes verzekerd als zijn geldschieter hem zeer grote steekproeven toestaat. Het zou verstandig zijn om tevoren te bedenken, welk verschil, of welke mate van correlatie, zo groot is dat het inhoudelijk belangrijk begint te worden, en de steekproefomvang zo te kiezen dat tegen het aldus geformuleerde alternatief een redelijk onderscheidingsvermogen wordt bereikt. Men raadplege bijvoorbeeld Elstrodt en Mellenbergh (1976).

De lezer moet zich maar eens afvragen, wat hij vindt van de socioloog A

die vindt dat bij acht van de tien door hem onderzochte echtparen de man de rekeningen betaalt, en de socioloog B die vindt dat dit bij negenenvijftig van de honderd door hem ondervraagde paren het geval is. Het voorbeeld komt terug in de volgende paragraaf.

Het is bekend dat de relevantie van statistisch vastgestelde verschillen of verbanden ook bedreigd wordt door de mogelijkheid dat zij door niet bedoelde externe factoren zijn veroorzaakt. Dit hangt samen met de netelige vraag van de causaliteit, waarop ik niet diep kan ingaan. Het volstrekt negeren van de mogelijkheid van externe invloeden is volgens mij even verkeerd als het net zo lang conditioneren of partialiseren tot de conclusie verdwenen is. Ter illustratie van mijn standpunt citeer ik een door mij opgesteld collegevraagstukje over een klassiek probleem.

Stel dat men in een survey-onderzoek bij een aselecte steekproef van stedelingen een significant grotere politieke belangstelling vindt dan bij een aselecte steekproef van plattelanders. Geef bij ieder van de volgende vier vraagstellingen aan, of dit onderzoeksresultaat zinvol gebruikt kan worden of niet.

a. Het gewestelijk verkiezingscomité van partij x vraagt zich af, of het budget besteed moet worden aan het organiseren van politieke debatavonden in de steden dan wel in de dorpen.
b. De planologische commissie van de provincie vraagt zich af, of het afremmen van de groei van de dorpen en het bevorderen van het wonen in de steden mede gemotiveerd kan worden doordat dit beleid de politieke participatie zou bevorderen.
c. De wenselijkheid van meer inspraak van de bevolking bij de vaststelling van uitbreidingsplannen wordt blijkens een opiniepeiling onder de bevolking van de gehele provincie door dertig procent van de respondenten onderstreept. Een lezer van het onderzoeksverslag vraagt zich af, of het van belang is dat de stedelingen nogal fors over-vertegenwoordigd waren bij de opiniepeiling.
d. Een politiek-socioloog poneert dat de concentratie van veel inwoners op een beperkt grondgebied de politieke belangstelling verhoogt.

6. Na schatten en toetsen

Veertig jaar geleden zou menig onderzoeker als uitkomsten alleen de steekproefwaarden van gemiddelden of correlatiecoëfficiënten vermeld hebben, zonder zich er veel om te bekommeren of het gevonden verschil, c.q. de gevonden correlatie, reproduceerbaar was dan wel een gevolg van toevalligheden bij de steekproeftrekking. Het was een grote stap vooruit, dat vijfentwintig jaar geleden dit gebruik op grote schaal vervangen bleek te zijn door de significantietoets.

Maar, zoals we in de vorige paragraaf gezien hebben, is de vervanging niet ten volle een verbetering. Op de vraag 'is er belangrijke positieve

correlatie in de populatie?' werd eerst zonder meer met de steekproefwaarde geantwoord; later werd de vraag omgebogen tot 'weet ik redelijk zeker dat de populatiewaarde positief is?' Kwaadwilligen zien hier hun visie bevestigd dat een wiskundige, geconfronteerd met een praktisch probleem van een onderzoeker, zijn cliënt nog al eens naar huis stuurt met een perfecte en elegante oplossing van een ander probleem dat er een beetje op lijkt. Het vervelende is, dat redelijke zekerheid in kleine steekproeven pas begint bij uitermate positieve steekproefcorrelaties, terwijl bij grote steekproeven ook onbelangrijke positieve correlaties tot deze zekerheid leiden.

Ik wil een tweetal mogelijke uitwegen signaleren voor ons dilemma van significantie en relevantie. Voor de eerste moet u bereid zijn om de interpretatie van een kans te verleggen van een frequentiequotiënt bij onbeperkt herhalen van het experiment naar een mate van vertrouwen in een bewering. De Bayesiaanse statistiek staat u dan toe, uw kennis na afloop van het steekproefexperiment in een a posteriori verdeling samen te vatten, waarvan de belangrijkste kenmerken gemakkelijk in voor leken toegankelijke vorm kunnen worden weergegeven. Deze verdeling laat schatting en toetsing toe. Omdat er een subjectief element in de statistische procedure wordt geïntroduceerd, lijkt de Bayesiaanse aanpak mij maar voor een beperkt aantal onderzoeksproblemen geschikt. Wel heb ik de neiging om maar een geringe geldigheid toe te kennen aan het gebruikelijke tegenargument, dat het vrijwel onmogelijk is de voorkennis eerlijk en zorgvuldig in een a priori verdeling samen te vatten (Molenaar, 1976).

Met behoud van de klassieke objectieve kansdefinitie kan men het dilemma oplossen door het rapporteren van betrouwbaarheidsintervallen. Het is merkwaardig dat die in de standaard-leerboeken en computerpakketten nog altijd met een klein hoekje genoegen moeten nemen, en in publikaties zelden worden aangetroffen. Ze geven schatting en toetsing tegelijk. De ligging van het interval stelt de lezer in staat om na te gaan of hij een populatiewaarde van deze orde van grootte relevant vindt. Door na te gaan of een in de nulhypothese opgenomen waarde wel of niet tot het interval behoort, stelt de lezer vast of die nulhypothese niet resp. wel behoort te worden verworpen. Kortom, hij krijgt tegelijkertijd informatie over relevantie èn over significantie.

De interpretatie van een betrouwbaarheidsinterval is voor beginners lastig, omdat het interval bekend, maar variabel, en de parameter onbekend, maar vast is: in een nieuwe steekproef uit dezelfde populatie liggen de zojuist berekende grenswaarden van het interval anders, terwijl de ons niet bekende populatiegrootheid uiteraard niet van waarde verandert.

Het is mijn ervaring dat dit aan een groep niet statistisch onderlegde cursisten vrij snel duidelijk gemaakt kan worden, door ieder lid van de groep zijn eigen interval te laten berekenen voor de door hem zelf getrokken steekproef, en de resultaten achteraf te vergelijken.

Ik licht het betrouwbaarheidsinterval toe aan de hand van enige al genoemde voorbeelden. Socioloog A uit de vorige paragraaf, niet tevreden met de schatting van 0,8 en ook niet met het nog nèt niet verwerpen van de hypothese dat de populatiefractie echtparen waarbij de man betaalt 0,5 of minder is (de eenzijdige overschrijdingskans van 8 uit 10 is 0,0546), kan nu rapporteren: schatting 0,8 en 95 procent betrouwbaarheidsondergrens 0,49, terwijl socioloog B, met schatting 0,59 en wel verworpen hypothese (overschrijdingskans 0,0285) nu schatting 0,59 en 95 procent betrouwbaarheidsondergrens $0{,}50^{+}$ meldt.

Als er geen duidelijke richting bedoeld is bij het onderzoek naar drink- en drug-gewoonten zou men daar een 95-procent-betrouwbaarheidsinterval van $-0{,}10$ tot $-0{,}01$ kunnen rapporteren voor de populatiecorrelatie. De toetsen op 5-procent-niveau zijn telkens impliciet meegegeven: er wordt alleen verworpen als de ondergrens hoger dan 0,5 is resp. de waarde 0 niet in het interval ligt.

7. Van het veld naar het lab

'In der Beschränkung zeigt sich erst der Meister'.
(Goethe, *Faust*)

Vooral in de sociologie, maar ook wel in de andere gedragswetenschappen, lijkt men vaak nog te luisteren naar het gebod van keizer Augustus dat heel de wereld beschreven dient te worden. Aan zeer veel respondenten worden zeer veel vragen gesteld, hetgeen buitengewoon veel antwoorden oplevert, die in een datamatrix worden opgenomen: per persoon een rij, per variabele een kolom. De datamatrix is nooit helemaal volledig en betrouwbaar, door weigeringen, 'niet van toepassing' en codeer- en ponsfouten. Meetniveau en aantal mogelijke waarden zijn per variabele verschillend. De gebruikelijke weg, factoranalyse van een matrix van alle produkt-momentcorrelaties, gevolgd door vindingrijk leuteren over de gevonden factoren, is dan ook met een laagje ijs bedekt, maar wie zonder stoppen of wenden rechtuit blijft gaan, merkt dat misschien niet op.

Dat vragenlijstgedrag beïnvloed is door taalkundige vaardigheid, sociale wenselijkheid en suggesties van de interviewer, is genoegzaam bekend. Bij gebrek aan beter wordt de vragenlijst toch voor veel massaal onderzoek

gebruikt: het meetinstrument is relatief goedkoop en weinig belastend, en de reproduceerbaarheid van de meting is vrij hoog. De validiteit zal, vooral bij emotioneel geladen onderwerpen, wel eens meer zorgen baren.

Zorgvuldiger bestudering van de proefpersonen, bijvoorbeeld door participerende observatie of een open gesprek, zou in een verkennende fase wel erg nuttig zijn. Het is tijdrovend, de meetomstandigheden gaan erg uiteenlopen en de resultaten zijn lastig te kwantificeren. Dat betekent niet dat het niet een waardevolle bron van kennis kan zijn. Maar wanneer statistische verwerking van de onderzoeksresultaten gewenst is, zal men naar beste weten moeten proberen, aan strijdige eisen te voldoen door een goedkoop, valide en betrouwbaar instrument te hanteren. Als buitenstaander heb ik de indruk dat veel te vaak een samenraapsel van hier en daar weggepikte en zelf verzonnen items voor zo'n instrument wordt aangezien. Door zowel vooronderzoek als gebruik van bestaande schalen af te wijzen, zadelt men zichzelf op met een meting die zo onbetrouwbaar is, dat zinvolle correlaties met andere gegevens bij voorbaat haast uitgesloten zijn.

Dat wij mensen meten en geen moleculen, heeft zijn charme, maar ook zijn bezwaren: niet alleen praat het gemetene terug, het heeft ook bezwaar tegen pijnlijke vragen en langdurige experimentatie. Over de ethiek van die experimentatie en over de communicatie tussen meter en gemetene is al veel geschreven; ik noem als voorbeelden Friedrichs (1972) en Hofstee (1974). Systematische bestudering van menselijk gedrag vraagt niet alleen om betrouwbare meetinstrumenten, minstens even nodig is het op één of andere manier onder controle houden van de talrijke storende invloeden die de voorspelbaarheid van dit gedrag verminderen. Naar mijn mening kunnen de fysicus en de meteoroloog alleen het gedrag van complexe systemen voorspellen, doordat zij eerst allerlei afzonderlijke invloeden hebben gemeten en bestudeerd. Het lijkt wel eens of de gedragswetenschappen die stap willen overslaan.

Partiële correlatie, uitsplitsing, covariantie-analyse, log-lineaire modellen, matchen en pad-analyse zijn bekende statistische hulpmiddelen voor de beoogde controle op allerlei verstorende invloeden. Ze worden gelukkig op vrij grote schaal toegepast, maar veronderstellen dat de storende variabele bekend is. Ik heb de indruk dat een andere mogelijkheid, nl. bewuste inperking van het probleem, van de proefgroep of van de meetomstandigheden, niet erg populair is. Uiteraard verlengt dat de tweede overspanning van Cornfield en Tukey, maar als de inperking zorgvuldig gekozen wordt, brengt de eerste overspanning ons misschien op een eiland waar we al heel wat inzicht over het onderzochte probleem kunnen opdoen. Voorbeelden van zo'n zinvolle beperking meen ik o.a. te hebben

aangetroffen bij de bestudering van macht, zowel in de experimentele sociale psychologie als in de sociologie.

Eigen vooronderzoek, literatuurstudie en secundaire analyse van de tegenwoordig vrij goed toegankelijke gegevens van eerder onderzoek zijn daarbij van onschatbare waarde. Exploratie door grafische technieken, misschien non-metrische schaalmethoden, en vooral zorgvuldig kijken zonder enige statistische bril, is in die fase geboden. Wij moeten ons blijven hoeden voor uiterst subtiele gegevensverwerking na uiterst grove gegevensverwerking.

8. Epiloog: de hand in eigen boezem, of de balk in eigen oog

Als de kerkgangers c.q. de kinderen zich ernstig misdragen, is een lange preek bepaald geen wondermiddel. De voorganger c.q. de ouder dient zich af te vragen, of zijn eigen optreden wel zo onberispelijk is geweest. Als beste stuurman aan de wal heb ik misschien de noden van de schepelingen in de storm onderschat. Als (afgedwaalde) vertegenwoordiger van het wiskundig ras wil ik graag onderkennen dat mijn soortgenoten soms in hun ivoren torentje leuk spelen met oneindig grote, perfect aselecte steekproeven uit perfect normale verdelingen met perfect gelijke varianties. Als een soort bemiddelaar wil ik proberen, de partijen rond de conferentietafel te krijgen. Ik hoop dat u de losse notities die u nu gelezen hebt, kunt beschouwen als een vredesvoorstel en niet als een oorlogsverklaring.

Noot

Na verzending van de kopij las ik de recensie door W. A. T. Meuwese in *M & M* 51 (1976) blz. 320-322, van *De samenhang van groepen* door Joh. Hoogstraten en H. C. M. Vorst. Meuwese schrijft: 'Erg sterk is het allemaal weer niet: het overzicht van de resultaten van drie experimenten op blz. 323 vermeldt acht verschillen tussen cohesieve en controlegroepen, waarvan echter maar vier significant. Dat de overige vier in het voordeel van de hc-conditie zijn, heeft niet zoveel betekenis.' De recensent wil blijkbaar niet ingaan op de methoden voor het gelijktijdig toetsen van een aantal soortgelijke hypothesen. Maar hij had minstens kunnen overwegen dat alle acht waargenomen verschillen ten gunste van de konditie zijn, terwijl bij het ontbreken van een systematisch verschil tussen konditie en controle verwacht mag worden vier in het voordeel en vier in het nadeel te vinden. Het citaat lijkt mij een schril voorbeeld van het gedachtenloos gelijkstellen van significantie en relevantie.

Cornfield, J., en J. W. Tukey, *Annals of Mathematical Statistics*, 27, 1956, blz. 912.
Elstrodt, M., en G. J. Mellenbergh, *Een minus de vergeten fout*, rapport Subfaculteit Psychologie, Universiteit van Amsterdam, 1976.
Ende, H. van den, en M. Verhoef, *Inductieve statistiek voor gedragswetenschappen*, Agon Elsevier, Amsterdam, 1973.
Friedrichs, R. W., 'Dialectical Sociology: toward a resolution of the current "crisis" in Western sociology', *British Journal of Sociology*, 23, 1972, blz. 263-274.
Gadourek, I., *Riskante gewoonten*, Wolters, Groningen, 1963.
Hemelrijk, J., 'Statistische proefopzetten: bewijs en detectie', *Statistica Neerlandica*, 12, 1958, blz. 111-118.
Hofstee, W. K. B., *Psychologische uitspraken over personen*, Van Loghum Slaterus, Deventer, 1974.
Knol, H. R., 'Het toepassen van statistiek voor enkelvoudig aselecte steekproeven terwijl de steekproef niet enkelvoudig aselect is', *Mens en Maatschappij*, 51, 1976, blz. 179-200.
Lieberman, D. (red.), *Contemporary problems in statistics*, Oxford University Press, New York, 1971.
Molenaar, W., 'Bijnen en de aard van het Bayes-je', *Sociale Wetenschappen*, 19, december 1976.
Morrison, D. E., en R. E. Henkel (red.), *The significance test controversy*, Aldine, Chicago, 1970.
Novick, M. R., en P. H. Jackson, *Statistical methods for educational and psychological research*, McGraw-Hill, New York, 1974.