KM 24(1987)
pag 7 - 21

DE ONGRIJPBARE ZEKERHEID

rede uitgesproken bij de aanvaarding
van het ambt van gewoon hoogleraar in de
statistiek aan de Rijksuniversiteit Limburg
op donderdag 27 juni 1985 door
dr. W. Albers

Zeer gewaardeerde toehoorders,

In mijn voordracht wil ik trachten u een indruk te geven van een aantal problemen waar de statistiek voortdurend mee te kampen heeft. Hierbij wil ik met name aandacht schenken aan een scala van relatief moderne technieken om aan de harde kern van deze problemen het hoofd te bieden.

Wellicht is het verstandig om eerst een zekere afbakening van het te beschouwen gebied aan te geven. Een standaard opdeling van de statistiek is die in beschrijvende statistiek enerzijds en mathematische statistiek anderzijds. Andere veel gebruikte adjectieven in dit verband zijn "descriptief" voor het eerste deelgebied, en, in opklimmende mate van lelijkheid, "verklarend", "inductief" en "inferentieel" voor het tweede. Bij de beschrijvende statistiek gaat het om zaken als het reduceren, ordenen en overzichtelijk weergeven van het waarnemingsmateriaal. De mathematische statistiek houdt zich bezig met het trekken van conclusies uit waarnemingen. Dit laatste nu is hetgeen waar ik hier in geïnteresseerd zal zijn.

Een van de oudste problemen van de statistiek is haar relatie tot de wiskunde. Op het eerste gezicht lijkt het misschien overdreven om van een probleem te spreken: we hebben immers zojuist gezien dat een der deelgebieden van de statistiek wordt aangeduid als "mathematische statistiek", waar uit volgt dat dit gebied "dus" tot de wiskunde behoort. Nu is deze conclusie wel juist, althans tegenwoordig vrij algemeen geaccepteerd, maar dat is meer geluk dan wijsheid, want het simpele formalistische argument waar zij op steunt, deugt uiteraard allerminst. Niets is eenvoudiger en verleidelijker dan om door een geschikt gekozen etiket de gewenste conclusie plotsklaps binnen handbereik te brengen. Waar dit toe kan leiden leert het beroemde en beruchte geval van de Gauss-Laplaceverdeling die toen ze eenmaal het predikaat "normaal" verworven had, niet meer kapot te krijgen was. "Alles hoort normaal verdeeld te zijn, want als het dat niet is, is het abnormaal, en dat hoort natuurlijk niet". Slechts met de grootste moeite is het de statistiek gelukt uit deze, door het ondoordachte gebruik van het woord "normaal" ontstane, vicieuze cirkel te ontsnappen.

Kortom, er is meer voor nodig geweest om de wiskundige wereld ervan te overtuigen dat statistiek niet slechts wikkunde maar ook wiskunde is. De hoofdrol in dit acceptatieproces is gespeeld door de waarschijnlijkheidsrekening. Aanvankelijk werd dit vak evenmin voor vol aangezien. Het werd wel afgedaan als "wat met dobbelstenen gooien". Dat dit allemaal nog niet zo vreselijk lang geleden is, moge blijken uit het volgende. In een recentelijk verschenen bundel van in de loop der tijd door de bekende Amerikaanse wiskundige Halmos geschreven opstellen wordt ter inleiding van zijn artikel "The foundations of probability" uit 1944 opgemerkt dat "it was written to convince mathematicians that probability is a branch of mathematics. The need may seem strange today, but it wasn't then". Het artikel zelf begint overigens met een zin die aan kernachtigheid niets te wensen overlaat. "Probability is a branch of mathematics". (Halmos heeft een voorliefde voor zulke bondige en prikkelende uitspraken. Een ander voorbeeld is "Applied mathematics is bad mathematics". Maar dit terzijde.)

Toen eenmaal de waarschijnlijkheidsrekening door de wiskundigen in genade was aangenomen, was het pad voor de statistiek geëffend. Er hoefde nog slechts aangetoond te worden dat een strenge behandeling van de statistiek met behulp van de waarschijnlijkheidsrekening mogelijk was, om haar acceptatie als tak van de wiskunde tot een feit te maken.

Na het probleem van de relatie tussen statistiek en wiskunde is het nu de beurt aan de relatie tussen statistiek en gebruiker. Weer is er op het allereerste gezicht alle hoop dat we slechts met een ingebeeld probleem te maken hebben. Immers, het grote publiek heeft weinig op met wiskunde en, zoals we hierboven zagen, de wiskunde had traditioneel weinig op met de statistiek. Wat ligt er in zo'n situatie meer voor de hand dan een mooie sfeer vol wederzijds begrip tussen de statistiek en haar gebruikers? Helaas, dit fraaie ideaal wordt maar zelden bereikt. Van enige vooruitgang is misschien wel sprake. Zo was vroeger een bekende titel "How to lie with statistics", terwijl onlangs een boekje verscheen met als titel "How to tell the liars from the statisticians", hetgeen tenminste op de mogelijkheid attendeert dat statistici en leugenaars niet per definitie hetzelfde zijn.

Maar toch, het blijft een feit dat de relatie tussen statistiek en gebruiker vaak te wensen overlaat. Hoe komt dit nu? Helaas is het niet mogelijk om één enkele boosdoener in dit verband aan te wijzen. Het is veeleer zo dat een complex van factoren hierbij een rol speelt. In wat nu volgt zal ik een aantal van deze factoren de revue laten passeren. Ik zal beginnen aan de buitenkant en langzamerhand naar de kern toewerken, in de zin dat de te behandelen factoren gaandeweg steeds meer met de eigenlijke mathematische statistiek te maken zullen hebben.

Als eerste punt noem ik het volgende: zoals we in het voorafgaande hebben vastgesteld, gaat het in de mathematische statistiek om het trekken van conclusies uit waarnemingen. De nadruk ligt daarbij op het zorg er voor dragen dat het de juiste conclusies zijn die getrokken worden. Het is echter eveneens onontbeerlijk dat het de juiste waarnemingen zijn waaruit getrokken wordt. Aan op zich juiste conclusies uit de verkeerde waarnemingen hebben we evenveel als aan onjuiste conclusies uit de goede waarnemingen. Dus ook niets, inderdaad. Om deze open deur nog wat verder in te trappen het volgende, letterlijk en figuurlijk kinderachtige voorbeeld. Als u een zak hebt met 8 appels en een zak met 6 peren, dan leert de rekenkunde ons dat u over 14 stuks fruit beschikt. Blijken er echter bij nader inzien 7 appels en 5 peren in de zakken te zitten, dan klopt het berekende totaal van 14 uiteraard ook niet meer. Dat het niet aangaat om vanwege dit fiasco de rekenkunde van leugenachtigheid te betichten snapt in dit infantiele voorbeeld gelukkig vrijwel iedereen. De enige denkbare verdachtmaking die nog rest is waarschijnlijk: "dat komt er nu van als je appels en peren probeert op te tellen".

Waarom dan zo door gehamerd op deze ingetrapte open deur? Wel, omdat we in de statistiek helaas keer op keer met precies hetzelfde fenomeen geconfronteerd worden, dat dan echter niet als zodanig herkend wordt. De voorbeelden liggen werkelijk voor het oprapen, ik zal u er eens een paar geven.

Het eerste ontleen ik aan David Moore. Hij haalt het geval aan van een columniste in een Amerikaans weekblad die haar lezers vroeg: "Als u het allemaal nog

eens over kon doen, zou u dan weer kinderen nemen?" Ze ontving bijna 10000 reacties waarvan er maar liefst 70% negatief waren. Op grond van deze cijfers kan men met behulp van elementaire statistische methoden tot de conclusie komen dat met hoge betrouwbaarheid geldt dat het percentage van de Amerikaanse ouders dat niet opnieuw kinderen zou willen vlak bij de 70 ligt (bijvoorbeeld tussen de 69 en 71 met 95% betrouwbaarheid). Echter, niets is minder waar, of in ieder geval, minder waarschijnlijk. Een vervolgens uitgevoerde "nette" steekproef onder zo'n 1400 ouders onthulde namelijk dat van hen 91% opnieuw kinderen zou willen hebben, hetgeen aangeeft dat het gezochte percentage in de buurt van de 9, en dus bepaald niet bij de 70, zal liggen. Dit resultaat illustreert nog een ander punt. Vaak wordt geredeneerd als volgt: "Goed, het onderzoek voldoet niet helemaal aan de eisen, dat snap ik ook best. Maar dat is toch niet zo erg, want ik hoef het antwoord ook niet preciés te weten, het gaat meer om een idee, een globale indruk als het ware." Welnu, het gegeven voorbeeld laat duidelijk zien dat dit ijdele hoop is en dat iedere overeenkomst met bestaande situaties en werkelijke uitkomsten geheel op toeval kan komen te berusten: de enige verdienste die het gevonden percentage van 70 nog toe te schrijven valt is dat het tenminste de tact heeft om zich tussen de 0 en de 100 te bevinden. Overigens, ik heb nog niet gezegd waarom het waarnemingsmateriaal niet deugde, om u de gelegenheid te geven daar zelf even over na te denken. Zoals u inmiddels gezien zult hebben, is de fout dat het hier vrijwillige antwoorden betreft. Met name ouders die problemen met hun kinderen hadden, zullen in de pen geklommen zijn, waardoor een in het geheel niet representatief beeld ontstond.

Zoals ik eerder reeds opmerkte, zijn er voorbeelden van het genoemde type te over en niets zou eenvoudiger zijn dan om mijn resterende tijd daarmee te vullen. Ik wil echter graag nog ettelijke andere punten aan de orde stellen en ik zal daarom verder minder in details treden en volstaan met het aangeven van een paar veel voorkomende categorieën. Als bijvoorbeeld bij een enquête of opinieonderzoek een niet te verwaarlozen deel van de te ondervragen personen niet bereikbaar is of weigert om mee te werken, kan ook dit de representativiteit van de verkregen resultaten op losse schroeven zetten. De groep uitvallers zal er immers vaak om allerlei redenen duidelijk anders uitzien dan de groep als geheel. Een andersoortig probleem doet zich voor als het onderzochte verschijnsel achteraf onlosmakelijk verweven blijkt te zijn met een of meerdere andere factoren. De enige conclusie die dan vaak nog getrokken kan worden is dat het experiment beter op had moeten worden gezet. Een standaard voorbeeld in deze categorie is de vraag of een patiënt nu opkikkert omdat het toegediende geneesmiddel zo goed helpt, of door de aandacht die de dokter aan hem besteedt en het feit op zich dat hij wat te slikken gekregen heeft. Een laatste type probleem dat ik wil noemen houdt verband met de vraag of er wel met de juiste maat gemeten is. Moeten het bijvoorbeeld geen percentages zijn in plaats van absolute aantallen die vergeleken worden? Moore stelt dit probleem even bondig als duidelijk aan de kaak: een tandarts zegt trots tegen zijn collega: "Mijn patiënten hebben maar

half zoveel gaatjes als de jouwe". Waarop de andere riposteert: "Ja, wat wil je, ze hebben ook maar half zoveel tanden en kiezen ."

We hebben nu vastgesteld dat als het begin al niet deugt, in de zin dat er met onjuist waarnemingsmateriaal gewerkt wordt, er voor de statistiek inderdaad weinig eer te behalen valt. Vervolgens wil ik laten zien dat het ook aan het andere einde nog fout kan gaan. In dat geval zijn dus uit de juiste waarnemingen met behulp van de statistiek weliswaar de juiste conclusies getrokken, maar deze worden dan vervolgens geheel verkeerd geïnterpreteerd. Ook dit werkt niet bevorderend voor de populariteit van de statistiek.

Berucht in dit opzicht is bijvoorbeeld het gebruik van de correlatiecoëfficiënt. Als statistisch aannemelijk is gemaakt dat twee variabelen een positieve correlatie vertonen, dan wordt daar vaak klakkeloos uit geconcludeerd dat de een "dus" de oorzaak van de ander is. Talloos zijn weer de voorbeelden die de roekeloosheid van deze conclusie overduidelijk maken. Het is bijvoorbeeld niet moeilijk aan te tonen dat er een duidelijke positieve correlatie is tussen het aantal brandweerlieden bij een brand en de omvang van de bij die brand ontstane schade. Niemand zal hier echter de conclusie aan willen verbinden dat de brandweer afgeschaft moet worden omdat statistisch zou zijn aangetoond dat de schade bij branden veroorzaakt wordt door brandweerlieden. Amusant is overigens dat sommigen van een dergelijk voorbeeld klaarblijkelijk zo schrikken dat zij vervolgens niet alleen het aanvankelijk gepostuleerde causale verband, maar ook de correlatie als "onzinnig" gaan betitelen. Dat nu is op zich weer onzinnig, want die correlatie is er wel degelijk. Ze is alleen niet te danken aan een causale relatie maar aan een gezamenlijke afhankelijkheid van de beschouwde variabelen van een onderliggende derde variabele. In bovenstaand geval is dat bijvoorbeeld de omvang van de brand.

Een andere situatie waarin op zich juiste statistische conclusies vaak verkeerd geïnterpreteerd worden, doet zich voor bij het toetsen van hypothesen. Men komt dan tot uitspraken als: "De uitkomst is statistisch significant op het 5%-niveau." Ook al is zo'n conclusie correct, ze is zo beknopt en cryptisch dat ze licht leidt tot een verkeerde voorstelling van zaken. Men vat haar namelijk vaak op als constatering van een aanmerkelijke afwijking van de nulhypothese. Dus als zo'n hypothese bijvoorbeeld stelt dat twee geneesmiddelen even effectief zijn of dat een bepaalde correlatiecoëfficiënt nul is, dan meent men te kunnen concluderen dat het ene geneesmiddel belangrijk beter is dan het andere of dat er een aanmerkelijke correlatie bestaat. Hier behoeft echter geen sprake van te zijn. Vastgesteld is slechts dat de nulhypothese hoogstwaarschijnlijk onwaar is. Hoe onwaar echter, valt nog geheel te bezien. Naarmate de gebruikte steekproeven groter zijn, is de statistische toets gevoeliger en zullen steeds kleinere afwijkingen al een gerede kans op verwerping van de nulhypothese opleveren. Dus bijvoorbeeld ook een feitelijke correlatiecoëfficiënt van 0.1 zal bij voldoende waarnemingen met grote kans leiden tot de boodschap dat "de uitkomst statistisch significant is op het 5%-niveau", terwijl aan een dergelijke correlatie toch moeilijk enig praktisch belang kan worden toegeschreven. Op de remedie voor dit soort perikelen zal ik hier verder niet ingaan, op één oplossing na, die ik overigens niet erg kan aanbevelen. Deze werd mij

gesuggereerd door een cliënt. Toen hij gewezen werd op de eerder genoemde stijgende gevoeligheid van toetsen bij toenemende aantallen waarnemingen, riep hij blijmoedig: "Nou, zeg maar hoeveel waarnemingen ik moet weggooien!"

In het voorafgaande hebben we gezien dat er van alles fout kan gaan, zowel vóór- als nadat de mathematische statistiek er aan te pas komt. Let wel, deze constatering is allerminst bedoeld om het vuil op andermans stoep te deponeren en vervolgens vergenoegd terug te keren naar de behaaglijke ivoren toren. Deze "voor-en-na-de-behandeling"-problemen verdienen zeker de volle aandacht van de statisticus. Zij vormen zelfs naar mijn idee een van de charmes van het vak, vandaar ook dat ik het niet heb kunnen laten er geruime tijd bij stil te blijven staan, in plaats van ze met een luchtige handbeweging weg te wuiven.

Niettemin wordt het nu de hoogste tijd om op onze speurtocht naar het grote ongenoegen de blik eens te richten op het middengedeelte. De stille hoop die we daarbij koesteren, is natuurlijk dat de reeds behandelde punten zó goed als bliksemafleider gefungeerd hebben, dat er geen klachten meer over zijn. Helaas, zoals u waarschijnlijk al aannam, zo makkelijk komen we er niet vanaf, er resteert nog minstens één klacht over de eigenlijke statistiek. Deze komt neer op het verwijt dat "je met statistiek nooit iets kunt bewijzen". Als u nu van mij een gloedvol betoog verwacht om deze aantijging te weerleggen, moet ik u teleurstellen. Ze valt namelijk niet te weerleggen, omdat ze juist is. Wat er niet deugt is niet de bewering zelf, maar haar verwijtend karakter. De duidelijkheid gebiedt mij om onomwonden te stellen dat uit zo'n karakter een puur onbegrip blijkt voor de aard van de statistiek. Immers, de statistiek houdt zich bezig met problemen die door het toeval beheerst dreigen te worden. Haar rol is het terugdringen en temmen van dit toeval en dat doet zij dan ook met overgave. Niet van haar verwacht mag worden dat zij het toeval,dat juist het kenmerkende aspect van de door haar beschouwde problemen vormt, als bij toverslag doet verdwijnen. Voor een dergelijke act moet men geen temmer, maar een goochelaar inhuren. Hiermee raken we overigens precies de kern van het probleem: diegenen die de statistiek verwijten dat zij niets bewijst, geloven meestal ook voor hun getallen een goochelaar nodig te hebben.

Laten we aan de hand van een simpel voorbeeld nog eens nagaan wat men wel en wat men niet van de statistiek mag verwachten. Met behulp van een experiment wil men nagaan of proefpersonen paranormaal begaafd zijn. De experimentator heeft een aantal, zeg 5, verschillende kaarten. Hij trekt hieruit willekeurig een kaart en laat de proefpersoon raden welke het is. Dit wordt een aantal keren, zeg 250, herhaald en er wordt vastgesteld hoe vaak de proefpersoon goed geraden heeft. Ligt dit aantal in de buurt van één vijfde deel van 250, dus 50, dan is er geen reden om aan te nemen dat er iets aan de hand is. Ligt het echter duidelijk hoger, dan is die reden er wel. Merk overigens op dat hier voorzichtig gesproken wordt over "aan de hand" en niet over "paranormaal begaafd". De statistiek kan nooit aantonen dat er van paranormale begaafdheid sprake is; ze kan echter wel helpen bij het uitsluiten van andere mogelijke verklaringen door middel van suggesties voor een goede opzet van het experiment. Zo'n andere verklaring hoeft bijvoorbeeld niet per se er op neer te komen dat het geheel doorgestoken kaart is. Ook onbewust en onbe-

doeld kan er een vertroebeling optreden die de bewijskracht van het experiment ondermijnt. Zo is het onder andere aanvechtbaar dat het trekken van de kaarten door de experimentator "willekeurig" plaatsvindt. Zonder dat de proefpersoon of hij het merken, kan hierbij toch licht iets van een patroon ontstaan dat het goed raden makkelijker maakt. Te denken valt bijvoorbeeld aan het vermijden van herhalingen vanuit de -onjuiste- gedachte dat de eerlijkheid gebiedt nu de andere kaarten eerst weer aan de beurt te laten komen. Een verbetering van het experiment in dit opzicht wordt eenvoudig bereikt door het vervangen van het willekeurig trekken door loten: de kaarten worden van één tot en met vijf genummerd en de experimentator kiest steeds de kaart die correspondeert met bijvoorbeeld het aantal ogen dat hij gooit met een zuivere dobbelsteen. (Als hij zes gooit, dan moet hij overgooien. Of hij kan zes in plaats van vijf kaarten gebruiken.)

Maar goed, na deze kleine excursie naar het eerder besproken gebied van het zorgen voor juiste waarnemingen om de conclusies op te baseren, keren we nu snel terug naar de opgeworpen vraag: "Wanneer is er iets aan de hand?" Bij 50 goed van de 250, duidelijk niet, en bij 250 goed van de 250, duidelijk wel. Waar onderweg van de 50 naar de 250 zullen we besluiten van mening te veranderen? Anders gezegd, tot welk aantal zijn we bereid het toeval het voordeel van de twijfel te geven in de zin dat we nog net bereid zijn zo'n aantal aan geluk bij het raden toe te schrijven? Wel, er wordt vaak gesteld "geluk valt niet te meten", maar toch is dat precies wat de statistiek, althans hier, voor ons kan doen. Voor ieder aantal kan zij berekenen wat de kans is om dit met louter geluk te overschrijden. Voor 61 is dat in dit geval bijvoorbeeld zo'n 5% en voor 66 zo'n 1%. Met dit soort gegevens gewapend zullen we vervolgens voor ons zelf moeten uitmaken wanneer we vinden dat het geluk te onwaarschijnlijk geworden is om nog langer als verklaring geloofwaardig te zijn. Dit is een keuze die valt buiten het bestek van de statistiek. De één zal bereid zijn een kans van 1/20 te accepteren op het ten onrechte beweren dat er iets aan de hand is, de ander kiest liever voor 1/100, enzovoort. Eén ding moet echter duidelijk zijn: zekerheid is niet voor het kiezen, of het moest zijn dat we pas besluiten het geluk af te schrijven als van de 250 pogingen er tenminste 251 goed zijn! Dan hebben we inderdaad zekerheid. Zowel dat we nooit een fout zullen maken, als ook dat we het nooit zullen opmerken als er wel iets aan de hand is. Want dat is natuurlijk de keerzijde van de medaille: hoe langer we wachten voor we het toeval als verklaring verwerpen, hoe groter de kans wordt dat een aanwezig effect onze aandacht ontsnapt. Ook de kans op dit soort fout kan de statistiek voor ons bepalen, waarna we zelf de afweging zullen moeten maken tussen deze twee soorten fout-kansen. Duwen we de eerste omlaag, dan veert de tweede omhoog, en omgekeerd. Ieder zal hier zijn eigen gulden middenweg moeten zien te vinden.

Zeer gewaardeerde toehoorders, laten we, op dit punt aangeland, een ogenblik stilstaan en nog eens terugblikken op het tot nu toe bestreken gebied. Hetgeen wij dan zien stemt hopelijk tot tevredenheid: met succes zijn alle aanvallen op de mathematische statistiek gepareerd. Uiteraard zal er in incidentele gevallen best nog wel eens wat aan te merken over blijven, maar de principes zijn ongeschonden uit de strijd gekomen. Van een aantal problemen is gedemonstreerd dat zij strikt

genomen niet op het terrein van de eigenlijke statistiek thuishoren. De meer wezenlijke klacht over het gebrek aan bewijskracht van de statistiek is ontzenuwd door uitvoerig aan te tonen dat redelijkerwijs geen zekerheid, maar slechts het meten van de onzekerheid van de statistiek verwacht mag worden.

Toch is er nog steeds geen reden om vol zelfgenoegzaamheid op de lauweren te gaan rusten. Er resteert namelijk een zeer wezenlijk probleem, vergeleken waarmee de in het voorafgaande behandelde aanvallen als ietwat speelse en frivole schermutselingen beschouwd kunnen worden. Ik zal daarom de nu volgende tweede helft van mijn voordracht in zijn geheel hieraan besteden. Daarbij wil ik trachten u een indruk te geven van een aantal tamelijk moderne tot zeer moderne ontwikkelingen die alle tot doel hebben dit probleem nader tot een oplossing te brengen.

Alvorens op het probleem zelf in te gaan, zal ik u eerst vertellen tot welke nare gevolgen het leidt. In het voorafgaande hebben we vastgesteld dat de statistiek geen zekerheid kan bieden, maar slechts kan meten hoe groot de onzekerheid is. Zo zagen we in het laatste voorbeeld dat niet met zekerheid gezegd kan worden of er wel of niet iets aan de hand is, maar dat slechts de kansen berekend kunnen worden op het nemen van elk der beide mogelijke verkeerde beslissingen, namelijk wel verwerpen als het niet moet en niet verwerpen als het wel moet. Bij een dergelijke stand van zaken verwachten we natuurlijk wel dat dit meten van de onzekerheid precies plaatsvindt, met andere woorden dat we op zijn minst zeker weten hoe onzeker we zijn. Helaas, dit is lang niet altijd het geval: soms is ook deze zekerheid van het tweede plan twijfelachtig en beginnen we ons af te vragen hoe ongrijpbaar de zekerheid is. Wat dit betreft vertoont de statistische zekerheid een treffende overeenkomst met onze beroemde sociale zekerheid: als je niet oppast, storten beide voor je ogen ineen.

Hoe komt dit nu? Om deze vraag te beantwoorden, moeten we ons verdiepen in de manier waarop in de klassieke statistiek problemen aangepakt worden. Vaak treft de wiskunde in het algemeen het volgende type verwijt: "Als je zo'n wiskundige een praktisch probleem voorlegt, kan hij het niet oplossen. Hij verandert het dan net zo lang tot hij iets heeft dat hij wel kan oplossen. Daar blijft hij dan verder verrukt mee spelen en laat jou met je nog steeds onopgeloste probleem zitten." Kortom, hij gedraagt zich als de dronkelap die 's avonds z'n verloren geld onder een straatlantaarn zoekt. Hij weet ook wel dat hij het een eind verderop al kwijt geraakt is, maar daar is het veel te donker om te zoeken. Voor de dronkelap is nog hoop, die is morgen weer nuchter, maar voor de wiskundige?

Gelukkig is de situatie bij lange na niet zo somber als hiervoor geschetst is: het is natuurlijk niet echt zo dat de klassieke statistiek alleen maar oplossingen aandraagt voor verkeerde problemen. De kern van waarheid die in dit gechargeerde beeld steekt is dat de problemen typisch wel in een nogal strak keurslijf van allerlei veronderstellingen gehesen worden, alvorens de dan ook vaak schitterende oplossing gegeven wordt. De vraag is nu hoe snel de schittering van zo'n oplossing verbleekt als een aantal van die veronderstellingen niet exact, maar slechts bij benadering juist blijken. De laatste jaren is deze vraag in steeds toenemende mate de statistiek gaan bezighouden. Ter illustratie vermeld ik een tweetal citaten uit recent

verschenen boeken. Het eerste is van Lehmann en luidt: ''The mathematical precision of the optimality results obtained tends to obscure the fact that statistically they are only rough approximations in view of the approximate nature of both the assumed models and the loss functions''. Het tweede is afkomstig van Hoaglin, Mosteller and Tukey en luidt: ''Classical statistical techniques are designed to be the best possible when stringent assumptions apply. However, experience and further research have forced us to recognize that classical techniques can behave badly when the practical situation departs from the ideal described by such assumptions''.

Samenvattend, blijkbaar kan de oplossing niet alleen te wensen overlaten als het onderliggende model, dat is het geheel van de gebruikte veronderstellingen, falikant verkeerd is, maar ook al indien het in plaats van exact slechts bij benadering goed is. Het eerste ligt in de lijn der verwachting, het tweede komt toch als een pijnlijke verrassing. Temeer daar bijna geen enkel model echt exact zal kloppen, er geldt veeleer dat ''all models are wrong, but some models are more wrong than others''. Om de eventueel resterende hoop dat het allemaal altijd wel meevalt de grond in te boren, zal ik een paar voorbeelden geven van het bedoelde verschijnsel. Het eerste is ontleend aan Scheffé. Stel we willen een uitspraak doen over de verwachting — populair gezegd het gemiddelde — van een van het toeval afhankelijke variabele waarvan we aannemen dat deze een normale verdeling heeft. We trekken dan een steekproef, waarvan we aannemen dat deze aselekt is, dat wil zeggen dat de erin voorkomende variabelen onafhankelijk zijn van elkaar. Binnen dit model kunnen we een betrouwbaarheidsinterval voor de verwachting afleiden. Dit houdt in dat we komen tot een uitspraak die er bijvoorbeeld zo uitziet: ''Met betrouwbaarheid 95% ligt de verwachting tussen 8.3 en 8.9''. Zoals alle aannamen, geldt de aanname van onafhankelijkheid echter slechts bij benadering. Vaak zal er sprake zijn van een lichte correlatie tussen opeenvolgende waarnemingen. Voor de betreffende correlatiecoëfficiënt werd in een aantal praktijkgevallen bijvoorbeeld als gemiddelde waarde 0.2 gevonden. Maar in dit geval is de feitelijke betrouwbaarheid niet langer 95%, maar slechts 90%. De kans om de ware waarde van de verwachting te missen is dus al verdubbeld bij een dergelijke geringe afhankelijkheid.

In plaats van de aanname van onafhankelijkheid kunnen we natuurlijk ook de veronderstelling dat de variabelen een normale verdeling hebben in twijfel trekken. Ten aanzien van betrouwbaarheidsintervallen voor de verwachting zijn de gevolgen hiervan, althans bij niet al te kleine steekproeven, niet al te ernstig. In dit geval is inderdaad vol te houden dat het bij kleine afwijkingen allemaal wel mee valt. Zijn we echter niet in de verwachting, maar in de variantie geïnteresseerd, dan verslombert het beeld weer drastisch en vertoont het sterke overeenkomst met het hiervoor geschetste beeld van betrouwbaarheidintervallen voor de verwachting onder afhankelijkheid: bij vrij geringe afwijkingen van de normaliteitsaanname kan de betrouwbaarheid al aanzienlijk afwijken van het nominale niveau.

Een ander bekend voorbeeld is afkomstig van Tukey. In dit ''dramatic example'', zoals hij het zelf noemt, beschouwen we weer een aselekte steekproef uit een normale verdeling en weer zijn we in de variantie geïnteresseerd, om precies

te zijn, in de standaardafwijking, dat is de wortel uit de variantie. Dit keer zoeken we geen intervalschatting zoals een betrouwbaarheidsinterval, maar zijn we al tevreden met een simpele puntschatting. Als het normale model precies klopt, is het optimaal om de bekende steekproefstandaardafwijking te gebruiken. Stellen we de kwaliteit hiervan op 100, dan is bijvoorbeeld de kwaliteit van een andere mogelijke schatter, de gemiddelde absolute afwijking, gelijk aan 88. Met de precieze definitie van "kwaliteit" in dit verband wil ik u niet vermoeien; laat ik volstaan met op te merken dat hoe sterker de schatter om de ware waarde geconcentreerd is, hoe hoger de kwaliteit. Stel nu echter eens dat er een lichte verontreiniging van het waarnemingsmateriaal heeft plaatsgevonden en dat enkele van de waarnemingen uit een andere normale verdeling, met een bijvoorbeeld drie keer zo grote variantie, afkomstig zijn. Dan gebeurt er inderdaad iets dat dramatisch genoemd kan worden: al bij 2 van zulke slechte waarnemingen in een steekproef van 1000 is de eerste schatter zijn hele kwaliteitsvoorsprong van 12% ten opzichte van de tweede kwijtgeraakt. Erger nog, bij 5% slechte waarnemingen is ondertussen de tweede schatter ruim tweemaal zo goed geworden als de eerste! Een voor de hand liggende repliek is uiteraard dat men dan ook maar niet met verontreinigd waarnemingsmateriaal moet werken. De ervaring leert echter dat in de praktijk een vervuilingsgehalte van enkele procenten geheel normaal is, ook bij redelijk zorgvuldig uitgevoerde steekproeven.

Na de genoemde voorbeelden zal het duidelijk zijn dat het probleem van de afhankelijkheid van de modelveronderstellingen niet genegeerd mag worden. Gelukkig is dat ook steeds minder het geval. Ondertussen zijn er allerlei technieken beschikbaar die alle op hun eigen wijze proberen daar iets aan te doen, bijvoorbeeld door het aantal veronderstellingen waarmee gewerkt wordt zoveel mogelijk te beperken of door de gevoeligheid voor afwijkingen van de veronderstellingen zoveel mogelijk tegen te gaan. In hetgeen nu volgt zal ik een aantal van deze methoden de revue laten passeren.

De waarschijnlijk oudste aanpak is die van het verwijderen van uitschieters uit het waarnemingsmateriaal. Zoals Barnett en Lewis het ietwat gedragen uitdrukken: "From the earliest gropings of man to harness and employ the information implicit in collected data as an aid to understanding the world he lives in, there has been a concern for "unrepresentative", "rogue" or "outlying" observations in sets of data". De gedachte is natuurlijk ook bijzonder aantrekkelijk. In één van de eerder genoemde voorbeelden zagen we dat een paar slechte waarnemingen al roet in het eten konden gooien. Wel, bekijk dan eerst het waarnemingsmateriaal eens, stel vast wat de boosdoeners zijn en gooi die eruit. Dan kan de analyse alsnog uitgevoerd worden op zuiver in plaats van verontreinigd materiaal.

Helaas, zo simpel is het nu ook allemaal weer niet, het bestuderen van uitschieters heeft zeker wel geleid tot een aantal nuttige en goed bruikbare technieken, maar het is bepaald niet zo dat ons probleem hiermee in één klap geheel de wereld uit is. Daarvoor zitten er toch te veel haken en ogen aan deze benadering. In de eerste plaats, als men niet uitkijkt, verwordt het vaststellen van wat uitschieters zijn en wat niet tot een hoogst subjectieve bezigheid. Wat de één een verrassende uitkomst

vindt, vindt de ander misschien nog gewoon. Maar, zoals Lehmann terecht opmerkt: "A procedure of this kind..... is legitimate only if a precise rule for discarding observations is formulated before any observations are taken. Discarding observations according to vague subjective criteria as to whether an observed value does or does not constitute a "normal" observation opens the way to the suppression of data that do not fit sufficiently well into preconceived ideas or do not produce desired results. It also deprives the statistical procedure of any basis for assessing its performance". Om een, niet geheel uit de lucht gegrepen, voorbeeld te geven, als een nieuwe behandelingsmethode bij een aantal patiënten in het algemeen leidt tot een beter resultaat dan de tot dan gebruikelijke methode, maar in één bepaald geval tot een dramatisch slechter resultaat, dan is de verleiding voor degene die de nieuwe methode bedacht heeft groot om deze ene waarneming maar liever weg te laten, als zijnde niet representatief, ongelukkig, of iets dergelijks. Toch dient, al was het maar om statistische redenen, aan deze verleiding krachtig weerstand geboden te worden!

Een tweede punt is vervolgens dat het echt uitschieten in de zin van onmiddelijk opvallen alleen optreedt in statistisch heel simpele situaties. Bij multivariate problemen is het veel lastiger om een idee te krijgen welke waarnemingen verdacht zijn, laat staan om vast te stellen of die verdenking terecht is. Nog een ander probleem is dat niet iedere afwijking van een veronderstelling behoeft te berusten op het vóórkomen van uitschieters. Behalve door het af en toe optreden van slechte waarnemingen kan de normaliteitsveronderstelling bijvoorbeeld ook aangetast worden door een verdeling met wat zwaardere staarten te kiezen. Het gevaar ontstaat dan dat men met uitschieter-technieken jaagt op iets dat er eigenlijk niet is. Tenslotte wijs ik er nog op dat het te optimistisch is om te menen dat na het verwijderen van de uitschieters simpelweg de gebruikelijke statistische analyse van de gezuiverde steekproef gegeven kan worden. Deze ingreep beïnvloedt wel degelijk het gedrag van de gebruikte statistische procedure en wel op een vaak zeer moeilijk te achterhalen wijze.

Het voorafgaande maakt duidelijk dat er redenen genoeg over blijven om nog eens een volgende categorie technieken onder de loupe te nemen. Dat zijn dan de verdelingsvrije of nietparametrische methoden. De verdienste van deze methoden is met name dat er voor hun geldigheid maar weinig aannamen nodig zijn. Dit tezamen met hun eenvoud en makkelijke toepasbaarheid heeft ze op een aantal gebieden tot bijzonder populaire concurrenten van de klassieke statistische methoden gemaakt. De harde kern bestaat uit procedures gebaseerd op rangnummers, zoals rangtoetsen. Voor de toepassing hiervan hoeft alleen de rangschikking van de waarnemingen bekend te zijn, de feitelijke waarden zijn overbodig. Bovendien hangt bij dergelijke toetsen de kans op ten onrechte verwerpen niet af van de onderliggende verdeling. Daarom ook worden ze "verdelingsvrij" of "nietparametrisch" genoemd, dat wil zeggen vrij van de aanname dat de verdeling tot een bepaalde parametrische familie, zoals de normale, behoort. Met name vanaf de vijftiger jaren heeft dit gebied een stormachtige ontwikkeling doorgemaakt, waaraan ook door Nederlandse statistici een belangrijke bijdrage werd geleverd. Aan-

vankelijk lag de nadruk op de snelheid, eenvoud en ruime toepasbaarheid van de rangtoetsen en ging men er van uit dat in ruil hiervoor wel een belangrijk kwaliteitsverlies voor lief genomen zou moeten worden ten opzichte van de klassieke toetsen onder optimale voorwaarden. Kortom, men beschouwde ze als "quick-and-dirty-methods". Gaandeweg werd echter steeds duidelijker dat rangtoetsen niet alleen een uitkomst bieden in moeilijke situaties, maar dat ze bovendien een vergelijking onder de klassieke aanname van normaliteit met de dan optimale standaard methoden eigenlijk verbluffend goed kunnen verdragen. Uiteraard is er dan sprake van een iets lagere kwaliteit, maar dit verschil is naar veler mening zó gering, dat het ruimschoots gecompenseerd wordt door de bescherming die geboden wordt tegen de gevolgen van afwijkingen van de normaliteitsaanname, afwijkingen waarvan we gezien hebben hoe gevoelig vele klassieke methoden daarvoor zijn. Veel van het meer recente onderzoek heeft zich beziggehouden met het verschaffen van meer en meer gedetailleerde informatie omtrent de feitelijke grootte van dit kwaliteitsverlies, om zo een steeds betere afweging tussen het gebruik van nietparametrische en klassieke procedures mogelijk te maken.

De zojuist geschetste situatie verschaft ons een haast onmerkbare overgang naar een volgende benadering, namelijk die der robuuste methoden. Hierbij gaat het om procedures die robuust zijn, dus een stootje kunnen velen, in de zin dat ze ongevoelig zijn voor kleine afwijkingen van de modelveronderstellingen. Onder dergelijke afwijkingen daalt hun kwaliteit slechts licht ten opzichte van het nominale niveau, en niet scherp, zoals bij de klassieke procedures vaak het geval is. In ruil hiervoor moet een gering kwaliteitsverlies ten opzichte van de optimale klassieke procedures voor lief genomen worden als het model onverhoopt wel exact klopt. Dit verlies kan dus opgevat worden als een premie die betaald wordt onder het model om verzekerd te zijn van een redelijke kwaliteit in de omgeving van het model. De eerder genoemde verdelingsvrije methoden passen aardig in dit stramien, maar dit wordt door Huber, de "godfather" van de robuuste statistiek, beschouwd als meer geluk dan wijsheid. Hij vindt dat de twee soorten methoden eigenlijk heel weinig met elkaar te maken hebben. Nu is dat misschien wat sterk gesteld, maar er is inderdaad een duidelijk verschil in benadering aanwezig. Bij verdelingsvrije methoden gaat het om het wéglaten van veronderstellingen, waardoor een zeer ruim gebied bestreken kan worden, terwijl de robuuste methoden letterlijk en figuurlijk veel sterker in de buurt blijven van het oorspronkelijke klassieke model. Er wordt als het ware een extra ring om dit oorspronkelijke model heen getrokken, waardoor een soort supermodel ontstaat waarin behalve de optimale situatie ook de mogelijke afwijkingen worden beschreven.

Ter illustratie kan dienen een situatie die we al bij de uitschieter-technieken hebben bekeken: onder het model komen alle waarnemingen uit eenzelfde normale verdeling, terwijl onder het supermodel een gegeven, vrij kleine, fractie van de waarnemingen slecht is in de zin dat ze uit een willekeurige andere verdeling stammen. De typerende aanpak van de robuuste statistiek is nu het bepalen van een minimax oplosssing. Dit betekent dat onder deze andere verdelingen diegene opgespoord wordt die het ergst is in de zin dat als de vervuiling van het waarnemingsmateriaal

uit deze verdeling afkomstig is, de best mogelijke oplossing in dit geval lager van kwaliteit is dan de best mogelijke oplossing voor elke andere mogelijke keuze van de vervuilingsverdeling. Vanwaar de interesse in deze slechtste beste oplossing? Wel, men kan laten zien dat deze oplossing de volgende minimax-eigenschap bezit: het maximale kwaliteitsverlies dat hierbij optreedt als we het hele supermodel doorlopen, dat wil zeggen alle mogelijke vervuilingsverdelingen uitproberen, is altijd kleiner dan het maximale kwaliteitsverlies dat we bij een dergelijke rondgang oplopen met ongeacht welke andere oplossing. Kortom, we beschermen ons hier nadrukkelijk tegen het ergste dat er kan gebeuren. Deze aanpak is karakteristiek voor dit gebied: het accent ligt heel duidelijk op de veiligheid.

Weer biedt deze schets van één groep technieken ons een natuurlijk aanknopingspunt voor een volgende categorie. Immers, in voornoemd voorbeeld mochten de slechte waarnemingen weliswaar uit elke willekeurige verdeling afkomstig zijn, maar we gingen er wel van uit dat we met een gegeven fraktie te maken hadden. Waarom eigenlijk? Onweerstaanbaar dringt zich de vraag op of het misschien niet beter zou zijn om eerst eens een blik op de waarnemingen te werpen om te zien hoe groot die fraktie eigenlijk is. Misschien is die wel nul, dan kunnen we gewoon de optimale standaard procedure gebruiken. En als de fraktie niet nul is, dan kunnen we tenminste zorgen dat we de bij deze fraktie behorende minimax-oplossing gebruiken. Dergelijke bespiegelingen leiden ons rechtstreeks naar de categorie der adaptieve of te wel zich aanpassende procedures. Grofweg gezegd, hierbij selekteert men eerst uit een klasse van mogelijke veronderstellingen diegenen die het meest overeenstemmen met het voorhanden zijnde waarnemingsmateriaal. Daarna past men de bij de gekozen veronderstellingen optimale procedure toe. Deze gang van zaken is door Lehmann treffend aangeduid als "a case of eating your cake and having it too". Meer op z'n Hollands zouden we dit waarschijnlijk kunnen aanduiden als "voor een dubbeltje op de eerste rang zitten". En, zoals we op grond van onze calvinistische voorgeschiedenis ragfijn zullen aanvoelen, zoiets kan natuurlijk nooit goed aflopen. Wel dat doet het dan ook niet, althans niet zonder meer.

In de eerste plaats, net als bij de uitschieter-technieken, geldt er weer dat we niet kunnen doen alsof er niets aan de hand is in de zin dat we gewoon de standaard statistische analyse kunnen uitvoeren. Net als het verwijderen van uitschieters beïnvloedt het kiezen in plaats van het als vaststaand aannemen van de modelveronderstellingen de statistische eigenschappen van de uiteindelijk toegepaste procedure. Daar moeten we dus rekening mee houden en als regel geldt ook hier weer dat dit aspekt de totale analyse aanzienlijk gecompliceerder maakt. Een tweede, minstens zo wezenlijk punt is het volgende. We kiezen uit de klasse van beschikbare modellen weliswaar degene die het meest in overeenstemming is met het waarnemingsmateriaal, maar dat garandeert natuurlijk niet dat we daarmee precies het juiste model hebben. In feite betreft het ook hier een schatting waarvoor zal gelden dat deze op den duur, dat wil zeggen naar mate het aantal waarnemingen toeneemt, naar alle waarschijnlijkheid steeds dichter en dichter bij de ware waarde, hier dus het juiste model, zal komen te liggen. En dat "op den duur" kan ons duur komen te staan: als de klasse van beschikbare modellen groot is, heeft de adaptieve proce-

dure zoveel keus dat het veel te lang duurt voordat hij zo dicht bij het ware model aangeland is dat zijn uiteindelijk optimale karakter zichtbaar wordt. Theoretisch gezien mag het tot innige voldoening stemmen als een bepaalde procedure op den duur altijd en overal beter is dan andere procedures, maar als daar eerst honderden of zelfs duizenden waarnemingen voor nodig zijn, wordt het praktisch nut op zijn zachtst gezegd wat twijfelachtig. Nog anders gezegd, het nobele streven naar de uiteindelijk hoogste kwaliteit kan ten koste gaan van de veiligheid in concrete gevallen. Precies om deze reden weigert bijvoorbeeld Huber om adaptieve procedures als een deelklasse van robuuste procedures te beschouwen.

De remedie tegen genoemde kwaal ligt voor de hand: men moet zich bij het aanpassen niet al te ambitieus betonen en de klasse van in aanmerking komende modellen niet al te groot kiezen. Te denken valt aan een beperking tot één of enkele parameters, zoals in het eerder genoemde voorbeeld, waar eerst aan de hand van de waarnemingen de vervuilingsgraad geschat wordt, alvorens de te gebruiken procedure gekozen wordt. Een dergelijke gematigde vorm van aanpassing kan naar de mening van vele experts ook al goed werken bij steekproeven van meer gebruikelijke omvang. Op deze manier wordt misschien een aanvaardbaar compromis bereikt tussen het toch wel wat sombere pessimisme van de robuuste minimax-oplossing enerzijds en het ongebreidelde streven naar het hogere van de adaptieve procedures anderzijds.

Zeer gewaardeerde toehoorders, in het voorafgaande heb ik getracht u een beeld te schetsen van de grote verscheidenheid aan problemen waar de statistiek mee geconfronteerd wordt. Eerst heb ik u langs enige randverschijnselen gevoerd, zoals aan het begin het toepassen van de statistiek op onjuist waarnemingsmateriaal en aan het eind het verkeerd interpreteren van op zich correcte statistische conclusies. Daarna heb ik u meegenomen naar het inwendige, waar we gezien hebben dat zekerheid echt ongrijpbaar blijft, maar dat de statistiek wel de pretentie heeft tenminste de mate van onzekerheid te bepalen. Het waarmaken van deze pretentie bleek bij nadere beschouwing vaak al moeilijk genoeg te zijn: als de modelveronderstellingen niet exact vervuld zijn, dreigt ook zekerheid over de mate van onzekerheid ongrijpbaar te worden. Ik hoop er althans enigermate in geslaagd te zijn u er van te overtuigen dat dank zij een skala van deels zeer recente technieken het ook in dergelijke theoretisch niet ideale maar in de praktijk vaak voorkomende situaties toch mogelijk geworden is om de onzekerheid redelijk in de greep te krijgen. Zo vormen deze technieken een nieuwe handreiking van de statistiek naar de zijde van de gebruiker. Zij stellen hem in staat op een aanmerkelijk ruimer gebied dan voorheen op verantwoorde wijze statistische analyses uit te voeren.

Aan het einde van mijn voordracht gekomen, betuig ik allereerst mijn dank aan Hare Majesteit de Koningin voor mijn benoeming tot hoogleraar aan deze universiteit (en in deze dank wil ik allen betrekken die aan het tot stand komen van deze benoeming hebben meegewerkt).

Dames en Heren leden van de wetenschappelijke en niet-wetenschappelijke staf, Ik hoop en verwacht tot een plezierige en vruchtbare samenwerking met velen van u te komen. Tevens hoop ik dat in mijn voordracht voldoende duidelijk naar voren gekomen is dat de problemen die de statistiek bij de gebruiker oproept, mij zeer ter harte gaan en dat ik gaarne bereid ben u met statistische adviezen terzijde te staan.

Hooggeleerde van Zwet,

U bent het die mij hebt ingewijd in de geheimen van de mathematische statistiek en vanaf deze plaats wil ik u daarvoor gaarne mijn erkentelijkheid betuigen. Veel heb ik van u geleerd, en niet alleen op statistisch gebied. Ook de wijze waarop u mijn begeleiding vorm wist te geven, dwingt nog steeds mijn respect af. Enerzijds bood u mij volop vrijheid en daarmee de mogelijkheid voor ontwikkeling tot zelfstandig onderzoeker. Maar anderzijds hield u zeer nauwgezet het oog op al hetgeen ik uitvoerde, zodat ik leerde mij wel tweemaal te bedenken voor ik iets opschreef dat net niet helemaal "af" was, net niet helemaal klopte, net niet helemaal correct Engels was, of onverhoopt op nog andere wijze net niet helemaal deugde. Ik kan slechts hopen dat deze training nog steeds zijn duidelijke sporen blijft nalaten.

Dames en Heren leden van de afdeling Toegepaste Wiskunde van de Technische Hogeschool Twente,

Met veel genoegen denk ik terug aan de jaren die ik in uw midden heb mogen doorbrengen in de vakgroep Stochastiek. De prettige onderlinge verhoudingen die al deze jaren bestonden en ook bleven bestaan toen de van buiten komende bedreigingen van universiteiten en hogescholen steeds verder begonnen toe te nemen, heb ik als zeer positief ervaren.

Dames en Heren studenten,

Ik hoop dat uit mijn voordracht gebleken is dat statistiek noch afschuwwekkende wiskunde, noch onbestemd gegoochel met getallen voor u behoeft te zijn. Met gezond verstand komt men niet door het ganse statistische land, maar wel een opvallend eind! Hiermee gewapend kunt u al snel leren zien of statistische conclusies kunnen kloppen, en zo ja, wat ze eventueel voor u mogen betekenen. Met nog iets meer inspanning kunt u leren hoe u niet alleen conclusies kunt begrijpen, maar ze ook in eenvoudige gevallen eerst zelf kunt trekken, hierbij al dan niet geholpen door de computer. Graag zal ik u bij het verwerven van dit begrip en dergelijke vaardigheden behulpzaam zijn.

Zeer gewaardeerde toehoorders, ik heb gezegd.