Sara van de Geer
*Mathematisch Instituut, Universiteit Leiden*
*Postbus 9512, 2300 RA Leiden*
*geer@math.leidenuniv.nl*

## Inaugurele rede

# Een zeker

**Deze rede is op 1 december 2000 uitgesproken door Sara van de Geer bij haar benoeming tot hoogleraar Kansrekening en Statistiek aan de Universiteit Leiden.**

Ik wil u vertellen over het nut van statistiek, en over wat mij aantrekt bij de beoefening van het vak. Het zal daarbij blijken dat deze twee zaken volkomen los van elkaar staan.

Statistische methoden zijn er voor om op een zinnige manier met gegevens om te gaan. Aan gegevens is op zich geen gebrek. De nieuwe kennis-economie draait om ICT: Information and Communication Technology. Het gaat daarbij bijvoorbeeld om het digitaal maken en op internet zetten van allerlei soorten informatie en desinformatie (reclame bijvoorbeeld). Met informatie worden hier ruwe gegevens bedoeld, zoals getallen, plaatjes, geluid, teksten, enzovoort. Zulke informatie consumeren is soms best leuk, maar schenkt vaak niet meer bevrediging dan een zak chips. Nog leuker is het om gegevens daadwerkelijk ergens voor te gebruiken, bijvoorbeeld om nieuwe structuren te ontdekken, tot conclusies en inzichten te komen, meningen te vormen of beslissingen te nemen. Veel informatie is al door de statistische molen gehaald, in andere gevallen wordt u geacht uzelf door de informatierijstebrij heen te eten. Statistische methoden en kennis van statistiek zijn hard nodig om de nieuwe economie niet te laten doldraaien. Hugo Battus [2] noemt statistiek dan ook 'het nuttig verliezen van informatie'.

De gegevensberg neemt sneller toe dan de geheugencapaciteit van computers. Voorbeelden zijn de gegevens verkregen met de klantenkaart van een grote kruidenier, of de gegevens over gen-expressieniveaus die nu bij het human genome project binnen stromen.

De toevloed van gegevens, de ontwikkelingen op het gebied van ICT, en ook andere aspecten van onze samenleving maken dat statistische methoden steeds meer een rol gaan spelen. Bij zaken van direct persoonlijk belang (bijvoorbeeld de veiligheid van onze leefomgeving, de toelaatbare grens voor blootstelling aan dioxine, PCB's of landbouwgif) is het vaak niet meer mogelijk zélf de voors en tegens af te wegen, maar moet men vertrouwen op statistische analyses uitgevoerd door experts. Ik citeer Verkijlen [21], die in *Filosofie Magazine* schrijft: "Met de onteigening van de waarneming is ook de individuele oordeelsvorming een gepasseerd station. Wie van statistiek en methodologie geen kaas heeft gegeten kan strikt genomen over de waarheidsaanspraken van de diverse wetenschappen niet oordelen."

We staan daarom voor de belangrijke taak mensen enthousiast te maken voor kansrekening en statistiek en een goede opleiding in deze vakken aan te bieden. Ik denk dat het nuttig is nu eerst mijn vermoeden over hoe in het algemeen over statistiek gedacht wordt met u te delen, en daarbij eventuele misverstanden aan de orde te stellen.

### Geen wiskunde

Onder leken, ook scholieren en studenten, is statistiek helaas niet zo populair. Het vak wordt bijvoorbeeld geassocieerd met saai boekhouden. De boekhoudkundigen onder u zullen nu misschien roepen dat ik boekhouden ten onrechte saai noem, hiervoor mijn excuses.

Vroeger was statistiek niet meer dan het weergeven van gegevens in tabellen en grafieken, het uitrekenen van gemiddelden en dergelijke. Eén ding kan ik alvast opmerken: tegenwoordig gaat het vaak niet alleen om statistische methoden voor 'saaie' rijtjes getallen, of om 'saaie' bevolkingsstatistieken, maar ook om de statistische analyse van meer exotische objecten, zoals plaatjes, geluid en films, en/of om kwalitatieve gegevens al of niet op een geordende schaal. Bij de analyse komen zaken aan de orde als patroonherkenning, compressiemethoden, neurale netwerken, beeldanalyse, filters, en driedimensionale grafische representaties. Het vak is veelomvattender geworden, en daarmee veel leuker!

Statistiek gaat in het algemeen uit van de impopulaire strategie de mens niet als individu te behandelen. De beroemde statistici Kendall en Stuart [11] zeggen inderdaad in hun standaardwerk: "The statistician, like Nature, is mainly concerned with the species and is careless of the individual." Het gaat hier om de tegenstelling tussen de individuele delen en het geheel: "Het specifieke en unieke tegenover het zich herhalende en het universele, het concrete tegenover het abstracte, voortdurende beweging tegenover rust, het innerlijke tegenover het uiterlijke, kwaliteit tegenover kwantiteit . . ." [3]. Met name het herhalende is een idee wat in mijn vak een grote rol is toebedeeld. Het is niet zo dat er van een werkelijke herhaling der gebeurtenissen

Sara van de Geer

# toeval

wordt uitgegaan, dat de geschiedenis zich herhaalt. Nee, een statisticus denkt meer in een metafysische trant en gaat er bij de theorievorming van uit dat het *in principe* mogelijk is de experimentele metingen willekeurig vaak te herhalen. De statisticus maakt zich er een voorstelling van wat er gebeurt bij oneindig vaak herhalen, zonder dat hij of zij van u zal verlangen dat die herhalingen daadwerkelijk uitgevoerd worden.

Ik noemde al het begrip gemiddelde. Het gemiddelde middelt individuele verschillen uit. (De gemiddelde Nederlander is overigens iemand anders dan Jan Modaal.) Men kan proberen zich zoveel mogelijk van de gemiddelde Nederlander te onderscheiden, maar die lastige gemiddelde Nederlander verandert met u mee! In een van de boeken van Gerrit Krol [14] vond ik een uitspraak: "van een bekend filosoof, die zegt dat het leven in zijn uitersten zijn waarden heeft, maar in het gemiddelde zijn behoud." Inderdaad, het verlangen gaat uit naar het extreme, niet naar het gewone dagelijkse. Aan de andere kant, die zelfde filosoof schrijft [4] "Een statisticus had eens uitgerekend dat een rivier die hij wilde oversteken een *gemiddelde* diepte van één meter had. Vol vertrouwen begaf hij zich te water en . . . verdronk."

De moderne statistiek maakt nog altijd gebruik van het uitmiddelen van individuele fluctuaties, al gebeurt het soms op een heel fijnmazig niveau. Het doel is namelijk om structuur te ontdekken, om zich niet te laten afleiden door toevallige afwijkingen van de onderliggende structuur. Een statisticus gaat er van uit dat de verschijnselen behept zijn met een zekere mate van toeval, en probeert het signaal (structuur) en ruis (toevallige afwijkingen) te scheiden. Het vak behelst heel wat meer dan een *black box* benadering van de verschijnselen. Als men zegt dat een gevonden samenhang tussen twee variabelen *slechts statistisch* van aard is, moet men dat dus niet opvatten als hét signaal om de statistische analyse te staken.

Statistiek kan een moeilijk vak zijn, niet gespeend van wiskunde. Desondanks komt het vak op het curriculum van zowat iedere studierichting voor. Vaak vormt het een struikelblok voor studenten. Nachtmerries over het statistiektentamen zijn bijna gemeengoed! Ik weet echter zeker dat het mogelijk is fascinatie, in plaats van frustratie, de boventoon te laten voeren. Het is mijn opdracht om studenten te laten inzien dat statistiek niet dat rigide, onbegrijpelijke vak van regeltjes, en $\chi$-kwadraattabellen is, en ze warm te maken voor het vak. Zie ook het verslag 'collegezweet' in *Mare* [13], waar u kunt lezen dat me dat niet altijd meevalt.

Ook het afgelopen najaar gaf ik les aan niet-wiskundigen. Ik herontdekte hoe lastig statistiek wordt als men in plaats van over het veilige wiskundige kader, het moet hebben over de implicaties in de werkelijke wereld. Op een gegeven moment had ik een numerieke illustratie gemaakt van de theorie. De getallen bleken echter niet te kloppen! Was de theorie dan fout? Ik besloot dat er verschillende waarheden bestaan: de numerieke waarheid (zoals $1 + 1 = 2$), de (wiskundig) theoretische waarheid, die vaak niet numeriek verifieerbaar is (bijvoorbeeld de waarheid dat er oneindig veel priemgetallen bestaan), de statistische waarheid (die waar is met grote kans), kansuitspraken (die een speciaal geval zijn van wiskundige uitspraken en die nooit in praktijk falsifieerbaar zijn), uitspraken die ongeveer waar zijn *en* uitspraken die meestal ongeveer waar zijn, of ongeveer waar zijn met grote kans. Het gegeven dat er zoveel verschillende waarheden zijn maakt dat ik, als ik statistiek probeer uit te leggen zonder de hulp van wiskunde, al gauw word verleid tot allerlei filosofische overpeinzingen.

Bij statistiek wordt meestal gedacht aan het toegepaste, ofwel technische deel. Er is dus ook een filosofische kant, een theoretische, ofwel wetenschappelijke kant. Op school is statistiek onderdeel van Wiskunde A, en dat is prima. Maar het vak is *geen* onderdeel van Wiskunde B. Dit geeft de indruk dat het alleen een hulpvak is, en weinig met echte wiskunde te maken heeft. De studierichting wiskunde heet in Nederland niet meer Wiskunde, maar Wiskunde en Statistiek. Ook daarmee wordt gesuggereerd dat statistiek niet onder wiskunde kan vallen. Of zou het zo zijn dat statistiek apart genoemd wordt omdat het zo'n belangrijk vak is? Wat mij betreft ligt hier een enorm spanningsveld. Aan de ene kant staat mijn persoonlijke motivatie om statistiek te bedrijven. Voor mij is het abstracte, echte wiskunde. Ik ben niet uit op maatschappelijk nut. Aan de andere kant is er een beweging die zegt dat de wiskunde in het algemeen zijn nut maar eens moet bewijzen. Dit idee is natuurlijk niet van vandaag of gisteren. Bij de oprichting in 1946 van het Mathe-

matisch Centrum (nu Centrum voor Wiskunde en Informatica) in Amsterdam was er ook een vraag vanuit de maatschappij dat wiskunde gericht moest zijn op maatschappelijke doelstellingen.

**Structuur en modellen**

Tot zover mijn worsteling met het imago van statistiek. Laten we het eens over die echte wiskunde hebben. Het mooie van wiskunde is voor mij dat het de mogelijkheid geeft een systeem op te zetten dat in zichzelf bestaat, en dus niet gehinderd wordt door storende factoren uit het werkelijke leven. Ik beweer hier niet dat wiskunde consistent is met zichzelf of iets dergelijks (want dat is niet zo, zie Gödel [10]), maar eerder het feit dat het om een abstractie gaat, los van de concrete realiteit. Het abstractie-ideaal leidt tot het beeld van de teruggetrokken wetenschapper, de echte beta, ook wel kortweg 'nerd' genoemd. Ik laat me dit abstracte speelgoed echter niet afnemen. Ik wil het wel graag met u delen!

Een abstractie is leeg, in die zin dat zij door de toepasser gevuld kan worden met een concrete betekenis. Statistiek gaat over de ontwikkeling van modellen die die concrete realiteit moeten beschrijven. De mathematische statistiek formuleert het modellen-bouwen in abstracte termen. Dat is dus een abstractie, die door de toepasser gevuld kan worden met een concreet model. Omdat zo'n model ook weer een abstractie is hebben we hier te maken met abstractie in de tweede graad!

Een van mijn drijfveren om me met statistiek bezig te houden is nu om de overeenkomsten in structuur van diverse modellen te begrijpen en te beschrijven. Het gaat mij daarbij niet zozeer om een gegeven model, maar om de overeenkomsten binnen een klasse van modellen. Hierbij komt het begrip entropie naar voren, ofwel chaos. Modellen kunnen met elkaar vergeleken worden op grond van de hoeveelheid entropie die toegelaten wordt. Hoe meer entropie, des te moeilijker het is om informatie uit de gegevens te halen. Dat begrip entropie is in feite een heel grove, maar algemeen bruikbare, kwantificatie van de hoeveelheid informatie die je uit de gegevens kan halen, ofwel van de hoeveelheid structuur, die het model toelaat.

De mathematisch statisticus onderscheidt diverse vormen van informatie: Fisher informatie, Kuhlback-Leibler informatie, Shannon informatie, et cetera. Deze begrippen formaliseren het idee dat gegevens *an sich* niet equivalent zijn met informatie, maar een bepaalde hoeveelheid informatie kunnen bevatten. Binnen een mathematisch model is precies aan te geven hoeveel informatie een bepaald type van gegevens bevat. Op grond van de daadwerkelijke gegevens kan men onder bepaalde voorwaarden de geobserveerde informatie uitdrukken in een getal. De geobserveerde informatie is wel wat anders dan de werkelijke informatie die in de gegevens zit: de geobserveerde informatie is een schatting van de werkelijke informatie. Het zal mij niet verbazen als het onderscheid voor velen van u niet overduidelijk is. Wat dit overigens wel illustreert is het volgende: het wiskundig formaliseren van begrippen uit het dagelijkse leven, zoals informatie, leidt er vaak toe dat men een onderscheid moet maken, daar waar dat in het dagelijks leven niet gebruikelijk is. In het mathematische leven zou men dan ook minder misverstanden of ruzie moeten hebben dan in het dagelijkse leven. Voor een deel is dat ook zo.

Een statisticus gebruikt gegevens niet alleen om er informatie uit te halen, maar ook om de kwaliteit ervan te beoordelen. Dit is een enigszins zelf-referente bezigheid. U kent vast het verhaal van de Baron van Münchhausen. Een statisticus is iemand die, niet ontmoedigd door Gödels waarschuwingen, zichzelf aan de haren uit het moeras probeert te trekken, en die dat nog lukt ook!

Statistiek gaat dus over het doen van uitspraken, maar vooral over in hoeverre men de uitspraken moet geloven. Voor post-normale wetenschappers (die de wetenschapsbeoefening meer op inspraak en participatie willen baseren) gaat dat niet ver genoeg. Zij vinden dat niet alleen rekening gehouden moet worden met de onzekerheid binnen het model, maar dat ook de geldigheid van het model zelf kritisch onder de loep moet worden genomen. Voor een mathematicus is dit geen wezenlijk vernieuwend idee: maak simpelweg het model onderdeel van een groter (meta)model. In praktijk kan het wel het een en ander aan discussies teweeg brengen. Zo wordt voorgesteld dat belangrijke problemen in de samenleving niet door de wetenschappers alleen opgelost kunnen worden [7]. Wetenschappers zouden namelijk alleen puzzeltjes kunnen oplossen, ik noem het maar speelgoedproblemen. Beslissingen over echte, complexe en vaak urgente problemen (klimaatverandering, afnemende biodiversiteit, enzovoort) zouden onder andere op grond van de uitkomsten van publiek debat genomen moeten worden. Eén van de argumenten hierbij is ook dat de wetenschappelijke aanpak gewoon te traag is om op de snelle maatschappelijke ontwikkelingen te kunnen reageren.

Men kan dus zeggen dat aan de ene kant de samenleving zo ingewikkeld wordt dat beslissingen, bijvoorbeeld over wat te eten en wat niet, aan de wetenschappelijke experts moeten worden overlaten, terwijl er aan de andere kant stemmen opgaan dat deze wetenschappelijke experts de zaak juist uit handen moeten geven, juist omdat het zo ingewikkeld is geworden en dus niet meer in een laboratorium of andere speelgoeddoos past.

Of de post-normale benadering een goede oplossing kan bieden voor ingewikkelde problemen is zeer de vraag. Volgens mij moeten we gewoon doorroeien met de wetenschappelijke riemen die we hebben. Misschien is post-normale wetenschap een uiting van de aloude botsing tussen twee culturen, de natuurwetenschappen en de menswetenschappen. Verder is de mathematische visie dat modelvorming op zich nooit een beperkende factor kan zijn, zo gek nog niet. Bijvoorbeeld, de veralgemenisering van het Newtoniaanse model, en daarmee van het determinisme, heeft een brug gelegd tussen natuur en cultuur: opeens is het idee van vrije wil weer mogelijk.

Ik heb het over modelvorming en structuur gehad. Statistiek houdt zich bezig met fundamentele vragen betreffende deze zaken, en is er niet wars van 'af te dalen' tot het aardse niveau. Misschien is dat laatste de reden dat statistici het onderwerp zijn van veel, meestal flauwe, grappen, getuige bijvoorbeeld de webpagina [22]. De volgende vond ik in het Informatisch Mathematisch Physisch Astronomisch Communicatie Tijdschrift van mei 2000 [9]: "A statistician is a person who draws a mathematically precise line from an unwarrented assumption to a foregone conclusion." Of, enigszins vrij vertaald, "Een statisticus is iemand die een wiskundig perfekte lijn trekt van een wankele veronderstelling naar een vérgaande conclusie." Inderdaad gebruiken statistici, en andere wetenschappers, vaak speelgoedmodellen die weinig realiteitsgehalte hebben. Model en realiteit moeten zeker niet met elkaar verward worden! Het lijkt erop dat er een misverstand is over deze kwestie. Sterrenkundigen stellen dat de zon een bol is en natuurkundigen dat er geen wrijving is, economen gaan uit van evenwichtssituaties en soms zelfs van rationeel gedrag van de mens! Dit zijn alle modelveronderstellingen, maar geen veronderstellingen over de werkelijkheid! Ik merk dat studenten hevig in opstand komen als ik op een college een aanname maak 'voor het wiskundig gemak'! Zelf vond ik het vroeger ook onbegrijpelijk hoe de docent het wist dat het verband tussen $x$ en $y$ lineair is, en hoe het moge-

lijk is dat concrete variabelen aan wiskundige wetten gehoorzamen. Het zou mooi zijn als er op het overvolle curriculum van school en universiteit nog plaats was om aandacht te besteden aan wat een model nu eigenlijk is, en hoe het zich verhoudt tot de werkelijkheid.

**Jongen of meisje**

Misschien bent u zolangzamerhand nieuwsgierig geworden naar de inhoudelijke kant van kansrekening en statistiek, en naar wat ze met elkaar te maken hebben.

De grondleggers van de kansrekening zijn Fermat en Pascal, die de fundamentele principes ontwikkelden in een briefwisseling. Op het eerste gezicht lijkt het begrip onzekerheid datgene te zijn wat zich per definitie niet laat onderwerpen aan wetten. Fermat en Pascal presteerden het toch om onzekerheid onder te brengen in een wiskundig systeem.

Het soort problemen die in de 17-de eeuw onder de loep werden genomen kunnen ook nu nog, in onze tijd, menigeen volkomen in verwarring brengen. Ik zal u met een voorbeeld plagen [17].

Laten we er van uit gaan dat bij een geboorte de kans op een meisje gelijk is aan de kans op een jongetje, dus gelijk aan $1/2$. U belt aan bij een gezin met twee kinderen, en een meisje doet open. Wat is de kans dat het andere kind ook een meisje is? Antwoord: $1/2$.

Andere situatie: van een gezin van twee kinderen is gegeven dat een van de kinderen een meisje is. Wat is de kans dat het andere kind ook een meisje is? Antwoord: $1/3$.

Stel nu u belt aan dat laatstgenoemde gezin, dat wil zeggen u weet van te voren dat een van de kinderen een meisje is. Een meisje doet open. Wat is de kans dat het andere kind ook een meisje is? Antwoord: $1/2$.

De ervaring leert dat de bovenstaande antwoorden vaak als nogal verrassend worden gezien. Omgaan met informatie om daarmee kansen in te schatten (voorspellingen te doen) lijkt de mens niet aangeboren te zijn. Misschien moeten er nog wat generaties over heen gaan voordat kansen net zo algemeen geaccepteerd zijn en begrepen worden, als $1 + 1 = 2$. Ik bedenk hierbij dat de Babyloniërs al min of meer het huidige systeem voor de notatie van getallen gebruikten, maar dat het '='-teken toch nog zo'n 3000 jaar op zich liet wachten [12]. Het kan dus best wel een tijdje duren voordat het muntje valt.

Om u een idee te geven hoe men van kansrekening in de statistiek geraakt, ga ik wat verder met historisch materiaal. Bij wat men kan noemen de eerste statistische analyses, ging het ook om de kans op meisjes of jongetjes. Ik veronderstelde in bovenstaande vragen dat een nieuwe wereldburger met kans $1/2$ een meisje is, en met kans $1/2$ een jongetje. Is dat nu wel zo? Het schatten van kansen op grond van gegevens is een van de onderwerpen binnen de statistiek. Eind 16-de eeuw waren er in Engeland nogal wat pestepidemiën en men besloot gegevens te gaan bijhouden over de toestand van de bevolking. Dit werden de 'Tables of Mortality' genoemd (zie [23]). John Graunt heeft begin 17de eeuw deze tabellen nader bekeken, en er allerlei statistische informatie uitgehaald. Hij kwam bijvoorbeeld tot de ontdekking dat er meer jongens dan meisjes geboren werden (ongeveer 13 jongens op 12 meisjes). De grootte van de dataset (het ging om gegevens van talloze jaren), deed Graunt concluderen dat het een statistisch significant verschil was, dat wil zeggen dat het verschil significant van toeval afweek. Er werd zelfs een overschrijdingskans uitgerekend, toen al. Dat is in dit geval de kans dat ieder jaar meer jongens dan meisjes worden geboren, als de kans op een meisje of jongetje gelijk aan $1/2$ zou zijn. Deze overschrijdingskans bleek $(1/2)^{82}$ (dat wil zeggen $2.068 \times 10^{-25}$) te zijn, vreselijk klein dus. Als een overschrijdingskans erg klein is mag je daar een conclusie aan verbinden, is een van de gouden regels van de statistiek. De conlusie van Graunt was dat polygamie niet God's wil kan zijn.

De overschrijdingskans, ook wel $p$-waarde genoemd, wordt gebruikt om een hypothese te toetsen. Meestal is de hypothese dat een geobserveerd verschijnsel toeval is. Laten we nog een voorbeeld bekijken. Stel we vinden dat in Parijs 70 procent van de geboortes een jongetje betreft. Zou dit dan aan het toeval te wijten kunnen zijn? De kans dat er door het toeval 70 procent jongetjes worden geboren is zo klein dat de hypothese van toevalligheid zeker kan worden verworpen. Het is echter verbazend te lezen hoeveel controverses statistische toetsen kunnen oproepen. Meehl [16] noemt het "a potent but sterile intellectual rake who leaves in its merry path a long train of ravished maidens but no viable scientific offspring".

Ryu Tajiri

**Toeval herkennen**

Wat is nu toeval? Bestaat toeval eigenlijk wel? Niet volgens David Hume [8], die zegt: "Men neemt algemeen aan dat er niets bestaat zonder een oorzaak voor zijn bestaan, en dat het toeval bij nauwkeurig onderzoek een zuiver negatief woord is en niet op een werkelijke kracht duidt, die ergens in de natuur voorkomt." Ondertussen zijn de inzichten wel wat veranderd. (Hoewel: nog steeds komt men van de middelbare school met een deterministisch wereldbeeld.)

We kunnen het begrip 'toeval', ofwel 'randomness' heel goed in een formeel wiskundig systeem vatten uitgaande van axioma's. Ik zal dat hier niet doen. Wel geef ik de volgende definitie, afkomstig uit de complexiteitstheorie. We bekijken getallenrijtjes. Sommige hebben minder structuur dan andere, zijn in die zin complexer. Bekijk bijvoorbeeld de rij $1, 2, 3, 4, 5, \ldots$ Deze is erg simpel, het volgende getal volgt uit het vorige door er 1 bij op te tellen. Hoe zit het met de rij $1, 2, 3, 5, 8, \ldots$? Na enig puzzelen herkennen we hier de Fibonacci getallen: het volgende getal volgt uit de twee vorige door ze bij elkaar op te tellen. Dan nu de rij $4, 3, 8, 5, 1, \ldots$ In deze rij lijkt weinig structuur te zitten. We kunnen nu de complexiteit van een rij getallen definiëren als de lengte van het kortste computerprogramma dat de rij getallen genereert. Een rij getallen van lengte $N$ is toevallig als de complexiteit van de rij gelijk is aan de lengte $N$ [1]. U ziet, het toeval vangen in een formele definitie is eigenlijk heel eenvoudig! De definitie wijkt trouwens nogal af van wat Fermat en Pascal voor ogen hadden, al was het alleen maar omdat de computer niet in hun gedachtenexperimenten kon figureren.

De zogenaamde toevalsgetallen die een computer genereert, en die bijvoorbeeld bij simulatiestudies worden gebruikt, zijn in ieder geval niet toevallig en verre van complex. Deze pseudo-toevalsgetallen worden gefabriceerd volgens een eenvoudig iteratieschema. In het geval van de multiplicatieve congruentiële random number generator gaat het om het volgende schema: neem twee constanten, bijvoorbeeld $a = 630360016$ en $m = 2^{31} - 1 = 2147483647$. Het volgende getal wordt uit het vorige verkregen door te vermenigvuldigen met $a$. Mocht dit groter dan $m$ uitpakken, trek er dan voldoend vaak $m$ vanaf. Deze pseudo-toevalsgetallen zijn dus verre van toevallig! We lopen hier aan tegen het verschil tussen toeval en bepaalde vormen van chaos: met een heel eenvoudige wet kan men een enorme chaos creëren. Een simpele wiskundige formule kan heel chaotisch gedrag genereren, maar die chaos heeft dus een lage complexiteit. Er treedt dan ook wel eens begrippenverwarring op: chaos wordt ook wel gedefinieerd als maximale entropie, ofwel de afwezigheid van structuur (zie bijvoorbeeld [18]).

Bovenstaande definitie, van de complexiteit van een rij getallen, is nauw gerelateerd aan het idee van datacompressie: gegevens zonder verlies van informatie opslaan in een samengevatte vorm. Bij een veelheid van verschijnselen proberen we de structuur, ofwel de orde in het systeem te ontdekken. De statisticus, in, bijvoorbeeld, zijn/haar pogingen signaal en ruis te scheiden, is niet anders bezig.

De ingewikkeldheid van een signaal/ruisprobleem moet men niet onderschatten! Natuurlijk is het niet de bedoeling om patronen te zien, daar waar ze niet zijn. In een inktvlek bijvoorbeeld, wordt altijd wel iets gezien, maar dat lijkt meer met de eigen verbeelding te maken te hebben dan met de vlek. Vroeger gebruikte men zelfs de Rorschach Inktvlek Test om iets te weten te komen over iemands onbewuste conflicten en motieven (vanuit een psychoanalytische invalshoek). De menselijke geest lijkt ervoor gemaakt te zijn om structuren en patronen te herkennen. Dat is de reden dat wij van muziek kunnen genieten, betekenis kunnen geven aan het algemene begrip 'stoel', en überhaupt kunnen overleven. De wetenschap is nu druk doende deze bekwaamheid ook computers aan te leren, en zelfs hierin beter te laten worden dan de mens. Denk bijvoorbeeld aan een computer die handschriften kan ontcijferen, of gesproken tekst correct kan omzetten in geschreven tekst. Automatische patroonherkenning (dat wil zeggen zonder gebruik te maken van het menselijk 'oog') is een belangrijk statistisch onderwerp.

**Herhaling**

In mijn betoog tot nu toe heb ik geprobeerd u mee te voeren langs een veelheid van statistische paden, en u opmerkzaam te maken op allerlei vergezichten en onverwachte doorkijkjes. Laat ik de hoofdwegen nog eens aangeven. Ik ga doorvoor wat terug in de tijd. Sir Ronald A. Fisher noemt in zijn boek *Statistical Methods for Research Workers* [6], drie onderwerpen van studie: *(i)* the study of populations, *(ii)* the study of variation en *(iii)* the study of methods of the reduction of data. Alledrie de onderwerpen zijn in mijn verhaal aan de orde gekomen.

Onderwerp (i) is de studie van het algemene, van de eigenschappen van het geheel, zoals de kinetische theorie van gassen, de theorie van natuurlijke selectie, en algemene theoriën voor populaties van individuen in bijvoorbeeld sociologische studies. Het principe van herhaalde experimenten speelt hier een belangrijke rol. Ik noemde dit principe al eerder, het wordt binnen de statistiek als bijna vanzelfsprekend aanvaard. Het is zelfs zo dat het nauwelijks expliciet wordt genoemd bij de theoretische afleidingen. Een belangrijke uitzondering is Le Cam, die zich in zijn artikelen om het herhalingsidee druk maakt, en zich excuseert dat hij het als benadering gebruikt, omdat het niet te operationaliseren is. Le Cam zegt over zichzelf: "[...] the author has followed the standard, though treacherous, practice of pretending that the problem considered is one of a sequence of analogous problems" [15]. Het herhalingsidee zit trouwens op sommige punten te krap in het vel. In praktijk is het nu eenmaal niet altijd mogelijk een experiment een aantal keren te herhalen. Niet alleen de geschiedenis laat weinig herhaling zien. De geschiedkundigen ontwikkelen ondertussen een eigen methodologie [19], er van uitgaande dat statistische methoden voor geschiedkundige gegevens niet geschikt zijn. Wat niet waar is, en wat de statistici niet over hun kant mogen laten gaan! Een ander voorbeeld: ook de enorme datasets met genexpressieniveau's bevatten weinig herhaalde experimenten. Ze gaan over enkele individuen (zeg 40) en enorme hoeveelheden variabelen(zeg 40 000 of meer). Vooralsnog is er geen bevredigende statistische methode voor dergelijke 'gekantelde' datamatrices. Natuurlijk blijft het constante en regelmatige, ofwel herhaling in ruime zin, een belangrijk element in onze zoektocht naar structuur.

Het tweede onderwerp, de studie van variaties, noemt Sir Ronald Fisher onder andere om de tegenstelling te benadrukken tussen "[...] the aims of modern statisticians and those of their predecessors." Ik citeer ook nog het vervolg: "For until comparatively recent times, the vast majority of workers in this field appear to have had no other aim than to ascertain aggregate, or average, values. The variation itself was not an object of study, but was recognized rather as a troublesome circumstance which detracted from the value of the average." Als ik zoiets lees voel ik mij gesterkt, maar ook enigszins ontmoedigd. Ontmoedigd, omdat is gebleken dat het idee dat statistiek alleen maar over gemiddelden gaat zo moeilijk is uit te roeien! Toevallige variaties vallen niet altijd onder de noemer 'ruis'. Ik denk daarbij ook aan genetische algoritmes, waarbij blijkt dat men door toeval toe

te laten tot complexe en betekenisvolle structuren kan komen. Het zou daarom niet ondenkbaar zijn dat de mens door het *toeval* is geëvalueerd tot wat ie nu is.

Onderwerp (iii), 'the study of methods of the reduction of data' gaat over het samenvatten van een berg gegevens in enkele representatieve getallen, en is in feite ook bijzonder veelomvattend. Ook hier zijn de ingrediënten weer structuur, toeval en complexiteit.

**Mannen en vrouwen**

Dan wil ik nu kort ingaan op een statistisch gegeven: slechts zo'n 6 procent van de hoogleraren in Nederland is vrouw. De hypothese dat dit toevallig is kan worden verworpen op het 5 procents-niveau. Een $p$-waarde zal ik maar niet noemen. Wat is nu de oorzaak van dit verschijnsel? Als echte statisticus houd ik het bij de statistische uitspraak, en laat ik het antwoord op het waarom over aan de experts. Misschien heeft onze bekende filosoof [4] gelijk. Hij beschrijft namelijk zo'n vrouwelijke hoogleraar, bij één van haar colleges, als volgt: "De toestand van aanhoudend gelijk te hebben, die aan de positie van hoogleraar verbonden is, is een onvrouwelijke situatie en zij gaf te kennen zich hiervan bewust te zijn door ons niet aan te kijken. Door die gêne werd zij weer vrouw."

Dekker [5] heeft een wetenschappelijk onderzoek gewijd aan de oorzaken van het kleine percentage vrouwen onder wetenschappers. Om vrouwen aan te trekken zou er een cultuuromslag nodig zijn in de universitaire wereld. Ik vind dat een alleszins redelijke gedachte, maar blijf zitten met de vraag waarom *mannen* een dergelijke omslag *niet* nodig schijnen te hebben. Er zit trouwens wel beweging in: het vrouwennetwerk in onze universiteit heeft onlangs haar taak beëindigd en zichzelf opgeheven, en het afgelopen jaar zijn vijf vrouwen benoemd als lid van de Koninklijke Nederlandse Academie van Wetenschappen.

Helaas, op het gebied van de wiskunde zijn de vrouwelijke wetenschappers nog steeds sterk in de minderheid. Ik was dan ook zeer verrast toen ik, als onervaren moeder, het Nieuw Medisch Gezinsboek [20] raadpleegde en daar las: "Het wegen van het kind voor en na iedere borstvoeding maakt moeders die wiskunde gestudeerd hebben rustig: ze leggen er een statistiek van aan. Alle andere moeders moeten ervan afzien en hun kind eenmaal in de week wegen."

**Dankwoord**

Aan het slot wil ik graag de mensen bedanken die bij mijn benoeming betrokken zijn geweest, en ieder die mij gesteund heeft, of mij gewezen heeft op mooie stukken in de wondere wereld van wiskunde en werkelijkheid.

Mijnheer de Rector Magnificus, leden van het College van Bestuur, leden van het Bestuur van de Faculteit der Wiskunde en Natuurwetenschappen, ik dank u voor het door deze benoeming in mij gestelde vertrouwen. Het is een bijzondere eer om hier in Leiden als opvolger van Prof. van Zwet aan te mogen treden. Het is een groot genoegen om op het Mathematisch Instituut te werken aan de verdere ontwikkeling van de Mathematische Statistiek, de relatie met andere takken van Wiskunde verder te verstevigen, en samen te werken met Leidse wetenschappers binnen en buiten onze faculteit. Ik dank ook de leden van het Mathematisch Instituut voor hun bijdrage aan deze benoeming. Ook dank ik mijn andere collega's in Nederland, en in het buitenland, voor hun steun.

Hooggeleerde Van Zwet, beste Willem, Ik dank je voor de eye-opener die je me aanreikte al tijdens de studie, en voor alle daarop volgende eye-openers. Ik heb enorm veel van je geleerd. Er zijn echter zaken waar jij een meester in bent en die voor mij altijd een beetje onwennig zullen blijven. Ik hoop dat ik ook in de toekomst bij jou te rade kan blijven gaan.

Hooggeleerde Gill, beste Richard, Ik ben je erg dankbaar. Jij zei op een dag, toen ik al een aantal maanden wanhopig op zoek was naar een promotieonderwerp, zo tussen neus en lippen door, dat de theorie van Vapnik en Chervonenkis misschien te gebruiken was bij de consultatie waar ik op dat moment mee bezig was. Door deze opmerking kwam mijn onderzoek in een stroomversnelling. Richard is zo ongeveer de meest aanstekelijke statisticus die men zich kan voorstellen. Iemand die hem kent moet wel door zijn enthousiasme worden meegesleurd. Richard, je hebt me steeds van die mogelijkheden aangereikt waardoor er af en toe iets bij mij van de grond kwam, en waardoor ik nu hier sta.

**Referenties**

1 Barrow, J.D. (1992). *Pi in the Sky: Counting, Thinking and Being*. Penguin Books, London.

2 Battus, H. (1983). *Rekenen op Taal*. Querido, Amsterdam.

3 Berlin, I. (1980). *Against the Current: Selected Writings*. Ed. H. Hardy, Viking Press, New York.

4 Bomans, G. (1977). *De wereld van Godfried Bomans: een keuze uit zijn beste werk*. Elsevier, Amsterdam/Brussel.

5 Dekker, R. (2000). *De wetenschappelijke mensch: Persooncultuurfit en Loopbanen van Vrouwelijke en Mannelijke Wetenschappers*. Universiteit Utrecht.

6 Fisher, Sir R.A. (1958). *Statistical Methods for Research Workers (13th edition)*. Oliver and Boyd, Edinburgh, London.

7 Funtowisc, S.O. en Ravetz, J.R. (1992). Three types of risk assessment and the emergence of post-normal science. In: *Social Science of Risk*, Eds. Krimsky en Golding, Greenwood Publishing Group, Chapter 11, 251–273.

8 Hume, D. (1748/1777). *An Enquiry concerning Human Understanding*, Cadell, London. Ned. Vertaling (1978): Het Menselijk Inzicht. Boom, Meppel.

9 *IMPACT* (2000), nr. 13, Universiteit Leiden.

10 Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme *I. Monatshäfte Für Mathematik und Physik*, 38, 173–198.

11 Kendall, M.G. en Stuart, A. (1958). *The Advanced Theory of Statistics (Volume I)*. Charles Griffin & Company Limited, London.

12 Kool, M.J.H. (1999). *Die Conste vanden Getale, Een Studie over Nederlandstalige Rekenboeken uit de Vijftiende en Zestiende Eeuw, met een Glossarium van Rekenkundige Termen*. Verloren BV, Hilversum.

13 Kortelever, W. (2000). *Collegezweet*. Mare nr. 20, Universiteit Leiden.

14 Krol, G. (1993). *Omhelzingen*. Querido, Amsterdam.

15 Le Cam, L. (1960). *Locally asymptotically normal families of distributions*. University of California Publications in Statistics 3, 37–98.

16 Meehl, P.E. (1967). *Theory testing in psycholgy and physics: A methodological paradox*. Phylosophy of Science 34, 103–115.

17 Tijms, H. (1999). *Spelen met kansen*. Epsilon Uitgaven, Utrecht.

18 Prigogine, I. en Stengers, I. (1985). Order out of Chaos. Ned. Vertaling (1990) *Orde uit Chaos: De Nieuwe Dialoog tussen de Mens en de Natuur*. Bert Bakker, Amsterdam.

19 Ragin, C.C. (1987). *The Comparative Method: Moving beyond Qualitative and Quantitative Strategies*. University of California Press.

20 Venzmer, G. (ed.) (1974/1975). *Das Neue Grosse Gesundheitsbuch*. Verlagsgruppe Bertelsmann GmbH/Bertelsmann Ratgeberverlag München, Gütersloh, Wien. (Ned. Uitgave 1976/1982, Zomer & Keuning Boeken B.V., Ede).

21 Verkijlen, A. (2000). Het massale streven naar een individuele levensstijl: 'We are all individuals'. *Filosofie Magazine 4*, 8–14.

22 www.ilstu.edu/~gcramsey/Gallery.html

23 www.fsw.leidenuniv.nl/www/w3_func/stathist/stathist.html