KANSRIJK

Dames en Heren, Meneer de Rector, Meneer de Dekaan, Vader en Moeder,

Zalig

Ongeveer vier en een half jaar geleden ben ik aan deze universiteit benoemd als hoogleraar in de stochastiek. De herhaalde herinnneringen aan mijn plicht een oratie te houden heb ik steeds beantwoord met twee argumenten: een gebrek aan tijd en een gebrek aan echt goede ideeen om zoiets aan te pakken. Uit het feit dat ik hier vandaag in deze mooie jas voor u sta mag u helaas niet afleiden dat beide argumenten inmiddels niet meer gelden. Te weinig tijd, om al die mooie plannen ten uitvoer te brengen, is eigen aan het hoogleraarschap. En om aan een groter publiek de fascinatie voor de wiskunde duidelijk te maken blijft een moeilijke opgave. Wiskunde heeft de reputatie moeilijk te begrijpen te zijn, en die reputatie is verdiend: wiskunde is heel moeilijk. Het wordt pas eenvoudig na een enorme inspanning om je een bepaald onderwerp eigen te maken. Dan valt alles keurig op zijn plaats en kun je die staat van zaligheid bereiken, waarvoor wiskunde ook een reputatie bezit. Helaas kan ik u vandaag, gegeven de geringe tijd die ter beschikking staat, nauwelijks deelgenoot maken van die zaligheid.

Math-envy

Vorig voorjaar woonde ik als lid van het Miller Institute aan de University of California een reeks interdisciplinaire lezingen bij. Een van de Miller postdocs vertoonde bij een lezing een dia van zichzelf in actie tijdens zijn onderzoek, staande in een enorme kuil, in grote laarzen en met een grote schep, op zoek naar een klein organisme waar ik inmiddels de naam weer van vergeten ben. De achtergrond van de dia toonde een landschap van een verbluffende schoonheid. Dit was, zei hij, om ons, het publiek, bio-envy, bio-jaloersheid, te geven, in plaats van de math-envy, wiskunde-jaloersheid, die veel biologen zouden bezitten. Ik weet niet of zoiets bestaat, math-envy. Vooral in Nederland lijken veel mensen er juist trots op te zijn niets van wiskunde te begrijpen. De wiskundige is zielig, niet zalig. In ieder geval kan ik u als wiskundige niet veel meer bieden dan het beeld achter mij, een glimlachende jongeman voor een schoolbord. Helaas kan ik u nauwelijks deelgenoot maken van de diepere reden van die glimlach.

Wat ik vanmiddag wel kan doen is een mogelijk misverstand uit de weg ruimen, volgens welke wiskunde alleen maar moeilijk is. Mijn eigen vakgebied, de stochastiek, de leer van toevalsverschijnselen, kent vele toepassingen. De ware uitdaging voor een kansrekenaar of statisticus is een staat van zaligheid te bereiken en tegelijkertijd wiskundige modellen te ontwikkelen voor heuse problemen. Mijn doel voor het komende half uur is u in een aantal voorbeelden te laten zien waar het toeval verstopt zit, en wat dat voor ons betekent, met name voor het onderzoek. Die voorbeelden zal ik kiezen uit projecten waarbij de stochastiek groep van deze universiteit betrokken is.

Toeval

De vraag of toeval bestaat werd overigens niet lang geleden door collega Meester hier gesteld, en bevestigend beantwoord. Hij gaf zelfs aan te denken dat er "oprecht toeval" bestaat, hoewel dat in zijn visie voor het belang van de kansrekening niet relevant is. Kansrekening heeft voor hem de functie een hanteerbare beschrijving te geven van systemen die vanwege een zeer grote complexiteit niet exact kunnen worden beschreven. Wat mij betreft bestaat toeval ook, al zou ik het niet "oprecht" willen noemen en beoog ik geen filosofische betekenis met de uitspraak: "toeval bestaat". Toeval zal eenvoudig weg het woord zijn waarmee ik tot u over een aantal verschijnselen wil spreken. Ik hoop dat u aan het eind van mijn lezing volledig met mij eens zult zijn dat toeval bestaat. Zalig toeval bestaat in ieder geval in het hoofd van de wiskundige.

PET

Achter mij ziet u een afbeelding van een Positron Emissie Tomograaf, een uitermate dure machine, waarvan deze universiteit er thans twee in huis heeft, onder beheer van mijn collega Lammertsma van de Faculteit Geneeskunde. Een PET machine is bedoeld om de binnenkant van, bijvoorbeeld, de hersenen in kaart te brengen. Op het scherm ziet u zo'n hersenkaart, een dwarsdoorsnede door de hersenen. De werkwijze bestaat eruit een radioactieve stof in het bloed te spuiten. De radioactieve stof zal naar de hersenen geleid worden en zich enigzins ophopen op de plaatsen waar zich veel bloed bevindt, zeg maar in het gedeelte van de hersenen die het het drukst hebben of een afwijkend deel. Veel radioactief bloed betekent veel straling. De straling dringt door de schedel naar buiten en wordt door de

PET machine gemeten. Op het scherm ziet u een dwarsdoorsnede door de tomograaf met daarop de ring van detectoren in zwart aangegeven. Ik wil u niet te veel vermoeien met de manier waarop de gemeten straling kan worden terug vertaald in een ruimtelijke afbeelding van het inwendige van de hersenen. Een zo'n reconstructie heb ik u al getoond. Het is een twee-dimensionale doorsnede van de hersenen waarop de radiactiviteit door kleuren is zichtbaar gemaakt. Het is toeval dat dergelijke reconstructies mogelijk zijn, op grond van metingen buiten de hersenen, en dat wil ik u graag uitleggen.

De radioactieve straling ontstaat doordat twee elementaire deeltjes, een positron en een electron, elkaar vinden, en vernietigen, daarbij overgaande in twee gamma deeltjes, die volgens een fysische wet in precies tegenovergestelde richtingen worden uitgezonden. U ziet dit achter mij gebeuren. Het eerste stuk toeval betreft de lijn waarop de twee gamma deeltjes zich bewegen: deze wordt puur toevallig gekozen: iedere mogelijke richting heeft dezelfde kans.

De plaats waar de annihilatie plaatsvindt, bevindt zich in de hersenen en is van buiten niet te zien. De tomograaf kan echter wel de lijn meten waarop een annihilatie heeft plaatsgevonden. Wordt immers op hetzelfde moment zowel in A als in B een gamma deeltje gedetecteerd, dan is het duidelijk dat ergens op de lijn tussen A en B een annihilatie heeft plaatsgevonden.

Het tweede stuk toeval betreft de manier waarop dit proces zich in de tijd afspeelt. Het samenkomen van een positron en een electron en de uitzending van twee gamma deeltjes is geen unieke gebeurtenis, maar herhaalt zich in de tijd, met een intensiteit die afhangt van de hoeveelheid aanwezige radioactieve deeltjes. De tussentijden tussen deze annihilaties zijn niet vast bepaald maar vormen een stochastisch proces. Het is een van de mooiste voorbeelden van een standaard model uit de kansrekening, het Poisson proces. Dat de tussentijden stochastisch zijn, helpt niet voor de reconstructie, maar maakt deze ook niet moeilijker, omdat we heel goed kunnen rekenen met het Poisson proces.

Dat de richting van de lijn per toeval is gekozen helpt wel voor de reconstructie. Stelt u zich maar voor dat zowel bij het paar (A,B) en het paar (C,D) veel gamma deeltjes worden gemeten. Dan is dit een aanwijzing dat op het snijpunt van de lijnen (A,B) en (C,D) veel radioactiviteit aanwezig is. Het is mogelijk dit principe op een precieze manier te kwantificeren, en zodoende uit alle metingen aan paren detectoren plaatjes terug te rekenen zoals ik u eerder heb laten zien. Zouden de gamma deeltjes altijd in dezelfde richting wegvliegen, dan zou een reconstructie onmogelijk zijn, omdat een gemeten annihilatie op de lijn (A,B) op iedere plaats op de lijn kan hebben plaatsgevonden.

Naast de twee genoemde structurele bronnen van toeval is er ook een derde toevalsfactor, die bij bijna alle metingen en experimenten voorkomt, en waarvoor dan ook vele namen bestaan: ruis, meetfout, storing, attenuatie, etc. Deze komt erop neer dat toevallige oorzaken de metingen verstoren. Denkt u hier bijvoorbeeld aan een gamma deeltje waarvan de rechte baan die we zouden willen dat het aflegt, wordt gebroken. We meten dan de verkeerde lijn. Of denkt u aan een gamma deeltje dat wordt vertraagd, waardoor we het niet kunnen paren aan zijn partner. Of denkt u aan een storing in de tomograaf. Tot op zekere hoogte kan voor deze afwijkingen van het model worden gecorrigeerd, maar er blijft altijd een zekere "ruis" of "meetfout" over. We kunnen hier discusseren of het om "oprecht toeval" gaat, of om onbekende oorzaken. Voor de analyse van de data maakt dit weinig verschil, en ik heb er daarom geen moeite mee over toeval te spreken.

Schoolonderzoeken

Het woord "toeval" is overigens niet altijd gelukkig. Zo analyseerde ik, als student psychologie lang geleden, cijfers voor schoolonderzoeken. Het idee was dat sommige leraren duidelijk hoger cijferden dan anderen en we wilden dit kwantificeren. Een moeilijke situatie om goed te modelleren, en het uiteindelijk model bevatte dan dan ook een grote "toevalsfout". Mijn begeleider wees erop dat de leraren waarschijnlijk weinig gelukkig zouden zijn met een beschrijving van hun gecijfer waarin de woorden "fout" of "toeval" prominent aanwezig waren. Dergelijke taal is echter goed gebruik in de statistiek. Mijn eigen ervaring met gecijfer suggereert overigens wel degelijk een zekere toevalscomponent, waarmee ik niet bedoel dat we er ons best niet op doen.

Beurskoersen

Op het scherm ziet u een grafiek van de beurswaarde van Hewlett Packard (hp), een financiele tijdreeks. Op de horizontale as bevindt zich de tijd en op de verticale as de prijs van een aandeel hp op de verschillende tijdstippen. Het betreft de periode van 1984 tot 1992. De grafiek ziet er nogal wild uit, u zou kunnen zeggen: tamelijk toevallig.

Het is goed gebruik zo'n reeks niet direct te beschouwen, maar de zogenaamde *log returns*, waarvan de definitie op het scherm is te zien. De log returns zien er nog toevalliger uit: ze varieren rond een waarde net boven 0, met hier en daar ook een flinke uitschieter. De waarde net boven 0 correspondeert met een exponentiele groei.

Het is opmerkelijk dat juist sommige van mijn geleerde wiskundige collega's het toevalskarakter van tijdreeksen ter discussie hebben gesteld. Zij hebben aangetoond, dat veel van dergelijke reeksen vrij behoorlijk kunnen worden beschreven door een zogenaamd *dynamisch systeem* van de vorm $X_{t+1}=f(X_t)$. Dat wil zeggen dat op grond van de waarde X_t van de return op tijdstip t de waarde op het tijdstip X_{t+1} kan worden berekend middels de functie f. Zo'n functie kunt u opvatten als een computerprogramma dat een getal $f(X_{t+1})$ kan berekenen gegeven het getal X_t. Daar komt dus totaal geen toeval aan te pas. Zijn economische tijdreeksen wel toevallig? We kunnen een filosofische discussie aan die vraag wijden, of we kunnen hem ontwijken. Ontwijken is sneller. Het probleem met de beschrijving door een dynamisch systeem is dat, zelfs al zou een dergelijke beschrijving mogelijk zijn, dan kunnen we nog steeds de functie f niet weten. Een perfecte voorspelling van de beurskoersen in de toekomst is immers onmogelijk. Niet alle zaligheid, zoals hier de theorie van de dynamische systemen is direct bruikbaar.

Ook het gebruik van een zogenaamd chaotisch systeem helpt niet echt. Bij zo'n systeem heeft een minieme variatie aan het begin een grote afwijking op termijn tot gevolg. Dit is echter niet genoeg voor de beschrijving van een beurskoers: de beurskoers van morgen is immers al onzeker. Aanhangers van dynamische systemen lossen dat probleem op door bij het systeem weer een toevalsvariabele op te tellen. Tenzij we de tijdseenheid korter maken dan een dag, moet deze toevalsvariabele behoorlijk domineren.

Het voordeel van denken in termen van toeval is de mogelijkheid om onzekerheid te kwantificeren. Dit gaat ongeveer als volgt. Gegeven de waargenomen werkelijke tijdreeks vormen we een kansmodel dat kan verklaren hoe de tijdreeks tot stand is gekomen. Dit model kunt u beschouwen als een scenario voor het genereren van alternatieve tijdreeksen, die ook de werkelijk opgetreden tijdreeks hadden kunnen zijn. Op het scherm ziet u een kleine selectie van zulke scenarios, gebaseerd op een standaard model voor de hp-koers; de variatie is opmerkelijk. Uiteraard kunnen we deze scenarios ook doortrekken naar de toekomst, daarbij de geschiedenis tot nu inbouwend, en bezitten zodoende een inschatting van hoe die toekomst eruit zou kunnen zien. Let wel, dit levert geen zekerheid, maar een oneindig aantal mogelijkheden, ieder met een bepaalde waarschijnlijkheid. Als voorbeeld laat ik u zien een histogram van 1000 mogelijke waarden, "voorspellingen", voor de hp-koers de dag volgend op de tijdreeks die ik u heb getoond (dus in 1992). De hoogte van de staven geeft aan hoeveel van de 1000 waarden zijn gevallen in het betreffende interval en drukken zo een kansverdeling uit. Omdat het hier gaat om het verleden, is de werkelijke waarde al bekend. Die ligt in het midden van de verdeling, maar dat is toeval; het had ieder van de 1000 waarden kunnen zijn. Dat het hier om meer gaat dan "puur" toeval kan ik illustreren met de 1000 mogelijke waarden volgend op de beurskrach in 1987; die zijn aanmerkelijk meer gespreid. De onzekerheid over de toekomst was op dat moment groter.

Inzicht in wat mogelijk is stelt ons in staat redelijke grenzen voor het toekomstige gedrag aan te geven. Dergelijke grenzen en scenarios zijn van groot praktisch belang. Banken zijn bijvoorbeeld wettelijk verplicht voldoende dekking te geven voor hun uitstaande risico's, zoals leningen en beleggingen. Interessant is dat het wettelijk dekkingscriterium voor deze zogenaamde *Value-at-Risk*, gesteld is in termen van frekwenties: banken mogen eens in de zoveel dagen een bepaalde minimale dekking onderschrijden. Frekwenties zijn natuurlijk kansen, en kansen gaan over toeval.

Overigens is het buitengewoon ingewikkeld economische verschijnselen te modelleren. Een collega voerde eens het volgende experiment uit. Hij genereerde negen scenarios van een bepaalde tijdreeks, op soortgelijke wijze als ik zojuist heb laten zien, gebruikmakend van een model dat volgens alle handboeken een uitstekend model was voor het verschijnsel onder studie. De negen gesimuleerde tijdreeksen en de echte gemeten tijdreeks had hij in willekeurige volgorde en in hetzelfde format op papier uitgeprint, en vervolgens aan zijn secretaresse voorgelegd. U raadt het al: de secretaresse, die de situatie totaal niet had bestudeerd, was feilloos in staat om de enige werkelijke tijdreeks uit de tien te kiezen. Dit was tien jaar geleden. Er zijn snelle ontwikkelingen geweest en de handboeken, met een groter arsenaal aan mogelijke modellen, zijn aangepast. Ik denk dat het u niet mee zal vallen de werkelijke tijdreeks uit de tien mogelijkheden te kiezen. De anecdote illustreert echter dat op het gebied van tijdreeksanalyse nog veel te doen is, en dat het wel nooit helemaal goed zal komen. Laat ik u de log returns van de tien tijdreeksen zien, dan kunt u de werkelijke er waarschijnlijk wel uitlichten. Toeval is gecompliceerd.

Financial engineering

Voordat ik het terrein van de economie verlaat, wil ik enkele opmerkingen te maken over een ander terrein waaraan de stochastiek veel heeft toe te voegen: de *financial engineering*. Zoals u weet is beleggen in de mode. Niet alleen voor particulieren, maar vooral ook voor banken, verzekeringsmaatschappijen, pensioenfondsen en andere instellingen die veel geld beheren. Aan beleggen zijn risico's verbonden en financiele instellingen proberen die zo goed mogelijk te beheersen, en tegelijkertijd een zo hoog mogelijke winst te behalen. Zogenaamde *opties* werden voor het eerst verhandeld in Chicago in 1973. Daarna zijn tal van andere slimme vondsten om risico's af te dekken, of te delen, of te speculeren, bijgekomen, met mooie namen als *swaps*, of *baskets*. Ook in 1973 verscheen een zeer invloedrijk artikel van twee econometristen, Black en Scholes, waarin werd betoogd hoe de juiste prijs van opties kan worden bepaald met behulp van het economisch principe van "*no-arbitrage*", en de aanname dat een beurskoers zich gedraagt volgens een toevalsproces, een Brownse beweging, een standaard model voor de toevallige beweging van een molecuul in een vloeistof. Black and Scholes hadden moeite hun artikel geaccepteerd te krijgen, maar daarna zeer veel succes, leidend tot een Nobelprijs voor Scholes drie jaar geleden (Black was toen al overleden).

Voor een wiskundige zijn veel economische artikelen moeilijk leesbaar: onduidelijke notatie, intuitieve beweringen zonder bewijsvoering, en wiskundige afleidingen die wij eerste jaars studenten proberen af te leren. Ver verwijderd van de zaligheid waarover ik eerder sprak. Het artikel van Black en Scholes valt in deze categorie. Het is echter de basis geworden voor wat sommigen wel een nieuwe tak van wiskunde noemen: financiele wiskunde. Wat niet zalig is, kan het blijkbaar wel worden. De relatie van Black en Scholes "no-arbitrage" principe met de stochastiek, het toeval, werd in het begin van de jaren 1980 expliciet gemaakt door Harrison en Pliska. Een financiele markt waarin geen arbitrage bestaat, de basisaanname van Black en Scholes, bleek te kunnen worden gemodelleerd in een wiskundig model waarin een zogenaamde "martingaalmaat" bestaat. Nu is martingaal theorie een van de centrale gebieden van de moderne kansrekening. Moeilijk voor de meeste economen en daarom bij uitstek een mooi gebied voor wiskundigen met interesse voor economie.

Overigens legt het “no-arbitrage” principe het toeval aan banden. Het is gebaseerd op het bestaan van beleggingstrategieen die zekere uitkomsten garanderen, ondanks het toevallige gedrag van beurskoersen. Niet altijd winst natuurlijk.

Missing data

Ontbrekende data waren vroeger de schrik van iedere statisticus. Hoe kun je conclusies trekken uit een gegevensbank waarin zich vele witte plekken bevinden? Voor de wiskundig statisticus is dit juist een aantrekkelijke uitdaging. Meestal verkeerd is om de witte plekken op een zogenaamd verstandige manier op te vullen, zelfs als je die praktijk met een mooi woord als *imputatie* aanduidt. Dit gebeurt nog wel, vooral omdat veel data-onderzoekers zich gebonden voelen aan standaard methoden en software, en die loopt vast op die witte plekken. Juist in de huidige tijd waarin berekeningen op snelle computers kunnen plaatsvinden en derhalve niet eenvoudig behoeven te zijn, is dat meestal niet nodig.

De computer heeft het vakgebied van de statistiek, ook de wiskundige kant, zeer veranderd. Als commentaar op een onderzoeksvoorstel dat ik een jaar of zes geleden schreef , en waarin ik nieuwe methoden voorstelde, merkte een referent op dat ik misschien wel wat al te kritisch was over de "klassieke statistiek". Het was nu eenmaal zo dat men in de jaren 1960 nog integralen berekende door de grafiek van de functie te tekenen, uit te knippen en het gewicht van het verkregen papier te bepalen. Nu had ik in het geheel niet bedoeld kritisch te zijn over "klassieke statistici", in het minst nog over deze referent, Lucien le Cam, vorig jaar overleden, wiens werk ik zeer bewonder. Wel zijn door de beschikbaarheid van computers andere wiskundige vraagstellingen interessant geworden, terwijl sommige klassieke vragen wat naar de achtergrond zijn gedrongen. Klassiek betekent hier overigens: van de jaren 1940-1970; statistiek is nog een relatief jong vak. De basisprincipes van de statistiek hebben echter niets van hun waarde verloren, zoals het principe van likelihood functies, of Bayesiaanse methoden, die juist door het uitvinden van nieuwe, computer-intensieve algoritmen voor hun berekening, rond 1990, aan populariteit hebben gewonnen.

Demografie

Als een extreem voorbeeld van de analyse van een situatie met missende data wil ik u presenteren het schatten van historische levensduren, het promotie onderwerp van Marianne Jonker, die enkele maanden geleden gepromoveerd is. Op het scherm ziet u een kopie uit het register van de parochie van Eyam, een dorpje in Engeland, uit het jaar 1667. De pastoor maakte hierin van de belangrijkste gebeurtenissen in de kerk melding, in chronologische volgorde: geboortes, sterftes, huwelijken, etc. Deze parochie registers zijn de enige bronnen van informatie over levensduren van mensen in deze eeuwen. Een registratie bij de burgelijke stand, zoals wij in Nederland hebben, bestond niet, en al helemaal niet een centraal statistisch bureau dat demografische informatie systematisch verzamelde.

De levensduren van mensen over de eeuwen is nochthans een interessant gegeven voor historici. Om deze te onderzoeken heeft men in Engeland van een aantal parochies die kleine, handgeschreven boeken doorgewerkt om per persoon geboren in die parochies de relevante data te verzamelen. Men noemt dit famile-reconstitutie. Ik heb mij laten vertellen dat dit monnikenwerk hoogopgeleide personen vereist, onder andere omdat de registers in het latijn zijn gesteld. Een demografisch verslag is te vinden in het getoonde boek.

Een heel klein gedeelte van het resultaat van dit werk ziet u op het scherm. Het gaat hier om een groep vrouwen in het dorp Alchester. Iedere groep van een of twee regels betreft een vrouw. Het eerste getal is een code voor het dorp en persoon. Het tweede getal is de geboortedatum van de vrouw, gemeten in dagen vanaf dag nul van onze jaartelling. Vervolgens de datum van huwelijk (alle vrouwen in deze selectie zijn getrouwd), en een aantal andere data, steeds gemerkt met een code. Code 2 betekent geboorte van een kind, code 3 sterfte van een kind, code 4 sterfte van de man, code 9 sterfte van de vrouw. De sterftedatum vind u steeds aan het eind van de regels, maar zoals u ziet niet altijd. Van ongeveer de helft van alle mensen in de data-bank kon de datum van sterfte niet worden achterhaald. Gedacht wordt dat dit vooral veroorzaakt is door het feit dat de persoon naar een parochie is verhuisd die niet in het onderzoek is opgenomen. Toeval. Dit keer werkt het ons tegen.

De historici zijn nu juist geinteresseerd in de totale levensduren, en die weten we dus maar voor de helft van de mensen. We kunnen nu een aantal dingen doen. Ten eerste zouden we de mensen waarvoor de sterftedatum onbekend is gewoon uit de database kunnen verwijderen. In totaal hebben we vele duizenden mensen, dus ook de helft is nog een mooi aantal. Dit is verkeerd. De mensen waarvan de sterftedatum ontbreekt, zijn namelijk niet te vergelijken met de mensen waarvoor dit niet zo is. Als het waar is dat ontbreken is veroorzaakt door verhuizen, dan zal immers voor mensen die langer hebben geleefd vaker de sterftedatum ontbreken, want die hebben dan meer tijd gehad om te verhuizen. Gewoon weglaten leidt dan tot een onderschatting van de levensduur. Een tweede mogelijkheid is de ontbrekende sterftedata in te vullen, op een verstandige manier natuurlijk. Dit is de oplossing die de historici tot nu toe zelf hebben gekozen. Een derde mogelijkheid is om het toevalsproces waardoor data ontbreken wiskundig te modelleren, en dit is natuurlijk onze geprefereerde methode. Een extra probleem bij het opstellen van een wiskundig model is dat we ook het tijdstip van verhuizing niet kennen. Verhuizing was immers geen kerkelijke gebeurtenis en werd niet in de parochie registers genoteerd. Op basis van een, naar ons idee, redelijk model, vond Marianne Jonker de schattingen die ik nu op het scherm laat zien. U ziet daarop bijvoorbeeld dat 50 % van de mensen een levensduur korter dan 40 had. Een kleine fractie mensen bereikte de leeftijd van 80. Zo'n 20 % van de kinderen overleed binnen 4 jaar.

Clinical trials

Dit voorbeeld heb ik ooit gebruikt als inleiding op een projectaanvraag in de wiskunde, met wat meer aandacht voor de zalige kant. Ik kreeg toen de raad dat het toch bepaald onverstandig was om te proberen mijn onderzoek te verkopen met toepassingen in de geschiedenis. Ouderwets en oubollig. Met statistiek kun je toch ook geld verdienen! Zelf vind ik het heel leuk een bijdrage te leveren aan een wetenschap waarvan de beoefenaren zeker geen math-envy zullen hebben.

Met soortgelijke modelleringsideeen wordt echter ook veel geld verdiend, bijvoorbeeld in de farmaceutische industrie. Voor het op de markt brengen van een nieuw medicijn moet dat medicijn uitgebreid worden getest. Bij dergelijke *clinical trials* wordt het toeval door de onderzoekers zelf geintroduceerd. Het voornaamste toeval betreft het splitsen van een groep proefpersonen in twee groepen, op een willekeurige, toevallige wijze. De ene groep mensen wordt behandeld met het medicijn; de andere groep krijgt geen behandeling. Is na de behandeling een verschil waarneembaar, dan schrijven we dat toe aan het medicijn. De twee groepen waren immers willekeurig. Het toeval uitgebuit.

Ook de vorming van de totale groep van proefpersonen zal, gedeeltelijk, door toeval tot stand komen. Dit is van belang voor de mogelijkheid het gevonden resultaat van toepassing te verklaren op een grotere groep mensen.

Semiparametrische statistiek

Dat clinical trials zeer veelvuldig voorkomen is mede de oorzaak dat een beroemd artikel van DR Cox in de Journal of the Royal Statistical Society van 1972 tot het meest geciteerde artikel in de wiskunde is geworden. Hij heeft daar een statistisch model geintroduceerd dat veelvuldig wordt gebruikt in de levensduur analyse. Dit is een zogenaamd semiparametrisch model en is een van de inspiraties voor het onderzoek in de *semiparametrische statistiek*, waaraan ik in de afgelopen tien jaar veel tijd heb besteed.

Het doel van die semiparametrische statistiek is het opstellen en hanteren van statistische modellen die zo min mogelijk a-priori veronderstellingen maken. Je kunt op verschillende manieren tegen deze ontwikkeling in de statistiek aankijken. Veel semiparametrische modellen zijn gemotiveerd door de statistische praktijk, met name in de medische statistiek en de econometrie. Routinematig gebruik van deze modellen is tegenwoordig mogelijk door betere rekenmethoden en grotere gegevensbanken. Anderzijds kun je ook zeggen dat de semiparametriek de voor de hand liggende stap in de ontwikkeling van de wiskundige kant van de statistiek is geweest, de weg naar verdere zaligheid. Dit laatste punt brengt mij weer terug bij de zaligheid waarover ik al eerder heb gesproken en waarover ik bij deze gelegenheid dus zal zwijgen.

Causaliteit

Het krantenknipsel op het scherm is afkomstig uit NRC Handelsblad van twee weken geleden. Het is een tamelijk willekeurige keuze, want de krant staat vol met boodschappen van dit type. Uit onderzoek is gebleken dat roken huidkanker veroorzaakt. Bij zo'n bericht wil ik altijd wel graag weten wie het onderzoek hebben uitgevoerd. Ik weet hoe moeilijk het is om goede conclusies uit statistische data te trekken en het beweerde effect wekt in eerste instantie verbazing. Roken veroorzaakt longkanker, ja, maar huidkanker? Het is niet mijn doel twijfel te zaaien over dit concrete onderzoek. Waar ik uw aandacht voor vraag is het feit dat, op grond van statistische data, een causaal verband wordt gelegd: roken veroorzaakt huidkanker. Dat het verband causaal bedoeld is, is misschien nog niet zo duidelijk uit de kop van het artikel, maar verder lezend lijkt die interpretatie onontkoombaar. Roken veroorzaakt huidkanker.

Causale uitspraken behoren tot het beste type wetenschappelijke uitspraken, maar roepen filosofische vragen op. Wat betekent het precies dat het een het ander veroorzaakt? Statistici wordt in hun opleiding ingeprent verschil te maken tussen fenomenen die samenhangen, en oorzaken en gevolgen. Zo is er een sterke samenhang tussen het aantal mensen in de zaal bij een oratie en het aantal bitterballen bij de receptie daarna. Echter, ook al was u vandaag met tien keer zoveel mensen hier aanwezig geweest, dan was het aantal bitterballen niet groter geweest; u bent niet de oorzaak van de bitterballen, want die moeten vooraf worden besteld. De samenhang tussen aantal aanwezigen en aantal bitterballen wordt hier veroorzaakt door de inschatting van de orator vooraf.

Het is goed gebruik de uitkomsten van een clinical trial zoals ik die al heb genoemd causaal te interpreteren: het medicijn veroorzaakt het effect (of niet). Nu zou een rechttoe-rechtaan clinical trial voor het onderzoeken van roken vereisen dat van een groep proefpersonen willekeurig de helft wordt aangewezen om stevig te gaan roken, of ze daar nu zin in hebben of niet. Dat gaat in de praktijk natuurlijk niet.

Toch zouden we graag causale conclusies trekken over roken. Het krantenknipsel meldt dat voor andere mogelijke oorzaken statistisch gecorrigeerd is. Tot op zekere hoogte is dat inderdaad mogelijk. Zo zal Judith Lok dit voorjaar promoveren op een proefschrift waarin beschreven wordt aan welke voorwaarden dan dient te worden voldaan en hoe zo'n "statistische correctie" kan plaatsvinden. Haar onderzoek is sterk gemotiveerd door ideeen van collega Robins aan Harvard University, waarbij onze bijdrage de wiskundige onderbouwing levert. Dezelfde martingalen die ik al noemde in verband met de financial engineering spelen ook hier een centrale rol, maar zijn te zalig om er vandaag over te praten.

Bio informatica

Vooruitziende geesten beweren dat we de eeuw van de levenswetenschappen zijn binnengetreden. Met enorme snelheid worden nieuwe gegevens verzameld over "leven", zoals de mechanismen van een cel of de opbouw en functie van ons erfelijk materiaal. Enkele maanden geleden is NWO een omvangrijk programma begonnen voor onderzoek in de bio-informatica, dat moeten leiden tot samenwerking tussen informatici, wiskundigen, biologen en andere wetenschappers om "leven" te ontrafelen. Het leven is erg toevallig en dus liggen hier ook voor stochastici kansen. Op dit moment zien we er zoveel dat we nauwelijks in staat zijn aan te geven in welke richting we, met onze beperkte menskracht, het beste kunnen bijdragen.

Voor de hand ligt een rol bij het onderzoeken van relaties tussen DNA en functie. Een bron voor toeval is hier gemakkelijk aan te wijzen en bekend van de biologie les: ons erfelijk materiaal bestaat uit paren chromosomen, 23 bij de mens, waarbij de ene helft van een paar afkomstig is van onze vader en de andere helft van onze moeder. Onze vaders en moeders hebben natuurlijk ook paren van chromosomen en van ieder paar geven ze de helft door. Een bekend model is dat die helft volgens toeval wordt gekozen; dit ligt ten grondslag aan de kanstafels achter de wetten van Mendel. De werkelijkheid is een stuk ingewikkelder. Chromosomen zijn lange ketens van nucleotidezuren, weergegeven door letters, en bij het doorgeven vinden recombinaties en mutaties plaats. Ook die processen zijn met kansen te modelleren. Het aantal letters waarmee we nu te maken hebben, in de orde van drie biljoen voor de mens, en de complexe manier waarop lettercombinaties tot uiting komen, verandert de vraagstellingen en de mogelijke oplossingen echter essentieel ten opzichte van de oude Mendeliaanse genetica.

DNA is pas het begin voor het begrijpen van leven. Onze divisie en faculteit zijn thans doende een onderzoeksgroep in de bio-informatica op te zetten, welke zich mogelijk zal richten op complexere, hogere structuren in levende cellen. Een voorbeeld van een bijdrage vanuit de stochastiek is het ontwikkelen van modellen voor de interactie van macro-moleculen in een cel. Op het scherm ziet u een eenvoudig model met vier typen moleculen. Alle moleculen bewegen zich, op een toevallige wijze, door de ruimte. De zwarte moleculen hebben geen interactie met de andere moleculen, buiten dan dat zij posities bezet houden die de andere moleculen niet kunnen innemen. De paarse en blauwe moleculen binden met een bepaalde kans als zij op eenzelfde plaats komen. Zo'n binding valt vervolgens weer uiteen volgens een toevalsproces met een bepaalde intensiteit. De vraagstelling hier, ons voorgelegd door collega Westerhoff, is hoe de concentratie van de zwarte moleculen de concentraties van de andere moleculen beinvloedt. Het getoonde model is te eenvoudig, maar een goed begin want wiskundig al moeilijk. Hier hebben we het toeval nog niet onder controle.

Ion kanalen

Tot de bio-informatica in ruime zin behoort ook het promotiewerk van Barry Schouten, uitgevoerd onder leiding van Mathisca de Gunst. Dit betreft het gedrag van ion-kanalen in celwanden. Ion-kanalen zijn ingewikkelde eiwitten, maar het is voor onze discussie gemakkelijker ze te beschouwen als deuren waardoor de cel met de buitenwereld communiceert. Zoals een goede deur betaamt kan een ionkanaal open of gesloten zijn, en zijn toestand is in de tijd te meten. Op het scherm ziet u het resultaat van zo'n meting. Op de horizontale as staat de tijd en op de verticale as ziet u in essentie twee niveaus, gesloten en open. Het zal u niet ontgaan dat de metingen ook een flinke mate van ruis vertonen. Dit bemoeilijkt de analyse aanzienlijk, want het soms niet goed te zien of het kanaal open of gesloten is, maar over deze bron van toeval wil ik niet spreken. De meer interessante bron van toeval zit in het patroon van openen en sluiten van de kanalen, de intervallen die ik nu aangeef. Deze vormen een toevalsproces. Waarom is niet duidelijk: de cel is gemeten in een stationaire toestand. Oprecht toeval, of geeft toeval gewoon een handige manier om wat we waarnemen te beschrijven?

Een voor de hand liggende aanname voor het openen/sluiten proces is *de Markov aanname*, naar de wiskundige Markov die in het begin van de vorige eeuw veel aan de theorie van de stochastische processen heeft bijgedragen. Deze aanname betekent ongeveer dat de cel bij haar beslissing om het kanaal te open of te sluiten niet in het verleden kijkt, maar alleen haar huidige toestand in de beslissing betrekt. Deze aanname, hoe eenvoudig ook, heeft vergaande consequenties. Zalige wiskunde bewijst bijvoorbeeld dat de tijdsintervallen tussen open en sluiten dan een zogenaamde exponentiele verdeling moeten bezitten. Welnu de metingen, onder andere verricht aan gerstcellen hier aan de VU, laten zien dat de tijdsintervallen deze verdeling niet bezitten. We kunnen nu ofwel de theorie van Markov terzijde schuiven, of een andere verklaring zoeken. Aangezien de Markov aanname biologisch plausibel is, kiezen we voor de tweede weg. Om precies te zijn nemen we aan dat de metingen niet de juiste intervallen opleveren, en wel omdat er meerdere typen van open en/of gesloten toestanden zijn. Denkt u bijvoorbeeld aan een doorgang bestaande uit twee deuren die allebei open moeten om de weg vrij te maken. Dit leidt tot een Markov model met niet twee maar met drie toestanden. In de metingen is echter geen verschil tussen gesloten of dubbel gesloten te zien.

Een ander geval van missende data, en ook hier zijn we in staat om op grond van wat we wel zien de werkelijkheid te reconstrueren, tenminste als de Markov aanname klopt. Mijn voorbeeld is daarbij te eenvoudig: het model dat door Schouten werd gevonden voor ion kanalen in gerstcellen heeft veel meer toestanden, zoals blijkt uit de figuur.

Spraak

De half-waarneembare Markov processen waarover ik zojuist sprak, worden ook gebruikt voor het modelleren van de letterreeksen van DNA, en voor het modelleren van spraak. Beide zijn voor ons relevant. Het tweede omdat we bij Eurandom, het Europees Instituut voor onderzoek op gebied van de stochastiek in Eindhoven, op het punt staan een spraakherkennings project te starten, in samenwerking met Philips.

Ivoren toren

Zo blijkt het probleem van ontbrekende data behalve in de levensduuranalyse ook in biologische experimenten op te treden. Verborgen Markov modellen zijn van belang voor zowel ion kanalen als spraak herkenners. Martingalen hebben zowel toepassingen in de economie als in het beschrijven van causaliteit in medische experimenten. Het basis model van Black en Scholes is een Markov model, net als het model voor het open en sluiten van ion-kanalen, en het model voor de interactie van macromoleculen in een cel. Radiactiviteit volgt een Poisson proces, een telproces, net als de processen die levensduren beschrijven. Het probleem van ontbrekende data ontmoeten we bij de analyse van PET gegevens, in de demografie, en bij de ionkanalen.

Zo hangen op het eerste gezicht zeer uiteenlopende verschijnselen sterk samen op het moment dat je naar de wiskundige modellering gaat kijken. Vanwege die samenhang is het zeer noodzakelijk wiskundigen de ruimte te geven zich nog regelmatig, en langdurig, want we denken geweldig langzaam, in hun ivoren toren terug te trekken. De toestand van zaligheid die we daar weten te bereiken heeft positieve gevolgen voor het begrijpen van de wereld om ons heen. Zaligheid brengt eenheid.

Profilering

Het college van bestuur van deze universiteit, in haar wijsheid, heeft in de afgelopen jaren gewerkt aan een omvattend plan voor de prioriteiten in het onderzoek. Een voorlopig resultaat ziet u schematisch weergegeven: de VU-ster, waarvan de verschillende punten de prioriteiten weergeven. Een vereenvoudigde en wiskundig zaliger ster is als volgt. Het eerste wat mij opvalt is dat wiskunde geen prioriteit is; zelfs een klein puntje voor de klassieke koning der wetenschappen kon er niet van af. Als u mij goed beluisterd heeft, dan valt echter snel op dat de onderwerpen die ik heb aangestipt uitstekend in een aantal van de punten vallen. Alleen voor cultuur en recht heb ik zo gauw niets weten te verzinnen. Prima in orde dus. De geldstroom van het college naar de wiskunde, de stochastiek in het bijzonder, zal na deze rede stellig flink aanzwellen. Zoals ik Rien Kaashoek, in zijn wijsheid en zijn handigheid, vaak heb horen beargumenteren: wij kunnen zo veel met maar een beetje geld. Behoudens een paar computers kosten we alleen ons salaris.

Toch ben ik geen liefhebber van dit soort prioritering. De reden dat de interesse en activiteiten van de stochastiek afdeling zo uitstekend binnen de VU-ster vallen, is eenvoudig dat dit gewoon de onderdelen van de wetenschap betreft die overal als prioriteit worden erkend. Aan de VU, in Leiden, in Utrecht, in Parijs, aan Harvard, of in Berkeley. Goede wetenschap is internationaal georienteerd en vernieuwd zich voortdurend, zowel onder invloed van interne ontwikkelingen als veranderde vragen uit andere wetenschappen of de samenleving. Niet alleen de VU gaat de eeuw van de levenswetenschappen binnen. Informatica is ook aan andere universiteiten prominent aanwezig. Die ster van het college van bestuur zegt mij weinig, behalve dat mijn hart krimpt omdat de wiskunde daarin ontbreekt.

Natuurlijk moet de universiteit keuzen maken voor het onderzoek, maar het is beter zijn als die keuzen op een zo laag mogelijk niveau plaats hebben. Uiteindelijk valt en staat het onderzoek met mensen. De kwaliteit van individuele wetenschappers, of kleine groepen, dient een groter gewicht te hebben, ten koste van de zogenaamde strategische keuzen. Goede mensen zorgen per definitie voor interessant onderzoek. De verkleining van het USF programma "Vernieuwing van het Wetenschappelijk Kader", waar ik zelf van heb mogen profiteren en waarvoor ik de universiteit heel dankbaar ben, ten faveure van strategische plannen, vind ik een mislukking. Misschien moet ik daarvoor echter eerder de minister, de vorige, met zijn top en breedte strategieen de schuld toeschuiven dan het college van bestuur. Een feit is dat de wiskunde slecht gedijt onder zulke strategieen. Wij zijn klein en kunnen wat betreft punten in een ster hoogstens aanhaken bij anderen, als bijwagens.

De eigenheid van de wiskunde gaat dan verloren. Ik zou het graag anders zien.

Dank

Tot slot spreek ik enige woorden van dank. Mijn collega's aan de VU, in het bijzonder bij wiskunde en informatica, dank ik zeer voor het goede werkklimaat. De afdeling stochastiek is onlangs versterkt met collega-hoogleraren Meester, Koole en Groeneboom; ik zie uit naar een hechte samenwerking. Met Geurt Jongbloed en Mathisca de Gunst heb ik de laatste jaren gezicht gegeven aan de statistiek aan de VU, een waar plezier, zowel in de wetenschappelijke als in de persoonlijke hoek.

Bij het afscheid van Willem van Zwet, vorig jaar, was ik bang dat er een groot gat zou vallen in statistisch Nederland, maar dat is niet gebeurt; Willem is er gewoon nog. Chris Klaassen heeft mij als promovendus begeleid in Leiden en ik ben hem daarna gevolgd naar Amsterdam, de andere universiteit; ik zie ernaar uit meer inhoud te geven aan samenwerking tussen de beide universiteiten, op ons terrein, binnen de gebruikelijke hartelijke verhoudingen. Net als voor iedereen is Richard Gill voor mij een grote bron van inspiratie; als hij dat al niet in de gaten heeft door de dwarse antwoorden die ik soms op zijn ideeen uit, dan verontschuldig ik mij daarvoor vandaag.

Kobus Oosterhoff heeft mij naar deze universiteit gehaald, hier ziet u het gebeuren, en mij hier vastgehouden, ondanks omzwervingen naar elders, en uiteindelijk heb ik hem mogen opvolgen. Ook hier speelt het toeval een rol; een gelukkig toeval. Voor het feit dat Kobus alle niet-toevalsredenen om hier weg te gaan systematisch heeft verwijderd, ben ik hem zeer erkentelijk. Evenals voor de bestuurlijke en wetenschappelijke raad die hij mij heeft gegeven. En natuurlijk voor het feit dat ook hij er nog steeds is; tot voor kort als voorzitter van onze divisie, op welke positie hij met visie vele saillante ontwikkelingen heeft gestart, en bovendien een enorm aantal administratieve karweien heeft geklaard. Het nadeel van zijn inzet is dat het gat wat hij heeft achtergelaten alleen met twee mensen te vullen is.

Ik heb gezegd.