STAtOR

periodiek van de VvS+OR jaargang 17, nummer 2, juli 2016

STAtOR
Jaargang 17, nummer 2, juli 2016

STAtOR is een uitgave van de Vereniging voor Statistiek en Operationele Research (VvS+OR). *STAtOR* wil leden, bedrijven en overige geïnteresseerden op de hoogte houden van ontwikkelingen en nieuws over toepassingen van statistiek en operationele research. Verschijnt 3 keer per jaar.

Redactie
Joaquim Gromicho (hoofdredacteur), Annelieke Baller, Ana Isabel Barros, Kristiaan Glorie, Johan van Leeuwaarden, Guus Luijben (eindredacteur), Richard Starmans, Gerrit Stemerdink (eindredacteur), Hilde Tobi en Vanessa Torres van Grinsven. Vaste medewerkers: Fred Steutel en Henk Tijms.

Kopij en reacties richten aan
Prof. dr. J.A.S. Gromicho (hoofdredacteur), Faculteit der Economische Wetenschappen en Bedrijfskunde, afdeling Econometrie, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, telefoon 020-5986010, mobiel 06-55886747, <j.a.dossantos.gromicho@vu.nl>.

Bestuur van de VvS+OR
Voorzitter: prof. dr. Fred van Eeuwijk <president@vvs-or.nl>
Secretaris: dr. Fetsje Bijma <bestuur@vvs-or.nl>
Penningmeester: dr. Ad Ridder <penningmeester@vvs-or.nl>
Overige bestuursleden: prof. dr. Eric Cator (SMS), prof. dr. Jeanine Houwing-Duistermaat (BMS), Maarten Kampert MSc., prof. dr. Albert Wagelmans (NGB), dr. Michel van de Velden (ECS), dr. Jelte Wicherts (SWS), Nynke Krol (Young Statisticians)

Leden- en abonnementenadministratie van de VvS+OR
VvS+OR, Postbus 244, 6700 AE Wageningen, telefoon 0317 - 419572, e-mail <admin@vvs-or.nl>.
Raadpleeg onze website over hoe u lid kunt worden van de VvS+OR of een abonnement kunt nemen op *STAtOR*.

VvS+OR-website
www.vvs-or.nl

Sociale media
Wilt u uw vakgenoten ontmoeten en wilt u discussiëren over actuele thema's, volg dan de VvS+OR en de Young Statisticians via LinkedIn, Facebook, Twitter en Flickr.
Sluit u aan bij de LinkedIn-groep van VvS+OR of Young Statisticians; bekijk foto's op <www.flickr.com/photos/vvs-or/sets>; like onze Facebook-pagina; volg de President van VvS+OR op <https://twitter.com/#!/dutchstat>.

Advertentieacquisitie
M. van Hootegem <hootegem@xs4all.nl>
STAtOR verschijnt in maart, juli en december.

Ontwerp en opmaak
Pharos, Nijmegen

Uitgever
© Vereniging voor Statistiek en Operationele Research
ISSN 1567-3383

INHOUD

Sport!

Als het goed is kunnen we in alle geledingen van de maatschappij toepassingen van ons vak tegenkomen. We hebben dat de afgelopen jaren al vele malen laten zien in interessante artikelen. Ons staat een bijzondere zomer te wachten, een waarin men met de beste wil ter wereld de sport niet kan ontlopen: EK voetbal, Wimbledon, Tour de France en Olympische Spelen. En dan hebben we de Giro en Roland Garros al achter de rug! Niets logischer dus dan dat *STAtOR* zich hieraan aanpast met een groot aantal artikelen over sportieve toepassingen.

Voor wie denkt dat deze aanpassing toeval is heeft Klaas Landsman de vraag gesteld of toeval wel bestaat. Gerard Kuper, Gerard Sierksma en Frits Spieksma laten daarna zien dat de selectieprocedure voor tennissers voor Rio allesbehalve op toeval berust, maar op degelijke data-analyse.

Niet alles in dit nummer is sport, Maarten Scholtes-Timmerman, Sarina Bijlsma en Jack Vogels beschrijven de zoektocht naar een methode waarbij diabetespatiënten niet meer dagelijks in hun vinger hoeven te prikken. Maar dan slaat de sport weer toe: Miriam Loois analyseert het verschil tussen binnen- en buitenbaan op de 500 meter schaatsen. Veel van de STAtOR-lezers zullen het werk van John Nash kennen en nog meer lezers zullen de film *A Beautiful Mind* kennen die over zijn leven is gemaakt. Harold Houba vraagt zich af of topsporters bij hun beslissingen gebruiken maken van zijn speltheoretische inzichten.

Ten slotte nog een niet aan sport gewijd artikel van Joost van Sambeeck, Mart Janssen, Peter Ligthart, Wim de Kort en Nico van Dijk. Zij vertellen over erfelijke eigenschappen bij bloedverwanten: hoe groot is de kans dat je zelf Rhesus-D-negatief bent als je broer of zus dat is.

Verder vindt u bijdragen van onze columnisten Henk Tijms over De Mooiste Formule en Fred Steutel over Ontmoetingen. Johan van Leeuwaarden heeft een aantal tv-colleges gegeven in de Universiteit van Nederland. In dit nummer een bijdrage van hem, gebaseerd op die colleges, over Einstein, Sociale Netwerken en waarom mensen net moleculen zijn. De Young Statisticians hebben veel nieuws te melden.

We besluiten met bijdragen van de scheidende én de nieuwe voorzitter van de VvS+OR. Jacqueline Meulman geeft op de valreep een kijk op de stand van zaken in ons vak en Fred van Eeuwijk vertelt over zijn visie waarmee hij de vereniging de komende vier jaar gaat leiden.

De redactie heeft onlangs afscheid genomen van Hilde Tobi. Jarenlang heeft zij met grote creativiteit ideeën aangedragen, we zullen haar inbreng missen. Heel veel dank voor alle prachtige artikelen die je binnenhaalde! Met ingang van het volgende nummer treedt Joep Burger toe tot de redactie. Hij werkt als methodoloog bij het CBS in Heerlen; we zochten al langere tijd naar een redactielid dat ons in contact kan brengen met het vele en veelzijdige onderzoek bij het CBS. Welkom Joep!

Wij wensen al onze lezers een warme zomer toe, met veel sport en veel sportief leesplezier in *STAtOR*.

De redactie

Het laatste nummer van dit jaar zal gewijd zijn aan **Smart Societies**. Steeds vaker worden beslissingen in de samenleving genomen op basis van verzamelde gegevens. En die gegevens komen in steeds grotere aantallen beschikbaar. Tijd voor *STAtOR* om daar aandacht aan te besteden. Als redactie hebben we al enkele ideeën en een aantal artikelen is al toegezegd. Daarnaast hopen we op veel aanmeldingen vanuit onze lezerskring, u kunt contact opnemen met een van de redacteuren. Wij vatten de term Smart Societies zéér breed op, alleen dan doen we recht aan de volle breedte van ons vak. Help ons om dit nummer tot een succes te maken! U kunt uw bijdragen en ideeën sturen naar Joaquim Gromicho <j.a.dossantos.gromicho@vu.nl>.

BESTAAT TOEVAL?*

Bedoelen we niet ‘onverwacht’ of ‘vooralsnog onverklaarbaar’ wanneer we ‘toevallig’ zeggen? Klaas Landsman, hoogleraar Mathematische Fysica aan de Radbouduniversiteit Nijmegen en een van de initiatiefnemers van de Week van het Toeval (16 – 20 mei 2016), buigt zich over kleine en grote vragen rondom toeval. Eerder dit jaar verscheen *The Challenge of Chance*, onder redactie van Klaas Landsman en Ellen van Wolde, met artikelen waarin vooraanstaande onderzoekers hun visie op het thema toeval geven.

Klaas Landsman

De Hoornse piraat Dirkie de Veenboer vergaarde vanaf 1607 een vermogen door in de Middellandse Zee niet-Hollandse handelsschepen aan te vallen, waarbij hij met name voor Spanjaarden geen genade kende. Zijn talloze successen en latere benoeming tot Grootadmiraal van het Ottomaanse Rijk suggereren dat zijn zeemanskunst en militair inzicht niet onderdeden voor dat van onze latere zeeheld Michiel de Ruyter, die bovendien op dezelfde wijze aan zijn einde kwam: een kanonskogel sloeg op 10 oktober 1620 een van Dirkies benen af, waarna hij aan bloedverlies overleed (en navenant De Ruyter 56 jaar later).

Waarom vertel ik dit? Een van de ontroerendste momenten in mijn leven vond plaats op 7 maart 2009, toen ik in het Westfries Museum te Hoorn samen met mijn vader (op dat moment 83) en mijn jongste zoon Felix (destijds zeven) als bewezen nazaten van Dirkie het eerste exemplaar van het boek *Piraten van de Lage Landen* van auteur Peter Smit in ontvangst mocht nemen. Maar wat als deze Dirkie al vóór de verwekking van zijn nageslacht zou zijn gedood? Alle kans!

Sterker nog: wat als de afstand tussen de aarde en de zon een tikje groter of kleiner was geweest, waardoor het op onze planeet te koud of te warm zou zijn geweest om leven sowieso een kans te geven? Wat als de zon er niet eens was geweest, bijvoorbeeld omdat de dichtheid van materie in het vroege heelal net iets groter was geweest, zodat het heelal na de oerknal al snel weer was ingestort, of juist een tikje kleiner, waardoor het te snel zou zijn uitgedijd om stervorming mogelijk te maken?

Gaat het hier om toeval? Hoeveel toeval kan een mens verdragen? Bestaat toeval überhaupt? Wat is toeval eigenlijk? Laten we met die vraag beginnen.

Een ontmoeting in Zwitserland

In het dagelijks leven betekent ‘toevallig’ meestal ‘onverwacht’: vrijwel alle voorbeelden die mensen desgevraagd te berde brengen dragen dit karakter, dikwijls in combinatie met het idee dat de onverwachte gebeurtenis bovendien een kleine kans had om op te treden (hetgeen de gebeurtenis nog onverwachter maakt). Een neef van mij (die eveneens van Dirkie de Veenboer afstamt) heeft bijvoorbeeld als slechts één van de twee medisch bekende personen in Nederland zijn hart rechts. De andere kwam hij toevallig tegen op een wintersportvakantie in Zwitserland.

In de (natuur)wetenschap betekent ‘toevallig’ echter iets anders dan onverwacht, namelijk: ‘zonder (aanwijsbare) oorzaak’. Een alledaags voorbeeld is de weerkaatsing van licht in een ruit: het grootste deel wordt doorgelaten, maar een aanzienlijke fractie wordt weerkaatst. Hoe kan dit? Deze vraag krijgt extra lading als we beseffen dat licht uit individuele deeltjes bestaat, waarvan we ook één enkel exemplaar op een ruit kunnen schieten. Dan blijkt het met onze huidige kennis van de natuurkunde volstrekt onvoorspelbaar wat er gebeurt: het lichtdeeltje wordt zonder aanwijsbare oorzaak doorgelaten of niet. Je kunt de uitkomst dan niet meer afdoen als ‘onverwacht’, zeker als de kans op elk van beide mogelijkheden 50 procent is!

Aan de andere kant overlapt het wetenschappelijke taalgebruik toch enigszins met het alledaagse: als je iemand toevallig tegenkomt, bedoel je ook meestal dat er geen speciale reden of oorzaak was om die persoon dan en daar tegen te komen. De hamvraag is echter of je, met de beste kennis van de wereld, *had kunnen weten* dat je hem of haar zou tegenkomen. En dat is vrijwel altijd het geval. Als mijn neef had geweten dat zijn lotgenoot net als hij van skiën hield en dan in het bijzonder altijd in de kerstvakantie in dat ene Zwitserse plaatsje, dan was hun ontmoeting helemaal niet meer zo toevallig geweest, of hun hart nu rechts of links of ondersteboven zit.

Zo is het ook met dobbelen of het opgooien van een munt: precieze kennis van de details van de worp en het oppervlak waar de munt op landt maken het wel degelijk mogelijk om de uitkomst van de worp te voorspellen, want deze is van tevoren door dergelijke details bepaald. En dat is bij een puur toevallige gebeurtenis niet het geval: we spreken dan van *zuiver toeval*, wat dus per definitie een vorm van toeval is die *niet* het gevolg is van onwetendheid (inclusief gebrek aan rekenkracht, kennis van de natuurwetten, en dergelijke).

De achterkanten van 2 van de 4 panelen van *De vier Visioenen van het hiernamaals* (1505–1515) van Jheronimus Bosch. Collectie: Museo di Palazzo Grimani in Venetië. Wim Pijbes in *NRC Handelsblad* van 30 april 2016: ‘De schilderingen doen denken aan een marmerimitatie. Of is het een sterrenhemel?’

Cruijff

De vraag ‘bestaat toeval?’ betekent in strikt wetenschappelijke zin dus: bestaan er gebeurtenissen die gegeven de begintoestand en de natuurwetten in principe ook anders hadden kunnen aflopen? Toeval bestaat dus *niet* als (bij gegeven natuurwetten) dezelfde begintoestand altijd tot hetzelfde vervolg leidt. De beroemde islamitische geleerde Omar Khayyam (1048-1131) drukte dit idee van determinisme als volgt uit: ‘*De eerste dag van de Schepping schreef wat de Dag des Oordeels zal lezen.*’ In dat geval zou bijvoorbeeld niet alleen bij de geboorte van Johan Cruijff op 25 april 1947 al hebben vastgelegen dat hij op 2 januari 1972 met een briljante lob zou scoren tegen ADO Den Haag, maar zou zelfs de precieze baan van die bal al miljarden jaren geleden vast hebben gelegen. Dit idee lijkt bizar, maar was lange tijd de norm en is door uiteenlopende denkers als Calvijn, Spinoza en Einstein verdedigd. Tot de twintigste eeuw kon het determinisme bovendien door de klassieke natuurkunde van Newton (die met name beweging beschrijft, zowel op aarde als van planeten) worden gerechtvaardigd.

Rond 1925 werd de discussie vanuit de natuurkunde echter heropend door een nieuwe theorie van atomen en licht, genaamd de kwantummechanica. In navolging van Bohr denken de meeste fysici dat de kwantummechanica een theorie is die zuiver toeval toelaat en daarmee radicaal breekt met de klassieke natuurkunde. Dit standpunt is echter met name betwijfeld door Einstein (‘God dobbelt niet’), recenter ook door de Nederlandse Nobelprijswinnaar ’t Hooft, en is ondanks de *communis opinio* dan ook allerminst bewezen. Het eerder genoemde voorbeeld van de weerkaatsing van licht is weliswaar volgens de kwantummechanica een zuiver toevallig proces, maar het is niet uit te sluiten dat een diepere theorie wel degelijk een bepaalde uitkomst voorspelt (wel is het zo dat een dergelijke theorie dan geen maximale signaalsnelheid kan hebben, zoals de lichtsnelheid dat in de huidige fysica is, en daarmee juist voor Einstein niet acceptabel zou zijn). Het antwoord op de vraag of zuiver toeval bestaat is dus niet bekend en het is zelfs de vraag of dat in principe wel gegeven kan worden.

Wat zegt deze analyse over het al dan niet toevallige karakter van ons leven, of zelfs hét leven? Het zogenaamde *fine-tuning argument* houdt in dat de kans op leven a priori zó klein is, dat er wel iets achter móet zitten. Dat ‘iets’ is volgens sommige gelovige geleerden een schepper, en volgens anderen een zogenaamd *multiversum*, ofwel een giga-hoeveelheid heelallen met steeds weer andere natuurconstanten, waar het onze er slechts eentje van is. En in dit ene heelal zouden dan toevallig, zoals in een serie worpen van een miljoen zessen, de condities voor leven zijn gerealiseerd.

Cirkelredenering

Klopt de premisse van het *fine-tuning argument*? De eerste vraag is of het ogenschijnlijke toeval in de precieze afstemming (*fine-tuning)* van alle boven genoemde grootheden (en nog andere) zuiver is, of een gevolg is van onze onwetendheid. Dit is niet duidelijk, omdat onbekend is of natuurconstanten en begincondities gevarieerd kunnen worden, in de zin dat andere waarden in principe mogelijk zouden zijn (hier zijn zowel goede argumenten voor als tegen). Als dit niet het geval is, en er dus slechts één mogelijke natuur is (zoals Spinoza dacht), is *fine-tuning* niet toevallig maar noodzakelijk. Dan is ook verder niets meer toevallig, inclusief het bestaan van Dirkie de Veenboer en diens nazaten. Ook daar kun je al dan niet iets achter zoeken, maar niet op grond van toeval.

Stel omgekeerd dat er meerdere mogelijkheden waren voor de waarden van de natuurconstanten en dergelijke. Dan weten wij met onze huidige kennis niets over het mechanisme dat aan de bepaling van de werkelijke waarden ten grondslag lag. Bovendien is de mate van potentiële variatie onbekend, waardoor ook al geen kansverdeling over alle mogelijkheden kan worden gegeven. Het is daarom onzinnig om te beweren dat de kans op de huidige waarden zeer klein zou zijn. Ook dan vervalt het *fine-tuning argument* voor zowel een schepper als een multiversum.

Hoeveel toeval kan een mens verdragen? Zelfs als het ooit mocht lukken om aan ons bestaan een kleine kans toe te kennen, beweer ik dat toevallige gebeurtenissen met een kleine kans juist gekoesterd moeten worden. Willen we echt liever een onderdeeltje van een aflopend uurwerk zijn? Een veel pijnlijker vraag is daarom hoeveel determinisme een mens kan verdragen!

*Eerder verschenen in *Vox, 16*(2016)8
E-mail: <np.landsman@math.ru.nl>

LITERATUUR
Landsman, N. P., & Wolde, E. J. van (Eds.). (2016). *The challenge of chance; A multidisciplinary approach from science and the humanities.* Cham: Springer International Publishing (ook digitaal te lezen via open access).

FRED STEUTEL column

Ontmoetingen

In de loop van zo'n tachtig jaar kom je heel wat mensen tegen. Ik had het al eens over Arnon Grunberg en Eugene Lukacs. Deze keer een rijtje mensen uit de wetenschap.

Norbert Wiener, grondlegger van de cybernetica, bezocht eind jaren vijftig het herseninstituut in Amsterdam. Wij van de afdeling statistiek van het Mathematisch Centrum (nu CWI) werden daarheen gedirigeerd om voor de nodige zaalvulling te zorgen. Heb ik Wiener werkelijk 'ontmoet'?

In Enschede kwam ik Wiebe Draijer tegen, nu topman bij de Rabobank. Hij zat op schoot bij zijn vader, Wiebe Draijer senior, hoogleraar stromingsleer bij de afdeling Werktuigbouwkunde aan de THT (nu UT).

In 1982 liep ik de marathon van Eindhoven. Vlak naast mij in een van de achterste startvakken zag ik Pieter Winsemius, toen minister van huisvesting. Hij liep ook een hele marathon; hij deed er wat langer over dan ik.

Op een congres in San Antonio 1979 sprak de beroemde Hongaarse wiskundige Paul Erdős over 'Drie problemen die ik graag opgelost zou zien voordat ik ga'. Veel later was hij te gast op de statistiekbijeenkomst Lunteren; hij was dus nog niet 'gegaan'.

Ook in Lunteren ontmoette ik Perci Diaconis. Vandaar reed hij met me mee naar Museum Kröller-Müller. Hij was erg geïnteresseerd in een schilderij met kaartspelers. Langgeleden was hij in verband met winstkansen bij kaarttrucs, waarmee hij de kost verdiende, in aanraking gekomen met het kansrekeningboek van William Feller. Omdat hij dat niet kon begrijpen, is hij wiskunde gaan studeren, Snel daarna werd hij hoogleraar in Harvard.

De Lenstra's. Alle vier wiskundigen, net als hun vader Hendrik. De bekendste van de vier is Hendrik jr., die onder de naam Hosia W. Labbers publiceert over 'sommetjes'. Jan Karel was decaan aan de TUE en daarna algemeen directeur van het CWI. Arjen werkt in Lausanne en Andries werkt aan de UvA. Zijn vrouw, Regina Albrink, is pianiste; zij gaf jarenlang – en misschien nog wel – op tweede kerstdag een recital in de kleine zaal van het Koninklijk Concertgebouw. Mijn vrouw en ik zijn daar vaak getuige van geweest.

Later mogelijk meer ontmoetingen.

FRED STEUTEL is emeritus hoogleraar kansrekening aan de TU Eindhoven.
E-mail: <fsteutel@xs4all.nl>

Kiki Bertens. Foto: Steven Pisano (wikimedia commons)

TENNISSELECTIE Olympische Spelen Rio 2016
rankings als basis voor selectieprocedures

In augustus van dit jaar zijn de Olympische Spelen en het ziet er naar uit dat de Nederlandse afvaardiging groter zal zijn ooit, terwijl de kwalificatienormen van het NOC*NSF voor deze Spelen nog nooit zo scherp zijn geweest. In het geval van tennis: het lijkt erop dat de prestaties van Kiki Bertens tijdens Roland Garros leiden tot haar selectie – maar wat zijn eigenlijk de precieze criteria?

GERARD KUPER, GERARD SIERKSMA & FRITS SPIEKSMA

In 2014 lanceerde chef de mission Maurits Hendriks van het NOC*NSF de contouren van de 'Nieuwe normen en limieten voor Rio 2016', waarbij 'een reële kans op een Top 8-notering bij de komende Olymische Spelen centraal dient te staan' (NOC*NSF, 2014). Maar wat moet worden verstaan onder 'reële kans op een Top 8-notering'? Dat dit goed moest worden vastgelegd was van meet af aan duidelijk, ook al omdat het met name de tenniss(t)ers waren die in het verleden bezwaren maakten tegen de selectieprocedures. De door de Rijksuniversiteit van Groningen (RUG) en ORTEC-Sports ontwikkelde schaatsselectieprocedure voor de Winterspelen van Sochi 2014, waarbij de kansen op olympische medailles van de schaats(st)ers zijn berekend in de zogenoemde Prestatiematrix, was een succes. Vandaar dat het NOC*NSF de RUG wederom vroeg een dergelijk systeem te ontwerpen voor de olympische tennisselectie. Een belangrijke randvoorwaarde was dat de rekenmethode voor de tennissers transparant is. Daarom is besloten de ATP- en de WTA-rankings te gebruiken als basis voor het rekenmodel, waarbij voor elke positie van die rankings de kans wordt geschat dat een speler op die positie doorgaat naar de kwartfinale (een kwartfinale komt immers overeen met een Top 8-notering).

De ontwikkelde methodiek is niet tennisspecifiek en kan worden gebruikt voor elke sport met adequate wereldranglijsten, waarbij vanzelfsprekend elke sport zijn eigen selectiecriteria gebruikt. Zo is ook voor de olympische badmintonselectie een dergelijk systeem ontworpen. Binnenkort maakt het NOC*NSF, mede op grond van onze bevindingen, bekend of er, en wellicht welke, tenniss(t)ers naar Rio worden afgevaardigd. Net als voor Sochi worden er deze keer geen vervelende gerechtelijke procedures verwacht.

ATP- en WTA-rankings als selectiegereedschap

Nationale olympische organisaties hebben te maken met het lastige probleem om atleten te selecteren, die het land gaan vertegenwoordigen op aankomende Olympische Spelen. Zowel de bonden als de atleten hebben belang bij een heldere, eerlijke en eenduidige selectieprocedure. Subjectieve en onduidelijke formuleringen leiden vrijwel zeker tot kostbare gerechtelijke procedures en kunnen een ongewenst negatief effect hebben op de prestatie van de atleet. Dit artikel focust op het olympische tennistoernooi van 2016 in Rio de Janeiro en is gemotiveerd door het hierboven geformuleerde verzoek van het NOC*NSF. Het NOC*NSF is de overkoepelende organisatie van georganiseerde sporten in Nederland met een totaal van 88 aangesloten sportbonden, die zo'n 28000 sportclubs omvatten met meer dan vijf miljoen sporters. Het NOC*NSF is verantwoordelijk voor het formuleren van de selectiecriteria en voor de feitelijke selectie van de atleten.

Samen met het NOC*NSF kwamen we tot de volgende probleemstelling. Bereken voor elke positie van de ATP- en de WTA-ranking de kans dat de speler op die positie op een vooraf vastgestelde datum (de referentiedatum) de kwartfinale bereikt. Het resultaat van de procedure is dat elke tennisser ruim op tijd weet welke positie op de ranking nodig is om geselecteerd te worden.

Rankings van tennisspelers zijn in de literatuur intensief gebruikt om resultaten van individuele wedstrijden te voorspellen. Zo wordt in Del Corral & Prieto-Rodríguez (2010) beargumenteerd dat het verschil in positie op een ranking een goede voorspeller is van Grand Slam-resultaten. In Klaassen & Magnus (2003) worden logitmodellen gebruikt om winstkansen te berekenen vóór en tijdens de match. In Clarke & Dyte (2000) wordt verschil in positie op de ranking gebruikt om voor elke toernooironde de winnaar te voorspellen en om de kansen op toernooiwinst van de deelnemers te bepalen. Ander onderzoek richt zich op de fysiologie van tennisprestaties (Kovacs, 2006) en op de verbetering van de rankingsystemen (Ruiz, et al., 2013; Irons, et al., 2014). Het gebruik van rankings voor kansbepalingen ontbreekt in de literatuur. De enige verwijzing die we hebben kunnen vinden, betreft een passage van M. Reid en C. Morris(2013, p. 350): '*Future work should focus on the change in Top 100 demographics over time as well as on the evaluation of the interaction between rankings and tournament plays.*' Het eerste deel van deze passage komt aan de orde in de publicatie van Reid et al. uit 2014; het tweede deel is het onderwerp van het onderliggende artikel.

Datapooling; gender- en toernooiverschillen

We gebruiken de resultaten van de drie meest recente Olympische tennistoernooien, tezamen met de resultaten van alle Grand Slams in de periode 2004–2014. Hoewel er grote verschillen zijn tussen beide toernooien, zijn de verschillen tussen de door ons berekende positiekansen van de twee toernooien statistisch niet significant. Omdat ook de genderverschillen niet-significant blijken te zijn, is het mogelijk de mannen- en de vrouwentoernooien te poolen, wat de nauwkeurigheid van de resultaten ten goede is gekomen.

Dit artikel is verder als volgt opgezet. We beginnen met de presentatie van de gebruikte data en het poolen van de resultaten van de Spelen van 2004 (Athene), 2008 (Beijing) en 2012 (London), zowel voor de mannen als de vrouwen. Vervolgens voegen we de Grand Slam-toernooien toe en presenteren we de gepoolde kansresultaten. Daarnaast geven we de resultaten van de toetsen waarmee is bepaald of de kansen op het bereiken van de kwartfinales statistisch significant verschillend zijn voor de Olympische Spelen en de Grand Slams. Ten slotte toetsen we de invloed van genderverschillen en de invloed van de verschillende baansoorten op de positiekansen.

We beschouwen de periode augustus 2004 tot januari 2014. In deze periode zijn 38 Grand Slams en drie Olympische tennistoernooien georganiseerd; zie tabel 1. Grand Slam-toernooien beginnen met 128 deelnemers en Olympische met 64. Onze database bevat de namen van de 64 spelers van de drie meest recente Olympische toernooien, plus de namen van de 64 spelers die de eerste ronde van de 38 Grand Slams hebben gewonnen. Daardoor bevat de database (41 × 64=) 2.624 observaties voor zowel mannen als vrouwen, een totaal van 5.248 observaties. Ook hebben we de posities op de rankings van al deze observaties op de referentiedata van de diverse toernooien verzameld (Stevegtennis, 2014). De referentiedata voor de Olympische toernooien worden vastgesteld door de International Tennis Federation (ITF); voor de Grand Slams hebben we de datum twee weken voor de start van het betreffende toernooi genomen. Tabel 1 geeft de toernooidata, inclusief het type ondergrond en de referentiedata.

TOERNOOI	LOCATIE EN PERIODE	ONDERGROND	REFERENTIEDATUM
OLYMPISCHE SPELEN	Londen 2012, 27/7–12/8	gras	11/6/2012
	Beijing 2008, 8/8-24/8	hardcourt	9/6/2008
	Athene 2004, 13/8–29/8	hardcourt	14/6/2004
GRAND SLAMS	Australian Open 2005–2014, januari	hardcourt	twee weken voor toernooi
	Roland Garros 2005–2013, mei-juni	gravel	twee weken voor toernooi
	Wimbledon 2005–2013, juni-juli	gras	twee weken voor toernooi
	US Open 2004–2013, augustus-september	hardcourt	twee weken voor toernooi

Tabel 1. Toernooien in de database

Tabel 2 geeft de verdeling van alle spelers (mannen en vrouwen) in de database. We hebben ons beperkt tot de Top 100 van beide rankings, waarbij we ons gesteund voelden door een citaat uit het artikel van Reid et al. (2014): '*Reaching the Top 100 can be seen as an important goal with more than just a symbolic value; it may result in an automatic qualification for the next Grand Slam tournament.*' Tabel 2 laat zien de meeste spelers zich bevinden in de posities 1–10: spelers met een hoge positie op de ranking hebben ook een hoge kans de eerste ronde te winnen van een Grand Slam, terwijl het aantal spelers daalt met het afnemen van de rankingpositie. Daarbij is er voor olympische deelname een landenlimiet (zie hierna), wat mede het aantal spelers met een lage ranking verklaart.

Puntschattingen en Betrouwbaarheidsintervallen

We beginnen met een eenvoudige puntschattingstechniek. Als voorbeeld vergelijken we Top 32-spelers met spelers buiten de Top 32 (aangeduid met Bot 33). Daarna introduceren we het probitmodel, dat de bijbehorende betrouwbaarheidsintervallen oplevert. Deze intervallen worden gebruikt voor de significantietesten van de genderverschillen en de verschillen tussen de Olympische en Grand Slam-toernooien. De methode is toegepast op de data van de Olympische tennistoernooien van 2005, 2008 en 2012 voor zowel mannen als vrouwen. In totaal hebben (3 × 64=) 192 vrouwen deelgenomen aan deze drie Spelen, waarvan 72 in de Top 32 en 120 in de Bot 33; zie de tweede kolom van tabel 3. Deze tabel geeft voor beide categorieën de aantallen die doorgingen naar de betreffende volgende ronde. Het bovenste deel van tabel 3 laat het volgende zien:

- Na de eerste ronde gingen (3 × 32=) 96 spelers door naar Ronde 2, waarvan 52 uit de WTA Top 32;

RANKINGPOSITIE	VROUW	MAN	TOTAAL
1 – 10	344	353	697
11 – 20	325	308	633
21 – 30	293	277	570
31 – 40	234	238	472
41 – 50	171	182	353
51 – 60	190	195	385
61 – 70	172	164	336
71 – 80	154	148	302
81 – 90	137	140	277
91 – 100	131	124	255
101+	473	495	968
TOTAAL	2624	2624	5248

Tabel 2. Aantallen spelers in de database verdeeld over rankingposities (64 spelers in de eerste ronde van de Olympische toernooien en 64 in de tweede ronde van de Grand Slams)

- Na de tweede ronde gingen (3 × 16=) 48 vrouwen door naar Ronde 3, waarvan 40 uit de Top 32;
- 22 spelers uit de Top 32 gingen door naar Ronde 3;
- (22/72=) 30,6% spelers uit de Top 32 gingen door naar de achtste finale;
- Slechts (2/120=) 1,7% spelers uit de Bot 33 haalden de laatste acht.

Het onderste deel van tabel 3 geeft de resultaten in de eerste drie rondes voor de mannen. In totaal (3 × 64=) 192 mannen namen deel aan de drie meest recente Spelen, waarvan er 74 in de Top 32 en 118 in de Bot 33 zaten. Uit tabel 3 kunnen we de volgende conclusies trekken:

- Van de 192 spelers zaten er 74 in de Top 32 en 118 in de Bot 33 van de ATP-ranking;
- (21/74=) 28,4% spelers uit de Top 32 bereikten de laatste acht, terwijl (3/118=) 2,5% in de Bot 33 zat.

Dit voorbeeld laat zich eenvoudig generaliseren voor andere rondes van andere toernooien. Echter, deze methode kent een aantal tekortkomingen. Ten eerste is de betrouwbaarheid van de berekende kansen onbekend. Die betrouwbaarheid is nodig om te kunnen bepalen of de kans dat een vrouw uit de Top 32 de kwartfinale bereikt (30,6%), statistisch verschillend is van de corresponderende 28,4% kans voor mannen. Een tweede punt betreft de beide clusters Top 32 en Bot 33 in tabel 3, die beide tamelijk groot zijn. Wat zouden de kansen zijn als we kleinere clusters zouden hanteren, bijvoorbeeld clusters van vier (1–4, 5–8, 9–12, et cetera)?

Een veel gebruikte methode voor het berekenen van puntschattingen en betrouwbaarheidsintervallen is regressieanalyse. Als de afhankelijke variabele, y, binair is wordt veelal gekozen voor een probitmodel, waarmee een S-vormige kromme wordt getransformeerd in een rechte lijn, die vervolgens wordt geanalyseerd met *maximum likelihood*. We gebruiken het volgende probitmodel:

$$P(y = 1|x, \beta) = 1 - \Phi(-x'\beta) = \Phi(x'\beta). \tag{1}$$

Hierin is $\Phi(.)$ de cumulatieve distributiefunctie van de standaardnormale distributie en x de verklarende variabele. $y = 1$ betekent dat de betreffende speler Ronde 3 wint en doorgaat naar de kwartfinale; $y = 0$ betekent verlies in de Ronde 3 van die speler. De vector x is een vector met uitsluitend enen, genoteerd als ι (de constante term, ofwel de intercept). Dus β is de schatter van deze intercept. Vergelijking (1) kunnen we nu vereenvoudigen tot:

$$P(y = 1|\iota, \beta) = \Phi(\beta). \tag{2}$$

Met deze specificatie van het probitmodel berekenen we niet alleen het aantal spelers dat de kwartfinale bereikt, maar ook de bijbehorende betrouwbaarheidsintervallen.

Tabel 4 geeft probitschattingen van Top 32-spelers die doorgingen naar de 'laatste acht'. De kans voor Top 32-vrouwen om de 'laatste acht' te bereiken is berekend als het marginale effect van de interceptkans van het probitmodel. Dit marginale effect is berekend door de betreffende coëfficiënt uit formule (2) te transformeren met behulp van de standaardnormale-distributiefunctie $\Phi(-0{,}508) = 0{,}306$. Tabel 3 laat zien dat 22 van de 73 Top 32-vrouwen de kwartfinale hebben bereikt en wel met een kans van (22/72=) 30,6%. Het corresponderende 95%-betrouwbaarheidsinterval is $(\Phi(-0{,}817); \Phi(-0{,}200)) = (0{,}207; 0{,}421)$. Voor de mannen betekent dit een kans van (21/74=) 28,4% met een 95%-betrouwbaarheidsinterval gelijk aan $(\Phi(-0{,}880); \Phi(-0{,}263)) = (0{,}189; 0{,}396)$. Echter voor nauwkeuriger kansen, met name voor de lagere clusters (bijvoorbeeld 85-88, 89-92, et cetera), hebben we meer waarnemingen nodig. Daarom hebben we Grand Slam-toernooien toegevoegd en een 'gepoold' probitmodel geschat voor meerdere clustergrootes van zowel de WTA- als de ATP-ranking.

VROUWEN	Aantal in Ronde 1	Aantal in Ronde 2	Aantal in Ronde 3	Aantal winnaars van Ronde 3
Totaal aantal spelers	(3 × 64=)192	(3 × 32=)96	(3 × 16=)48	(3 × 8=)24
WTA 1–32	72	52	40	22
WTA 33+	120	44	8	2
MANNEN	**Aantal in Ronde 1**	**Aantal in Ronde 2**	**Aantal in Ronde 3**	**Aantal winnaars van Ronde 3**
Totaal aantal spelers	(3 × 64=)192	(3 × 32=)96	(3 × 16=)48	(3 × 8=)24
ATP 1–32	74	51	34	21
ATP 33+	118	45	14	3

Tabel 3. Prestatietabel van Top 32- en Bot 33-spelers voorafgaand aan de Spelen van 2004, 2008 en 2012

Valkuilen van data-pooling

Zoals we zagen moeten er om de kwartfinale te bereiken van een Grand Slam vier wedstrijden worden gewonnen; voor de Spelen zijn dat er drie. Daarom hebben we in de dataverzameling alleen spelers opgenomen die de eerste Grand Slam-ronde overleefden. Een ander verschil tussen beide toernooien is het feit dat er voor Olympische tennistoernooien een deelnamelimiet is van vier mannen en vier vrouwen per land (zie Australian Olympic Committee, 2014). We hebben getoetst of de schattingen verschillend zijn voor mannen en vrouwen, voor verschillende baansoorten, en voor de verschillen tussen Olympische Spelen en Grand Slams. De resultaten van deze toetsen zijn beschreven door Kuper et al. (2014). Vergelijkbare toetsen laten zien dat de verschillen tussen baansoorten klein zijn en alleen statistisch verschillend zijn voor spelers in de clusters 9–12 en 41–44. Wat betreft de Spelen en de Grand Slams, blijkt (zie Kuper et al., 2014) dat de kansen om de kwartfinales te bereiken niet statistisch verschillend zijn voor de hoge clusters; alleen voor het cluster 81–84 verschillen de resultaten.

De Gepoolde Dataset in Actie

We passen tenslotte het probitmodel toe op de gehele database met 41 toernooien, drie Olympische toernooien en 38 Grand Slams, mannen en vrouwen samengenomen. Dit betekent een steekproefgrootte van 5.248 observaties. Tabel 5 geeft de gepoolde schattingen voor clusters ter grootte vier. Het aantal waarnemingen per cluster staat in deze tabel, evenals het aantal enen in de steekproef. De laatste drie kolommen van tabel 5 geven de puntschattingen en de 95%-betrouwbaarheidsintervallen. Uit de tabel kunnen we afleiden dat voor een Top 4-speler geldt dat de kans om de kwartfinale te bereiken 0,722 is met een 95% betrouwbaarheidsinterval van (0,669; 0,771). Wanneer de positie op de ranking niet relevant zou zijn, is de kans om drie wedstrijden achter elkaar te winnen gelijk aan $0{,}5^3 = 0{,}125$, wat overeenkomt met 12,5%. Dus geldt dat voor Top 4-spelers de positie op de ranking er zeker toe doet, omdat de ondergrens van het 95% betrouwbaarheidsinterval groter is dan 0,125. Ook voor de clusters 17–20, 25–28 en lagere, geldt dat het 95% betrouwbaarheidsinterval de waarde 0,125 niet bevat. Dus voor deze clusters doet de positie op de ranking er zeker toe.

Er zitten kleine anomalieën in deze tabel. Merk op dat onder meer voor spelers met een positie op de ranking uit cluster 41–44 een grotere kans hebben om door te gaan naar de kwartfinale dan spelers in cluster 37-40. Echter, de 95% betrouwbaarheidsintervallen van deze clusters overlappen elkaar, terwijl het formeel toetsen ervan laat zien dat het verschil statistisch niet verschilt van nul op een significantieniveau van 5%.

Zoals gezegd mogen, in tegenstelling tot de Grand Slams, aan het enkelspel van het Olympisch tennistoernooi slechts ten hoogste vier tennissers en vier tennissters van hetzelfde land deelnemen. We zouden daarom Grand Slam-observaties moeten weglaten uit de dataset. Dat dit niet nodig is, blijkt uit de volgende drie argumenten. Ten eerste is het zo dat de kans om drie rondes achterelkaar te winnen niets heeft te maken met de nationaliteit van de speler. Wat wel telt is de ATP/WTA-rankingpositie van de speler. Ten tweede geldt dat derde-rondereresultaten van Olympische tennistoernooien en vierde-rondereresultaten van Grand Slamtoernooien statistisch gelijk zijn. Ten derde gaat het weglaten van meer dan 800 waarnemingen ten koste van de nauwkeurigheid.

VROUWEN	Aantal observaties: 72 (waarvan 22 met waarde 1) Top 32-spelers					
					95% betrouwbaarheidsinterval	
Variabele	Coëfficiënt	Std. dev.	*z*-statistiek	*p*-waarde	Laag	Hoog
Constante	−0,508	0,155	−3,284	0,001	−0,817	−0,200
MANNEN	**Aantal observaties: 74 (waarvan 21 met waarde 1) Top 32-spelers**					
					95% betrouwbaarheidsinterval	
Variabele	Coëfficiënt	Std. dev.	*z*-statistiek	*p*-waarde	Laag	Hoog
Constante	−0,572	0,155	−3,696	<0,001	−0,880	−0,263
Afhankelijke variabele: (0,1), dummy-variabele: 1 = derde ronde winnen, 0 = anders						

Tabel 4. Probitschattingen van Olympische kwartfinalisten uit de Top 32 (Spelen van 2004, 2008 en 2012)

Conclusie

In 2014 heeft het NOC*NSF de kwalificatie-eisen voor deelname aan de Olympische Spelen aangescherpt. Voor wat betreft de tenniskwalificatie voor de Spelen van 2016, hanteert het NOC*NSF het in dit artikel gepresenteerde model, dat gebaseerd is op de relatie tussen de positie op de ATP- of de WTA-ranking en de kans om een Olympische kwartfinale te halen. We gebruikten een database met 41 toernooien voor zowel de mannen als de vrouwen in de periode 2004–2014. We hebben aangetoond dat gender- en toernooipooling statistisch verantwoord is, waardoor we in staat waren kleine rankingclusters van slechts vier posities te gebruiken. De gepoolde analyse laat zien dat de positie op de ATP- en de WTA-ranking er toe doet. Voor een Top 4-tennis(t)er geldt dat de kans de Olympische kwartfinale te bereiken gelijk is aan 0,722 met een 95%betrouwbaarheidsinterval van (0,669; 0,771), terwijl deze waarden voor een positie tussen 5 en 8, respectievelijk, 0,476 en (0,416; 0,536) zijn. Voor de posities 1–20 gelden de kansen >12,5% en voor de posities 25–100 zijn die kansen <12,5%. Het is nu aan het NOC*NSF de kans te kiezen die nodig is de kwartfinale te halen. De corresponderende posities op de ATP- en de WTA-rankings op de referentiedatum bepalen dan de 2016 Olympische tennisselectie.

				95% betrouwbaarheidsinterval	
cluster	observaties	met waarde 1	kans	laag	hoog
1–4	299	(216)	0,722	0,669	0,771
5–8	267	(127)	0,476	0,416	0,536
9–12	256	(78)	0,305	0,251	0,363
13–16	263	(45)	0,171	0,129	0,221
17–20	245	(44)	0,180	0,135	0,232
21–24	224	(21)	0,094	0,061	0,138
25–28	231	(16)	0,069	0,042	0,109
29–32	223	(18)	0,081	0,050	0,123
33–36	186	(14)	0,075	0,044	0,121
37–40	178	(4)	0,022	0,008	0,055
41–44	144	(9)	0,062	0,032	0,113
45–48	143	(10)	0,070	0,037	0,123
49–52	153	(4)	0,026	0,009	0,064
53–56	152	(5)	0,033	0,013	0,073
57–60	146	(5)	0,034	0,013	0,076
61–64	130	(3)	0,023	0,007	0,065
65–68	134	(2)	0,015	0,003	0,052
69–72	135	(4)	0,030	0,010	0,072
73–76	123	(2)	0,016	0,004	0,057
77–80	116	(4)	0,034	0,012	0,084
81–84	112	(2)	0,018	0,004	0,062
85–88	106	(2)	0,019	0,004	0,066
89–92	111	(3)	0,027	0,008	0,075
93–96	98	(3)	0,031	0,009	0,085
97–100	105	(2)	0,019	0,004	0,066

Tabel 5. Probitschattingen en 95% betrouwbaarheidsintervallen voor de kansen om de kwartfinale te bereiken

LITERATUUR

Australian Olympic Committee, 2014. http://corporate.olympics.com.au/files/dmfile/Rio2016QualificationSystem-Tennis.pdf.

Clarke, S., & Dyte, D., 2000. Using official ratings to simulate major tennis tournaments. *International Transactions in Operations Research, 7*(6), 585–594.

Del Corral, J. & Prieto-Rodríguez, J., 2010. Are differences in ranks good predictors for Grand Slam tennis matches? *International Journal of Forecasting, 26*(3), 551–563.

Irons, D. J., Buckley, S., & Paulden, T., 2014. Developing an improved tennis ranking system. *Journal of Quantitative Analysis in Sports, 10*(2), 109–118.

International Tennis Federation (ITF), 2014. http://www.itftennis.com/olympics/ players/ qualification.aspx,.

Klaassen, F., & Magnus, J., 2003. Forecasting the winner of a tennis match. *European Journal of Operational Research, 148*(2), 257–267.

Kovacs, M. S., 2006. Applied physiology of tennis performance. *British Journal of Sports Medicine, 40*(5), 381–386.

Kuper, G., Sierksma, G., & Spieksma, F. C. R., 2014, *Using tennis rankings to predict performance in upcoming tournaments,* University of Groningen, SOM Research Report, vol. 14034-EEF.

NOC*NSF, 2014. Nieuwe normen en limieten voor Rio 2016. Reële kans op top 8-notering staat centraal. *Lopend vuur,* (2014)6.

Reid, M., & Morris, C., 2013. Ranking benchmarks of Top 100 players in men's professional tennis. *European Journal of Sports Science, 13*(4), 350–355.

Reid, M., Morgan, S., Churchill, T., & Bane, M. K., 2014. Rankings in professional men's tennis: a rich but underutilized source of information. *Journal of Sports Sciences, 32*(10), 986–992.

Ruiz, J. L., Pastor, D., & Pastor, J. T., 2013. Assessing professional tennis players using data envelopment analysis (DEA). *Journal of Sports Economics, 14*(3), 276–302.

Stevegtennis, 2014. www.stevegtennis.com, 20 oktober 2014.

We danken Laurens den Ouden (NOC*NSF) voor zijn commentaar en het NOC*NSF voor het beschikbaar stellen van data.

GERARD KUPER is als Universitair Hoofddocent werkzaam bij de Faculteit Economie en Bedrijfskunde van de Rijksuniversiteit Groningen. E-mail: <g.h.kuper@rug.nl>

GERARD SIERKSMA is emeritus hoogleraar Operations Research en Sport Statistiek aan de Faculteit Economie en Bedrijfskunde van de Rijksuniversiteit Groningen. E-mail: <g.sierksma@rug.nl>

FRITS SPIEKSMA is als hoogleraar verbonden aan het Research Centre for Operations Research and Business Statistics (ORSTAT) van de KU Leuven. E-mail: <frits.spieksma@kuleuven.be>

Glucosesensor

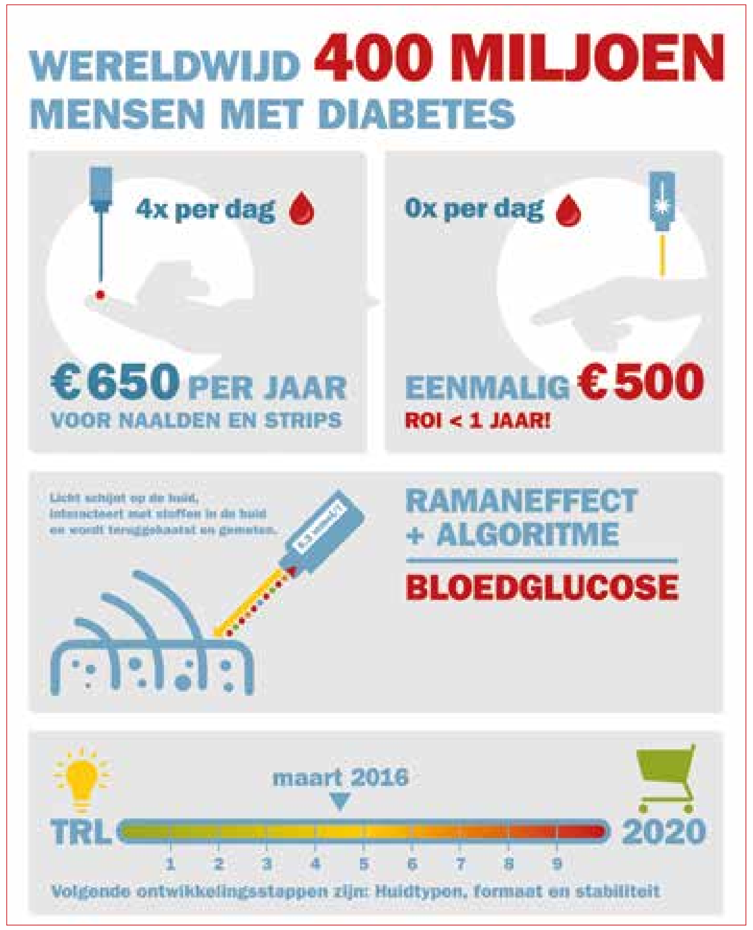
Pijnloos meten van glucose voor diabetespatiënten:
DE UITDAGINGEN EN OPLOSSINGEN

MAARTEN SCHOLTES-TIMMERMAN, SABINA BIJLSMA & JACK VOGELS

Diabetes Mellitus (letterlijk, *zoete vloed*), of suikerziekte, is een aandoening waar wereldwijd honderden miljoenen mensen aan lijden. Het kenmerkt zich door het niet in staat zijn van het lichaam om het glucoseniveau in het lichaam te reguleren. Te hoge glucosewaarden in het bloed kunnen leiden tot nierproblemen, problemen met de ogen, wondjes die niet genezen, et cetera. Te lage waarden kunnen leiden tot concentratieverlies of zelfs bewusteloosheid. Ruwweg is diabetes op te delen in twee verschillende types: type I, waarbij het lichaam geen insuline kan aanmaken (insuline is het hormoon dat de opname van glucose door cellen regelt), en type II, waarbij het lichaam wel insuline produceert maar er niet of minder gevoelig voor is.

Voor een goede regulering van het glucoseniveau bij diabetespatiënten is het noodzakelijk om regelmatig (o.a. voor en na maaltijden) het suikerniveau in het bloed te bepalen. Indien het niveau te laag is moeten er koolhydraten ingenomen worden; is het te hoog, dan dient de diabetespatiënt zichzelf insuline toe. Het bepalen van glucose in het bloed gebeurt al tijden op dezelfde manier: met een fijn (schiet)naaldje prikt de patiënt in zijn/haar vingertop, om een druppeltje bloed te laten analyseren met een draagbaar glucosemetertje. Bij kleine kinderen met diabetes is dat elke keer in de vingers moeten prikken vanzelfsprekend vervelend, maar op lange termijn is het voor elke diabetespatiënt een vervelende handeling, die vaak leidt tot gevoelloze vingertoppen. Daarnaast is bloedprikken niet pijnloos waardoor sommige patiënten minder vaak hun bloedwaarden testen dan ze zouden moeten doen.

Al jaren lang probeert de wetenschappelijke wereld



Figuur 1. De TNO-aanpak

het voor elkaar te krijgen een technologie te ontwikkelen waarmee *niet*-invasief, dus zonder prikken, glucosegehaltes kunnen worden bepaald. Tot op heden is dat echter nog niet echt gelukt. Bij onderzoeksinstituut TNO worden nu belangrijke stappen gezet naar de uiteindelijke marktintroductie van een niet-invasieve glucosemeter. TNO gebruikt een optische technologie, gekoppeld aan signaal-processing, om glucosewaarden te kunnen voorspellen op basis van een eenvoudig te meten spectrum. Bijkomend voordeel van deze aanpak is dat er geen kostbare meetstrippen en naalden meer nodig zijn. Zie voor de schematische weergave van de TNO-aanpak figuur 1.

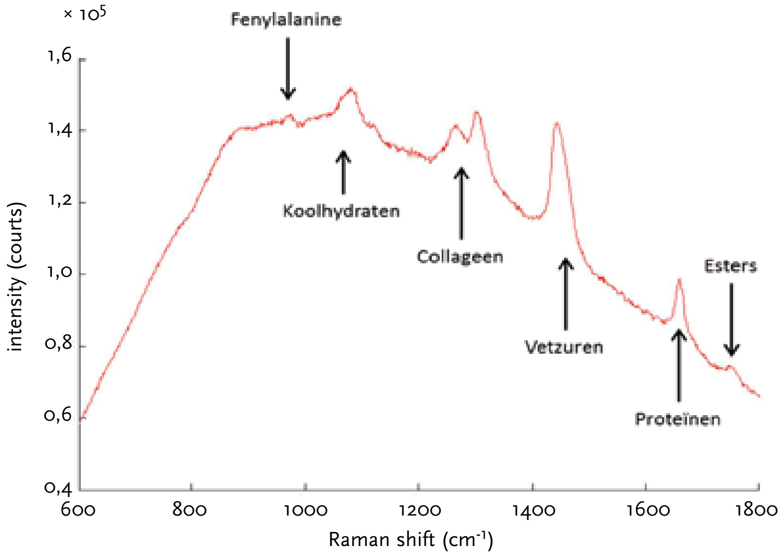
Glucose meten in het bloed met glucosemeter

De technologie

TNO gebruikt Ramanspectroscopie om informatie over de samenstelling van weefsel te verkrijgen (Scholtes-Timmerman et al., 2014). Met behulp van een speciaal ontwikkelde sensor, die maar even op de huid van de onderarm hoeft te worden geplaatst, worden metingen gedaan. De technologie is gebaseerd op verstrooiing van licht afkomstig van een laser. Het licht levert een spectrum op waarin informatie is verborgen over alle moleculen in het gemeten stukje huid. Het spectrum bevat onder meer informatie over vetten, proteïnen (eiwitten) en intra- en extracellulaire componenten. Óók koolhydraten, waaronder glucose, maken deel uit van dit complexe signaal. Het probleem met glucose echter, is dat het niet als één uniek deel van het signaal aanwezig is; de bijdrage van glucose is als het ware uitgesmeerd over het hele signaal.

Het idee is, dat een generieke database wordt opgebouwd door vele verschillende proefpersonen te meten. Deze database bevat dan Ramanspectra en bijbehorende glucosegehalten. Via een algoritme worden deze met elkaar verbonden zodat, op basis van het gemeten signaal, de glucosewaarde kan worden bepaald.

Een fundamentele hobbel die hierbij moet worden genomen is dat glucose maar een heel kleine fractie van het gehele signaal uitmaakt; terwijl andere delen van het signaal van nature sterk variëren tussen verschillende proefpersonen. Het is dus essentieel om bij de verwerking van de metingen de veel sterk variërende (niet-glucose-specifieke) informatie te scheiden van de weinig wél-glucose-specifieke informatie (figuur 2).



Figuur 2. Een typisch Ramanspectrum met globale identificatie van delen van het signaal

De TNO-aanpak: EROS

Er zijn verschillende manieren om met ongewenste variaties in complexe signalen om te gaan. Er is voor gekozen om op zoek te gaan naar een methode waarbij de ongewenste variantie wordt afgetrokken van de signalen, gebaseerd op het indelen van metingen in groepen. De rationale hierachter is dat alle variantie in een set Ramanspectra met ongeveer hetzelfde glucosegehalte niet echt interessant is voor de bepaling van het glucosegehalte maar het verschil tussen de groepen wel. De methode die hier voor gebruikt wordt heet Error Removal by Orthogonal Subtraction (EROS) (Zhu et al., 2008).

In een dataset met Ramanspectra en bijbehorende glucosewaarden worden alle spectra, die behoren bij een afgeronde integere glucosewaarde, per groep ‘gepooled’. Bijvoorbeeld alle spectra die behoren bij een glucosewaarde van 3,50 tot en met 4,49 worden bijeen genomen in één groep. Door vervolgens van elke groep het gemiddelde spectrum af te trekken wordt de relatie tussen de groepen en het glucosegehalte grotendeels vernietigd. Aangenomen wordt dat dan alle resterende informatie gerelateerd is aan andere factoren; bijvoorbeeld verschillen in geslacht, leeftijd en BMI van de proefpersonen. Als vervolgens ook nog wordt aangenomen dat deze factoren modelleerbaar zijn, kan met behulp van een techniek zoals bijvoorbeeld Principale Componenten Analyse (PCA) een model worden gebouwd waarmee de originele spectra kunnen worden gecorrigeerd. Verwijderen van deze invloed van het model zou zo dan theoretisch een concentratie van de spectrumsignalen op de uitgemiddelde factor, i.e. het glucosegehalte, moeten geven. Dit noemen we een EROS-filter.

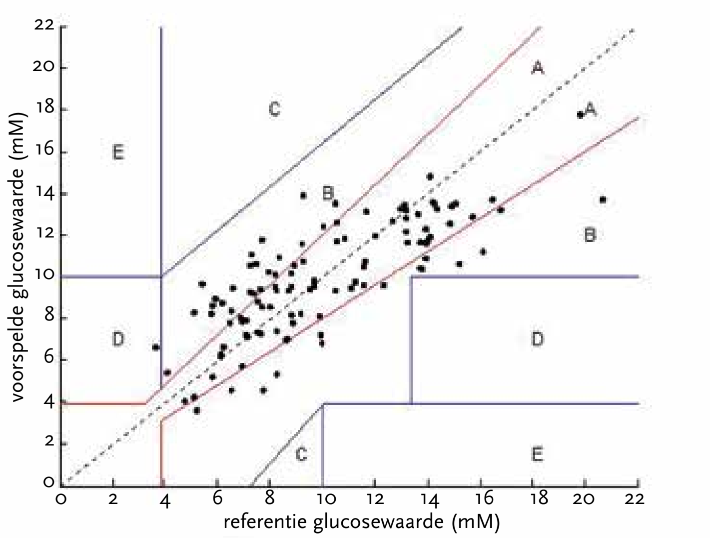
Theorie

Na verzamelen van de spectra in discrete glucosegroepen en verwijderen van de groepsgemiddelden wordt een PCA uitgevoerd op alle resterende signalen (het EROS-model). PCA is een techniek die een *n*-dimensionale ruimte projecteert op een *m*-dimensionale ruimte waarbij geldt dat $m<n$. De (hoog) *n*-dimensionale ruimte wordt hierbij dus vereenvoudigd tot een (laag) *m*-dimensionale ruimte waarbij er naar gestreefd wordt zo weinig mogelijk informatie te verliezen. De informatie die niet in het model terechtkomt wordt beschouwd als ruis die door meetfouten en variatie in de metingen veroorzaakt kan zijn.

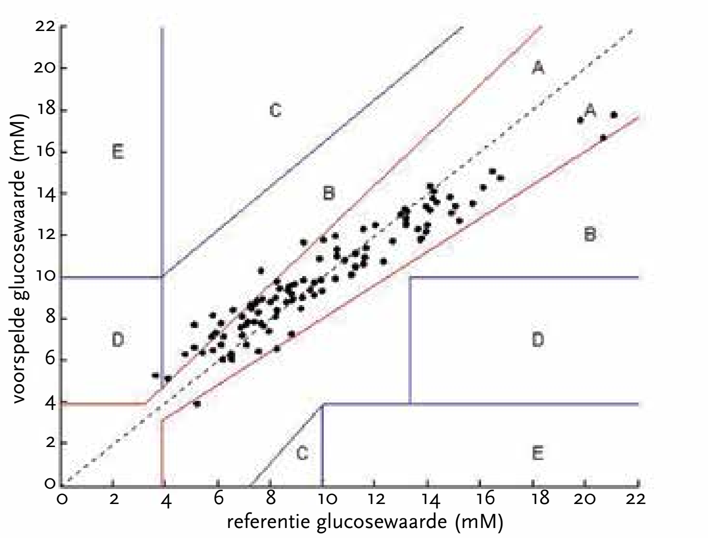
Vervolgens worden de originele spectra geprojecteerd op de PCA-ruimte van het EROS-model. Aan de hand van de scores van de individuele spectra in dit model, wordt een variantie-gecorrigeerd spectrum berekend:

$$RS_{EROS} = RS - \sum_{i=1}^{n} s_i \times l_i$$

Hierbij is RS_{EROS} het spectrum ná correctie, *RS* het spectrum vóór correctie, *n* het aantal relevante componenten



Figuur 3. Glucosevoorspellingen zonder EROS; Clarke Error Grid van niet-EROS-gecorrigeerde data



Figuur 4. Glucosevoorspellingen met EROS; Clarke Error Grid van wél-EROS-gecorrigeerde data

in het PCA model, s_i de score van het op het model geprojecteerde spectrum *RS* in dimensie *i*, en l_i de loadingvector van dimensie *i*.

Partial Least Squares als regressietechniek

Partial least squares (PLS) (Geladi & Kowalski, 1986; Martens & Naes, 1989) is een regressietechniek waarmee een bepaalde kwantitatieve waarde (bijvoorbeeld de glucoseconcentraties *G*) door middel van een model beschreven wordt als een lineaire combinatie van Ramansignalen. De waarde van *G* wordt dan beschreven als:

$$G=b1x1+b2x2+b3x3+\ldots+bnxn$$

Waarbij *b* de gewichts- of regressiefactoren van het model zijn; en *x* de verschillende meetkanalen waaruit het Ramanspectrum bestaat). Trainen van het model op een basis-set van spectra resulteert in een set van gewichtsfactoren waarmee vervolgens andere spectra die niet in het model zaten kunnen worden voorspeld. De validiteit van het gevonden model kan worden beoordeeld door verschillende modellen met sets van verschillende samenstelling te berekenen en te beoordelen. Een methode die hierbij vaak wordt gebruikt is de zogenaamde dubbele kruisvalidatie waarbij een set van spectra wordt voorspeld die niet is betrokken bij de optimalisatie van het model (Smit et al., 2007). Ter verbetering van het resultaat kan het PLS-algoritme, in ons geval, ook nog worden gecombineerd met een variabelenselectiemethode om ruis uit het totale Ramansignaal te filteren en zo de modellering nog specifieker voor glucose te maken.

De studie

Om de toepasbaarheid van de TNO-aanpak te illustreren wordt een dataset gebruikt die verkregen is uit een klinische studie. Hiervoor gebruiken we een dataset van 111 metingen aan evenzovele patiënten, uitgevoerd in het najaar 2013 bij het ISALA-ziekenhuis in Zwolle. Hierbij zijn Ramanspectra gemeten, terwijl op hetzelfde moment via een bloedbepaling het glucosegehalte van de proefpersoon bepaald is. Het doel van de studie is aantonen dat met behulp van geavanceerde statistische algoritmes de glucosewaarde van een proefpersoon (diabetespatiënt) kan worden afgeleid uit zijn/haar Ramanspectrum; daarmee is immers de stap gezet om te komen tot een nieuwe methode van bloedglucosebepaling zonder prikken.

Meestal worden technieken om bloedglucosemeters te vergelijken weergegeven in een zogeheten Clarke error grid (Clarke et al., 1987). In deze grafiek staan de 'werkelijke' glucosewaarden (bepaald door de klinisch geaccepteerde gouden standaard referentiemethode) op de x-as, en de door de nieuwe techniek bepaalde waarden op de y-as. In het meest ideale geval, zullen alle meetpunten zich bevinden op de lijn x=y. In het geval van bloedglucosemeters is het gebied rondom de lijn x=y met een foutmarge van 20% het relevante gebied. In het Clarke-grid wordt dit gebied het 'A'-gebied genoemd. Het 'B'-gebied is buiten dit gebied; slechts een beperkt percentage (5 á 10%) van de metingen mogen in dit gebied vallen om de methode als klinisch acceptabel aan te merken. De gebieden 'C', 'D' en 'E' dienen geheel te worden vermeden.

Het Clarke-grid is daarmee een goede maat om te kijken of een bepaalde methode binnen klinisch acceptabele grenzen presteert in vergelijking met de klinische standaard. Nadat door middel van EROS de niet-specifieke glucosevariantie is verminderd, blijkt het mogelijk met behulp van PLS glucosewaarden binnen klinisch acceptabele grenzen te kunnen voorspellen. Als we proberen met PLS een voorspellend model te maken zonder EROS, krijgen we voorspellingen zoals weergegeven in figuur 3.

Doen we echter datzelfde mét een EROS-stap er tussen, dan krijgen we voorspellingen zoals in figuur 4 is weergegeven.

Het is duidelijk te zien dat het PLS-algoritme door het wegschalen van de niet-relevante biologische varianties, een veel betere inschatting kan maken van het glucosegehalte. In tabel 1 is samengevat hoe de bepaling verbetert.

ITEM	ZONDER EROS	MET EROS
N in 'A'	70 / 111 (63,1%)	97 / 111 (87,4)
N in 'B'	40 / 111 (36,0%)	13 /111 (11,7)
N in 'C'/'D'/'E'	1 / 111 (0,9%)	1 / 111 (0,9%)
Gemiddelde absolute fout	1.8 mM	1.0 mM
Gemiddelde relatieve fout	18,5%	10,5%
Pearson R2 *	0,638	0,911

* De Pearson R^2 op basis van dubbele kruisvalidatie

Tabel 1. kentallen van de glucosebepalingen zonder en met EROS

Conclusie

Diabetespatiënten zouden veel baat hebben bij een systeem waarmee bloedglucosewaarden kunnen worden bepaald zonder dat daarbij bloed hoeft te worden geprikt en dus kostbare meetstrippen/naalden niet meer nodig zijn. De aanpak van TNO laat zien, dat Ramanspectroscopie een techniek is waarmee dit in combinatie met slimme data-processing en -analysemethodieken mogelijk lijkt.

LITERATUUR

Clarke, W.L., Cox, D., Gonder-Frederick, L. A., Carter, W., & Pohl, S. L., 1987. Evaluating clinical accuracy of systems for self-monitoring of blood glucose. *Diabetes Care, 10*, 622–628.

Geladi, P., & Kowalski, B. R., 1986. Partial Least Squares Regression: A Tutorial. *Analytica Chimica Acta, 185*, 1–17.

Martens, H., & Naes, T., 1989. *Multivariate Calibration*. Chichester: John Wiley & Sons.

Scholtes-Timmerman, M. J., Bijlsma, S., Fokkert, M. J., Slingerland, R., & Veen, S.J.F. van, 2014. Raman Spectroscopy as a Promising Tool for Noninvasive Point-of-Care Glucose Monitoring. *Journal of Diabetes Science and Technology 8*(5), 974–979.

Smit, S., Breemen, M. J. van, & Hoefsloot, H. C. J., 2007. Assessing the statistical validity of proteomics based biomarkers. *Analytica Chimica Acta, 592*, 210–217.

Zhu, Y., Fearn, T., Samuel, D., Dhar, A., Hameed, O., Bown, S. G., & Lovat, L. B., 2008. Error removal by orthogonal subtraction (EROS): A customized pre-treatment for spectroscopic data. *Journal of Chemometrics, 22*, 130–134.

MAARTEN SCHOLTES-TIMMERMAN studeerde Analytische chemie aan de Vrije Universiteit in Amsterdam. Sinds 2008 is hij onderzoeker bij TNO waar hij zich richt op ontwikkeling van sensoren gebaseerd op spectroscopische analysetechnieken.
E-mail: <maarten.scholtes@tno.nl>

SABINA BIJLSMA is gepromoveerd in de Scheikunde aan de Universiteit van Amsterdam met een specialisatie in chemometrie en procesanalyse. Sinds 2000 is zij werkzaam bij TNO als onderzoeker. Zij houdt zich bezig met (multivariate) data-analyse & statistiek.
E-mail: <sabina.bijlsma@tno.nl>

JACK VOGELS is een analytisch chemicus die na een studie scheikunde aan de universiteit Utrecht is gepromoveerd aan de universiteit van Leiden op het gebied van multivariate data analyse van NMR-data. Sinds 1986 is hij werkzaam bij TNO als onderzoeker. Hij houdt zich tegenwoordig vooral bezig met de organisatie en analyse van complexe 'Big' data.
E-mail: <jack.vogels@tno.nl>

Jorien Voorhuis

HET BINNENBAAN-BUITENBAAN EFFECT OP DE 500 METER

SCHAATSEN EN HET BELANG VAN EEN GOEDE LOTING

De 500 meter schaatswedstrijd is een spel van honderdsten en soms zelfs duizendsten van seconden. Daarom is er onderzoek gedaan naar de vraag of het voordelig is om te starten in de binnenbaan of de buitenbaan, of dat dit geen verschil maakt. 'Ja' zegt het ene onderzoek, en dus moet de 500 meter twee keer gereden worden. 'Nee' zegt het andere onderzoek, en dan is één omloop voldoende. In dit artikel werpen we nieuw licht op deze vraag door de rol van de tegenstander mee te nemen.

Miriam Loois

Bijna elke schaatsfan zal zich de 500 meter kunnen herinneren op de Olympische Spelen in Sotsji. Toen Jan Smeekens over de finish kwam waande hij zich Olympisch kampioen. Maar nadat de tijd van Michel Mulder naar beneden werd bijgesteld ging deze er met het goud vandoor. Zijn voorsprong was 0,012 seconde na twee ritten. De discussie laaide op of het wel mogelijk is om op basis van duizendsten van seconden te zeggen wie de winnaar is. Als de mogelijke fout in de tijdwaarneming groter is dan het verschil in tijd, is het niet eerlijk om een van de twee als winnaar uit te roepen, en zou er eigenlijk een gedeelde eerste plek moeten zijn. Er is al eerder onderzoek gedaan naar de eerlijkheid van de opzet van de 500 meter. Dat onderzoek ging over de vraag of de 500 meter eenmaal of tweemaal gereden zou moeten worden. Nils Hjort toonde in 1994 aan dat het uitmaakt of een schaatser in de binnenbaan of in de buitenbaan start. Schaatsers die de laatste buitenbocht reden waren in het voordeel. Het idee hierachter is dat schaatsers in de laatste bocht zo'n hoge snelheid hebben, dat het makkelijker is om de buitenbocht te rijden dan de binnenbocht. Het gebeurt dan ook regelmatig dat een schaatser in de laatste binnenbocht 'uit de bocht vliegt' en in de helft van de tegenstander terecht komt. Op basis van dit onderzoek werd besloten om de 500 meter twee keer te rijden op de Olympische Spelen. Elke schaatser start een keer in de binnenbaan en een keer in de buitenbaan.

In 2010 toonden Richard Kamst, Gerard Kuper en Gerard Sierksma aan dat er na de introductie van de klapschaats geen significant verschil meer was. De klapschaats biedt betere grip op het ijs, waardoor schaatsers minder moeite hebben met de bochten. De Internationale Schaatsunie heeft dan ook besloten om vanaf 2018 nog slechts één omloop te houden op de Olympische 500 meter. Machiel Smit liet echter zien dat het overall klassement van een 500 meter omloop significant vaker wordt gewonnen door een schaatser die de laatste binnenbocht heeft. In de periode 2010 tot 2014 had de winnaar van een omloop in maar liefst drie kwart van de keer de laatste binnenbocht. In het eerste onderzoek is dus de laatste buitenbocht een voordeel, later is dit effect niet significant, en het derde onderzoek vindt juist een voordeel van de laatste binnenbocht. We werpen nieuw licht op deze op het eerste oog tegenstrijdige resultaten door de rol van de tegenstander mee te nemen. Ook laten we zien dat een goede loting het verschil tussen winst en verlies kan maken.

Het model

Langebaanschaatsen gebeurt op een 400-meter baan. De 500 meter is dus een volle ronde, plus nog 100 meter. Schaatsers rijden in tweetallen, waarbij de ene schaatser in de binnenbaan start, en de andere in de buitenbaan. Na de eerste bocht wisselen ze van baan, zodat ze beiden dezelfde afstand afleggen. De schaatser die start in de buitenbaan komt, als beide schaatsers even snel openen, achter zijn tegenstander de buitenbocht uit. Hij legt immers een langere afstand af. Op de kruising kan hij dan toerijden naar zijn tegenstander. In het algemeen wordt verondersteld dat dit kunnen toerijden naar de tegenstander een voordeel oplevert. Dit voordeel kan zowel mentaal van aard zijn, als fysiek, omdat de luchtweerstand afneemt als je achter je tegenstander rijdt. Tot nu toe is dit effect niet meegenomen in onderzoeken. In dit artikel laten we zien dat dit effect de tegenstrijdige conclusies van eerder onderzoek kan verklaren.

Als uitgangspunt van de analyse gebruiken we het model van Nils Hjort. Hij verklaart het verschil tussen de eerste en de tweede 500 meter van schaatser i op toernooi k, ΔY_{ik} aan de hand van het verschil in 100 meter tijd ΔX_{ik}, een toernooi specifieke factor c_k (bijvoorbeeld slechtere weersomstandigheden op dag twee) en de volgorde van starten W_{ik} (eerst binnen en dan buiten of andersom). Dit laatste effect was in zijn analyses nog significant, maar sinds de introductie van de klapschaats niet meer. Wij voegen aan dit model een extra variabele toe, die meet of je naar je tegenstander toe kunt rijden. We gaan er vanuit dat een schaatser naar zijn tegenstander toe kan rijden als hij de laatste binnenbocht rijdt, en zijn opening maximaal 0,3 seconde sneller en 0,1 seconde langzamer is dan de opening van zijn tegenstander. Als een schaatser een te grote voorsprong heeft, kruist hij over zijn tegenstander heen. Opent hij langzamer, dan ligt hij te ver achter om te kunnen profiteren van zijn tegenstander. Om uit te sluiten dat we een effect vinden omdat sneller openen op zich voordelig is, voegen we twee termen toe. De eerste term $\alpha_k(S_{2ik}-S_{1ik})$ meet het effect van een opening die sneller of slechts beperkt langzamer is dan de opening van de tegenstander. De tweede term, $\beta_k(I_{S_{2ik}=1} I_{W_{1ik}=1} - I_{S_{1ik}=1} I_{W_{ik}=-1})$ meet of de schaatser naar zijn tegenstander toe kan rijden. Kan hij dit in de eerste rit, omdat hij de eerste rit de laatste binnenbocht rijdt, en zijn opening in de goede range ligt, dan resulteert een effect $-\beta_k$. Kan hij dit in de twee rit, omdat hij de tweede rit de laatste binnenbocht heeft en zijn opening in de goede range ligt, resulteert een effect β_k.

We schatten daarom het volgende model:

$$\Delta Y_{ik}=c_k+b_k\Delta X_{ik}+d_kW_{ik}+\alpha_k(S_{2ik}-S_{1ik})+\beta_k(I_{S_{2ik}=1}I_{W_{ik}=1}-I_{S_{1ik}=1}I_{W_{ik}=-1})+\varepsilon_{ik}$$

- $h = 1$ (eerste rit) of 2 (tweede rit);
- ΔY_{ik} = Tweede 500 meter tijd min eerste 500 meter tijd, door schaatser *i* op toernooi *k;*
- ΔX_{ik} = Tweede 100 meter tijd min eerste 100 meter tijd, door schaatser *i* op toernooi *k;*
- W_{ik} = 1 (schaatser heeft de tweede rit de laatste binnenbocht) of -1 (schaatser heeft de eerste rit de laatste binnenbocht);
- S_{hik} = 1 als de 100 meter tijd in rit *h* maximaal 0,1 seconde langzamer, en maximaal 0,3 seconde sneller is dan de 100 meter tijd van zijn directe tegenstander, anders 0;
- ε_{ik} = Toernooi-specifieke normaal verdeelde storingsterm.

Dit model is geschat voor alle WK-afstanden, WK-sprints en Olympische Spelen tussen 2004 en 2015 uit de A-divisie. We nemen een toernooi alleen mee als minimaal 20 races beschikbaar zijn waarin zowel de 100 meter als de 500 meter tijd van de rijder en zijn tegenstander beschikbaar zijn. Dit resulteert in 27 toernooien. Het model wordt eerst geschat via lineaire regressie (ordinary least squares). Vervolgens worden outliers verwijderd via de methode van Nils Hjort (races met een T-waarde groter dan 2,75) en wordt het model opnieuw geschat. Dit resulteert in toernooispecifieke schattingen voor de parameters *c, b, d, α en β* en de bijbehorende standaardfouten. Deze toernooispecifieke schattingen worden gewogen met de variantie geaggregeerd tot één schatting. Voor de parameter *b* gebeurt dat op de volgende manier (en voor de andere parameters analoog):

$$\hat{b}=\frac{\Sigma_k\frac{\widehat{b_k}}{\widehat{\sigma_{bk}}^2}}{\Sigma_k\frac{1}{\widehat{\sigma_{bk}}^2}}$$

$$\widehat{\sigma_b}^2=\frac{1}{\Sigma_k\frac{1}{\widehat{\sigma_b}^2}}$$

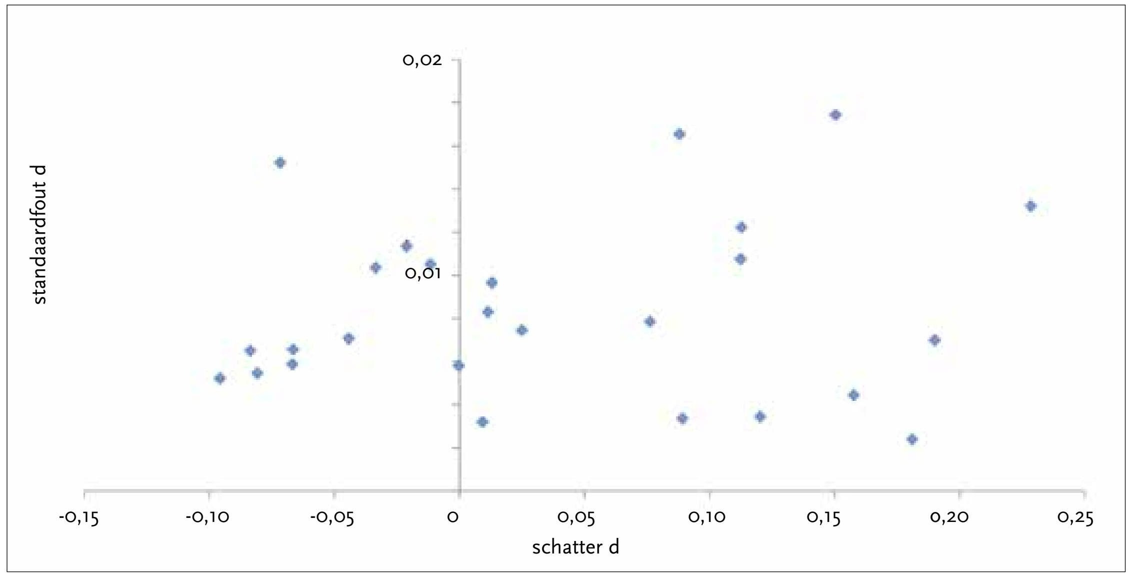
Uitkomsten

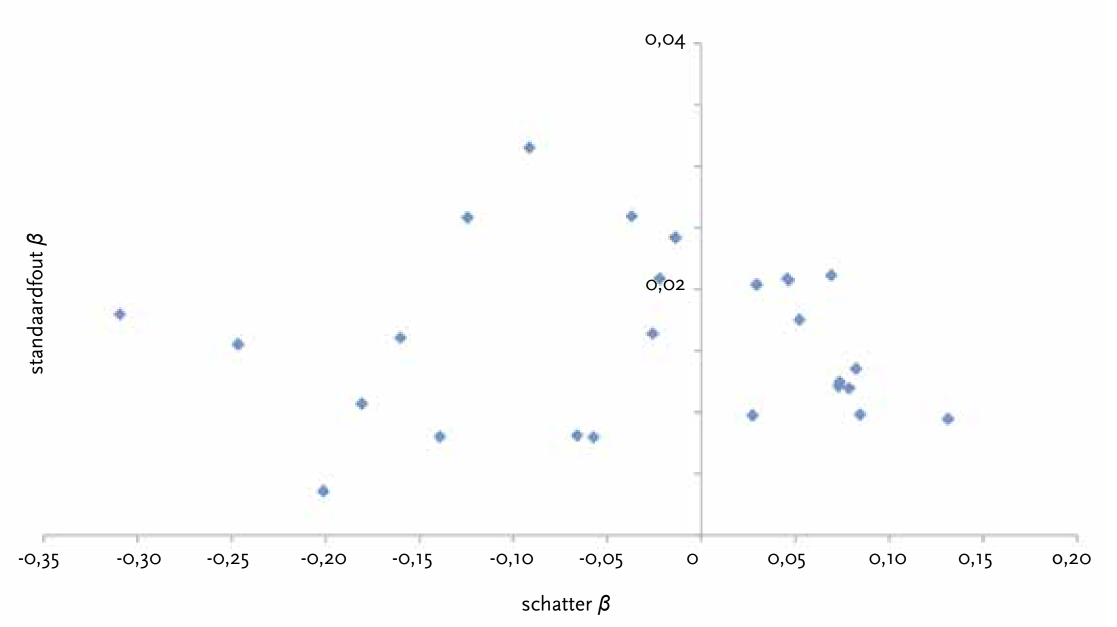
De uitkomsten zijn samengevat in tabel 1. We zien dat net als bij de andere onderzoeken, de 100 meter tijd een (zeer) significante factor is. Een opening die 0,1 seconde sneller is, resulteert in een 0,15 seconde snellere eindtijd. De schaatser profiteert dus ook in de ronde na de eerste 100 meter van zijn hogere snelheid. Het effect van een opening tussen de 0,3 seconde sneller en 0,1 seconde langzamer dan de tegenstander wijkt niet significant af van 0; de standaardfout is twee keer zo groot als de geschatte waarde.

Dan de parameters waar het om draait. Het binnenbaan-buitenbaan effect, en het effect van het kunnen toerijden naar je tegenstander. Hier vinden we, net als klapschaats en los van de tegenstander, dat het nadelig is om de laatste binnenbocht te hebben. Het geschatte effect *d* is 0,05 seconde, en wijkt meer dan twee standaardfouten af van 0. Ook het effect van het kunnen toerijden naar je tegenstander als je de laatste binnenbocht hebt is significant. Dit levert een ongeveer even grote winst op, *β*=-0,05. Het nadeel van de laatste binnenbocht wordt dus ongeveer opgeheven door het kunnen toerijden naar je tegenstander. Verklaart dit de tegenstrijdige onderzoeken? Ja en nee. Het verklaart waarom het onderzoek inclusief klapschaats geen significant effect vond. Het netto effect van kunnen toerijden naar je tegenstander, en het nadeel van de laatste binnenbocht is ongeveer 0. Echter, om te kunnen verklaren dat het overall klassement van een omloop significant vaker wordt gewonnen door een

PARAMETER	SCHATTING	STANDAARDFOUT
b (100 meter tijd)	1,480	0,070
d (effect laatste binnenbocht)	0,049	0,015
α (effect opening max 0,3 sneller en 0,1 langzamer dan tegenstander)	-0,008	0,017
β (effect kunnen toerijden naar je tegenstander in de laatste binnenbocht)	-0,049	0,021
d+*β* (netto effect van toerijden naar je tegenstanden en de laatste binnenbocht)	-0,001	0,011

Tabel 1. Toernooispecifieke schattingen voor de parameters *c, b, d, α en β* en de bijbehorende standaardfouten





Figuur 1. De toernooispecifieke schattingen voor d en *β* en de bijbehorende standaardfout

ONDERGRENS	BOVENGRENS	d	σ_d	β	σ_β	d+β	$\sigma_{(d+\beta)}$
-0,3	0,1	0,049	0,015	-0,049	0,021	-0,001	0,011
-0,4	0,1	0,061	0,016	-0,065	0,021	-0,003	0,010
-0,3	0,0	0,009	0,012	0,021	0,021	0,027	0,015
-0,2	0,1	0,043	0,013	-0,057	0,021	-0,011	0,012
-0,3	0,2	0,045	0,021	-0,032	0,027	0,008	0,010

Tabel 2. Resulaten van de gevoeligheidsanalyse

schaatser die de laatste binnenbocht heeft, zou het voordeel van het toerijden groter moeten zijn dan het nadeel van de laatste binnenbocht. Toch is het aannemelijk dat het kunnen toerijden naar de tegenstander de oorzaak is van het gevonden voordeel van de laatste binnenbocht. Het is moeilijk om op basis van een 100 meter tijd precies te bepalen of je voordeel hebt van je tegenstander. De 0,3 en 0,1 seconden zijn een inschatting. Gemiddeld is dit effect 0,05 seconde, maar het zal uitmaken of je 1, 2 of 5 meter achter je tegenstander rijdt. Bij de 'ideale' tegenstander kan het effect groter zijn dan 0,05 seconde, en net het verschil maken tussen winst of verlies.

Dit model kan dus aannemelijk maken waarom de 500 meter vaak wordt gewonnen door iemand die de laatste binnenbocht heeft, maar het is geen statistisch bewijs. We kunnen echter wel een andere conclusie trekken, die wel ook statistisch significant is: Een goede loting is van cruciaal belang. Als een schaatser de laatste binnenbocht heeft, maar niet naar zijn tegenstander toe kan rijden, heeft hij een nadeel van 0,05 seconde ten opzichte van iemand die wel naar zijn tegenstander toe kan rijden. Op de 500 meter gaat het om honderdsten van seconden, dus de juiste tegenstander kan het verschil tussen winst en verlies betekenen.

Gevoeligheidsanalyse

De 0,3 en de 0,1 seconde zijn een inschatting, maar zijn moeilijk objectief vast te stellen. Daarom voeren we een gevoeligheidsanalyse uit, waarin we deze parameters wat verhogen of verlagen. Zie tabel 2.

We zien dat bij de meeste gevoeligheidsanalyses het beeld gelijk blijft. Alleen de combinatie -0,3 en 0 geeft geen voordeel van het kunnen toerijden naar je tegenstander. De combinatie -0,2 en 0,1 is het interessantst. Hier is het voordeel van het toerijden naar je tegenstander groter dan het nadeel van de laatste binnenbocht. Het netto effect is een winst van ruim 0,01 seconde. Dit kan dus verklaren waarom een 500 meter omloop vaker gewonnen wordt door een rijder die de laatste binnenbocht heeft. Het effect is echter ongeveer even groot als de standaardfout, en dus statistisch niet significant.

Conclusies

Moet de 500 meter nu wel of niet twee keer gereden worden? Als je alleen het binnenbaan-buitenbaan effect modelleert, zonder tegenstander, is hier een eenduidig antwoord op te geven. Is het effect significant, dan moet de 500 meter twee keer gereden worden om een zo eerlijk mogelijk toernooi te krijgen, anders niet. Maar bij dit model ligt dat wat gecompliceerder. Het effect van de tegenstander is aanzienlijk. Maar de tegenstander is niet te sturen. Je kunt ervoor zorgen dat elke rijder een keer in de binnenbaan start en een keer in de buitenbaan start. Maar je kunt er niet voor zorgen dat de invloed van de tegenstander bij iedereen precies even groot is. De loting heeft grote invloed. De oplossing zou zijn om zonder tegenstanders te rijden. Maar dat maakt de sport een stuk minder aantrekkelijk om naar te kijken. De meest praktische oplossing lijkt om de 500 meter wel twee keer te rijden, en daarbij ervoor te zorgen dat de paren zo ingedeeld worden dat de openingstijden redelijk aan elkaar gewaagd zijn. Dat is in de praktijk vaak het geval, maar het lijkt erop dat we zullen moeten accepteren dat een bepaalde mate van geluk altijd een rol zal spelen.

Dank aan Gerard Sierksma en Gerard Kuper voor hun kritische blik en nuttige tips.

LITERATUUR

Hjort, N., 1994. *Should the Olympic sprint skaters run 500 meter twice?*. Oslo: Institute of Mathematics, University of Oslo.

Kamst, R., Kuper, G. H., & Sierksma, G., 2010. The Olympic 500m speed skating; the inner–outer lane difference. *Statistica Neerlandica, 64*(2010)4, 448–459.

Smit, M., 2014. *Analyse van de 500 meter zeges vanuit de binnen- en buitenbaan*. http://www.schaatsstatistieken.nl/analyse_500_meter_juni2014.pdf

MIRIAM LOOIS heeft de Master Theoretische Natuurkunde gevolgd aan de Universiteit Utrecht en de Master Actuarial Science and Financial Mathematics aan de Universiteit van Amsterdam. Ze werkt nu bij pensioenuitvoerder PGGM en houdt zich bezig met Asset Liability Management en Big data. Als hobby schrijft ze daarnaast artikelen over statistiek op miriamenstatistiek.wordpress.com.
E-mail: <miriamloois@gmail.com>

SPELEN TOPSPORTERS NASH?

HAROLD HOUBA

Veel sporten kennen elementen die zich lenen voor een speltheoretische beschrijving, onder andere schaken, bridge, poker, voetbal en tennis. Voor de meeste van deze spelen geldt dat een volledig speltheoretische analyse niet haalbaar is omdat het aantal mogelijke bordposities in het schaakspel of het aantal mogelijke handen in kaartspelen gigantisch is. Dit neemt niet weg dat bepaalde deelsituaties wel degelijk aan een grondige speltheoretische analyse onderworpen kunnen worden: penalty's in voetbal en de service en eerste return in tennis. Daarnaast maakt de grote beschikbaarheid van wedstrijden op YouTube het mogelijk om zelf data te verzamelen om te onderzoeken of de topsporters zich gedragen in overeenstemming met de speltheorie.

Penalty's en services worden binnen de speltheorie vaak aangehaald als praktijkvoorbeelden. De penaltynemer heeft de keuze om naar de linker- of rechterhoek van de keeper te schieten en de serveerder kan op de forehand of backhand serveren. Gezien de snelheid van de bal, is de ontvanger niet in staat adequaat te reageren op de richting van de bal en rest hem of haar niets anders dan blind op links of rechts te anticiperen. Binnen de speltheorie staat dit bekend als *matching pennies*. An en Bob kiezen ieder tegelijkertijd en onafhankelijk van elkaar kop of munt. Als beiden hetzelfde hebben gekozen, dan wint An één punt, anders wint Bob één punt. Tabel 1 geeft deze situatie weer, waarbij An een rij kiest, Bob een kolom, en de cijfers 1,0 geven aan dat An één punt wint en Bob geen, etc. Een hele tenniswedstrijd is een reeks van matching pennies na elkaar.

	KOP	MUNT
KOP	1,0	0,1
MUNT	0,1	1,0

Tabel 1. Matching pennies

Kop of munt

Matching pennies is de canonieke vorm voor situaties waarin de ene speler (An) baat heeft om het gedrag van de ander (Bob) correct te voorspellen en de ander er juist baat bij heeft dat dit niet gebeurt, bijvoorbeeld in inspectie situaties van grenscontrole. Matching pennies is een constante-som-spel waarbij in iedere cel van het spel één punt tussen de spelers wordt verdeeld. De minimax stelling, die de wiskundige John von Neumann al in 1928 bewees, zegt dat in ieder constante-som-spel minimax en maximin strategieën samen vallen. Maximin strategieën zijn met behulp van lineair programmeren te berekenen (Taha, 2011). In matching pennies kiezen beide spelers met gelijke kansen kop of munt. De minimax interpretatie is dat de andere speler op een verwachting van maximaal een ½ punt wordt gehouden. De maximin interpretatie is dat iedere speler zichzelf een verwachte uitbetaling van ½ punt kan garanderen. In matching pennies spelen risicohoudingen van de spelers geen enkele rol en kunnen buiten beschouwing worden gelaten.[1] De intuïtie is dat gelijke kansen de tegenstander in het ongewisse laat wat er gaat gebeuren, en dat bij herhaaldelijk spelen de ene speler geen systematisch voordeel op de andere speler kan behalen. Binnen de speltheorie spreekt men tegenwoordig van een Nash-evenwicht.[2] Maximin strategieën in constante-som-spelen zijn equivalent aan Nash-evenwichten en kunnen als normatieve aanbeveling worden beschouwd.

De experimentele economie onderzoekt welke economische keuzen mensen (meestal studenten) maken en hoe deze keuzen te verklaren zijn, waarbij ook veel spel-situaties worden getest. In matching pennies wordt redelijk volgens het unieke Nash-evenwicht gespeeld (Binmore, Swierzbinski en Proulx, 2001), maar in andere spelen kiezen mensen vaak anders dan Nash-evenwichten voorspellen en faalt dit evenwichtsconcept als beschrijvende theorie. Bijvoorbeeld zoals in tabel 2 die een asymmetrische variant van matching pennies representeert (matching pennies is equivalent met 80 in plaats van 320), waarin het constante-som-karakter is losgelaten.

In dit geval kiest 96% van de deelnemers in een experiment die een rij kiezen voor Kop, en kiest 84% de kolom Munt (Goeree en Holt, 2001). Als beschrijvende theorie is het Nash-evenwicht vaak niet overtuigend. Een reden hiervoor kan zijn dat het evenwicht hier niet intuïtief is: iedere speler kiest zijn kansen zo dat de *andere* speler indifferent wordt tussen kop en munt, waarbij de eigen punten (vooral de 320) wordt genegeerd. In Tabel 2 zal volgens Nash de rij speler An nog steeds met gelijke kans kop of munt kiezen om Bob indifferent te laten zijn; Bob zal met kans 12,5% Kop kiezen en met 87,5% Munt om An indifferent te laten zijn.[3] Evenwichtsconcepten als quantal response evenwicht en Noisy Introspection kunnen de observaties wel verklaren en zijn beide potentiële opvolgers van het Nash-evenwicht als beschrijvende theorie (Goeree en Holt, 2001).

John Nash ging ervan uit dat introspectie door de spelers tot het door hem geïntroduceerde evenwichtsconcept zou leiden. Men kan zich afvragen of de spelers een evenwicht moeten leren spelen. Links of rechts houden in het verkeer kent twee Nash-evenwichten, of iedereen houdt links zoals in Groot Brittannië, of iedereen houdt rechts, en kinderen worden opgevoed volgens het gangbare Nash-evenwicht. Dit maakt topsporters interessant omdat zij veel trainings- en wedstrijdervaring hebben. En volgens de intuïtie van matching pennies lijken degenen die niet het Nash-evenwicht volgen in het nadeel om de wereldtop te bereiken. Spelen topsporters maximin strategieën c.q. een Nash-evenwicht bij penalty's en services?

Die vraag is door Ignacio Palacios-Huerta, Mark Walker en John Wooders onderzocht (Palacios-Huerta, 2003, Walker en Wooders, 2001)’ maar alvorens naar de resultaten kijken is het interessant enige statistische aspecten voor het voetlicht te brengen. In een experimentele setting van matching pennies wordt de puntenverdeling aan de deelnemers bekend gemaakt en is deze ook bij de statisticus bekend. De data van de individuele keuzen kunnen vervolgens met standaardmethoden worden getoetst met als nulhypothese of zij overeenkomen met de Nash-evenwichtskansen. Echter bij penalty's en bij services is de succeskans van iedere combinatie van keuzen onbekend. Hoeveel kans heeft de penaltynemer op succes als de keeper naar de goede hoek gaat? Wat is de kans dat de serveerder na de eerste geslaagde return de rally die volgt wint? Succeskansen zijn niet 100% of 0% en de statisticus kent deze kansen niet. De statisticus dient uit te gaan van tabel 3.

	KOP	MUNT
KOP	320,40	40,80
MUNT	40,80	80,40

Tabel 2. Asymmetrische variant van matching pennies

	LINKERHOEK	RECHTERHOEK
LINKERHOEK	P_{LL}, 1- P_{LL}	P_{LR}, 1- P_{LR}
RECHTERHOEK	P_{RL}, 1- P_{RL}	P_{RR}, 1- P_{RR}

Tabel 3. Theoretische succeskansen bij penalty's

Bij penalty's, kan worden vastgesteld of de bal doel trof of niet, zodat eenduidig interpreteerbare data over succes kunnen worden verkregen waaruit succeskansen zijn te schatten. In tennis interpreteren Mark Walker en John Wooders succes als degene die de rally uiteindelijk wint en zo verkrijgen zij data om succeskansen te kunnen schatten. Het simultaan schatten van zowel succeskansen als het toetsen van gedrag volgens maximin strategieën die uit de schattingen volgen is bepaald geen sinecure.

Volgens Ignacio Palacios-Huerta treffen 80,1% van alle penalty's doel en is er voldoende statistische onderbouwing voor de bewering dat profvoetballers penalty's nemen volgens maximin strategieën. Zijn schattingen leverden tabel 4 met succeskansen op. De bijbehorende maximin strategieën laten de penaltynemer met kans 38,5% in de linkerhoek schieten en de keeper met kans 42,0% naar de dezelfde hoek gaan. De geobserveerde frequenties waren 40,0%, respectievelijk, 42,3%.

Ook Mark Walker en John Wooders concluderen dat er voldoende statistische onderbouwing is voor de bewering dat proftennissers openen volgens maximin strategieën. Zij geven geen tabel met succeskansen. Interessant is dat zij rapporteren dat in reeksen van services *serial independence* wordt verworpen. Het is bekend dat het voor mensen moeilijk is een willekeurig lijkende reeks van kop of munt te verzinnen. Waarschijnlijk is dit ook voor topsporters in een tennismatch het geval.

Samenvattend, uit experimenteel gedrag blijkt dat onervaren spelers vaak van het Nash-evenwicht afwijken. Het Nash-evenwichtsconcept als beschrijvende theorie staat hierdoor onder druk. Hier tegenover staat dat er statistische onderbouwing is voor de bewering dat ervaren topsporters maximin strategieën volgen cq. een Nash-evenwicht spelen. De besproken literatuur laat zien welke boeiende analyses de experimentele economie met de grote beschikbaarheid van sportdata kan doen. Er breken boeiende tijden aan voor enerzijds topsporters en profclubs om met deze methoden hun prestaties te verbeteren, en anderzijds voor docenten, studenten en iedere andere belangstellende om zelf data te verzamelen en aan de slag te gaan om hypothesen op te stellen en te toetsen. Zo is de linker- of rechterhoek kiezen misschien een te eenvoudige weergave van penalty's en kunnen uitbreidingen worden overwogen waarin hoog versus laag schieten of hard met veel risico versus minder hard met minder risico worden geanalyseerd. Alle artikelen zijn via www.jstor.org gratis toegankelijk en ik wens iedere geïnteresseerde veel plezier met verdere studie en het uitvoeren van eigen experimenten.

	LINKERHOEK	RECHTERHOEK
LINKERHOEK	58,3% , 41,7%	95,0% , 5,0%
RECHTERHOEK	92,9% , 7,1%	69,9% , 30,1%

Tabel 4. Geschatte succeskansen bij penalty's

Noten

[1] Dit geldt voor zowel expected utility theory als prospect theory (Kahneman en Tversky, 1979).

[2] John Nash (1928-2015) ontving in 1994 een Nobelprijs Economie voor zijn in 1950 gepubliceerde evenwichtsconcept voor spelen in normale vorm, waaraan later zijn naam werd verbonden (Nash, 1950).

[3] Hier wordt risico neutraliteit verondersteld, maar andere risico houdingen veranderen de evenwichtskansen.

Literatuur

Binmore, K., Swierzbinski, J., & Proulx, C., 2001. Does Minimax Work? An Experimental Study. *The Economic Journal*, 111, 445–464.

Goeree, J., & Holt, C., 2001. Ten Little Treasures of Game Theory and Ten Intuitive Contradictions. *American Economic Review*, 91, 1402–1422.

Kahneman, D., & Tversky, A., 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47, p. 263–292.

Nash, J., 1950. Equilibrium Points in N-person Games. *Proceedings of the National Academy of Sciences*, 36, 48–49.

Palacios-Huerta. I., 2003. Professionals Play Minimax. *Review of Economic Studies*, 70, 395–415.

Taha, H., 2011. *Operations Research: an Introduction* (9th ed.). Prentice Hall.

Walker. M., & Wooders, J., 2001. Minimax Play at Wimbledon. *American Economic Review*, 91, 1521–1538.

Harold Houba is UHD Wiskundige Economie bij de Afdeling Econometrie en OR van de Vrije Universiteit.
E-mail: <harold.houba@vu.nl>

HEB JIJ DAT OOK VAN JE BROER OF ZUS?

Bloedtransfusies zijn onmisbaar en levensreddend, maar kunnen ook nadelige gevolgen voor de ontvanger met zich meebrengen. Risico's kunnen geminimaliseerd worden door de best passende donors te selecteren. Elementaire statistiek in combinatie met OR biedt hiervoor hulp.

Joost van Sambeeck, Mart Janssen, Peter Ligthart, Wim de Kort & Nico van Dijk

In Nederland wordt bij elke vrouw die zwanger is rond de 12de week bloed afgenomen en onderzocht op bloedgroepantistoffen. Als blijkt dat er antistoffen in haar bloed aanwezig zijn, kan dit voor problemen zorgen tijdens of vlak na de zwangerschap. Deze antistoffen kunnen namelijk in de bloedsomloop van de foetus terecht komen en daar leiden tot afbraak van rode bloedcellen. Het bekendste voorbeeld hiervan is 'het rhesuskindje', waarbij een vrouw met een rhesus-D (RhD) negatieve bloedgroep en met RhD-antistoffen zwanger is van een RhD-positief kind.

Behalve dat het bloed van een zwangere vrouw onderzocht wordt op antistoffen, wordt er ook vastgesteld of ze voor bepaalde bloedgroepen negatief is. Het kan namelijk voorkomen dat er tijdens een bevalling zoveel bloed verloren gaat dat het noodzakelijk is om een bloedtransfusie toe te dienen. Bij voorkeur wordt er dan een donor geselecteerd die negatief is voor alle bloedgroepen waarvoor de vrouw ook negatief is. In dat geval zullen er namelijk geen antistoffen gevormd worden die bij een volgende zwangerschap tot problemen zouden kunnen leiden. Dergelijke donors zijn echter schaars, omdat het met zo'n 300 verschillende bloedgroepen praktisch gezien onmogelijk is om met elke bloedgroep rekening te houden. Daarom wordt er bij een bloedtransfusie tijdens een bevalling alleen gematcht op een beperkt aantal bloedgroepen. Toch blijft het voor sommige bloedgroepprofielen lastig om voldoende geschikte donors te vinden, zelfs met dit restrictieve matchingsbeleid.

Omdat bloedgroepen erfelijk zijn, is de kans op het vinden van een vergelijkbaar bloedgroepprofiel bij naaste familieleden veel groter dan bij een willekeurig persoon. Voor selectieve werving van potentiële donors is het dus interessant om precies te weten hoe groot deze kans als functie van een familierelatie is. Ter illustratie zal daarom de volgende vraag beantwoord worden:

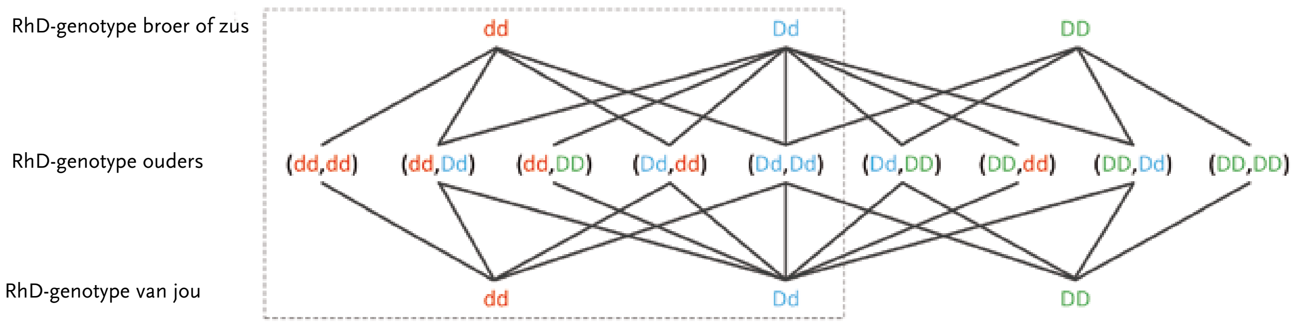
Wat is de kans dat je met een RhD-negatieve broer of zus zelf ook RhD-negatief bent?

Een ogenschijnlijk eenvoudige vraag die gemakkelijk met de regel van Bayes opgelost kan worden. De a priori kansen echter zijn niet bekend. Deze dienen berekend te worden door het bepalen van stationaire kansen. In dit artikel wordt daartoe een methodiek ontwikkeld die elementaire statistiek en OR combineert. Door de generieke formulering is het mogelijk om hiermee ook een breder scala aan gelijksoortige vragen te beantwoorden.

Overerving van genen

Voordat we de kans op een RhD-negatieve broer of zus gaan berekenen is het belangrijk om eerst naar de erfelijkheid van bloedgroepen te kijken. Een belangrijk punt daarbij is het verschil tussen de begrippen genotype en fenotype. Een bloedgroep genotype bestaat uit een combinatie van twee genen die samen de uiteindelijke bloedgroep, het fenotype, bepalen. De genen erft een kind van beide ouders, één van de vader en één van de moeder. De RhD-bloedgroep bijvoorbeeld, wordt bepaald door twee genen, namelijk *D* (grote D) en *d* (kleine d). Hierbij bestaat het d-gen feitelijk niet, maar staat dit voor het ontbreken van het D-gen. De genen *D* en *d* kunnen leiden tot drie verschillende genotypen: *dd*, *Dd* en *DD*. Iemand met genotype *dd* heeft fenotype *d* en wordt RhD-negatief genoemd. Iemand met genotype *Dd* of *DD* heeft fenotype *D* en wordt RhD-positief genoemd.

Aangezien de ouders elk twee genen hebben die eenzelfde kans hebben om doorgegeven te worden (uitgaande van Mendeliaanse overerving), kunnen er uit een ouderpaar vier verschillende gen-combinaties (genotypen) ontstaan, elk met kans één vierde. In het onderste gedeelte van figuur 1 is de overerving van genen van ouderpaar naar kind voor de RhD-bloedgroep weergegeven.



Figuur 1. Kansboom ter beantwoording van de vraag 'wat is de kans dat je met een RhD-negatieve broer of zus zelf ook RhD-negatief bent?'

Stationaire verdeling van genotypen

In het algemeen is de verdeling van genotypen in een populatie niet eenvoudig te bepalen, maar kan wel verkregen worden op basis van de fenotypen verdeling (welke wel eenvoudig te bepalen is) en de manier waarop genen doorgegeven worden van ouder naar kind. In de literatuur wordt hiervoor vaak een kwadratische stochastische operator (KSO) gebruikt.[1] Dit is pas in zeer beperkte mate toegepast bij het bepalen van de kans op overerving van bloedgroepen.[2,3] In deze context is de KSO een overervingsmatrix Q, met Q_{ijk} de kans dat twee ouders met genotypen i en j een kind krijgen met genotype k. De verdeling van genotypen in de populatie wordt beschreven door de vector $\boldsymbol{x}=(x_1,\ldots,x_{|G|})$, waarbij x_i de frequentie van voorkomen van genotype i is. Om de verdeling van genotypen te bepalen zal het volgende stelsel van kwadratische vergelijkingen opgelost moeten worden (zie ook figuur 3a):

$$x_k=\mathbf{x}^{\mathrm{T}}Q_k\mathbf{x} \quad \forall\, k \in G$$

Voor de RhD-bloedgroep is het mogelijk om dit stelsel van kwadratische vergelijkingen, onder de voorwaarde dat de som van de genotypefrequenties optelt tot 1, analytisch op te lossen (zie figuur 2). Ook weten we dan dat 15% van de populatie een RhD-negatief fenotype heeft. Aangezien *dd* het enige genotype is dat tot dit fenotype leidt is de genotypefrequentie gelijk aan de fenotypefrequentie: $x_{dd}=0{,}15$. Dit geeft voor de Nederlandse bevolking de volgende stationaire verdeling van RhD-genotypen: $x^*_{dd}=0{,}15$, $x^*_{Dd}=0{,}47$ en $x^*_{DD}=0{,}38$.

Voor complexere bloedgroepen kunnen we het stelsel van kwadratische vergelijkingen oplossen door middel van een iteratieve benadering:

$$x_k^{(n+1)}=\sum_{i,j\in G} Q_{ijk}x_i^{(n)}x_j^{(n)} \quad \forall k \in G.$$

Als $n\to\infty$ zal $\boldsymbol{x}$ convergeren naar een stationaire verdeling $\boldsymbol{x}^*$. Hierbij is het belangrijk om een goede initiële verdeling $\boldsymbol{x}^{(0)}$ te kiezen, omdat er meerdere stationaire oplossingen kunnen zijn. De fenotypefrequenties gebruiken we om deze initiële verdeling te bepalen.

Conditioneren op de donor

We weten de kans dat een kind een bepaald genotype heeft, gegeven de genotypen van de ouders (Q_{ijk}). We zijn echter ook geïnteresseerd in de kans dat een ouderpaar een bepaalde genotype combinatie heeft, gegeven het genotype van het kind. Dit is een klassiek voorbeeld van de regel van Bayes. Eerst bekijken we welke ouderparen een kind met een bepaald genotype kunnen hebben. Vervolgens berekenen we door middel van de stationaire verdeling $\boldsymbol{x}^*$ en de regel van Bayes de voorwaardelijke kans op het genotype van de ouders (zie figuur 3b).

Kans genotype broer gegeven genotype donor

Inmiddels zijn alle ingrediënten aanwezig om de vraag te beantwoorden: 'Wat is de kans dat je met een RhD-negatieve broer of zus zelf ook RhD-negatief bent?'. Dit is het eenvoudigst uit te leggen door middel van het omkaderde gedeelte in figuur 1. Eerst bepalen we de kans dat een RhD-negatieve donor een ouderpaar met een bepaalde genotypeverdeling heeft. Vervolgens wordt voor elk ouderpaar de kans dat ze een RhD-negatief kind krijgen bekeken. Door voor elk ouderpaar deze twee kansen te vermenigvuldigen en te sommeren over alle mogelijke ouderparen wordt het getal wat antwoord geeft op de

$$\begin{cases} x_{dd} &= x_{dd}^2 + x_{dd}x_{Dd} + \frac{1}{4}x_{Dd}^2 \\ x_{Dd} &= x_{dd}x_{Dd} + 2x_{dd}x_{DD} + \frac{1}{2}x_{Dd}^2 + x_{Dd}x_{DD} \\ x_{DD} &= \frac{1}{4}x_{Dd}^2 + x_{Dd}x_{DD} + x_{DD}^2 \\ 1 &= x_{dd} + x_{Dd} + x_{DD} \end{cases} \quad \Rightarrow \quad \begin{cases} x_{dd}^* &= x_{dd} \\ x_{Dd}^* &= 2\left(\sqrt{x_{dd}} - x_{dd}\right) \\ x_{DD}^* &= 1 + x_{dd} - 2\sqrt{x_{dd}} \end{cases}$$

Figuur 2. Stationaire verdeling van genotypen voor de RhD-bloedgroep

$$\begin{aligned} x_k &= P(\text{kind } k) \\ &= \sum_{i,j\in G} P(\text{kind } k \mid \text{vader } i \text{ en moeder } j)\cdot P(\text{vader } i \text{ en moeder } j) \\ &= \sum_{i,j\in G} P(\text{kind } k \mid \text{vader } i \text{ en moeder } j)\cdot P(\text{vader } i)\cdot P(\text{moeder } j) \\ &= \sum_{i,j\in G} Q_{ijk}x_ix_j \end{aligned}$$

Stationaire verdeling van genotypen

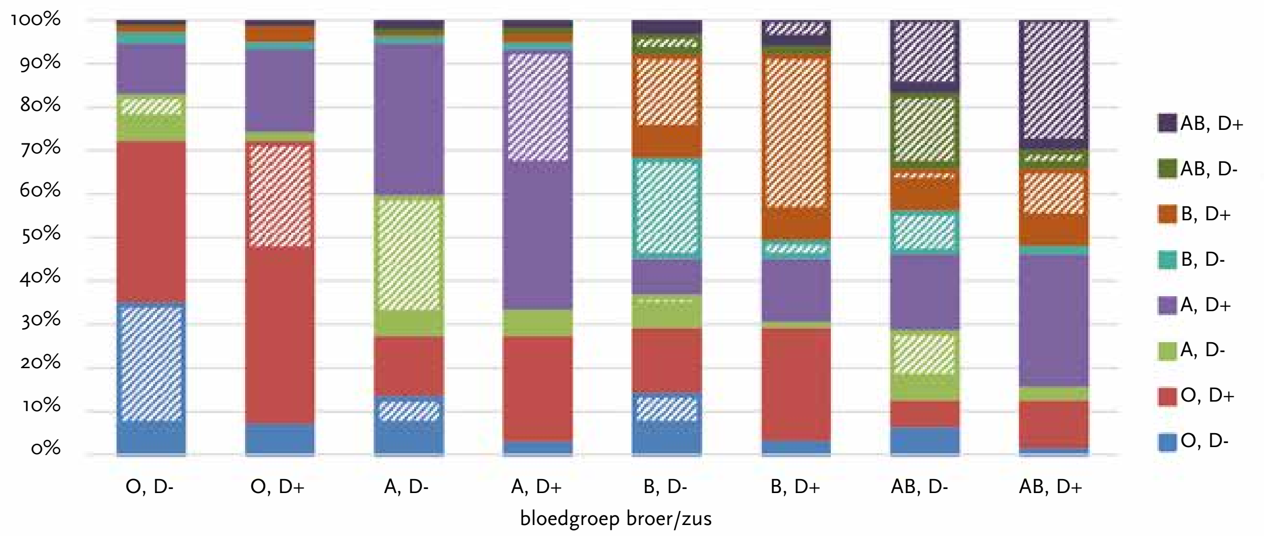
$$\begin{aligned} P(\text{vader } i \text{ en moeder } j \mid \text{kind } k) &= \frac{P(\text{vader } i, \text{moeder } j \text{ en kind } k)}{P(\text{kind } k)} \\ &= \frac{Q_{ijk}x_i^*x_j^*}{x_k^*} \end{aligned}$$

Regel van Bayes

$$\begin{aligned} P(\text{jij } l \mid \text{broer of zus } k) &= \sum_{i,j\in G} P(\text{vader } i \text{ en moeder } j \mid \text{kind } k)\cdot P(\text{kind } l \mid \text{vader } i \text{ en moeder } j) \\ &= \sum_{i,j\in G} \frac{Q_{ijk}x_i^*x_j^*}{x_k^*}\cdot Q_{ijl} \\ &= \frac{1}{x_k^*}\sum_{i,j\in G} Q_{ijk}Q_{ijl}x_i^*x_j^* \end{aligned}$$

Kans dat jij een bepaald genotype hebt, gegeven het genotype van je broer of zus

Figuur 3a, 3b, 3c. Generieke methodiek ter bepaling van de kans dat de broer/zus van een donor een bepaald genotype heeft



Figuur 4. Kans dat je een bepaalde (ABO, RhD) bloedgroep hebt, gegeven de bloedgroep van je broer of zus. De massieve balken zijn kleiner of gelijk aan de frequentie van een bloedgroepprofiel in de Nederlandse bevolking. De gearceerde balken geven de verhoogde kans op het vinden van een bepaald bloedgroepprofiel weer

vraag gevonden: 0,481. In vergelijking met het voorkomen van het RhD-negatieve fenotype in de Nederlandse populatie (0,15) is dit dus een factor 3 hoger.

Generalisatie

Door de generieke methodiek is het mogelijk om ook veel meer gelijksoortige vragen te beantwoorden. Ter illustratie is in figuur 4 de combinatie tussen de ABO en RhD-bloedgroep weergegeven. Daarnaast is het ook mogelijk om, met behulp van de ontwikkelde methodiek, de kans dat een broer of zus genotype l heeft – gegeven dat een donor genotype k heeft, te beschrijven (zie figuur 3c). Wat opvalt, is dat er in deze methodiek alleen gebruik gemaakt wordt van een bepaalde set van genotypen G zonder specificatie van de precieze bloedgroep. Het is daarom mogelijk om deze methodiek ook voor andere bloedgroepen, of zelfs combinaties van bloedgroepen te gebruiken. Het enige wat zal veranderen is de set van mogelijke genotypen en de stationaire verdeling $\boldsymbol{x}^*$ van deze genotypen.

Het meest voor de hand liggend is om ervan uit te gaan dat genen met een kans van een half worden doorgegeven van ouder op kind. Echter, we zien in figuur 3c dat de overervingsmatrix Q_{ijk} ook een input-parameter voor het model is, net als de genotypeverzameling. Dat betekent dat de methodiek dus ook gebruikt kan worden voor genen waarvoor Mendeliaanse overerving niet geldt, wat voor sommige bloedgroepen inderdaad ook het geval is.

Matchen bij zwangerschap

Bij een bloedtransfusie wordt slechts voor een beperkt aantal bloedgroepen gematcht. In het geval van bloedtransfusies bij de bevalling zou je graag een donor willen die negatief is voor aantal van deze bloedgroepen (O, RhD-neg, RhE-neg, K-neg). Met behulp van de beschreven methodiek is het eenvoudig om te berekenen wat de kans is dat een broer of zus van die donor ook negatief is voor al deze bloedgroepen: dat is voor deze combinatie van bloedgroepen 0,3340, wat een factor 5 hoger is dan de kans op de aanwezigheid van deze specifieke combinatie in de algemene bevolking. Ook voor willekeurig andere combinaties van bloedgroepen kunnen deze kansen eenvoudig worden berekend.

Conclusie

Bovenstaande analyses laten zien dat de in dit artikel beschreven methodiek kan worden toegepast om de kansverdeling van bloedgroepen te berekenen van een bloedverwant van een donor met een bepaalde combinatie van bloedgroepen.

Zwangere vrouwen zijn niet de enige patiëntengroep waarbij het wenselijk is om uitgebreid te matchen. Zo zijn er bijvoorbeeld ook patiënten die maandelijks een bloedtransfusie krijgen toegediend. Ook voor hen geldt dat de vorming van antistoffen zoveel mogelijk moet worden voorkomen. Afhankelijk van de bloedgroep(en) waarvoor gematcht moet worden kan de beschreven methodiek gebruikt worden voor het werven van donors met bepaalde combinaties van genen in families of etnisch verwante groepen.

LITERATUUR

1. Ganikhodzhaev, R., Mukhamedov, F., & Rozikov, U., 2011. Quadratic stochastic operators and processes: results and open problems. *Infinite Dimensional Analysis, Quantum Probability and Related Topics*, *14*(02), 279–335.
2. Ganikhodjaev, N., Saburov, M., & Jamilov, U., 2013. Mendelian and non-Mendelian quadratic operators. *arXiv preprint arXiv*:1304–5471.
3. Ganikhodjaev, N., Daoud, J. I., & Usmanova, M., 2009. Linear and nonlinear models of heredity for blood groups and Rhesus factor. *Journal of Applied Sciences*, *10*, 1748–1754.

JOOST VAN SAMBEECK is promovendus bij Sanquin Research en verbonden aan SOR en CHOIR, Universiteit Twente.
E-mail: <j.vansambeeck@sanquin.nl>

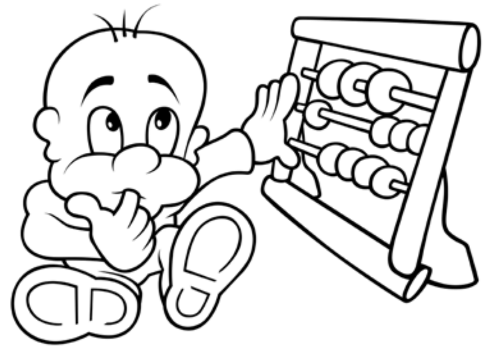
MART JANSSEN is hoofd van de afdeling Transfusion Technology Assessment bij Sanquin Research en werkzaam als onderzoeker bij het Julius Centrum van het UMC Utrecht.
E-mail: <m.p.janssen@umcutrecht.nl>

PETER LIGTHART is senior analist immunohematologie bij Sanquin Diagnostiek.
E-mail: <p.ligthart@sanquin.nl>

WIM DE KORT is hoofd van de afdeling Donor Studies bij Sanquin Research en hoogleraar donorgeneeskunde aan de Universiteit van Amsterdam.
E-mail: <w.dekort@sanquin.nl>

NICO VAN DIJK is hoogleraar bij SOR en CHOIR aan de Universiteit Twente en bij Kwantitatieve Economie aan de Universiteit van Amsterdam.
E-mail: <n.m.vandijk@utwente.nl>

Young Statisticians



STATISTICAL PUBQUIZ 2.0

March second, on a drizzly winter evening, an illustrious bevy gathered in a pub on the crossing of two roads, of which one of them is known to lead to Leiden's Mathematical Institute. One of the subsets of this bevy, that likes to call themselves 'The Informative Priors', included yours sincerely. All attendees were marshalled by a marvelous team of Young Statisticians to revel in a legendary statistical pubquiz. After a cracking night with assiduous teams and brilliant questions, the hiatus in the Informative Prior's cumulative knowledge (dates of birth of Fischer, Pearson and henchmen) could not be compensated with astuteness (number of lakes in Scandinavia, R-programming), and the team missed out on the triumph, though secured the second place. Needless to say after such a delightful pastime, The Informative Priors cannot wait to enter the battle next year. (*Rianne de Heide | The Informative Priors*)

PUBQUIZ 2.0 | EXAMPLE QUESTION

Starting with version 2.14.0, R versions received nicknames. 3.2.1 (2015-06-18) is World-Famous Astronaut, 3.2.2 (2015-08-14) is Fire Safety. What is the nickname of 3.2.3 (2015-12-10)?

A. Wooden Christmas-Tree
B. Golden Menorah
C. Santa's Fluffy Beard
D. Rudolph's Shiny Nose

Antwoord: A. Wooden Christmas-Tree

SCIENCE CAFÉ: TO P OR NOT TO P?

The Young Statisticians gathered on April 20th in Utrecht to join in a discussion on the (ab)use of p-values. First, Prof. Peter Grünwald gave an overview of common misuses of p-values, and demonstrated how they can be abused to even support the claim of clairvoyance! Luckily, he and Prof. Eric-Jan Wagenmakers also introduced us to two methods that might replace p-values in the future. Dr. Don van Ravenzwaaij led a very interesting discussion afterwards, with loads of critical questions from the Young Statisticians. The discussion even extended quite a bit over drinks following the event.
Thank you Peter, Eric-Jan and Don for your help and all participants for being there!

NEW BOARD MEMBERS!

- Nynke Krol and Sanne Willems have left the Young Statisticians board after three years of service.
- Kees Mulder (PhD candidate at University of Utrecht) will continue as board member and he will be joined by five new members!
- Annette Emerenciana (master student at Leiden University)
- Jonas Haslbeck (PhD candidate at University of Amsterdam)
- Erik-Jan van Kesteren (master student at University of Utrecht)
- Laura Verkerk (master student at Leiden University)
- Machiel Visser (master student at Leiden University)

This huge board already shared some of their great ideas with us. Our advice is to keep a close look at the Facebook page <www.facebook.com/young.statisticians?-fref=ts> and so you don't miss the new activities!

ACTIVITIES

→May 30: Company visit Rabobank
The Young Statisticians have visited Rabobank on May 30th. A review about this visit will follow in the next *STAtOR* edition!

Stay updated on future events and **SIGN UP TO THE MAILINGLIST** info@youngstatisticians.nl.
Please let us know if you are interested in becoming an active member!

Einstein, sociale netwerken en waarom mensen net moleculen zijn[*]

Johan van Leeuwaarden

Stel je het volgende experiment voor. Je staat met heel veel anderen in een volgepakt stadion, als bij een concert. Dan wordt er een enorme strandbal, met een doorsnede van 10 meter, uitgerold over het publiek, en alle mensen die de bal aanraken delen een klein tikje uit. We volgen de crowdsurfende bal met een camera die 100 meter boven het stadion zweeft. Vanaf die hoogte lijkt de bal kleine schokkerige bewegingen te maken, maar blijft wel lange tijd ongeveer in het midden van het publiek.

Wat de camera registreert, lijkt op wat Robert Brown zag toen hij in 1827 met een microscoop naar een stuifmeelkorrel in een bakje water keek. De stuifmeelkorrel, toch echt dode materie, maakte net als de strandbal een zenuwachtige beweging. Hoe kon dat? Leefde de stuifmeelkorrel dan toch? Nee, de korrel dreef tussen bijna ontelbaar veel watermoleculen. De korrel was weliswaar 250.000 keer groter dan een molecuul, maar het resultaat van alle tikjes van alle moleculen samen deed de korrel bewegen.

Naar deze verklaring kon Robert Brown alleen maar gissen. In 1827 was de techniek nog niet ver genoeg om moleculen te kunnen zien. Sommigen geloofden in het bestaan van moleculen en anderen niet. Totdat Albert Einstein in 1905 met een weergaloze ingeving kwam. Hij dacht: ik neem gewoon aan dat de moleculen bestaan. En met de grootte van de korrel en de moleculen, in combinatie met de klassieke mechanicawetten van Isaac Newton, wist Einstein wiskundig te voorspellen hoe de korrel in de tijd en ruimte zou moeten bewegen. Die beweging, die nu Brownse beweging heet, gaf dus een verklaring voor wat Brown en velen na hem onder de microscoop hadden gezien. Zonder moleculen te kunnen zien had Einstein aangetoond dat ze bestonden!

Terug naar ons experiment. Mensen in een stadion zijn dus net moleculen die allemaal tikjes tegen een grote strandbal geven. De strandbal volgt daardoor, ongeveer, een Brownse beweging. Maar die Brownse beweging is op veel meer manieren waargenomen dan onder een microscoop of in een stadion. Het bekendste voorbeeld uit de twintigste eeuw: aandelenkoersen. Wiskundige modellen die aandelenkoersen beschrijven, gaan ervan uit dat de koers een Brownse beweging volgt. Waarom? Nou, stel je een aandelenkoers voor als een strandbal. Dan geven alle mensen die een mening hebben over dat aandeel, of het aandeel kopen of verkopen, eigenlijk kleine tikjes tegen die koers. Omdat het om heel veel mensen en evenzoveel meningen gaat, beweegt het aandeel als resultaat van vele kleine tikjes. En hoewel een aandelenkoers geen natuurkunde of exacte wetenschap is, kun je wiskundig wel onderbouwen waarom dit moet leiden tot een Brownse beweging. Het gedrag van mensen kun je daarmee beschrijven met een soort van natuurwetten, alleen is het niet de natuur maar zijn wij het zelf die de wetten creëren.

Nu we ons dat realiseren, kunnen we dan niet iets nuttigs doen met die zelfgemaakte natuurwetten? Dingen die wij mensen doen en die gezien kunnen worden als de optelsom van heel veel kleine tikjes, kunnen we die dan beter begrijpen en liefst ook sturen?

Een bekend voorbeeld hiervan is de pot met snoepjes en de prijsvraag: hoeveel snoepjes zitten er in de pot? Iedereen waagt een gok en dan blijkt vaak dat het gemiddelde van de gokjes onwaarschijnlijk dicht bij het werkelijke aantal snoepjes ligt. Dat komt omdat het gemiddelde gedrag van heel veel tikjes/gokken/meningen een goede voorspelling oplevert. Met de optelsom van alle meningen heb je waardevolle informatie in handen. De eenvoud van het experiment met de snoepjes maakt het een geweldig voorbeeld om het fenomeen van Wisdom of the Crowd te illustreren, het idee dat wij samen slimmer zijn dan ieder van ons afzonderlijk. Ook een hoopgevende gedachte. En het laat bovendien zien hoe je analoge informatie kan omzetten in digitale, bruikbare data. Ieder van ons kijkt naar de pot, naar de hoogte en de breedte, naar de grootte van een snoepje. Ieder van ons pijnigt de hersenen, laat de 100 miljard neuronen vuren, om met een verstandige gok te komen. En hoewel niemand begrijpt wat er in onze hersenen gebeurt, komen we met een helder antwoord: 624 snoepjes. Hoe meer mensen je vraagt naar het aantal snoepjes, hoe meer breinen je aan het werk zet – als een soort supercomputer met heel veel processoren – en hoe accurater het antwoord.

Ook dat is een aantrekkelijke gedachte, dat we onze breinen kunnen koppelen om samen een superbrein te vormen. Maar dan moet het natuurlijk wel om belangrijker zaken gaan dan snoepjes. Terug naar de koers van een aandeel. Als wij dan toch samen die koers veroorzaken, kunnen we dan ook niet samen de koers voorspellen? Dat moet haast wel, als je zou kunnen blootleggen wat wij zoal denken over dat aandeel. En dat kunnen we, door te kijken naar alles wat we e-mailen, twitteren en bloggen. Honderden miljarden stukjes informatie sturen we elke dag weer de wereld in. We moeten de kennis halen uit alles waarover we praten. Filter bijvoorbeeld op tweets die gaan over een bedrijf en probeer gevoelens om te zetten in meetbare informatie. Zo zullen combinaties van woorden als ‘Ik heb geen goed gevoel over…’, ‘Ik hoop dat …’, ‘Ik maak me zorgen…’ duiden op positieve of negatieve tweets. Waardeer dan een positieve tweet met een +1 en een negatieve tweet met een -1. De optelsom van alle kleine plussen en minnen geeft dan een beeld van het sentiment en dus waarschijnlijk een goede voorspelling voor de koers. De woorden vertalen in plussen en minnen is een vorm van datamining. En wees gerust, datamining is geen rocket science. Je mag een keer fout zitten en een positieve tweet verkeerd inschatten met een -1. Ook bij de snoepjes zullen er mensen zijn die er vreselijk naast zitten. Maar kleine foutjes middelen uit, zolang het maar niet te vaak gebeurt en je de denkkracht van de massa voldoende in stelling weet te brengen. Ben je dus behoorlijk precies en vooral heel snel met dataminen, dan kun je (waarschijnlijk) de koers voorspellen!

Vormen we eenmaal samen een machtig brein, dan moet dat brein wel gaan handelen naar een geweten, vind je ook niet? We moeten dan verder kijken dan ons eigenbelang. Waar het bij snoepjes en aandelen gaat om winnen, moeten we nu kijken of we iets voor iedereen kunnen betekenen. Een treffend maar beangstigend voorbeeld is de verspreiding van een virus. Iedereen beseft hoe belangrijk het is om een virus snel te detecteren en het in de kiem te smoren. Maar zijn we te laat, dan kan een virus uitgroeien tot een epidemie, waarbij velen van ons ziek worden. Om een virus snel te kunnen ontdekken moeten we terug naar de strandbal. Ditmaal is de strandbal de toestand van de epidemie. De strandbal beschrijft dus alle mensen die het virus hebben. Wij geven tikjes tegen die strandbal, want we ontmoeten de besmettelijke mensen en krijgen het virus mogelijk ook zelf. Heel veel kleinschalige interacties met mogelijk grote gevolgen. En, kunnen we voorspellen welke kant het opgaat met de strandbal? Nu gaat het opeens niet meer om wat geld verdienen, maar om de gezondheid van alle mensen. En we denken terug aan de plussen en minnen. We moeten dus van alle mensen afzonderlijk te weten komen of ze ziek zijn of niet. Maar wachten we op de huisarts of de overheid, dan zijn we waarschijnlijk te laat. Mensen melden zich vaak pas bij de dokter als ze al behoorlijk ziek zijn. Willen we op tijd zijn dan moeten we opvangen wat er in de lucht hangt, net als bij aandelenkoersen.

En met alle social media die we onderling uitruilen moet dat mogelijk zijn. Voor dat virus hebben we al iets bedacht. Neem het griepvirus dat ons land ieder jaar weer fors raakt tijdens de donkere wintermaanden. Hoe sneller je het detecteert, hoe meer je preventief zou kunnen doen. Het idee is simpel: kijk wat er in zoekmachines wordt ingetikt. Tikt iemand ‘hoofdpijn’, ‘hoestsiroop’, ‘apotheek’ of ‘huisarts in Bussum’ in, dan weet je genoeg. Deze informatie kan van iedereen, op elk moment, worden verzameld. Zie je dus in een bepaalde regio een verhoogde activiteit in virusgerelateerde zoekopdrachten – en wordt het zelfs een trending topic – dan weet je waar je (niet) moet wezen. Met datamining ben je dan vast sneller dan de huisarts.

De pot met snoep, de aandelen en het virus zijn slechts voorbeelden van wat mogelijk is zodra we de informatie die wij samen genereren aan elkaar koppelen. Bedenk zelf maar eens een voorbeeld. Zo zijn niet de minste bedrijven ook ooit begonnen.

[*] Deze tekst is gebaseerd op een van de vijf colleges die Johan van Leeuwaarden heeft gegeven bij de Universiteit van Nederland. Deze colleges van 15 minuten elk zijn te zien op <www.universiteitvannederland.nl>.

Johan van Leeuwaarden is hoogleraar wiskunde aan de TU Eindhoven en lid van De Jonge Akademie van de KNAW.
E-mail: <j.s.h.v.leeuwaarden@TUE.nl>

Op de voorzijde van het 10-markbiljet staat Carl Friedrich Gauss afgebeeld; op de achterkant de door Gauss ontworpen sextant

DE 5 MOOISTE FORMULES UIT DE KANSREKENING

Wat zijn de vijf mooiste formules in de kansrekening? Op deze vraag is uiteraard geen eenduidig antwoord te geven. Echter op de vraag wat de mooiste wiskundige formule is, zal vrijwel iedere wiskundige het erover eens zijn dat de toekenning uitgaat naar Eulers identiteit:

$$e^{i\pi}+1=0,$$

waarbij i de wortel uit -1 is. Deze formule verbindt de vijf meest belangrijke constantes uit de wiskunde ('het droomteam van getallen') en drie belangrijke wiskundige bewerkingen – optellen, vermenigvuldigen en machtsverheffen.

Terug naar de vraag wat de vijf mooiste formules in de kansrekening zijn. De onderstaande vijf formules zijn mijn persoonlijke favorieten.

1. De Gauss-curve

Aristoteles was al van mening dat symmetrie één van de voornaamste elementen is van het universele idee van schoonheid. In de kansrekening wordt symmetrie het mooist tot uitdrukking gebracht door de formule voor de normale kansdichtheid

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}.$$

De curve $f(x)$ is symmetrisch rond het punt $x = \mu$. Ongeacht de waarden van de verwachtingswaarde μ en de spreiding σ, de totale oppervlakte onder de curve is 1, waarbij ongeveer 68% van de oppervlakte tusen $\mu - \sigma$ en $\mu + \sigma$ ligt, ongeveer 95% tussen $\mu - 2\sigma$ en $\mu + 2\sigma$, en ongeveer 99,7% tussen $\mu - 3\sigma$ en $\mu + 3\sigma$. De normale curve wordt ook wel Gauss-curve genoemd naar de beroemde wiskundige Carl Friedrich Gauss (1777 – 1855) die deze curve ontdekte bij zijn foutenanalyse voor astronomische waarnemingen. Op het oude Duitse 10-markbiljet wordt Gauss geëerd met naast zijn portret een afbeelding van de normale curve en de wiskundige formule daarvoor.

2. De formule van Bayes in odds vorm

De formule van Bayes luidt:

$$\frac{P(H\mid E)}{P(\overline{H}\mid E)}=\frac{P(H)}{P(\overline{H})}\times\frac{P(E\mid H)}{P(E\mid \overline{H})}.$$

De formule laat zien hoe een (subjectieve) kanstoekenning die vooraf gedaan is aan een bepaalde hypothese aangepast moet worden in het licht van nieuwe evidentie. De *posterior odds* van de hypothese H wordt gevonden door de *prior odds* te vermenigvuldigen met de Bayes-factor. Deze factor is gelijk aan de conditionele kans dat evidentie E optreedt gegeven dat de hypothese H waar is gedeeld door de conditionele kans dat evidentie E optreedt gegeven dat de hypothese H niet waar is. De schoonheid van de formule van Bayes is dat het rationeel denken in een simpele wiskundige formule vastlegt. De formule van Bayes heeft talloze praktische toepassingen: zoekprocedures voor vermiste objecten (zie *STAtOR*, juni 2013), onderbouwing van gerechtelijke uitspraken, genetica en ziektes, spamfilters, Google's zoekmachine, etc. In de Tweede Wereldoorlog werd door de beroemde wiskundige Alan Turing de formule van Bayes gebruikt om geheime nazi-codes te breken. Het belang hiervan voor het beëindigen van de Tweede Wereldoorlog kan moeilijk worden overschat.

3. De gokkers formule

De gokkersformule die teruggaat tot Christiaan Huygens luidt als volgt:

$$\frac{1-(q/p)^a}{1-(q/p)^{a+b}},$$

waarbij de formule gelezen moet worden als $a/(a+b)$ als $p = q$. Voor het geval van twee spelers A en B met beginkapitalen van a euros en b euros, geeft de formule de kans dat speler A uiteindelijk al het geld wint als in iedere uitvoering van het spel elk van de spelers 1 euro inzet en speler A met kans p de inzet wint en speler B met kans $q = 1 - p$. Uit de gokkers formule kan worden afgeleid dat in het casino je beter grotere bedragen kunt inzetten dan kleinere bedragen wanneer je enige doel zou zijn om met zo groot mogelijke kans een van te voren vastgelegd eindbedrag te bereiken. Intuïtief is dit duidelijk: het best is om je geld zo kort mogelijk bloot te stellen aan het huisvoordeel van het casino. Deze wijsheid werd in april 2012 wel erg letterlijk door de Engelsman Ashley Revell toegepast in een casino in Las Vegas. Revell had al zijn bezittingen verkocht, inclusief zijn kleren, en reisde naar Las Vegas met 135.300 dollar op zak. Zonder te aarzelen zette hij in één keer het gehele bedrag in op rood bij roulette en keek kalm toe hoe het balletje zou vallen. Het viel op rood! Revell had zijn doel bereikt en zijn kapitaal was verdubbeld. Weinigen zouden dit optreden, dat puur volgens de kanswetten het beste is om te doen, aangedurfd hebben. Overigens Revell was in 1980 vooraf gegaan door William Lee Bergstrom. Deze Texaanse paardenhandelaar was een professioneel gokker die bekend werd als de 'kofferman' nadat hij in 1980 een casino in Las Vegas was binnengegaan met twee koffers, waarvan één leeg en de ander gevuld met 777 duizend dollar. Hij zette dit gehele bedrag in op een eenmalige weddenschap bij het casinospel craps en verliet even later het casino met twee gevulde koffers. Zowel met Bergstrom als met Revell is het later slecht afgelopen. Ze konden het gokken niet laten!

4. De wortelformule voor de standaarddeviatie

De formule luidt:

$$\sigma(X_1 + \ldots + X_n) = \sigma\sqrt{n}.$$

De formule stelt dat de standaarddeviatie van een som van n onafhankelijke stochastische variabelen $X_1,\ldots, X_n$ die elk afzonderlijk standaarddeviatie σ hebben slechts met een factor $\sqrt{n}$ toeneemt bij groter wordende n. De wortelformule wordt soms ook wel De Moivre's vergelijking genoemd naar de beroemde kansrekenaar Abraham de Moivre die de wortelformule in 1730 vond. Deze formule had direct impact op de wijze waarop het gewicht van van de door de London Mint geslagen gouden munten werd gecontroleerd. De toegestane afwijking in het beoogde gewicht van 128 grains (destijds eenheid van gewicht, gelijk aan 0,0648 gram) van een munt was 1/400 daarvan, oftewel 0,32 grains. Een steekproef van 100 munten werd periodiek genomen van de geslagen munten en het totale gewicht van deze munten werd vergeleken met het gewicht van een standaard van 100 munten. De waakhonden van de Engelse koning hadden al eeuwenlang een afwijking van 100 × 0,32 = 32 grains toegestaan in het gewicht van de 100 geïnspecteerde munten. Het goud waaruit de munten geslagen werden, was afkomstig van de Engelse koning. Direct na De Moivre's publicatie van de wortelformule werd de toegestane afwijking in het gewicht van 100 munten gewijzigd in $\sqrt{100} \times 0{,}32 = 3{,}2$ grains, maar inmiddels had de onwetendheid over de wortelformule de Engelse koningen al een fortuin aan goud gekost. De wortelformule heeft vele toepassingen en verklaart ook waarom de grootte van een stad of de grootte van een ziekenhuis er toe doen in de statistieken van de misdaadpercentages of de sterftepercentages bij operaties. Kleinere ziekenhuizen zullen eerder bovenaan of juist onderaan in ranglijstjes verschijnen dan grote ziekenhuizen. Intuïtief is dit duidelijk door te bedenken dat bij het werpen met een zuivere munt de kans dat meer dan 70% of minder dan 30% van de worpen kop geeft veel groter is bij 10 worpen dan bij 100.

5. De inzetformule van Kelly

De inzetformule van Kelly heeft betrekking op een serie weddenschappen waarin het voordeel aan de kant van de speler ligt. De speler zet dan niet elke keer zijn volledige kapitaal met dan het risico in één keer alles kwijt te zijn, maar zet elke keer eenzelfde vaste fractie van zijn

$$\frac{pf-1}{f-1},$$

huidige kapitaal in. Deze fractie wordt gegeven door wanneer bij elke weddenschap met kans p een bedrag gelijk aan f keer de inzet van de speler wordt uitbetaald en met kans $1 - p$ de inzet verloren gaat. De aanname is dat $pf > 1$, oftwel de speler is in het voordeel bij de weddenschap. De Kelly-fractie $(pf - 1)/(f - 1)$ heeft de fraaie interpretatie van zijnde het quotiënt van de netto verwachte opbrengst van een inzet van één geldseenheid gedeeld door de *payoff odds*. Bij het volgen van deze zogenaamde Kelly-strategie maximaliseert de speler de groeivoet van zijn kapitaal op de lange duur. Het idee van de Kelly-strategie is niet alleen bruikbaar voor weddenschappen bij bijvoorbeeld paardenraces, maar ook bij investeringsbeslissingen. Verschillende beroemde beleggers waaronder Warren Buffett hebben met succes gebruik gemaakt van investeringsstrategieën die op het Kelly-systeem gebaseerd zijn. Het Kelly-systeem heeft vooral zijn bekendheid gekregen door het werk van Edward O'Thorp die als eerste het Kelly-inzetsysteem in casino's gebruikte bij zijn succesvolle strategie voor blackjack die de speler een voordeel geeft ten opzichte van het casino. Wie zegt daar dat je met wiskunde niet rijk kan worden?

HENK TIJMS is emeritus hoogleraar operations research aan de Vrije Universiteit en auteur van diverse leerboeken over operations research en kansrekening.
E-mail: <tijms@quicknet.nl>

When it comes to data, size isn't everything[1]

Op 17 maart vond in Utrecht de jaarlijkse Dag voor Statistiek en OR plaats met als thema Big Data. Jacqueline Meulman – voorzitter van de VvS+OR – belichtte de ontwikkeling van Data Science en de belangrijke rol die de statistiek hierbij speelt.

JACQUELINE J. MEULMAN

The theme of this year's Annual Meeting was *The Role of Statistics in Data Science*. For years we happily cited Google's chief economist Hal Varian, who said in 2009 'the sexy job in the next ten years will be statisticians'. In the same year, *The New York Times* published an article with the headline 'For Today's Graduate, Just One Word: Statistics'. However, it's not even ten years later yet, and we seem already to have lost the qualification 'sexy job', since *Harvard Business Review* declared in 2012 'Data Scientist: The sexiest job of the 21th Century'.

Now there are a number of objections to the idea that the role of statistics is overtaken by a new discipline, called 'data science' (much like what happened before, when 'data mining' was the hype, and more recently, when people started to replace the name 'statistics' with 'analytics'). The first objection is of course that Statistics = Data Science, and as such has been around for quite some time. John Tukey, a true visionary, was at the same time Professor of Statistics in Princeton as Associate Executive Director-Research Information Sciences at Bell Labs. As David Donoho writes in his paper *50 years of Data Science*, Tukey called for a reformation of academic statistics, pointing in his famous 1962 paper *The future of data analysis*, to the existence of an as-yet unrecognized science called 'data analysis', whose subject of interest was learning from data.[2] According to Donoho, Tukey's new discipline 'data analysis' had four driving forces:

1. The formal theories of statistics;
2. Accelerating developments in computers and display devices;
3. The challenge, in many fields, of more and ever larger bodies of data;
4. The emphasis on quantification in an ever wider variety of disciplines.

Tukey[3] himself said, 'As I see it, data analysis [is] a science, one defined by a ubiquitous problem rather than by a concrete subject'.

Two other early data scientists, John Chambers and Bill Cleveland (like Tukey, both working at Bell Labs), also urged academic statistics to expand its boundaries beyond the classical domain of theoretical statistics.[2] John Chambers, co-developer of the S language for statistics and data analysis, elaborated in 1993 ideas for a new discipline called Greater Statistics. Bill Cleveland, well-known for his statistical methods and data displays, even suggested the name Data Science for his envisioned field: 'An Action Plan for Expanding the Technical Field of Statistics' (at the 1999 meeting of the International Statistical Institute). Leo Breiman, a famous probabilist, left university to work as an independent consultant, eventually returned to Berkeley, where he developed 'Classification and Regression Trees' (CART), together with Friedman, Olshen, and Stone, and Random Forests. As early as 1977, Breiman organized a conference on the *Analysis of Large and Complicated Data*. Jerome Friedman, a high energy particle physicist, encouraged by Tukey to become a statistician, developed together with Tukey 'Projection Pursuit', and 'PRIM9', as well as many other data analysis methods. If we would not consider all this work part of Data Science, we would make a historical mistake. Be that as it may, according to Donoho in his 50 years of Data Science article, 'the statistics profession is caught at a confusing moment: the activities which preoccupied it over centuries are now in the limelight, but those activities are claimed to be bright shiny new, and carried out by (although not actually invented by) upstarts and strangers.' He continues with quoting various statisticians active in professional statistics organizations:

- 'Aren't we data science?' (ASA President Marie Davidian in AmStat News);
- 'A grand debate: is data science just a "rebranding" of statistics?' (Martin Goodson, co-organizer of the Royal Statistical Society meeting May 11, 2015 on the relation of Statistics and Data Science);
- 'Let us own data science.' (Bin Yu in her IMS Presidential address, and also others);
- 'Why do we need data science when we've had statistics for centuries?' (Irving Wladawsky-Berger, CIO report, 2014);
- 'Data science is statistics: If you're analyzing data, you're doing statistics. You can call it data science or informat-

> ics or analytics or whatever, but it's still statistics. […] You may not like what some statisticians do. You may feel they don't share your values. They may embarrass you. But that shouldn't lead us to abandon the term "statistics".' (Karl Broman, University of Wisconsin).

Some, however, would argue that Data Science differs from statistics because it deals with Big Data. Now this is a point that we should vehemently object to, as succinctly worded by Donoho: 'We can immediately reject Big Data as a criterion for meaningful distinction between statistics and data science […] statisticians deal with data however it arrives – big or small.' In a very worth reading article in the Financial Times, under the tile *Big Data: are we making a big mistake?* Tim Harford[1] summarizes (and debunks) what he calls 'Big Data's four articles of faith', and this comes to:

1. Data analysis without any theory or hypothesis can produce uncannily accurate results. However, this is usually only true if we simply ignore false positives.
2. When every single data point can be captured, old statistical sampling techniques become obsolete. We do not need scientific or statistical models anymore, because 'with enough data, the numbers speak for themselves' (The End of Theory, Wired, 2008). This naïve idea simply ignores the fact that in large data sets spurious patterns vastly outnumber genuine discoveries. With Big Data, the multiple-testing problem is also very big, as is the danger of a high False Discovery Rate.
3. A big data set is one where 'N = All, and when 'N = All' there is no issue of sampling bias because the sample includes everyone (Mayer-Schönberger). This is of course not true, because in most cases Big Data are not random samples, and this is one of the major problems using Big Data.
4. With Big Data, causality won't be discarded, but it is being knocked off its pedestal as the primary fountain of meaning (Mayer-Schönberger & Kenneth Cukier). Predictions, however, are not made in a stable environment, but in a world that is changing all the time. As Spiegelhalter remarks: there are a lot of small data problems that occur in Big Data, and these don't disappear because you have a lot of data. In fact, they get worse.

At best, the four beliefs stated above are overly optimistic, oversimplifications (Harford); at worst, they can be complete bullocks, absolute nonsense (Spiegelhalter). As Harford concludes 'Big data do not solve the problem that has obsessed statisticians and scientists for centuries: the problem of insight, of inferring what is going on, and figuring out how we might intervene to change a system for the better'.

In 1997, Jerry Friedman wrote a stimulating paper called *Data Mining and Statistics: What's the Connection?* It's very worthwhile to reread, since it deals with the very same issues that are discussed nowadays with respect to Data Science. Friedman writes, for instance, the following about Data Mining (DM).[4] 'If statisticians and data miners are to join together to address the data analysis challenges of the future […] the DM community may have to moderate its romance with 'big'. A prevailing attitude seems to be that unless an analysis involves [huge amounts of] data, it cannot possibly be worthwhile. A dominant theme of many presentations at the 1977 Dallas Conference [the one on *Analysis of Large and Complex Data Sets*, organized by Leo Breiman] has been 'My data set is larger than your data set'. It seems to be a requirement that all of the data that has been collected must be used in every aspect of the analysis […] However, it is often the case that the questions being asked of the data can be answered to sufficient accuracy with less than the entire data base. Sampling methodology, which has a long tradition in Statistics, can profitably be used to improve accuracy while mitigating computational requirements […] a powerful computationally intense procedure operating on a subsample of the data may in fact provide superior accuracy than a less sophisticated one using the entire data base'. It is very tempting to replace Data Mining (DM) by Data Science (DS) in Friedman's article, rendering a perfectly up-to-date paper.

As for the relation between statistics and data science, we should embrace data science as part of statistics, and not leave it to computer science or other fields that seem to be claiming it. On the contrary, when statistics and computer science join forces, for example in Data Science graduate programs, we can have the best of both worlds. Perhaps we should call it 'Statistical Data Science'. As for 'Big Data', we should welcome the challenge to develop new strategies for data analysis. Given the many traps that analysis of 'Big Data' without using statistical principles may fall into, the existence of 'Big Data' makes the role of statistics even more important than ever.

LITERATUUR

1. Harford, T. (2014). Big data: are we making a big mistake? *Financial Times, March 28.*
2. Donoho, D. L. (2015). *50 Years of Data Science.* Based on a presentation at the Tukey Centennial workshop, Princeton NJ, September 18.
3. Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics,* 33, 1-67.
4. Friedman, J. H. (1997). Data Mining and Statistics: What's the Connection. In 2001 published as 'The role of Statistics in the Data Revolution'. *International Statistical Review,* 69, 5-10.

JACUELINE MEULMAN is hoogleraar Toegepaste Statistiek aan het Mathematisch Instituut van Universiteit Leiden. Van 2011 tot 2016 was ze voorzitter van de VvS+OR.

EEN ZICHTBARE TOEKOMST VOOR DE STATISTIEK IN NEDERLAND EN DE VVS+OR

Op 17 maart droeg Jacqueline Meulman de voorzittershamer over aan Fred van Euuwijk, de nieuwe voorzitter van VvS+OR. In zijn bijdrage ontvouwt hij zijn visie op de toekomst van de vereniging.

FRED VAN EEUWIJK

Beste *STAtOR*-lezer, op de laatste algemene ledenvergadering van de VvS+OR, op 23 maart 2016, is mij de eer te beurt gevallen de komende jaren voorzitter te mogen zijn van de VvS+OR. In deze bijdrage wil ik een eerste idee geven van de richting waarin ik als voorzitter de vereniging wil loodsen. Deze richting ligt min of meer in het verlengde van de strategische visie die onder de aftredende voorzitster Jacqueline Meulman werd ontwikkeld. Ze sluit nauw aan bij de visies van zusterorganisaties als de American Statistical Association en de Royal Statistical Society.

Allereerst wil ik Jacqueline Meulman bedanken voor haar inzet als voorzitster in de periode 2011-2016. Ik wil drie verdiensten van haar noemen op thema's die in de komende jaren verder uitgewerkt zullen worden. Als een eerste belangrijke verdienste kan genoemd worden dat zij de jaarlijks Statistische Dag tot een boeiende en bruisende bijeenkomst voor statistici heeft weten te maken, met zowel aandacht voor inhoudelijk hoogwaardige statistische bijdragen alswel relevante beschouwingen over de rol van statistici en statistiek in wetenschap en samenleving. Daarnaast bieden de statistische dagen een goede ontmoetingsgelegenheid voor een grote diversiteit aan personen die op vele manieren met statistiek bezig zijn. Een andere belangrijke verdienste van Jacqueline ligt in haar bijdrage aan de vorming van een internationale federatie van statistiekverenigingen met als doel het belang en de inhoud van statistiek beter uit te dragen naar wetenschap en maatschappij (FENStatS). Een verdere noemenswaardige verdienste betreft haar streven en inzet om *Statistica Neerlandica* een betere positie te geven binnen de markt van statistische bladen.

Op alle drie de bovenstaande punten, Statistische Dag, FENStatS en *Statistica Neerlandica* zal de komende jaren voortgegaan worden op de door Jacqueline ingeslagen weg. Wat betreft de verdere ontwikkeling van Statistische Dag en FENStats, beide activiteiten dienen wat mij betreft geïnterpreteerd te worden als bijdragen aan de centrale opdracht die ik voor mezelf en de VvS+OR zie: het vergroten van de zichtbaarheid van de statistiek en de statistici in Nederland. *Statistica Neerlandica* zou hier eventueel ook een rol in kunnen vervullen, maar waarschijnlijk is *STAtOR* hiervoor een meer geschikt blad. *Statistica* kan bijdragen aan zichtbaarheid in een meer beperkte wetenschappelijk zin.

Waarom zoveel drukte over zichtbaarheid? Opvallend is dat alle grotere statistische associaties en verenigingen in de afgelopen jaren tot een min of meer gelijkluidende diagnose zijn gekomen ten aanzien van de voorwaarden waaronder in de toekomst statistiek zich als een succesvolle wetenschappelijke discipline kan blijven handhaven en hopelijk ontwikkelen: statistici zullen veel nadrukkelijker dan ze gedaan hebben of gewend waren te doen statistiek moeten verkopen als een wetenschap die essentieel is voor vooruitgang op een veelheid van terreinen rakend aan wetenschap, industrie en beleid. Ik onderschrijf deze diagnose van de noodzaak tot een grotere zichtbaarheid van statistici en statistiek als de voorwaarde voor een gezonde toekomst van de statistiek. Inderdaad, statistiek wordt tegenwoordig meer gerespecteerd dan in het verleden en wordt gezien als een activiteit die kan bijdragen aan verantwoorde en onderbouwde voorspellingen en beslissingen. Een studie tot statisticus wordt tegenwoordig steevast aangemerkt als hoogst aantrekkelijk vanwege het gegarandeerde beroepsperspectief. Desalniettemin blijkt elke keer opnieuw dat een groot gedeelte van het 'publiek' statistiek eerder ziet als een verzameling gereedschappen en trucs dan als een volwassen wetenschap die belangrijk is voor het opzetten en analyseren van reproduceerbaar onderzoek en het formuleren van gefundeerde beslissingen in het licht van toevalsvariaties. Het door Jacqueline Meulman geschetste grensgeschil tussen statistiek en data science elders in dit blad is een goed voorbeeld van hoe een onderzoeks- en toepassingsterrein dat door vrijwel iedere statisticus als behorend tot de statistiek gezien wordt, in handen kan vallen van wetenschappers die zichzelf in eerste instantie niet als statistici zullen betitelen. Het gevolg van een dergelijke identiteitsverwarring is dat anderen dan statistici met de eer en de fondsen aan de haal zullen gaan. Bovendien is er het risico van suboptimale oplossingen voor problemen waarvoor statistici de geëigende expertise in huis hebben. Soortgelijke grensgeschillen als die met data science, deden zich in het recente verleden voor met bioinformatica en epidemiologie.

Wat kunnen we doen om onze zichtbaarheid te vergroten? De American Statistical Association (ASA) heeft nadrukkelijk nagedacht over deze kwestie (zie mei-2015-nummer van *The American Statistician*). Als eerste punt noem ik de lancering van de campagne *This is Statistics* gericht op het vergroten van bekendheid en interesse onder studenten aan het begin van hun studies (*high school, undergraduate*). Vier kernboodschappen worden geformuleerd: statistiek is niet wat je denkt dat het is (gericht op verkeerde interpretaties); het veld van statistiek is breder en dieper dan je denkt (toont diversiteit van statistiek en zijn toepassingen); weinig loopbanen zijn zo veelbelovend als die in statistiek (goed beroepsperspectief); statistische geletterdheid is essentieel in het dagelijkse leven (in persoonlijk- en beroepsleven). ASA richt zich ook op de docenten die inleidende statistiekcursussen moeten verzorgen en wil deze docenten helpen door een codificering van de statistische vakinhoud voor deze cursussen en het beschrijven van de bijbehorende pedagogiek en didactiek.

Naast de bovenstaande studentgerelateerde acties neemt ASA ook officiële posities in op onderwerpen gerelateerd aan klimaatverandering, auditen van verkiezingen, onderwijs en andere zaken. ASA produceert verder een serie van korte documenten die belangrijke bijdragen beschrijven van statistici aan de samenleving (gezondheidszorg, economie, veiligheid). Naar aanleiding van een verzoek van onderzoeksfondsenbeheerders heeft ASA enkele white papers geschreven die tonen hoe statistici kunnen bijdragen aan onderzoeksinitiatieven en aan beleidsprioriteiten van de regering. Prijzenswaardig is een ASA initiatief om journalisten te helpen statistiek-gerelateerde onderwerpen fatsoenlijk te beschrijven. Ten slotte noem ik de pogingen van ASA om bedrijven en organisaties op te zoeken en aan te spreken die aan analyses van *big data* werken om er zodoende achter te komen welke vaardigheden en kennis verondersteld worden nodig te zijn. ASA hoopt daarmee de big data en data science gemeenschap binnen zijn werkgebied te kunnen opnemen. ASA investeert sterk in een imago als *Big Tent for Statistics* om recht te doen aan de brede diversiteit van statistiekbeoefenaars.

De initiatieven van ASA om de statistiek als discipline te stimuleren en de beoefenaars van de statistiek ten dienste te zijn vormen interessant materiaal voor een discussie over de toekomst van de statistiek in Nederland en de toekomst van de VvS+OR. Niet alle ASA-voorstellen zijn één op één overdraagbaar naar de Nederlandse situatie. In de komende maanden zal ik samen met enkele andere VvS+OR bestuursleden de besturen van de individuele VvS+OR-secties bezoeken om van gedachten te wisselen over welke rol zij zien voor de VvS+OR in de komende jaren. Daarbij zal een belangrijke vraag zijn wat de secties zelf als hoofdtaken van beleid zien en wat zij als taken voor de VvS+OR zien. De gedachten van de ASA over de toekomst van de statistiek in hun vereniging kunnen als een soort leidraad gelden in dergelijke discussies. Vanuit het ‘Big Tent’ denken zal de VvS+OR ook gaan praten met statistische verenigingen binnen Nederland die momenteel nog geen deel uitmaken van de VvS+OR. Ik hoop de resultaten van al deze besprekingen te kunnen beschrijven in een toekomstige aflevering van *STAtOR*.

Fred van Eeuwijk is hoogleraar Wiskundige en Statistische Methoden bij de afdeling Biometrie van Wageningen UR. Daarnaast is hij voorzitter van de VvS+OR.
E-mail: <fred.vaneeuwijk@wur.nl>

KLASSIEKERS

In dit nummer geen bijdrage in de rubriek Klassiekers. We hebben enkele verzoeken uitstaan, maar graag nodigen we alle lezers uit om een bijdrage te leveren voor deze rubriek. Vertel aan uw medelezers welk boek of artikel voor u een beslissende rol heeft gespeeld in uw loopbaan. U kunt contact opnemen met Richard Starmans <R.J.C.M.Starmans@uu.nl>; hij vertelt u graag meer over deze serie artikelen.