‘Big data heeft echt alles veranderd in de statistiek’

Nieuwe kansen

De hoeveelheid data die wordt opgeslagen op computers groeit exponentieel. De dataverzamelingen zijn tegenwoordig zo groot geworden, dat er nieuwe statistische methoden nodig zijn om ze te analyseren. Hoogleraar stochastiek Aad van der Vaart doet onderzoek hiernaar. In 2015 krijgt hij de NWO-Spinozapremie van 2,5 miljoen euro.

TEKST: ANOUSCHKA BUSCH / FOTOGRAFIE: ADRIE MOUTHAAN

Was deze premie een verrassing?
‘Dat was het zeker. Je weet dat je in aanmerking komt gegeven je cv. Je kent je eigen *standing* in je vakgebied. Maar dat je daadwerkelijk een Spinozapremie krijgt, blijft een enorme verrassing.’

Je bent stochastiekprofessor. Wat houdt dat in?
‘Stochastiek is gewoon een ander woord voor kansrekening. Maar waar ik me mee bezighoud, gaat meer in de richting van statistiek: het ontwikkelen van methodes om grote en ingewikkelde verzamelingen gegevens te analyseren.’

Wat is er zo leuk aan statistiek?
‘Een van de dingen die ik er leuk aan vind is dat het wiskunde is met toepassingen. Je bent in de eerste plaats wiskundige, maar je komt ook in contact met de medische wetenschap, economie en sociale wetenschappen, en met de wiskundige problemen die er spelen. Statistiek zit in ongeveer alle andere wetenschappen verpakt. Je krijgt een beetje mee wat er in andere vakgebieden speelt en je ziet dat al die wiskunde ergens goed voor is, dat het niet alleen abstracte theorie is maar dat het in de praktijk werkt.’

Wat zijn interessante toepassingen van jouw onderzoek?
‘Genoomonderzoek is een goed voorbeeld. Vijftien jaar geleden kregen medici ineens metingen op heel veel genen tezamen tot hun beschikking. Maar ze wisten niet goed wat ze ermee aan moesten. Er bestond een lange traditie van statistiek toepassen op medische data. Maar deze data waren totaal anders dan wat ze gewend waren.’

Wat was er zo anders aan die data?
‘Vooral dat ze zo groot waren. Pas de laatste drie of vier jaar spreekt men van ‘big data’, maar genoomonderzoek was in feite het eerste voorbeeld daarvan. Het gaat om metingen aan 30.000 genen tegelijk. En dat is fundamenteel anders dan de beperkte hoeveelheid medische data waar men daarvoor mee te maken had, zoals geslacht, bloeddruk of cholesterolgehalte.
Big data heeft echt alles veranderd in de statistiek. Vroeger kon je gewoon door een statistische dataset scrollen op de computer. Tegenwoordig zijn de bestanden soms te groot om ze zelfs maar op de gewone manier te openen. We hebben een tijd gehad dat computers zo veel krachtiger waren geworden dat we dachten: nu kunnen we echt alles. Maar inmiddels is de hoeveelheid data zo groot geworden dat we een stapje terug hebben moeten doen. We moeten berekeningen heel simpel houden, omdat het anders niet werkt. De berekening zou oneindig lang doorgaan.’

Het probleem zit niet alleen in de hoeveelheid?
‘Nee, het zit ook in de complexiteit. Stel dat je wilt weten welke genen verantwoordelijk zijn voor een ziekte. Een manier van onderzoeken is om de genen van twee mensen die die ziekte hebben met elkaar te vergelijken en te kijken of je een bepaald patroon vindt. In zeldzame gevallen gaat dat meteen goed. Dan heb je al snel een bepaald gen te pakken. Maar meestal ligt het ingewikkelder. Dan is het niet één gen dat de ziekte veroorzaakt, maar meerdere genen of interacties tussen bepaalde genen. Dat maakt het heel ingewikkeld om te vinden. Ik probeer nieuwe methodes te ontwikkelen om die verbanden er toch uit te halen.’

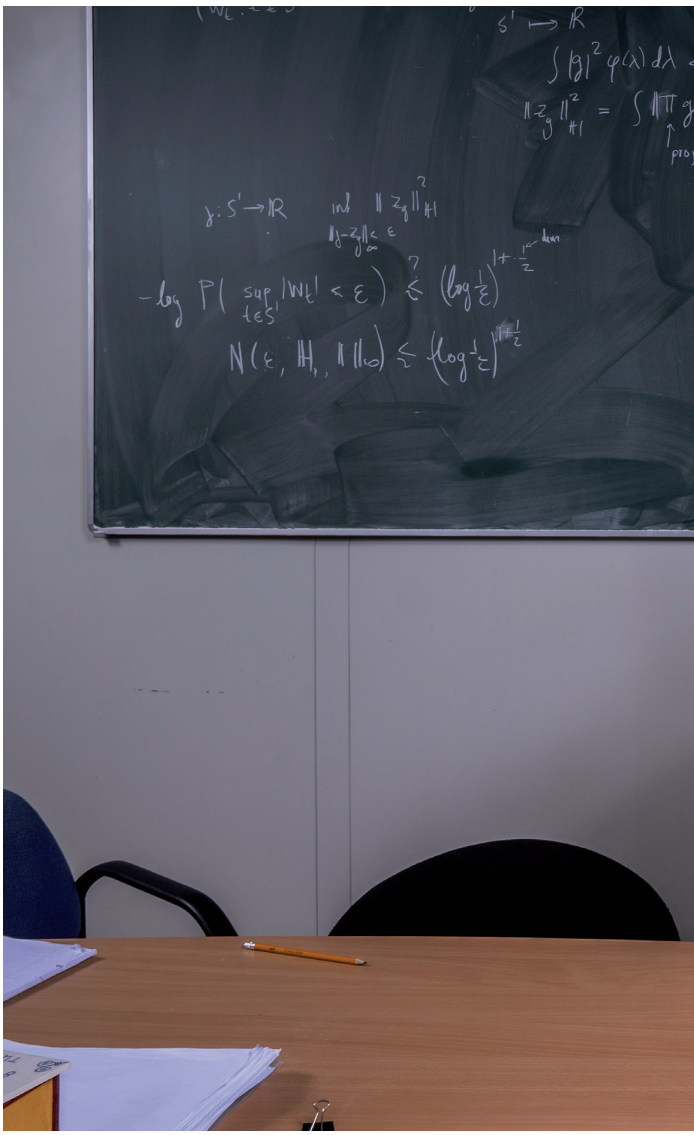
Er was de laatste tijd veel te doen om onderzoek dat statistisch niet deugde, vooral in de sociale wetenschappen. Speelt dat hier ook?
‘Niet direct. Maar als je veel metingen doet, wordt de kans groot dat je verbanden vindt die op het eerste gezicht statistisch significant lijken, maar die het bij nader inzien toch niet zijn. Als je de genenverzameling van patiënten vergelijkt met die van gezonde

AAD VAN DER VAART,
hoogleraar stochastiek aan de Universiteit Leiden, ontvangt de NWO-Spinozapremie 2015.

mensen, is er altijd een kans dat je een afwijkend gen vindt, maar dat die vondst puur toeval is. Dat heet de 'p-waarde': de kans dat het verband dat je vindt toeval was. We proberen die kans klein te houden, minder dan vijf procent. In dat geval noemen we een verband statistisch significant. Maar als je vaak zo'n vergelijking doet, moet je die vijf procenten telkens bij elkaar optellen. Als je twintig keer twee genenverzamelingen met elkaar vergelijkt, dan weet je bijna zeker dat je een keer een overeenkomst vindt, die puur toevallig is. Ik heb in de begintijd van het genenonderzoek wel eens bij lezingen gezeten waar medici trots vertelden: 'We hebben 10.000 genen getoetst en daarvan waren er vijftig significant.' Tja, dat is precies wat je volgens de foutenmarge kunt verwachten. Dan heb je dus eigenlijk niks gevonden. Misschien zit er een gen tussen dat verantwoordelijk is voor de ziekte. Maar het is goed mogelijk dat het vijftig missers geweest zijn. Tegenwoordig weet een onderzoeker wel dat als hij naar 30.000 genen kijkt, dat hij daar een statistische correctie op moet toepassen. Maar er worden heel veel van dat soort onderzoeken uitgevoerd en eigenlijk zou je daar dezelfde correctie op moeten toepassen. Als je heel veel onderzoekers hebt die allemaal die p-waarde van vijf procent halen, dan weet je vrij zeker dat er ook veel bij zijn die onzinresultaten hebben gevonden. Kortom: zelfs onderzoeken die statistisch goed uitgevoerd worden en die netjes op een p-waarde van vijf procent komen, geven niet altijd betrouwbare conclusies.'

Zijn er meer statistische valkuilen waar wetenschappers voor moeten oppassen?
'Als wetenschapper wil je vaak iets zeggen over wat het effect is van iets wat iemand doet of eet. Is het ▶

‘De echte wiskunde, het echte nadenken moet je ook zelf doen’



gezond om veel broccoli te eten? Om dat te bepalen deel je in het ideale geval je proefpersonen willekeurig in twee groepen in. De ene groep laat je broccoli eten, de andere niet. Leven de broccoli-eters gemiddeld langer, dan kun je concluderen dat de broccoli het verschil maakte. Maar in de praktijk werkt het zo niet. Zo’n experiment doen is te ingewikkeld. Wetenschappers kijken naar gegevens die ze al hebben: ze vergelijken de data van groepen mensen waarvan bekend is dat de ene groep broccoli eet en de andere niet. Maar als je dan ontdekt dat de broccoli-eters gemiddeld langer leven kun je daar niet direct conclusies uit trekken, want de groepen zijn niet willekeurig gekozen. Er kunnen allerlei andere verschillen zitten tussen de mensen die wel en niet broccoli eten. Misschien eten broccoli-eters ook meer andere groente, misschien sporten ze meer. Dat is waar de statistiek meedoet: corrigeer voor die alternatieve verklaringen. Staat er in de krant bij een onderzoek: ‘de resultaten zijn gecorrigeerd voor leeftijd of inkomen’, dan is dit waar het om gaat.’

En als je zo’n correctie uitvoert, weet je dan zeker dat je conclusies kloppen?
‘Nee, dat niet. Je gaat na welke andere variabelen van invloed kunnen zijn op de uitkomst die je hebt gezien en dan probeer je te modelleren hoe die twee groepen verschillen in die andere variabelen. Het probleem is: vaak heb je veel variabelen die een rol zouden kunnen spelen. Ook het corrigeren zelf is niet zo gemakkelijk. Er wordt veel onderzoek naar gedaan. En het zal waarschijnlijk nooit perfect worden. Ik ben er zelf mee bezig met een epidemioloog uit Harvard. Hij zit vooral in de modellering en ik kijk naar de wiskundige kant van zo’n probleem.’

Het idee dat hoogleraren managers zijn die geen tijd hebben voor onderzoek, gaat voor jou niet op?
‘In de wiskunde gaat dat in zijn algemeen niet op. Ik manage het liefst niemand. De mensen in mijn groep zijn heel zelfstandig. Ze zijn meer partners dan mensen die ik aanstuur. Als hoogleraar heb je iets meer overzicht, dus je stelt vragen: waarom doe je dit zo, of zou je het niet anders aanpakken? Maar je kunt geen wiskundigen ‘managen’ zonder zelf ook het basiswerk te doen. Als je zelf niet inhoudelijk bezig bent, ben je binnen drie weken uitgerangeerd. Een slimme medewerker weet dan al zo veel meer. Dan kun je niet meer meepraten. En dat basiswerk is ook het leukste om te doen. Alleen het uiteindelijke rekenwerk kun je wel uitbesteden. Het programmeren is heel tijdrovend en dat zijn taken die je kunt verdelen. Dan ga je uiteindelijk samen kijken: wat is er uitgekomen en hoe interpreteren we dat? Maar de echte wiskunde, het nadenken moet je toch ook zelf doen.’

Jij denkt het grootste deel van de dag dus na?
‘Klopt. Het gaat vooral om stellingen bedenken en bewijzen. Je wilt heel precies een bewering keihard kunnen maken, zoals in iedere vorm van wiskunde.’ Om zijn punt te illustreren pakt Van der Vaart een artikel dat hij heeft geschreven, vol formules. ‘Kijk, zo ziet het er uiteindelijk uit: stellingen en bewijzen’.

Zie je nu ook direct wat er in die formules staat? Lees je dit zoals een ander een roman leest?
‘In dit geval wel. Maar voor mij staat hier niets nieuws. Als je lang in het vak zit dan verbaas je je soms ook als eerstejaarsstudenten aan je vragen: ‘Moeten we echt al die ingewikkelde formules leren?’ Voor mij is het allemaal vanzelfsprekend, het zijn geen ingewikkelde formules. Maar als ik nieuwe wiskunde moet lezen, lees ik maar drie pagina’s per dag. Met de pen erbij, aantekeningen maken en stapje voor stapje proberen er meer van te begrijpen.’

Wat is de grootste uitdaging op het gebied van statistiek op het moment?
‘Ik denk toch wel *uncertainty quantification*, dat heeft te maken met de onzekerheid die altijd gepaard gaat met een schatting. Als arts zou je bijvoorbeeld graag kunnen voorspellen wat voor effect een bepaalde behandeling heeft. Daar zit altijd een mate van onzekerheid in. In de klassieke statistiek is dat het probleem van de opiniepeiling. Stel dat je aan 1000 mensen vraagt wat ze gaan stemmen, en dat 500 daarvan antwoorden: ‘partij A’. Dan is je beste schatting dat vijftig procent van de bevolking op partij A zal stemmen. Maar omdat je dit maar aan 1000 mensen hebt gevraagd, is er altijd een marge waarin je fout kunt zitten. Wat die marge is, kunnen we in zo’n geval vrij goed berekenen. Maar een opiniepeiling is een eenvoudig model. De uitdaging zit hem daar vooral in het kiezen van 1000 willekeurige mensen. Want

je loopt al snel het gevaar dat je steekproef niet representatief is voor de hele bevolking van een land. Maar de berekening zelf is vrij simpel.
Als je een groep patiënten in het ziekenhuis hebt, dan is dat heel anders. Er zijn een heleboel gegevens die een rol kunnen spelen bij het berekenen van het effect van een behandeling. En met al die gegevens wil je rekening houden, ook in de heel ingewikkelde gevallen waar je heel veel gegevens tot je beschikking hebt. Op dit moment weten we hoe we de foutenmarge moeten berekenen met eenvoudige datasets, maar nog niet hoe dat met heel grote of ingewikkelde sets moet. Je zou allerlei verschillende dingen willen kunnen meten en iemand dan kunnen vertellen: 'Dit is de beste therapie en dit is de kans dat hij succesvol is."

Wat ga je met het geld doen?

'Ik weet pas net dat ik de premie gekregen heb, dus ik heb er nog niet heel lang over nagedacht. Maar het eerste waar ik aan zou denken is het onderwijs in de statistiek verder te ondersteunen en proberen om een wat meer zelfstandig geheel te maken van de statistiek.
Er zijn redelijk wat statistici in Nederland. Maar ze zitten heel erg verspreid. Er zitten er een paar bij wiskunde, bij de ziekenhuizen, bij sociale wetenschappen. En bij economie natuurlijk: econometrie is voor een groot deel statistiek. Natuurlijk hebben we contact. Er is ook een vereniging voor statistiek waar iedereen bij zit. En sinds vijf jaar hebben we hier ook een master statistiek bij wiskunde, in samenwerking met het LUMC, het Leidse ziekenhuis, en sociale wetenschappen. Maar ik zou graag een stap maken waarbij statistiek in Nederland ook echt een vak wordt. De uitwisseling die nu plaatsvindt is nog te beperkt. Het is heel nuttig als iedereen van elkaar leert. Want wat er in de sociale wetenschappen gebeurt is toch voor een groot deel hetzelfde als wat er in de medische wetenschappen gebeurt. Hoe ik dat wil doen, weet ik nog niet precies. Maar het geld zal vermoedelijk vooral opgaan aan het inhuren van heel slimme mensen.' ■

anouschka.busch@quest.nl

Wie is Aad van der Vaart?

1959: wordt op 12 juli geboren in Vlaardingen.
1983: studeert cum laude af in de wiskunde aan de Universiteit Leiden.
1987: promoveert in Leiden op statistische schattingen in hoogdimensionale ruimtes.
1987: docent aan de Vrije Universiteit Amsterdam.
1988: wint de C.J. Kok-prijs voor uitzonderlijke begaafdheid op het gebied van onderzoek.
1990: docent aan de Texas A & M University (VS).
1996: wordt hoogleraar stochastiek aan de Vrije Universiteit Amsterdam.
1995: hoogleraar aan de Université Paris XI (Frankrijk).
2000: Miller Fellow aan de University of California in Berkeley (VS).
2000: wint de 5-jaarlijkse Van Dantzig-prijs voor de beste Nederlandse onderzoeker onder de 40 op het gebied van de statistiek en operations research.
2003-2007: President Netherlands Society for Statistics and Operations Research.
2003-2011: gastonderzoeker aan Harvard School of Public Health (VS).
2009: wordt lid van de KNAW.
2012: wordt hoogleraar stochastiek aan de Universiteit Leiden.
2013: krijgt een Advanced Grant van € 2,5 miljoen van de EU voor grensverleggend onderzoek.