‘Statistiek is zo veelzijdig. Ik snap eigenlijk niet dat er niet veel meer statistici zijn’

WETENSCHAPPELIJK ONDERZOEK KLOPT VAAK NIET

Een wirwar van datapunten, onmogelijke grafieken en lastige kansberekeningen: statistiek kan zelfs de beste wetenschappers intimideren. Velen van hen maken dan ook statistische denkfouten, zegt Casper Albers. Of erger nog: ze gebruiken de nevelen van de statistiek om collega’s en het brede publiek om de tuin te leiden.

Foto: Bob Bronshoff

Ans Hekkenberg & Jim Jansen

Hoe groot is een grote kans eigenlijk? Stel dat je op de radio hoort: er is een grote kans dat vanmiddag de zon schijnt. Ga je dan vol vertrouwen naar het strand? Of blijf je thuis, omdat die voorspelling je onzeker in de oren klinkt en je een nat pak wilt vermijden? ‘Een grote kans’ betekent niet voor iedereen hetzelfde, ontdekte de Groningse statisticus Casper Albers met een aantal collega’s. Waar sommige mensen uit de bovenstaande zin concluderen dat de kans op een zonnige middag 60 procent is, menen anderen dat de zon 99 procent zeker schijnt. Verschillende interpretaties van kans-woorden zijn natuurlijk niet zo belangrijk wanneer er een middagje strand op het spel staat. Een ander verhaal wordt het wanneer wetenschappers, artsen en juristen communiceren over risico’s. Een ‘grote’ kans op kanker, een ‘onwaarschijnlijke’ bijwerking of een ‘vermoedelijke’ dader – bij zulke zaken kan een Babylonische spraakverwarring levensgevaarlijk zijn. En de interpretatie van kanswoorden is maar een van de vele manieren waarop mensen dagelijks worstelen met statistiek. Ook hebben we moeite met het interpreteren van grafieken en het herkennen van statistische valkuilen. Een voorbeeld: in mei rapporteerde de Belastingdienst zich schuldig te maken aan etnisch profileren. Mensen met twee nationaliteiten werden vaker geselecteerd voor de belastingaangiftecontrole. Het probleem: met zo’n selectiecriterium vind je wat je zoekt – altijd. Als je meer linkshandigen controleert, vind je veel fraude onder linkshandigen, gewoon doordat zij vaker zijn geselecteerd. ‘Een methodologische doodzonde’, tweette Albers. Toch verbazen zulke blunders hem niet. ‘Ook wetenschappers maken fouten in hun statistische methodes.’ En dat is waar zijn werk begint. ‘Als statisticus ben ik de assistent: ik wil de statistiek van andermans werk zo solide mogelijk maken.’ Betere wetenschap én betere voorlichting, daarvoor maakt Albers zich sterk.

De onderzoeken waar u aan meewerkt, lopen uiteen van het oplossen van problemen tot het helpen van jongeren met psychische stoornissen. Wat is de rode draad in uw werk?

‘Het puzzelen met getallen. De toepassingen wisselen steeds, maar wat hetzelfde blijft, is dat er een onderzoek wordt gedaan met uitdagende statistiek. Aan dat deel draag ik bij. Bij elk project ga ik in gesprek met een expert die verstand heeft van de inhoud; van het verkeer of de mentale gezondheidszorg bijvoorbeeld. Ik heb zelf geen kaas gegeten van die onderwerpen. Wel kan ik helpen ervoor te zorgen dat het onderzoek goed wordt uitgevoerd. Ik adviseer bijvoorbeeld hoe je het beste meetgegevens verzamelt en hoe je die vervolgens analyseert. Eigenlijk vertel ik onderzoekers steeds hoe ze zo goed mogelijk hun eigen vragen kunnen beantwoorden. Dat maakt mijn werk leuk: statistiek is zo veelzijdig dat ik steeds weer bij een ander vakgebied kan aansluiten en nieuwe dingen leer. Ik snap eigenlijk niet dat er niet veel meer statistici zijn. Momenteel ben ik betrokken bij onderzoeken die menselijk gedrag in kaart brengen, bijvoorbeeld van jongeren met psychische problemen. Dat is uitdagend, want menselijk gedrag en emoties zijn ongelofelijk moeilijk te meten. Hoe meet je bijvoorbeeld hoe depressief iemand is? Dat is statistisch gezien een veel grotere uitdaging dan bijvoorbeeld verkeerstoepassingen.

Hoe ziet de adviesrol van een statisticus eruit?

In de ideale situatie ben je van begin tot eind betrokken bij een onderzoeksproject; zelfs voordat de data verzameld worden. Dan kun je namelijk meedenken over welke en hoeveel meetgegevens er eigenlijk nodig zijn om een onderzoeksvraag te beantwoorden. Maar ik heb ook weleens meegemaakt dat collega’s al data verzameld hadden en toen naar mij toe kwamen met de vraag hoe ze met die meetgegevens hun onderzoeksvraag konden beantwoor-

den. En dan moest ik zeggen dat dat met die gegevens helemaal niet kon. Daarvoor had je bijvoorbeeld op een andere manier metingen moeten doen, of aanvullende informatie moeten verzamelen. De Britse statisticus Ronald Fisher zei ooit zoiets als: “Een statisticus om hulp vragen wanneer het experiment al is uitgevoerd, is alsof je de dokter binnenroept nadat de patiënt is overleden. Het beste dat je dan nog kunt doen, is achteraf vaststellen waar het is misgegaan.”

U noemde al even het onderzoek naar jongeren met psychische problemen. Wat wilt u over deze mensen weten?

‘Het project waarbij ik betrokken ben, is van hoogleraar psychologie Maaike Nauta. Zij wil weten of je jongeren die je eerder hebt behandeld voor een psychische stoornis kunt behoeden voor een terugval. Daarvoor moet je achterhalen welke vroege signalen erop duiden dat iemand risico loopt op zo’n terugval. Een slecht slaappatroon bijvoorbeeld, of somberheid. Hoe beter je die signalen kent, hoe eerder je kunt ingrijpen en hoe beter je iemand kunt helpen. Voor dit project heb ik bedacht welke proefpersonen we moesten rekruteren en hoe we metingen moeten doen.

Waarom is het zo lastig om die risicosignalen in kaart te brengen?

‘Omdat je bij mensen veel meer geïnteresseerd bent in het proces dat een individu doorloopt dan in een gemiddelde. Stel, een slecht slaappatroon is een mogelijke risicofactor. Dan wil je niet weten hoeveel uur iemand gemiddeld slaapt. Iemand die elke nacht acht uur slaapt, heeft een heel ander slaappatroon dan iemand die één nacht drie uur en de nacht daarna dertien uur slaapt. Beide personen slapen gemiddeld acht uur, maar die tweede persoon is veel vatbaarder voor een onregelmatig ritme. Dat geldt niet alleen voor slaap, maar ook voor emoties. Als je veel stemmingswisselingen hebt, kun je de dag ’s avonds alsnog waarderen met een gemiddelde zeven. Dat ziet er goed uit, maar sommige delen van je dag waren misschien vreselijk. Aan gemiddelden heb je dus niets. Je wilt eigenlijk live, op elk moment kunnen meten. Daarom werken wij met een app op de telefoons van de jongeren. Die app laat vier keer per dag een piepje horen en stelt dan een aantal vragen. Dat levert hele rijke, nuttige data op. Bij dit project was ik vanaf het begin betrokken. Dan kun je meedenken over elke stap. Hoe vaak moet zo’n alarm afgaan? Kunnen we beter vier weken lang vijf keer per dag meten, of vijf weken lang vier keer per dag? In beide gevallen krijg je twintig metingen, maar waar heb je statistisch gezien het meest aan? Dat soort zaken.

Foto: Bob Bronshoff

Casper Albers (1975) is adjunct-hoogleraar toegepaste statistiek en datavisualisatie aan de Rijksuniversiteit Groningen. Zijn onderzoek richt zich op de ontwikkeling van statistische modellen voor de sociale wetenschappen. Albers maakt zich sterk voor de verbetering van de wetenschap, onder meer als ambassa-deur voor open science, de beweging die pleit voor het openlijk delen van data, analyses en onderzoeksartikelen. Daarnaast schrijft hij columns voor *de Volkskrant* en voor de Groningse universiteitskrant *UKrant*.

Waarom kun je bij dit soort onderzoek niet gewoon de traditionele statistiekmethodes gebruiken?

‘Het feit dat je veel metingen doet gedurende een periode voegt een ingrediënt toe aan de meetgegevens waar de ‘oude’ statistiek niet mee kan omgaan. Normaal gesproken maak je een model: je voorspelt bijvoorbeeld bij studenten hoe goed ze zullen scoren bij een tentamen op basis van metingen van hun inzet en motivatie. Zo’n model werkt prima, maar is nooit perfect. Je maakt altijd wat voorspellingsfouten. Het model houdt er bijvoorbeeld geen rekening mee dat Jim niet kon slapen, of dat Ans plots hard ging leren. In dit geval is dat niet erg, want de voorspellingsfouten zijn bij Jim anders dan bij Ans; ze zijn onafhankelijk van elkaar en stapelen niet op. Maar als je veel metingen na elkaar doet, zoals wij, dan kun je een meetfout die er bij de ene meting insloop opnieuw meenemen in de volgende meting. Om daarvoor te corrigeren, heb je een ander statistisch model nodig.’

Welke verbeteringen voert u dan door in zo’n statistisch model?

‘Eigenlijk wil je altijd weten hoe groot de meetfout is die je meeneemt. Voor hoeveel procent telt die mee in je voorspellingen? De uitdaging is om dat getal zo goed mogelijk in te schatten. Het probleem van veel modellen is dat ze aannemen dat het percentage gelijk blijft. Vaak is dat ook zo. Bijvoorbeeld als je beurskoersen voorspelt. Als de koers gisteren steeg, zal dat vandaag vast ook nog zo zijn. Maar in klinische context, wanneer je patiënten aan het behandelen bent, geldt dat niet. Die patiënten veranderen constant, zeker als hun behandeling werkt. Om daar ruimte voor te maken, moet het getal dat omschrijft voor hoeveel procent je meetfout meetelt mee veranderen. In ons nieuwe model gebeurt dat. Het getal kan daarbij geleidelijk veranderen, maar er is ook ruimte voor wat we een sudden shock noemen. Soms gebeurt er plots iets. Stel, een broer van een patiënt overlijdt. Als zo’n shock plaatsvindt, moet je model dat opmerken. We hebben dat nu getoetst in een simulatie, met succes. Nu gaan we het testen in de praktijk.’

Niet elke sociale wetenschapper heeft een statisticus aan zijn of haar zijde. Gebeurt het vaak dat wetenschappers in statistische valkuilen stappen?

‘Heel vaak. Er gebeurt zoveel bij onderzoeken dat eigenlijk niet mag. Zo zie je bijvoorbeeld weleens dat wetenschappers een groep proefpersonen onderzoeken en geen interessant verband vinden. Maar als ze een paar ongebruikelijke meetpunten weglaten, dan verschijnt zo’n verband wél. Dan is het verleidelijk om te zeggen: ach, die paar afwijkende metingen, die nemen we niet mee. Maar dat mag natuurlijk niet. Als wetenschapper wil je verder dat de zogeheten p-waarde van je onderzoek onder de 0,05 ligt. (Dat wil zeggen dat het gevonden effect zó groot is dat het erg onwaarschijnlijk is dat je het op basis van toeval gevonden zou hebben – red. *New Scientist*). Dan is je resultaat significant. Wat je nog weleens ziet, is dat een onderzoek nét niet onder die 0,05 uitkomt. Wetenschappers voegen dan soms wat extra proefpersonen toe aan hun test, om die lage p-waarde alsnog te krijgen. Ook dat klopt statistisch niet, want zo raken de resultaten vertekend. Hier ligt een taak voor statistici. Ik geef bijvoorbeeld lezingen voor collega’s om ze erop te wijzen dat je dit soort dingen niet mag doen. Dan zie je mensen weleens schrikken. Zo van: “Oh, maar dat doe ik al jaren!” Die collega’s wilden de boel niet verdraaien; ze wisten gewoon niet dat dit niet oké is.

Gebeurt zoiets ook weleens met opzet?

‘Absoluut! Er is een anoniem onderzoek gedaan onder wetenschappers, waarbij ze vragen kregen als: ‘Hoe vaak heb je data-punten aangewezen als uitschieter en weggelaten om de p-waarde van je onderzoek te verbeteren?’ Wat blijkt: meer dan de helft van de wetenschappers maakt zich weleens schuldig aan dubieuze onderzoekspraktijken. En dat geldt voor alle vakgebieden. Je kunt ook bewust valsspelen met randvoorwaarden. In medische studies wordt bijvoorbeeld vaak de leeftijd van proefpersonen gevraagd. Die proefpersonen deel je vervolgens op in bijvoorbeeld drie groepen: jong, midden en oud. Maar met de grenzen van die groepen kun je sjoemelen. Als je bijvoorbeeld geen significant verschil vindt tussen ‘jong’ en ‘midden’, kun je de grens van jonge mensen verplaatsen van 30 naar 35 jaar. Kijken wat er dan gebeurt. Dat gesjoemel is soms overduidelijk terug te zien in een onderzoeksrapportage. Ligt de grens tussen ‘jong’ en ‘midden’ op bijvoorbeeld op 32 jaar en 2 maanden? Dat is een rode vlag. Niemand kiest zomaar zo’n willekeurige grens.

Waarom spelen zoveel wetenschappers vals?

‘Dat is voor een groot deel te wijten aan hoe de weten-

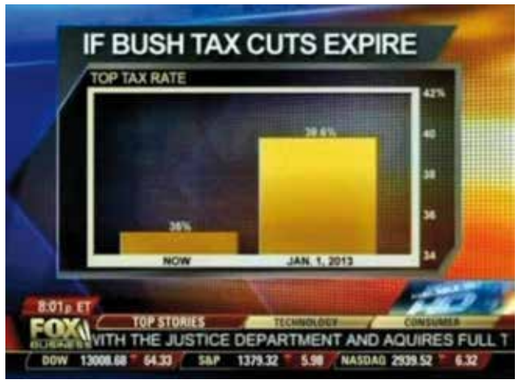
schap werkt. Als wetenschapper doe je onderzoek, schrijf je daar een artikel over en wil je dat vervolgens publiceren in een belangrijk wetenschappelijk blad. Die bladen accepteren eigenlijk alleen artikelen met interessante, tastbare resultaten. Dus móét je er eigenlijk wel voor zorgen dat er dat soort resultaten in je onderzoek zitten. Alleen dan maak je kans op een publicatie. En zo'n publicatie is cruciaal: als je niet publiceert, maak je minder kans op financiering. En als je geen financiering krijgt, wordt je contract bij de universiteit niet verlengd. Je moet eigenlijk wel valsspelen om een baan aan de universiteit te krijgen. Gelukkig zijn steeds meer mensen ervan overtuigd dat dit zo niet langer werkt. Je zou tegen zo'n wetenschappelijk blad moeten kunnen zeggen: als ik deze onderzoeksvraag beantwoord, willen jullie mijn publicatie dan hebben? Zo ja, dan zouden ze op dat moment moeten garanderen dat je werk gepubliceerd wordt – ook als de resultaten niet interessant blijken te zijn. Op die manier vallen de perverse prikkels weg om je onderzoeksdata te masseren.

Hoe groot is de impact van die perverse prikkels?

'Erg groot. Door dit soort trucjes vinden onderzoekers veel vaker significante resultaten dan wanneer het werk eerlijk zou gebeuren. Je ziet dat terug als je de p-waardes van veel onderzoeken op een rijtje zet. Dan blijkt dat er vrijwel geen studies zijn met een p-waarde van net boven de 0,05. Bij die onderzoeken masseren mensen die waarde dus naar net onder de 0,05. We zien dat dit gebeurt bij minimaal eenderde van de onderzoeksartikelen. Als je drie wetenschappelijke artikelen leest, zit er daardoor gemiddeld één tussen waarbij op een verkeerde manier met data is omgegaan. En je weet natuurlijk niet welke.

Niet alleen wetenschappers worstelen met statistiek. Welke valkuilen zijn er voor 'de gewone mens'?

'Veel mensen vinden het lastig om data goed te interpreteren. Dat wordt nog veel lastiger wanneer je geconfronteerd wordt met misleidende grafieken. Daar zijn er veel van. Zo deelde *Fox News* ooit een grafiek hoe de belastingtarieven zouden veranderen zodra maatregelen van Bush werden afgeschaft (figuur 1). De verandering leek enorm: na de afschaffing leek het tarief vijf keer zo hoog als ervoor. Maar Fox had de y-as van een grafiek gemanipuleerd. Die begon niet bij 0, maar bij 34 procent. Het tarief steeg in werkelijkheid maar met een paar procentpunt. Een ander voorbeeld is nog erger: het omkeren

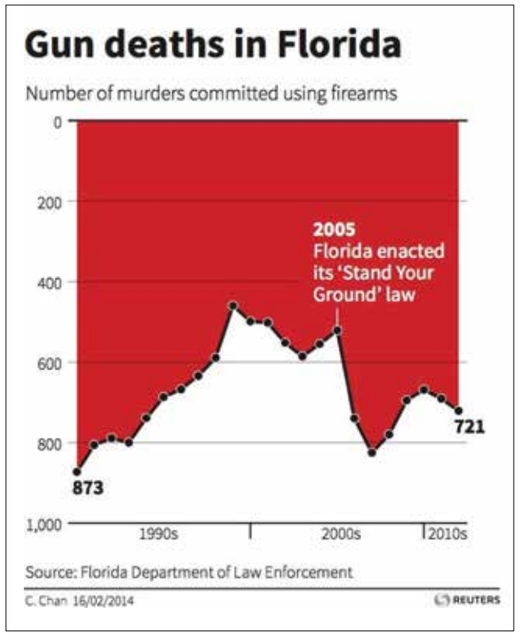


Figuur 1. In de grafiek lijkt het verschil tussen de linker- en rechterstaaf enorm. dat komt vooral doordat de y-as niet bij 0 begint maar bij 34

van de y-as. Dat gebeurde bijvoorbeeld bij een grafiek van Amerikaanse veiligheidsautoriteiten (figuur 2). Die grafiek liet zien hoe het aantal moorden met vuurwapens in Florida veranderde met de tijd. Er was een punt aangegeven waarop er een nieuwe wet was ingevoerd. Het aantal moorden leek daarna te kelderen. Maar de y-as stond verkeerd om: hoe hoger de grafiek, hoe mínder doden. Geen wonder: de geïntroduceerde wet was de *stand your ground law*. Die zegt dat als iemand op je terrein komt, je hem eerst mag neerschieten en daarna pas hoeft te vragen wat hij daar deed. Dit waren bewuste pogingen om het publiek te misleiden. Maar niet elke slechte grafiek is misleidend bedoeld. Vaak zie ik grafieken waarvan ik denk, tja, als je in Excel werkt en op de standaardgrafiekknop drukt, dan is dit wat je krijgt. Dat zijn niet altijd goede grafieken.

Valt hier iets aan te doen?

Het begint met goed nadenken over hoe je je onderzoeksresultaten wil laten zien. Ik zoek altijd manieren om meetgegevens duidelijk weer te geven. Zo heb ik onlangs geholpen bij een project van een collega die de aardbevingsproblemen in Groningen had gemeten. Het rapport liet zien wanneer er bevingen waren, hoe sterk die waren en waar de schademeldingen vandaan kwamen. We hebben die data samengevat in een filmpje. Daarin zie je op een kaart bevingen verschijnen die gevolgd worden door meldingen. Je ziet soms ook geen beving, maar wel meldingen. Of juist wel een beving en geen meldingen. Dat



Figuur 2. In de grafiek lijkt het aantal doden door vuurwapens af te nemen na 2005. De y-as staat echter verkeerd om; er was dus juist sprake van een toename

filmpje maakt de informatie heel inzichtelijk.

Meer structureel: ik wil een project beginnen om uit te vinden welke mensen er beter in zijn om grafieken te snappen. Ligt dat bijvoorbeeld aan het opleidingsniveau? Beeldgevoeligheid? Hoe introvert of extrovert iemand is? Kortom: met welke persoonlijkheidskenmerken hangt 'grafiekgeletterdheid' samen? Als je dat weet, kun je ook beter informatie geven, toegespitst op je doelgroep. Stel bijvoorbeeld dat vrouwen taartdiagrammen beter kunnen lezen en mannen staafdiagrammen. Dan kun je medische flyers voor vrouwen beter op een andere manier ontwerpen dan voor mannen.'

Veel mensen kijken nu meer dan gebruikelijk naar grafieken, bijvoorbeeld over het aantal coronagevallen. Wat adviseert u deze mensen?

'Om niet naar grafieken te kijken. Een grafiek laat misschien zien hoeveel zieken er nu zijn, maar dat vertelt je niets over hoe dat aantal zich gaat ontwikkelen. Je kunt die grafieken beter laten voor wat ze zijn en je laten voorlichten door experts. Kijk, als je met een jumbojet wilt reizen en iemand doet de motorkap open, dan kun je daar ook best naar kijken. Maar het heeft geen zin. Als er een technicus naast staat die zegt dat alles in orde is, dan ben ik geneigd om dat te geloven. Vertrouw dus de mensen die er verstand van hebben.

Gebeurt statistiekverwarring alleen door grafieken? Of geldt het ook voor gesproken en geschreven taal?

'Mensen interpreteren kanswoorden heel verschillend. Woorden zoals 'soms', 'waarschijnlijk' en 'mogelijk'. Neem bijvoorbeeld de zin: "Ik kom soms te laat." In hoeveel gevallen kom je dan te laat? De een denkt aan 5 procent van de gevallen, de ander aan 50 procent. Bij sommige woorden loopt dat enorm uiteen. Bij andere kanswoorden is de interpretatie veel consistenter. Als je over kansen spreekt, is het goed om je daarvan bewust te zijn. Ik geef graag zowel het kanswoord als de kwantitatieve informatie. Dus "soms, in 18 procent van de gevallen..." Zo neem je de onzekerheid weg.

Op Twitter bemoeit u zich vaak met maatschappelijke kwesties, van politieke stemmingen tot oneerlijke controles bij de Belastingdienst. Waar komt die drang vandaan?

'Ik denk dat ik een zendingsdrang heb, een missie. Ik hoop mensen te kunnen helpen dingen op waarde te schatten. Om fouten of onjuistheden te zien. Om ingewikkelde concepten te begrijpen. Als ik kritiek lever, is dat op wetenschappelijke gronden. Als ik dan iets zeg over klimaatverandering of Forum voor Democratie, dan gaat het op Twitter los. Sommige mensen beginnen meteen te schelden. Nou ja. Als je geen inhoudelijke argumenten hebt, lig ik daar niet wakker van; dan heb je de discussie al verloren.

Het interview verscheen eerder in *New Scientist* (2020)79 en is met dank overgenomen.

Ans Hekkenberg is natuur- en sterrenkundige en redacteur bij de Nederlandstalige editie van het populairwetenschappelijk tijdschrift *New Scientist*. Ze spreekt regelmatig als studiogast op radio en televisie over de laatste ontwikkelingen in de wetenschap. Je vindt Ans als @GirlforScience op Twitter en Instagram.

Jim Jansen is hoofdredacteur bij *New Scientist* en wetenschapscoördinator bij *Het Parool*. Op zaterdag verzorgt hij de rubriek 'Eureka' voor het *Algemeen Dagblad*.