Toevalstreffers

INAUGURELE REDE DOOR PROF. DR. J.J. (JELLE) GOEMAN

change perspective

Radboud Universiteit

INAUGURELE REDE
PROF. DR. J.J. (JELLE) GOEMAN

Regelmatig is in de media te lezen dat de wetenschap in crisis verkeert. Veel wetenschappelijke resultaten blijken bij herhaling van het experiment niet reproduceerbaar, en de vraag is of niet een meerderheid van alle wetenschappelijke resultaten onjuist is. Over de rol van de statistiek bij dit probleem zijn twee meningen te beluisteren. Sommigen vinden dat onderzoekers jarenlang door statistici de verkeerde statistische methoden onderwezen hebben gekregen. Anderen zeggen juist dat de onderzoekers de teugels voor zichzelf te veel hebben laten vieren. Iedereen zou weer netjes de statistische regels moeten volgen, gecontroleerd door de statistici. Statistici zelf waren tot nu toe opvallend stil in dit debat. Prof. Jelle Goeman belicht de crisis in de wetenschap vanuit statistisch perspectief en pleit ervoor dat onderzoekers methodologie, en dus ook statistiek, niet zien als een stel regels dat gevolgd moet worden, maar als de wetenschappelijke kern van hun vakgebied.

Prof. Jelle Goeman (1976) is sinds 1 oktober 2013 hoogleraar Biostatistiek aan de Radboud Universiteit/het Radboudumc. Hij is als statisticus betrokken bij veel verschillende soorten medisch onderzoek, met name op het gebied van *genomics*. Zijn belangrijkste aandachtsgebied is het ontwikkelen van nieuwe statistische methoden op het gebied van meervoudig hypothesetoetsen.

Radboud Universiteit

TOEVALSTREFFERS

Toevalstreffers

Rede uitgesproken bij de aanvaarding van het ambt van hoogleraar Biostatistiek aan de Radboud Universiteit/het Radboudumc op vrijdag 20 juni 2014

door prof. dr. J.J. (Jelle) Goeman

Vormgeving en opmaak: gloed*communicatie*, Nijmegen
Fotografie omslag: Bert Beelen
Drukwerk: Van Eck & Oosterink

Geachte rector magnificus,
hooggeleerde collegae,
beste toehoorders,
lieve vrienden en familie,

In 2010 voerden vier breinonderzoekers een opmerkelijk experiment uit. Zij legden een proefpersoon in een hersenscanner en boden deze een aantal stimuli aan in de vorm van emotioneel geladen beelden. Op basis van dit experiment konden zij onderdelen van het brein aanwijzen die reageerden op de aangeboden stimulus, en die dus waarschijnlijk een functie hadden in emotionele verwerking. Het experiment werd uitgevoerd en geanalyseerd volgens de geldende normen van het vakgebied, en zou goede kans hebben gehad om geaccepteerd te worden door een vooraanstaand wetenschappelijk tijdschrift. Op één detail na, en dat was de proefpersoon. Dit was niet, zoals gebruikelijk, een mens, maar een Atlantische zalm, enige dagen daarvoor uit de oceaan gevist, en diezelfde morgen nog in de supermarkt gekocht. Morsdood.

De studie werd gepubliceerd in het *Journal of Serendipitous and Unexpected Results,* zeg maar het Tijdschrift voor Toevalstreffers, en won in 2012 de *IgNobel prize*, een prachtige prijs voor onderzoek dat je eerst aan het lachen maakt, en je vervolgens achter je oren doet krabben. Wat leert deze studie ons? Niet zoveel over de gedachtenspinsels van een Noorse zalm, pijnzend over de fjorden. Des te meer zegt het over onderzoeksmethoden. De ondertitel van het artikel luidt, vertaald, 'Een pleidooi voor goede correctie voor meervoudig toetsen.' De onderzoekers pleiten daarmee heel direct voor meer gebruik van het soort statistische methoden dat ik in mijn onderzoek ontwikkel.

Maar ik wil het vandaag hebben over de bredere vraag die door deze onderzoekers impliciet wordt opgeworpen: Hoe betrouwbaar zijn de resultaten van wetenschappelijk onderzoek eigenlijk? Deze vraag is zo oud als de wetenschap zelf, maar is tegelijkertijd ook heel actueel. Een verhitte discussie hierover heeft zich de laatste jaren ontsponnen in de wetenschappelijke en in de populaire literatuur. Een van de eersten die de noodklok luidden was John Ioannidis, die in 2005 met theoretische argumenten betoogde dat de meeste medische onderzoeksresultaten onjuist zijn. Sindsdien verschijnen er met enige regelmaat ook in de populaire media doemverhalen over de toestand van de wetenschap, zoals bijvoorbeeld in 2013 een groot pessimistisch omslagverhaal in de Economist, getiteld, *How science goes wrong*. Over het probleem zijn de meeste auteurs het wel eens. Veel te veel wetenschappelijke resultaten lijken niet stand te houden als het onderzoek nog eens wordt overgedaan. De statistici Leek en Storey hebben in 2014 geprobeerd in te schatten welk percentage van de bevindingen in vooraanstaande medische tijdschriften onjuist is. Hun schatting van 14 procent (1 op 7) werd door de meeste commentatoren als een sterke onderschatting gezien. Een op 3 of 4 lijkt op theoretische gronden realistischer en er zijn aanwijzingen dat het percentage onjuiste bevindingen in de toptijdschriften nog hoger ligt dan in minder vooraanstaande bladen. Om echt uit

te vinden hoe groot het percentage vals positieve resultaten is, zal grootschalig replicatieonderzoek moeten worden uitgevoerd. De psychologie loopt hierin voor op de geneeskunde. Dit jaar verscheen een special issue van *Social Psychology* met pogingen tot replicatie van 13 recente psychologische studies. In 11 van de 13 gevallen lukte het opnieuw om een effect te vinden, hoewel het gevonden effect vaak wel kleiner was dan in de oorspronkelijke studie. In 2 van de 13 gevallen was er bij replicatie helemaal geen effect meer te vinden.

Dit zijn hoe dan ook schokkende cijfers. Wetenschappers, voor wie het hoogste doel is om uit te vinden hoe het echt in elkaar zit, publiceren vrij vaak resultaten die achteraf de toets der kritiek toch niet kunnen weerstaan. Hoe kan dit zo zijn?

Onjuiste conclusies in wetenschappelijk onderzoek zijn nooit helemaal uit te sluiten. Toevallige variatie in onderzoeksgegevens kan de onderzoeker op het verkeerde been zetten, of belangrijke systematische factoren die van invloed zijn op de metingen kunnen over het hoofd gezien zijn. In het ergste geval trekt een onderzoeker als gevolg van dergelijke problemen de verkeerde conclusie uit het experiment. Hij of zij concludeert bijvoorbeeld dat een medicijn wel werkt, terwijl dat in werkelijkheid niet het geval is, of vindt belangrijke hersenprocessen in een dode zalm. Dit noemen we een vals positief resultaat. Of andersom, de onderzoeker concludeert dat het medicijn niet werkt, terwijl dat in werkelijkheid wel zo is: een vals negatief resultaat.

Vals positieve en vals negatieve resultaten zijn allebei vervelend, maar hebben niet dezelfde consequenties. Vals negatieve resultaten zijn vooral een probleem voor de onderzoeker zelf. Die heeft geld, tijd en prestige geïnvesteerd in een onderzoek waar geen belangwekkend resultaat uit gekomen is, en loopt daardoor een mooie publicatie mis. Een vals positief resultaat daarentegen is juist een probleem voor de wetenschap als geheel. Vals positieve resultaten leiden tot investeringen in vervolgonderzoek dat waarschijnlijk nergens toe zal leiden. Het grote risico is dat de resultaten in de handboeken terecht komen en deel uit gaan maken van de canon van een vakgebied.

Op de lange duur, zo wordt gezegd, zullen zowel vals positieve als vals negatieve resultaten worden opgeschoond als gevolg van verdere ontwikkeling van het wetenschapsgebied. De kracht van de wetenschap ligt voor een belangrijk deel in dit zelfreinigend vermogen. Inderdaad zou je mijns inziens wetenschapsgebieden kunnen rangschikken op hun wetenschappelijkheid op basis van precies dit criterium. Hoe gemakkelijker het is voor een junior onderzoeker om overtuigend aan te tonen dat een gevestigde naam in het veld het bij het verkeerde eind heeft, hoe wetenschappelijker het gebied, en hoe sneller dat wetenschapsgebied vooruit zal gaan. Toch zijn er ook grenzen aan dit zelfreinigend vermogen. Hoeveel vals positieve resultaten kan het systeem aan? In de methodologie geldt bovendien net als in de geneeskunde: voorkomen is beter dan genezen.

Om vals positieve resultaten te voorkomen gebruiken onderzoekers vaak statistische methoden. Hier zijn we bij mijn eigen vakgebied aanbeland. Het mooiste vakgebied van de wereld. Statistiek is toegepaste wiskunde, maar tegelijkertijd ook methodologie, en

dus toegepaste wetenschapsfilosofie. Het vakgebied gaat over de vraag hoe wetenschappelijke experimenten het best kunnen worden opgezet. Het gaat over de vraag welke conclusies uit onderzoeksgegevens kunnen worden getrokken, en hoe zeker we kunnen zijn over die conclusies. Methodologie gaat over de manier waarop wetenschap bedreven kan en moet worden, en vormt daarmee het kloppend hart van het wetenschappelijk bedrijf.

Statistiek moet onderzoekers dus ook helpen vals positieve resultaten te voorkomen. Het is daarom niet verwonderlijk dat statistiek steeds terugkomt in de discussie over de reproduceerbaarheid van wetenschappelijk onderzoek, of het gebrek daaraan. Volgens sommigen is juist de statistiek schuldig aan de geconstateerde problemen. De klassieke statistiek met zijn p-waarden en hypothesetoetsen zou hoogst onbetrouwbaar zijn en onderzoekers op het verkeerde been zetten, zo konden we dit jaar in *Nature* lezen. Volgens deze critici moet statistiek dan ook over een geheel andere boeg. Een radicale omslag naar statistiek op zuiver Bayesiaanse grondslag zou dan wenselijk zijn. Volgens anderen ligt de schuld juist bij de onderzoekers. Onder druk gezet door hoge publicatiedruk nemen zij het niet zo nauw met de statistische criteria. Volgens deze auteurs moeten onderzoekers door tijdschriften goed in het gareel gehouden worden. Statistici spelen in hun visie de rol van de accountant die controleert of alle statistische recepten wel naar behoren zijn uitgevoerd. Een voorbeeld van deze opvatting is het beleid van het *Journal of the American Medical Association*, dat enige tijd voor alle submissies eiste dat een statisticus tekende voor de correctheid van de uitgevoerde analyses, zelfs als er zich onder de auteurs geen statisticus bevond. Van dit beleid zijn zij gelukkig inmiddels teruggekomen.

Ik ben ervan overtuigd dat statistici niet de oorzaak zijn van de huidige crisis in de wetenschap, maar ik ben er helemaal niet blij mee als statistici als politiemensen onderzoekers moeten controleren. Ik zal in het komende half uur uitleggen hoe ik denk dat vals positieve resultaten voorkomen kunnen worden, en wat de rol van statistiek daarbij kan zijn.

Omdat ik naast statisticus ook historicus ben wil ik eerst met een historisch perspectief beginnen. Met ons competitieve beloningssysteem, onze voortsnellende technologie en ons exponentieel groeiend aantal publicaties, denken we misschien dat het wetenschappelijk onderzoek een ongekende graad van efficiëntie heeft bereikt. Ik ben echter juist vaak onder de indruk van de enorme wetenschappelijke vooruitgang die wetenschappers in de periode voor de Tweede Wereldoorlog voor elkaar wisten te krijgen. Met beperkte middelen en beperkte menskracht werd op veel wetenschapsgebieden een solide theoretisch fundament gelegd.

Hoe gingen deze mensen te werk? Hoe voorkwamen zij vals positieve resultaten? Laten we als voorbeeld eens kijken naar het werk van Luigi Galvani en Alessandro Volta in de achttiende eeuw. In die tijd wist men dat het mogelijk was via ontladingen van statische elektriciteit de ledematen van dode en levende mensen en dieren te laten

bewegen. De vraag was hoe dat verklaard kon worden. Galvani, in Bologna, geloofde dat elektriciteit vanuit de hersenen naar de spieren werd gestuurd, en daar werd opgeslagen als statische elektriciteit. Deze elektriciteit leverde vervolgens de energie voor de beweging. Door de elektrische prikkeling door de onderzoeker werd deze energie volgens hem vrijgemaakt. Volta, in Pavia, was het hier niet mee eens. Volgens hem was het de elektriciteit die de onderzoeker aanbracht, niet de intern aanwezige elektriciteit, die de beweging van de ledematen veroorzaakte. Volta en Galvani correspondeerden uitgebreid over deze kwestie, maar geen van beiden wist de ander te overtuigen. In 1781 bedacht Galvani wat hij zag als het definitieve experiment. Hij hing een dode kikker aan een ijzerdraadje waaraan ook een koperen draad bevestigd was. Als hij met die koperdraad vervolgens de opengewerkte zenuwen van de kikker beroerde, bewogen de poten van de kikker op dezelfde manier als wanneer ze met een elektrische lading geprikkeld werden. Het bewijs was voor Galvani geleverd: de elektriciteit voor de beweging kwam uit de kikker zelf. Volta herhaalde de experimenten van Galvani, maar was desondanks nog niet overtuigd van diens verklaring. Hij hield vol dat de elektriciteit toch van elders kwam. Veel later, in 1800, wist hij met een ander experiment aan te tonen dat contact tussen verschillende metalen, zoals het koper en het ijzer in het experiment van Galvani, een klein elektrisch stroompje kon opwekken. Vanuit dit inzicht ontwikkelde Volta vervolgens de allereerste accu.

De uitwisseling tussen Volta en Galvani was uiterst productief, en heeft zowel de fysiologie als de natuurkunde een flinke stap vooruit geholpen. De uitwisseling is ook vrij typisch voor die tijd en laat een manier van wetenschapsbeoefening zien die niet meer zo vaak voorkomt. Wat valt op? We zien twee wetenschappers die ieder sterk gedreven waren door een eigen theorie. Ze waren het onderling heftig oneens, maar bleven desondanks voortdurend in dialoog. Die dialoog bestond eruit dat ze elkaar uitdaagden met experimenten die de eigen theorie ondersteunden, en die van de ander probeerden te ontkrachten. Het lijkt voor mij een beetje op een *hip hop battle* tussen twee rappers, maar ik kan me voorstellen dat niet iedereen deze associatie heeft. Binnen deze dialoog zijn pogingen tot replicatie van succesvolle experimenten vanzelfsprekend: Volta geloofde het experiment van Galvani pas nadat hij het met eigen ogen gezien had. Desondanks had hij een eigen verklaring voor de uitkomsten. Replicatie is niet genoeg. Het gaat tenslotte om de implicaties van het experiment voor de theorie.

Een competitieve samenwerking tussen twee onderzoekers met diametraal tegengestelde theoretische ideeën, zoals we bij Volta en Galvani zien, is een ideale voedingsbodem voor methodologisch hoogstaand onderzoek. Voor Galvani was Volta de beroepsscepticus, altijd alert op onjuiste aannames, verkeerd opgezette experimenten of overhaaste conclusies. Hij kon erop rekenen dat Volta al zijn belangwekkende experimenten onmiddellijk zou herhalen, en de zwakke punten ervan zou aanvallen. Tegelijkertijd gaf de competitie met Volta voor Galvani richting aan zijn experimenten. Zijn doel was tenslotte om Volta's theorieën te ontkrachten. Zijn experimenten moesten

er op gericht zijn om precies de zwakke punten van de theorie van Volta bloot te leggen. Alleen experimenten waarvan Galvani dacht dat Volta een ander resultaat verwachtte dan hijzelf, waren in de context van hun onderlinge strijd zinvol. Dat is wetenschap op het scherpst van de snede.

Vanuit het inzicht dat samenwerking tussen wetenschappers met concurrerende visies uiterst productief is, ontwikkelde de Groningse hoogleraar psychologie Willem Hofstee het zogenaamde weddenschapsmodel voor wetenschappelijk onderzoek. Hoe gaat dit in zijn werk? Een onderzoeker die een experiment wil uitvoeren zoekt allereerst contact met een andere wetenschapper die een andere theorie aanhangt en vanuit die theorie een ander resultaat van het experiment verwacht. Laten we hem de scepticus noemen. Als een dergelijke wetenschapper niet te vinden is, is het niet nodig om het experiment uit te voeren: er is dan toch niemand die van de resultaten zou opkijken. Dat voorkomt alvast een heleboel onnodige experimenten. Als de onderzoeker zijn scepticus gevonden heeft gaan ze samen om de tafel zitten om precies af te spreken hoe het experiment uitgevoerd zal worden. Alle methodologische haken en ogen worden vanuit twee perspectieven onder de loep genomen. Een onafhankelijk statisticus kan hierbij nog helpen in de rol van arbiter en ervoor zorgen dat toevalsprocessen niet de onderzoeker of juist de scepticus bevoordelen. Het symmetrische besliskundige hypothesetoetsmodel van Neyman en Pearson past heel goed bij deze opzet. Als de onderzoekers het eens zijn geworden kan het experiment, eventueel in tweevoud op twee labs, worden uitgevoerd. Het eindresultaat zou dan een gezamenlijke publicatie zijn waarin het experiment met zijn resultaten beschreven en geïnterpreteerd wordt. Mochten de onderzoekers het uiteindelijk oneens zijn over de interpretatie, dan kan ieder zijn eigen discussie schrijven waarin de implicaties van het experiment vanuit de eigen theorie belicht worden.

Dit prachtige model wordt in de praktijk eigenlijk zelden of nooit gebruikt. Ik vind dat jammer, maar ik begrijp dat ook heel goed. Wetenschap is mensenwerk en wetenschappers hebben vaak grote ego's. Ik ben bang dat het uitvoeren van dit model minstens zo vaak een knallende ruzie zal opleveren als een baanbrekende publicatie. Volta en Galvani had je ook niet samen aan één tafel gekregen om het eens te worden over een gezamenlijk experiment.

Het model van wetenschapsbeoefening van Galvani en Volta en het weddenschapsmodel hebben allebei aandacht voor methodologie die op een natuurlijke manier is ingebouwd in het model. Beide modellen doen dit via een wetenschappelijke concurrent die een expliciete rol krijgt toebedeeld, en over de schouder van de onderzoeker meekijkt. Dit extra paar kritische ogen zorgt ervoor dat vals positieve resultaten veel minder kans hebben om de wetenschappelijke literatuur binnen te sluipen.

Hoe zit dat bij de huidige onderzoekspraktijk? Bij verreweg de meeste wetenschappelijke artikelen is geen kritische scepticus betrokken. In contrast met het weddenschapsmodel zouden we naar het huidige model van wetenschappelijk publiceren kunnen verwijzen als het 'Wedden dat…'-model, genoemd naar de bekende spelshow met Jos

Brink en Sandra Reemer. Dit vindt u misschien een heel vreemde vergelijking, maar ik zal nog uitleggen waarom ik deze vergelijking passend vind.

Hoe werkt dit model? Laten we het eerst illustreren met een geval waarin het model goed werkt. Neem het bekende Blikkie onderzoek uit 2013 van de voedingshoogleraar Martijn Katan. Katan wilde aantonen dat dagelijkse consumptie van suikerhoudende frisdrank kinderen doet aankomen. Dit lijkt misschien een open deur, maar er waren onderzoekers (en frisdrankmaatschappijen) die dachten dat kinderen voor die suikerhoudende frisdrank compenseerden door meer te bewegen of door van andere dingen minder te eten. Katan gaf 650 kinderen op verschillende scholen een jaar lang ofwel een suikerhoudende frisdrank, ofwel een gelijk smakende frisdrank met zoetstof. Welke drank de kinderen steeds kregen werd door het lot bepaald en niet aan de kinderen of hun ouders verteld. Na een jaar werd gemeten of de kinderen die de suikerhoudende frisdrank dronken meer waren aangekomen dan de kinderen die de zoetstofvariant gekregen hadden. Dit was inderdaad het geval. Een jaar wel of niet suikerhoudende frisdrank drinken bleek een verschil van wel een kilo lichaamsgewicht te betekenen. De bevinding werd naar het tijdschrift *Contemporary Clinical Trials* gestuurd. Anonieme reviewers beoordeelden het stuk op kwaliteit en nieuwswaarde en gaven groen licht voor publicatie.

Dit is serieus, origineel en belangwekkend wetenschappelijk onderzoek. Waarin, zult u misschien vragen, lijkt dit op een spelshow? De show 'Wedden dat…' had het volgende format. Een kandidaat dient zich aan met een bijzondere vaardigheid. Hij kent bijvoorbeeld het hele telefoonboek van Haarlem uit zijn hoofd of kan een auto laten rijden over twee strakgespannen touwen. Deze vaardigheid wordt tijdens de show, live, op de proef gesteld. De telefoonboekkenner krijgt bijvoorbeeld vijf willekeurige namen te horen en moet daarbij de bijbehorende telefoonnummers reproduceren. Slaagt hij voor deze proef, dan wordt de vaardigheid overtuigend bewezen geacht. Eeuwige roem volgt.

Wat is nu de parallel met wetenschappelijk publiceren? Welnu, de claim van Katan was niet dat hij een bijzondere vaardigheid bezat, maar de wetenschappelijke claim dat kinderen door het drinken van suikerhoudende drank aankomen. Net als in 'Wedden dat…' werd deze claim aan een proef onderworpen, in dit geval in de vorm van een experiment dat een jaar duurde. De rol van Jos Brink werd gespeeld door de reviewers: zij beoordeelden het experiment als geslaagd, en als spectaculair genoeg om uit te zenden. De claim van Katan werd overtuigend bewezen geacht.

In het voorbeeld van het onderzoek van Katan is het gemakkelijk om te zien dat het model hier goed werkt om vals positieve resultaten te voorkomen. Allerlei elementen in zijn onderzoeksopzet van klinische trials zijn herkenbaar als elementen die in de spelshow de prestatie extra overtuigend zouden maken. Het is heel precies gedefinieerd wat Katan wil laten zien, en de manier waarop dat gemeten wordt is eenduidig en van tevoren vastgelegd. Katan doet maar één poging, en tijdens die poging wordt niet flexibel met de criteria omgegaan. De criteria zijn bovendien zo gesteld dat het onwaarschijnlijk is dat Katan door stom geluk aan deze criteria zou voldoen.

Vergelijken we de opzet die Katan gevolgd heeft met het weddenschapsmodel van Hofstee, dan kunnen we vaststellen dat het onderzoek onder dat model nauwelijks anders zou zijn uitgevoerd. Katan zou dan een onderzoeker hebben gevonden die zou denken dat het niets uitmaakt of kinderen hun calorieën als frisdrank of als aardappelen binnen krijgen. Zo'n scepticus had zich waarschijnlijk prima kunnen vinden in de gebruikte onderzoeksopzet.

We zouden dus kunnen zeggen dat Katan zelf de rol van scepticus overtuigend gespeeld heeft. In feite wordt deze rol expliciet gemaakt in de statistische methoden die hij gebruikt heeft. De mening van de scepticus heet in statistisch jargon de nulhypothese. Op basis van deze nulhypothese wordt een p-waarde berekend, die we zouden kunnen omschrijven als een maat in hoeverre deze virtuele scepticus na het onderzoek nog steeds achter zijn oorspronkelijke mening staat. We hebben afgesproken dat de virtuele scepticus pas overtuigd is als deze p-waarde onder de 5 procent ligt, wat betekent dat de scepticus maar 1 van de 20 keer dat hij gelijk heeft, dat gelijk niet krijgt.

Het is heel moeilijk om sceptisch te staan ten opzichte van eigen werk, en niet alle onderzoekers zijn daartoe in staat. Laten we ter illustratie naar een ander voorbeeld uit het voedingsonderzoek kijken. Grote krantenkoppen in 1995 kondigden aan dat het eten van tomaat de kans op prostaatkanker met tientallen procenten zou verminderen. Het ging dan niet om verse tomaten, maar juist om geconcentreerde tomaten in de vorm van ketchup, pizza, tomatensoep en zelfs chips met ketchupsmaak. Dat had Edward Giovannucci uit Harvard uitgevonden. Het kwam volgens hem vanwege het stofje lycopeen, dat de vrije radicalen afving die de kanker veroorzaakten. Het artikel heeft de afgelopen twintig jaar grote invloed gehad, met meer dan duizend citaties in de wetenschappelijke literatuur. Maar was het eigenlijk een overtuigend resultaat om mee te beginnen? Lezen we in detail wat Giovannucci gedaan heeft om tot zijn conclusie te komen, dan zien we dat hij voor 46 verschillende voedingsmiddelen tegelijk heeft onderzocht of er een relatie was met prostaatkanker. Bij 4 van de 46 wist hij een relatie te vinden. Die vier waren allemaal gerelateerd aan industrieel verwerkte tomaat, maar dat was niet het idee waarmee Giovannucci met zijn onderzoek begonnen was.

Vinden we dit overtuigend? Als we deze vraag willen beantwoorden helpt het om erover na te denken in termen van het 'Wedden dat…'-model of het weddenschapsmodel. Als het onderzoek van Giovannucci een 'Wedden dat…' -aflevering was geweest, hadden we een kandidaat gezien die beweerde dat hij 46 verschillende dingen kon, en vervolgens 42 keer voor de test faalde. Als deze kandidaat achteraf een mooi verhaal vertelt dat zijn echte kwaliteiten liggen bij de gemeenschappelijke noemer van de vier succesvolle pogingen, zou u hem geloven? Ikzelf was allang weggezapt. Als we ons voorstellen dat Giovannucci een scepticus bij zijn onderzoek had betrokken die niet gelooft in relaties tussen voeding en prostaatkanker, zou deze zich hebben kunnen vinden in de gebruikte onderzoeksopzet? Ik betwijfel het.

Het onderzoek van Giovannucci is voor mij een typisch voorbeeld van een soort onderzoek waarbij methodologie en statistiek niet meer lijkt te zijn dan het volgen van regeltjes. Onderzoekers hebben meegekregen dat nulhypotheses geformuleerd moeten worden en p-waarden moeten worden uitgerekend, maar hebben niet altijd oog voor de plek van nulhypothesen en p-waarden in het wetenschappelijk discours, de dialoog met de scepticus. Er wordt dan over statistiek gedacht in termen van bewijs, niet in termen van overtuiging, een gedachtengang die sterk wordt bevorderd door het 'Wedden dat…'-model.

Twee grote gevaren liggen op de loer bij een dergelijke mechanische benadering van statistische analyse. Het eerste probleem is dat van de stroman. Het is een bekende debattechniek om de standpunten van de tegenstander veel radicaler af te schilderen dan ze zijn, om ze zo des te gemakkelijker te kunnen neersabelen. Precies hetzelfde gebeurt, bewust of onbewust, met nulhypothesen. Het is gemakkelijk om een nulhypothese te kiezen die eenvoudig te verwerpen is, maar waarin geen enkele wetenschapper gelooft. Dan zal het verwerpen ervan ook geen enkele wetenschapper werkelijk ergens van overtuigen. Het tweede probleem is dat van selectie. Als we, zoals Giovannucci, veel verschillende vragen tegelijk stellen, als er verschillende manieren zijn om een statistisch model of een nulhypothese te formuleren, of als er allerlei subgroepen van patiënten zijn waarnaar we kunnen kijken, is er grote flexibiliteit mogelijk in de precieze manier waarop we naar onze onderzoeksgegevens kijken. De verleiding is dan groot om de beslissing hoe de analyse precies aan te pakken uit te stellen, en soms zijn daar ook goede wetenschappelijke redenen voor. De overtuigingskracht van de analyse jegens de scepticus wordt hierdoor natuurlijk sterk ondermijnd.

Het is ook de taak van reviewers om zich in te leven in de rol van de scepticus en daarmee diens belangen te behartigen. De ervaring leert echter dat reviewers hiertoe niet goed in staat zijn. Reviewers komen pas in een laat stadium in beeld, zodat een groot deel van de relevante selectie buiten hun gezichtsveld blijft. Belangrijker nog, reviewers lijken eerder te kijken of ze het inhoudelijk eens zijn met de conclusies, dan of het onderzoek goed is uitgevoerd. In 1998 deed Fiona Godlee, editor van het *British Medical Journal*, hier onderzoek naar. Zij stuurde een artikel met acht substantiële methodologische fouten naar meer dan tweehonderd reguliere reviewers van haar tijdschrift. Gemiddeld werden per reviewer slechts twee van de acht fouten opgemerkt. Wel 33 procent suggereerde om het artikel met minimale wijzigingen te plaatsen, terwijl maar 30 procent afwijzing aanraadde.

Waar de reviewers het laten afweten ligt de verantwoordelijkheid voor het voorkomen van vals positieve resultaten in het 'Wedden dat…' model geheel bij de onderzoeker zelf. Deze moet in dialoog blijven met de interne scepticus, en twijfel blijven cultiveren over de eigen resultaten. Dat is moeilijk, maar des te meer van belang in een tijd waarin deze wetenschappelijke grondhouding onder druk staat. Het streven naar meetbare excellentie vanuit universiteiten vraagt om veel en snel publiceren. Subsidieverleners selecteren het liefst onderzoekers die vol vertrouwen de prachtigste resultaten

beloven. Dergelijke prikkels eroderen de wetenschappelijke grondhouding, omdat ze onderzoekers juist stimuleren hun interne scepticus te onderdrukken. Excellentie is niet altijd hetzelfde als kwaliteit. Ik heb daarom grote sympathie voor initiatieven als *Science in Transition* en *Slow Science*, die pleiten voor meer rust in de *ratrace* van de wetenschap. We denken te veel over wetenschappelijke publicaties als persoonlijke prestaties van wetenschappers. Ook dat is een uitwas van het 'Wedden dat…'-model. Te vaak hoor ik alleen zeggen dat iemand een publicatie in *Nature of Cell* heeft, zonder dat ik te horen krijg wat daar dan precies in stond. Veel liever zou ik horen wat de doorbraak was, en niet waar die in gepubliceerd werd.

De statisticus is de ideale samenwerkingspartner voor iedere wetenschapper die geïnteresseerd is in de methodologische kwaliteit van zijn onderzoek. Vanuit onze expertise in wetenschappelijk redeneren, en vanuit een grotere distantie ten opzichte van het onderwerp, kunnen wij als geen ander de onderzoeker helpen de wetenschappelijke grondhouding vast te houden en de rol van de scepticus goed te spelen. Zo kunnen juist die experimenten en analyses gedaan worden die het verschil maken in het wetenschappelijk discours. Methodologie is de grammatica van dat discours en vormt daarmee het hart van de wetenschap. Het beste middel om vals positieve resultaten uit de wetenschappelijke literatuur te houden is dan ook grotere aandacht voor methodologie en beter begrip ervan. Het verminderen van de publicatiedruk is hiervoor een noodzakelijke voorwaarde. Deze maatregelen zijn mijns inziens essentieel om de wetenschap op lange termijn gezond te houden.

Nog een toevalstreffer ter afsluiting. De website *careercast.com* kwam kort geleden uit met een lijst van de beste beroepen van 2014. Statisticus stond in dit lijstje op 3. Dat was al erg mooi. Op 1 stond wiskundige. Dat ben ik ook! En op 2? Daar stond het beroep van hoogleraar. Zo verenig ik vandaag de drie beste beroepen van 2014 in één persoon. Daar kan ik alleen maar dankbaar voor zijn. Ik zal u op deze feestelijke dag niet vermoeien met mijn methodologische kanttekeningen bij het achterliggende onderzoek.

DANKWOORD

Ik wil het college van bestuur en het stichtingsbestuur van de Radboud Universiteit, de raad van bestuur van het Radboudumc en mijn afdelingshoofd prof. Bart Kiemeney graag hartelijk danken voor het in mij gestelde vertrouwen. Ik kijk ernaar uit om mij vol enthousiasme in te zetten voor aantoonbare kwaliteit van het wetenschappelijk onderzoek in het hele huis.

Mijn collega statistici ben ik dank verschuldigd voor alles wat ik heb geleerd. In de eerste plaats natuurlijk prof. Hans van Houwelingen. Hans, de manier waarop jij wiskundige elegantie weet te combineren met een fijn oog voor de echte problemen uit de praktijk zal voor mij altijd een voorbeeld zijn. Jouw enthousiasme voor data, zoals ik recent weer heb gezien in ons project met *Glaxo Smith Kline* vind ik enorm inspirerend. Ik ben uit Leiden uitgevlogen, maar ik blijf altijd je leerling.

De andere Leidse en voormalig Leidse statistici hebben mij mede gevormd. Ik heb erg veel van jullie geleerd. Ik dank jullie daarvoor. Met name wil ik noemen mijn kamergenoot Erik, vanwege de vele stimulerende gesprekken over statistiek die we altijd aan het eind van de middag hadden, en Saskia die mij als eerste wegwijs maakte in de medische statistiek, toen ik vers van wiskunde kwam en nog niet wist wat een t-toets was.

Mijn nieuwe groep statistici in Nijmegen wil ik danken voor het hartelijk welkom dat ik bij jullie gevoeld heb, en voor de leuke discussies over statistiek op de dinsdagochtenden en op andere tijdstippen. My *PhD students and postdocs* Rosa, Mathijs, Nimisha, Jesse, Joanna, Jakub, René: 'I am honored that you chose to work with me, and the great ideas and interesting results that you come up with brighten my days. Thank you Rosa for translating my lecture for the benefit of the non-Dutch speakers present here. A particular welcome and thanks to Aldo Solari and Livio Finos, my two Italian collaborators. I am extremely happy that you are both here, and I hope we can always continue to work together as closely as we have. My ideas have profited immensely from your input.'

Ik wil ook alle medisch, biologisch en bioinformatisch onderzoekers danken die samenwerking met mij hebben gezocht om hun methodologische problemen op te lossen. Deze contacten zijn vaak de basis of inspiratie geweest voor mijn eigen onderzoek, en ik heb er en passant ook heel veel over biologie van geleerd. Uit deze grote groep wil ik een paar mensen special noemen, van wie ik de oprechte interesse voor methodologie steeds bijzonder heb gewaardeerd, en dat zijn Peter-Bram 't Hoen, Bas Heijmans, Judith Boer en Jan Oosting.

Christa, lieve moeder: het is niet moeilijk om te zien waar ik mijn interesse voor wetenschap en methodologie vandaan heb. Ik kom uit een wetenschappelijk nest. Dat heb ik altijd met me meegedragen en dat heeft me gebracht waar ik nu sta.

Lieve Matty, rondom mijn beslissing om naar Nijmegen te gaan is er een hoop veranderd in ons leven. Je helpt me telkens opnieuw om de wereld als een spel te blijven zien, en ik ben je heel dankbaar voor je steun in deze moeilijke tijd.

Ik heb gezegd.