Prof.dr. J.J. Goeman

# De Zoekende Onderzoeker

Universiteit Leiden

Bij ons leer je de wereld kennen

# De zoekende onderzoeker

Oratie uitgesproken door

prof.dr. J.J. Goeman

bij de aanvaarding van het ambt van hoogleraar in de

Analyse van Hoog-dimensionele Medische Data

aan de Universiteit Leiden

op vrijdag 17 november 2017


Universiteit
Leiden

*Mijnheer de rector magnificus, hooggeleerde collegae, beste toehoorders, lieve vrienden en familie,*

In 1992, ik zat nog op de middelbare school, werd een grote ontdekking gepresenteerd die door wiskundigen overal ter wereld met enthousiasme ontvangen werd: Andrew Wiles, onderzoeker uit Cambridge, had de laatste stelling van Fermat bewezen, een belangrijk wiskundig probleem waaraan al sinds de 17e eeuw door de allerslimste mensen was gewerkt en dat niemand al die tijd had weten op te lossen. Andrew Wiles had zichzelf hiervoor zeven jaar op zijn zolderkamertje afgezonderd en had in het diepste geheim aan dit probleem gewerkt. Een eenzame genie. Dit was academische romantiek van de bovenste plank. Ik, zestien jaar oud, besloot om wiskunde te gaan studeren.

In de documentaire die de BBC in 1996 over hem maakte beschrijft Wiles het proces van onderzoek doen. Ik vertaal. "Misschien kan ik de ervaring van het doen van wiskundig onderzoek het beste beschrijven als het binnengaan in een donker gebouw. Je gaat de eerste kamer binnen en het is donker, helemaal donker. Je struikelt rond en botst tegen het meubilair. Heel langzaam leer je waar elk meubelstuk staat. Uiteindelijk, na een maand of zes, vind je het lichtknopje. Je doet het licht aan en alles is helder: je kunt precies zien waar je bent.".[1]

De zoekende onderzoeker. Voor mij beschrijft dit citaat heel goed de aard van exploratief, fundamenteel[2] wetenschappelijk onderzoek. Het is onderzoek gericht op theorievorming, en dus op begrip. Het gevonden licht, de nieuwe theorie, verlicht niet alleen de plekken waar de onderzoeker al geweest is, maar de hele kamer, inclusief deuren naar volgende kamers waar het nu nog donker is. Dit licht gaat aan via een klein knopje: een nieuwe theorie is tenslotte een plotseling inzicht, vaak onverwacht. Op weg naar dat knopje heeft de onderzoeker geen keus dan op intuïtie rond te struinen en via vallen en opstaan zijn of haar weg te vinden: niemand weet van tevoren in welke hoek van de kamer het knopje zich zal bevinden. Zonder mezelf met Andrew Wiles te willen vergelijken herken ik veel in zijn beeldende beschrijving, en niet alleen voor mijn eigen wiskundig onderzoek, maar ook als ik kijk naar het onderzoek van de biomedische onderzoekers met wie ik samenwerk.

Biologisch onderzoek is natuurlijk op een aantal punten anders dan wiskundig onderzoek. In plaats van een eenzame wiskundige zie ik bijvoorbeeld meerdere groepjes biologen voor me die door dezelfde donkere ruimte tasten, in verwoede competitie wie het volgende meubelstuk als eerste ontdekt. Maar dat is niet zo'n essentieel verschil. Belangrijker is dat de manier waarop biologen hun ruimte verkennen verschilt van die van wiskundigen. Waar de wiskundige op basis van puur redeneren vooruitgang kan boeken, moet de bioloog werken vanuit de resultaten van experimenten. En experimenten zijn feilbaarder dan redeneringen.[3]

Door toeval, botte pech, kan een experiment een vertekend beeld geven van de werkelijkheid. Bijvoorbeeld, als we een nieuw middel onderzoeken om verkoudheid te behandelen kunnen we een groep verkouden mensen verzamelen. Dan geven we iedereen willekeurig ofwel het nieuwe middel ofwel een placebo, een nepmedicijn. Als nu de mensen die het nieuwe middel gekregen hebben flink sneller opknappen dan de mensen met de placebo, zullen we geneigd zijn om te concluderen dat het medicijn werkt. Als dat niet zo is, en de mensen met de placebo knappen even snel of sneller op, concluderen we dat het middel niet werkt. Maar, niet iedereen heeft een even goede weerstand, en we zouden per ongeluk precies het nieuwe middel kunnen hebben gegeven aan de mensen met een goede weerstand. Dan zouden we ten onrechte concluderen dat het nieuwe middel werkt, ook als het in werkelijkheid geen effect heeft. Omgekeerd kunnen we de pech hebben dat we juist het middel gegeven hebben aan de mensen met een slechte weerstand. Dan zouden we concluderen dat het middel niet werkt, terwijl het in werkelijkheid wel effectief is. Deze twee mogelijke uitkomsten van een experiment noemen we vals positieve en

vals negatieve resultaten. Een vals positief resultaat is als we denken iets nieuws gevonden te hebben (een werkend medicijn) terwijl dat in werkelijkheid niet zo is. Een vals negatief resultaat betekent dat we ten onrechte denken dat er niets te vinden was (het medicijn leek niet te werken).

Van deze twee zijn vals positieve resultaten erger dan vals negatieve. Dat kunnen we begrijpen vanuit Wiles' parabel van het donkere huis. Een vals negatief resultaat betekent het missen van een meubelstuk, een rondgang door de kamer zonder nieuwe informatie. Dat is vervelend, maar overkomelijk: het meubelstuk wordt later wel door iemand anders ontdekt. Een vals positief resultaat is te vergelijken met het verkeerd identificeren van een meubelstuk. Bijvoorbeeld het aanzien van een eettafel voor een tafeltennistafel. Dat is een veel groter probleem omdat het verwarring oplevert: na een vals positief resultaat lopen we de volgende keer met de verkeerde veronderstellingen dezelfde kamer binnen. Bij een vals negatief resultaat gaat het onderzoek niet vooruit; bij een vals positief resultaat gaat het de mist in.



De maatschappelijke impact van vals positieve resultaten is ook groter dan die van vals negatieve. Het is verwarrend voor het publiek als dat dagelijkse glaasje rode wijn eerst goed is en later toch weer slecht. Nog erger is het als vals positieve resultaten nooit gecorrigeerd worden, en in de handboeken en uiteindelijk in het publieke domein terecht komen. Om ons te realiseren hoe groot de maatschappelijke schade kan zijn van een vals positief resultaat, hoeven we alleen maar te denken aan het artikel van Andrew Wakefield, die een relatie meende te vinden tussen vaccinatie en autisme. Zijn bevinding, gepubliceerd in 1988[4] en uiteindelijk teruggetrokken in 2010, heeft zich in het bewustzijn van het grote publiek genesteld, en is nog steeds een grote stoorzender voor vaccinatieprogramma's wereldwijd.

Het is een belangrijke taak van mijn vakgebied, de statistiek, om onderzoekers houvast te bieden en de maatschappij te beschermen tegen vals positieve resultaten. Hierbij moeten we ons realiseren dat het onmogelijk is om vals positieve resultaten helemaal uit te bannen. Dat komt omdat vals positieve en vals negatieve resultaten werken als communicerende vaten. Vermijden we de een, dan krijgen we meer van de ander en omgekeerd. We kunnen bijvoorbeeld heel streng zijn en pas een resultaat rapporteren bij overweldigend bewijs. Dan hebben we weliswaar een kleine kans op een vals positief resultaat, maar ook een grote kans op een vals negatief resultaat, omdat overweldigend bewijs zeldzaam is. De onderzoeker laveert tussen vals positieve en vals negatieve resultaten als tussen Scylla en Charybdis.

We kunnen vals positieve resultaten dus niet helemaal uitbannen, maar wel de kans dat ze optreden proberen in de hand te houden. Hiervoor biedt de statistiek twee belangrijke en onafscheidelijke stukken gereedschap: de p-waarde en het betrouwbaarheidsinterval. Ik zal me in dit verhaal beperken tot de p-waarde. Die zal ik uitleggen aan de hand van een klassiek experiment van Ronald Fisher, een van de grondleggers van de statistiek, en beschreven in zijn boek *The design of experiments* uit 1935.[5] Aanleiding voor het experiment was de bewering van Muriel Bristol, een collega van Fisher, dat ze feilloos kon proeven of in een kop thee met melk eerst de melk was ingeschonken en dan de thee, of eerst de thee en dan de melk. U begrijpt, dit wordt een zeer Brits experiment. Fisher geloofde er niets van: hij dacht dat ze niet beter kon doen dan te gokken.[6] Om dit te testen bedacht hij het volgende experiment, dat bekend staat onder de naam "de dame proeft thee": acht koppen thee werden gezet. Bij vier daarvan werd de thee ingeschonken voor de melk, bij de andere vier de melk voor de thee. De theekopjes werden vervolgens goed geroerd en in willekeurige volgorde aangeboden, en aan Bristol werd gevraagd aan te wijzen welke bereidingswijze bij welk kopje hoorde.[7] Zij proefde en onderscheidde de kopjes foutloos.

In statistisch jargon noemen we Fisher's sceptische blik de nulhypothese. Bristol's bewering noemen we het alternatief. Als Fisher gelijk had, en Bristol kon slechts gokken, dan was de uitkomst van het experiment een heel verbazingwekkende:

door puur gokken zou ze slechts 1 op de 70 keer alle acht de kopjes goed raden.[8] Deze 1 op 70, of 1,4% noemen we de p-waarde. De p-waarde kun je zien als een maat van verbazing over de uitkomst van het experiment voor iemand die in de nulhypothese gelooft: hoe kleiner de p-waarde hoe groter de verbazing. In dit experiment geloofde Fisher dat Bristol niet beter kon doen dan gokken, dus hij was erg verbaasd: 1,4%. Als Bristol een van de vier melk-eerst-kopjes had aangezien voor een thee-eerst-kopje was Fisher veel minder verbaasd geweest: een p-waarde van 17 op 70 of 24,3%.

Het doel van het experiment was om uit te maken of Bristol of Fisher het bij het rechte eind had. Wanneer geeft Fisher zich gewonnen? De algemene afspraak is dat een p-waarde van 5% hiervoor de grenswaarde is. Ligt de p-waarde onder deze 5%, dan zeggen we dat de resultaten van het experiment te verbazingwekkend zijn om nog in de nulhypothese te kunnen geloven. Fisher concludeerde dus dat het inderdaad mogelijk is voor een mens om de volgorde van inschenken van melk en thee te kunnen onderscheiden: een nieuwe wetenschappelijke bevinding.[9]

Als algemene wetenschappelijke methodologie heeft deze manier van werken een belangrijke eigenschap: onderzoekers die slechts nieuwe wetenschappelijke bevindingen rapporteren als ze een p-waarde onder de 5% hebben gevonden, zullen in de loop van hun carrière in tenminste 19 van de 20 experimenten die ze uitvoeren geen vals positief resultaat genereren.[10] Dat is een fijne garantie voor de rondtastende onderzoeker, die de terechte populariteit van de p-waarde als methodologisch instrument kan verklaren.

Als we kijken naar het experiment dat Fisher heeft opgezet, zien we dat het heel strak ontworpen is. Er is precies vastgelegd hoeveel kopjes thee zullen worden geserveerd, hoe die worden bereid en aangeboden, en welke conclusies zullen worden getrokken bij welke uitkomst van het experiment. Zo gauw het experiment van start is gegaan is er geen ruimte meer voor flexibiliteit. Bijvoorbeeld, de onderzoeker kan niet, als de proefpersoon zes van de acht kopjes goed blijkt te hebben alsnog besluiten om een paar kopjes thee extra te zetten. De reden hiervoor is dat de benodigde wiskundige berekeningen die flexibiliteit niet aankunnen. Grootschalige klinische studies, bijvoorbeeld om de effectiviteit van nieuwe medicijnen aan te tonen, zijn daarom heel strak geprotocolleerd: de onderzoeker weet precies, voordat de eerste proefpersoon zijn of haar eerste pil slikt, hoe jaren later de uiteindelijke data zullen worden samengevat, en welke uitkomst van het experiment tot welke conclusie zal leiden. Precies als bij Fisher's experiment uit 1935. De onderzoeker moet dat zo doen, omdat de statistische methoden ervan uitgaan dat de onderzoeker zo te werk gaat, en dus ook alleen in dat geval de juiste garanties leveren.

Hoe helpt dit de zoekende onderzoeker die in het donker rondtast en nieuwe, onbekende werelden betreedt? Degelijke onderzoekers willen graag verrast worden, en vinden het moeilijk of zelfs onmogelijk om van tevoren vast te leggen wat voor conclusies ze in welke gevallen precies zullen trekken. Als je nog niet weet wat voor meubelstuk je tegen zult komen in die donkere ruimte kun je moeilijk van tevoren zeggen of je er links- of rechtsom omheen zult proberen te lopen.

Goede garanties tegen vals positieve resultaten zijn in exploratief onderzoek dan ook lastig. Dit probleem speelt vooral in het gebied waarin ik meestal mijn statistische methoden toepas. Mijn leeropdracht luidt "Analyse van hoog-dimensionele medische data". Hoog-dimensioneel betekent hier dat er veel waarnemingen zijn per proefpersoon. Dat gebeurt bijvoorbeeld in de genetica, waarin onderzoekers geïnteresseerd kunnen zijn in de expressie, zeg maar de activiteit, van genen. Zij meten dan die activiteit voor iedere proefpersoon van 20 duizend genen tegelijk. U kunt zich misschien voorstellen dat er niet één direct voor de hand liggende manier is om die brei van getallen te analyseren, en dat onderzoekers liefst eerst naar de data kijken voordat ze een keuze maken. Een ander voorbeeld is in het hersenonderzoek, waarbij in de MRI-scanner de

activiteit van kleine blokjes in de hersenen wordt gemeten. Het aantal blokjes waarnaar tegelijk gekeken wordt kan gemakkelijk in de tienduizenden lopen.

Degelijk onderzoek wijkt op belangrijke punten af van Fisher's ideaaltypische onderzoek naar de dame die thee drinkt. In plaats van één onderzoeksvraag zijn er vele onderzoeksvragen tegelijk. Bovendien zijn deze onderzoeksvragen nog niet zo precies gedefinieerd voordat het onderzoek van start gaat. Een hersenonderzoeker wil bijvoorbeeld op zoek naar breingebieden die reageren op stemgeluid. Er zijn ontelbare manieren om breingebieden op te bouwen uit blokjes, en er kan onmogelijk van de onderzoeker gevraagd worden dat hij of zij van tevoren precies vastlegt hoe de data eruit moeten zien om bepaalde conclusies te trekken. Er zijn eenvoudigweg teveel onbekenden: er kunnen veel of weinig gebieden te vinden zijn, groot of klein, grillig van vorm of regelmatig. Voor zo'n onderzoeker voelt precies vastleggen van een protocol als Odysseus die zich vastbindt aan de mast om de lokroep van de vals positieve sirene te kunnen weerstaan.

Toch is ook in dit type onderzoek het gevaar van vals positieve resultaten levensgroot. Dat werd op een ludieke manier geïllustreerd in 2009 door Craig Bennett en zijn collega's.[11] Ik heb daar in mijn Nijmeegse oratie in 2014 uitgebreid over verteld.[12] Bennett's team stopte een dode zalm in een MRI-scanner, en toonde die emotioneel geladen beelden. Ze lieten zien dat de analysemethoden die op dat moment in zijn vakgebied gangbaar waren prachtig breingebieden deden oplichten die leken te duiden op respons van de zalm op deze beelden. Ten duidelijkste een vals positief resultaat, en wat mij betreft prachtige wetenschap. Bennett's publicatie deed twijfels rijzen over de betrouwbaarheid van eerder gevonden resultaten in het breinonderzoek.

De vingerwijzing van de dode zalm heeft onderzoekers in het breinonderzoek heel voorzichtig gemaakt. Men weet nu dat het gevaarlijk is om dezelfde data twee keer te gebruiken: een keer om te selecteren welke breingebieden er interessant of veelbelovend uitzien, en dan nogmaals om te kijken hoeveel bewijs er is dat er inderdaad signaal in die gebieden te vinden is. Hier schuilt namelijk het gevaar van een cirkelredenering in: natuurlijk lijkt er signaal te zitten in die gebieden die we hebben uitgezocht omdat er signaal in lijkt te zitten. Dit soort cirkelredenering noemen ze in het breinonderzoek "double dipping": je afgekloven frietje opnieuw in de mayonaise steken. Dat mag niet. Onderzoekshygiëne. Deze boodschap is diep in het vakgebied doorgedrongen.

Was dit een overwinning voor de statistiek en de methodologie? Het lijkt misschien zo, maar ik vind het een Pyrrhusoverwinning. Het taboe op double dipping beperkt namelijk de creatieve vrijheid van de onderzoeker. Waarom doet het dat? Alleen maar omdat de huidige statistische methoden met die vrijheid niet om kunnen gaan. De onderzoeker heeft zijn gedrag aangepast aan de beperkingen van de statistische methoden. Dat is niet zoals het zou moeten. Ik vind dat methodologie onderzoekers moet helpen, niet beperken.

In mijn eigen onderzoek ontwikkel ik daarom methoden die de vrijheid van onderzoekers vergroten, zonder de bescherming tegen vals positieve resultaten te verliezen. In het breinonderzoek betekent dat, dat onderzoekers naar hun data mogen kijken zoveel ze willen voordat ze een keuze maken welk breingebied, of welke breingebieden, er veelbelovend uitzien. Over deze breingebieden kan ik dan een statistische uitspraak doen hoe actief die breingebieden zijn en wat het risico is op vals positieve resultaten. In de statistische berekeningen hiervoor houd ik er rekening mee dat de onderzoeker naar de data gekeken heeft voordat hij of zij de breingebieden heeft uitgezocht. Ik pas de statistiek aan de onderzoeker aan, in plaats van de onderzoeker aan de statistiek.

De wiskunde die dit mogelijk maakt heb ik al uitgedacht in 2009, tijdens een eigen bescheiden lichtknopjesmoment, toen ik inzag dat de bekende gesloten toetsprocedure voor dit doel gebruikt kon worden. Pas in de loop van de daaropvolgende

jaren heb ik de details uitgewerkt en heb ik de volle potentie van deze benadering ingezien. Samen met mijn collega Aldo Solari uit Milaan en mijn promovendi en voormalige promovendi Rosa Meijer, Jesse Hemerik, Ningning Xu en Mitra Ebrahimpoor ben ik bezig nieuwe verrassende toepassingen van deze flexibele methoden uit te werken. Bijvoorbeeld, dezelfde methode die voor breinonderzoek goed werkt, kan ook gebruikt worden om een oud probleem uit de genetica op te lossen hoe op een effectieve manier naar groepen genen te kijken, een probleem waaraan ik al in mijn proefschrift gewerkt heb. Ook hier geeft de methodologie meer flexibiliteit aan de onderzoeker dan ik zelf ooit voor mogelijk had gehouden.

De onderzoeker tast door een donkere ruimte. Ik probeer gereedschap aan te reiken dat de onderzoeker beschermt, zonder zijn of haar bewegingsvrijheid te beperken.

Overigens ligt een van de belangrijkste stukken statistisch gereedschap, de p-waarde, de laatste tijd onder vuur.[13] De p-waarde is een belangrijk ingrediënt van de benadering die ik zojuist heb beschreven, dus het is de moeite waard om even bij de kritiek stil te staan. P-waarden zijn in veel vakgebieden de norm. Ze worden inderdaad te pas en te onpas uitgerekend, vaak verkeerd gebruikt, en slecht begrepen. Dat laatste is helaas het lot van zo'n beetje alles wat we uitrekenen in de statistiek. Andere statistische maten, zoals de puntschatter, de posterior, en statistische voorspellingen, worden naar mijn ervaring net zo slecht begrepen en evengoed misbruikt. In de psychologie is veel onderzoek gedaan naar de manier waarop de menselijke geest met kansen omgaat, en de conclusie is eenduidig: onze intuïtie gaat hier steevast keihard de mist in.[14] De mens is niet voor statistiek in de wieg gelegd. Dat maakt het vakgebied juist zo boeiend. P-waarden zijn ingewikkeld omdat statistiek ingewikkeld is. We moeten de dingen zo eenvoudig maken als mogelijk, maar zeker niet eenvoudiger dan dat.[15]

De kritiek op het gebruik van de p-waarde is echter wel reëel, en gaat enerzijds over de beperkte zeggingskracht van de p-waarde, en anderzijds over de misbruikgevoeligheid ervan. Laten we deze twee kritiekpunten kort in detail bekijken.

Eerst de zeggingskracht. Ik heb de p-waarde uitgelegd als een maat van verbazing over de data voor iemand die gelooft in de nulhypothese. Het volgt daar onmiddellijk uit dat de p-waarde alleen maar betekenisvol kan zijn als er iemand is die in de nulhypothese gelooft. Er worden echter in de praktijk heel vaak p-waarden uitgerekend voor nulhypotheses waar helemaal niemand in geïnteresseerd is. Laat ik een voorbeeld geven. Stel we onderzoeken een nieuw cholesterolverlagend medicijn. Dan kunnen we de nulhypothese kiezen dat het medicijn het cholesterolniveau in het geheel niet beïnvloedt. Als we deze nulhypothese verwerpen hebben we bewezen dat het medicijn tenminste 'iets' doet. Dat is echter helemaal waar we in geïnteresseerd zijn: we willen dat het medicijn genoeg effect heeft om de moeite van het voorschrijven waard te zijn. Het effect moet, zoals we dat zeggen, klinisch relevant zijn. De gebruikte nulhypothese is dus een soort stroman: gemakkelijk te verslaan, maar van weinig waarde.

Een p-waarde is ten hoogste zo interessant als de nulhypothese waar zij over gaat. Een nulhypothese die niemand interesseert levert dus ook een p-waarde op die niemand interesseert. Als de p-waarde weinig zeggingskracht heeft moeten we dus een nulhypothese kiezen die wél interessant is. We kunnen bijvoorbeeld de nulhypothese formuleren dat het medicijn niet genoeg effect heeft om klinisch relevant te zijn. De p-waarde bij die nulhypothese geeft aan hoeveel bewijs er is dat het medicijn een klinisch relevant effect heeft.[16] Deze p-waarde heeft wel zeggingskracht. Het incorporeren van klinische relevantie in de definitie van de nulhypothese vergt een andere, minder mechanische blik op statistiek bij gebruikers. Statistiek wordt soms teveel als een kookboek, een stel voorschriften, benaderd, in plaats van als een taal waarin onderzoekers over onzekerheid praten. Maar dit was de boodschap van mijn Nijmeegse oratie uit 2014. Ik zal dit onderwerp op deze plek laten rusten.

Een ander kritiekpunt op p-waarden is dat ze vatbaar zijn voor gerommel. Dit is relevant omdat de afkapwaarde van 5% zo belangrijk wordt gevonden: onder de 5% mogen we een nieuwe wetenschappelijke bevinding claimen; daarboven niet. Dat betekent weer dat een p-waarde van 0.048 toegang kan geven tot een publicatie in een prestigieus wetenschappelijk tijdschrift als, zeg, de Lancet, maar een p-waarde van 0.051 niet. Dat voelt onrechtvaardig: zo'n klein verschil met zulke grote gevolgen. In de praktijk zie je daarom dat onderzoekers een beetje gaan foezelen. Misschien was de gekozen analysemethode toch niet optimaal? Laten we het net iets anders proberen. Uiteindelijk komt een onderzoeker dan uit bij een analysemethode met een p-waarde die net aan de goede kant van de 5% ligt. Inderdaad zien we in publicaties een onnatuurlijke piek van veel te veel p-waarden net onder die bewuste 5%.[17] P-waarden zijn het slachtoffer van de wet van Goodhart, die zegt dat wanneer een maat gebruikt wordt als beloningscriterium, deze ophoudt een goede maat te zijn.

Want de keuze van statistische analysemethoden kan grote invloed hebben op de conclusies van het onderzoek. Dat werd heel mooi geïllustreerd door Brian Nosek.[18] Nosek stuurde dezelfde dataset met dezelfde onderzoeksvraag naar 29 verschillende teams statistici. De dataset ging over internationaal voetbal in het seizoen 2012-2013. De onderzoeksvraag was of spelers met een donkere huidskleur vaker een rode kaart toegewezen kregen dan spelers met een lichte huidskleur, een maatschappelijk relevante vraag. De 29 teams statistici gingen ermee aan de slag en kwamen met de meest uiteenlopende antwoorden. Twintig teams vonden bewijs dat spelers met een donkere huidskleur inderdaad vaker een rode kaart kregen, en 9 teams vonden daar onvoldoende bewijs voor. De p-waarden besloegen zo'n beetje het hele spectrum tussen 0 en 1. Dit is een extreem voorbeeld omdat de dataset heel complex was en de expertise van de analyseteams nogal uiteen liep, maar het illustreert de invloed die de keuze van analysemethode kan hebben. Dat betekent weer dat een onderzoeker die twee verschillende methoden uitprobeert op dezelfde data twee nogal uiteenlopende antwoorden kan krijgen. Het kiezen van de analysemethode aan de hand van de data is ook een vorm van double dipping die de garanties van statistische methoden in de war kan schoppen.

Toch denk ik dat er, net als bij de breinwetenschappers, vaak goede wetenschappelijke redenen zijn voor de zoekende onderzoeker om keuzes uit te stellen en zich niet vast te willen leggen. Ik juich daarom het recente initiatief toe om in plaats van de gebruikelijke afkapgrens van 5 procent een nieuwe, strengere afkapgrens van 5 promille in te stellen voor de p-waarde.[19] In ruil voor deze strengere afkapgrens zou de onderzoeker wat mij betreft dan wat extra ruimte mogen nemen om een aantal verschillende analysemethoden uit te proberen.[20] Onderzoekers die wel bereid zijn om zich vast te leggen op één analysemethode zouden dan gewoon de gebruikelijke 5% aan mogen houden. Ook hier pas ik liever de statistiek aan de onderzoeker aan dan de onderzoeker aan de statistiek.

Dat laatste principe zou ik graag in een nog groter licht zien. Methodologie zou erop gericht moeten zijn om de wetenschap als geheel zo snel en betrouwbaar mogelijk vooruit te laten gaan. Dat betekent dat we niet alleen ieder individueel onderzoek moeten verbeteren, maar dat we ook oog zouden moeten hebben op de dynamiek van het gehele onderzoeksveld. Mijn droom is een veelomvattender blik op methodologie, die het sociale proces van wetenschappelijk onderzoek meeneemt en zich afvraagt hoe de wetenschap als geheel ingericht zou kunnen worden om kennis te bevorderen. Laten we dit macro-methodologie noemen, om het te onderscheiden van de klassieke micro-methodologie die ervoor zorgt dat ieder individueel artikel wetenschappelijk in orde is.

In het laatste deel van deze oratie wil ik exploreren hoe we naar de Nederlandse wetenschappelijke wereld kunnen kijken vanuit zo'n macro-methodologisch perspectief. Hoe kiest de wetenschappelijke kudde zijn richting?

Als ik om me heen kijk zie ik dat veel van mijn collega's een groot deel van hun tijd bezig zijn met het schrijven van subsidieaanvragen. Zo'n subsidieaanvraag is een bijzonder literair genre. Hierin kondigt een wetenschapper aan wat voor wereldschokkend onderzoek hij of zij van plan is te gaan doen, en klopt zichzelf op de borst over het fantastische onderzoek dat hij of zij in het verleden gedaan heeft. Subsidiegevers, zoals NWO, kiezen vervolgens de meest veelbelovende aanvragen uit en financieren die. Subsidiegevers zijn dus heel sturend wat betreft de richting waar de wetenschap opgaat, en daarmee grote spelers op macro-methodologisch gebied. Wat is de invloed van subsidiegevers op de methodologie?

Kijken we naar de beoordeling van subsidieaanvragen, dan valt allereerst op dat mooie beloftes bij de beoordeling zwaarder wegen dan degelijke methodologie, en bovendien dat er een groot appèl wordt gedaan op de statusgevoeligheid van wetenschappers. Dit heeft al een groot gevolg gehad voor de manier waarop we naar wetenschap en wetenschappers kijken. We zijn helemaal gewend geraakt aan het idee dat wetenschappers vooral met elkaar in competitie zijn om status. Wetenschap als topsport.[21] Dat beeld verdringt de alternatieve, macro-methodologische visie dat wetenschappers samen bouwen aan een groot gemeenschappelijk project dat wetenschap heet.

De competitiegerichte blik op wetenschap heeft allerlei ongewenste gevolgen. Als de eerste die een bepaalde ontdekking publiceert alle eer krijgt, kun je beter snel en slonzig onderzoek doen dan langzaam en grondig.[22] Bij latere subsidieaanvragen is reproduceerbaarheid van eerder onderzoek tenslotte geen criterium. Subsidiegevers belonen daarmee vals positieve resultaten. Dat effect is indirect, maar heel reëel.

De controverse tussen Keith Baggerly en Anil Potti illustreert voor mij de statusgerichte visie op wetenschap. Potti, een onderzoeker bij Duke University, had in 2006 gerapporteerd over een onderzoek waarmee op basis van genetische informatie kon worden voorspeld welke patiënten goed zouden reageren op chemotherapie. Op basis hiervan had hij een grote subsidie gekregen en was een grote klinische studie begonnen. Baggerly, een bioinformaticus, was meteen begonnen om te proberen de conclusies van Potti te reproduceren op basis van diens data. Hoe meer hij groef, hoe meer methodologische slordigheden en problemen hij tegenkwam. Als snel kwam hij erachter dat het resultaat van Potti gebaseerd was op een vergissing, en bewijsbaar geen enkele basis had in de verzamelde gegevens, en dat er misschien zelfs sprake was van fraude. Dat betekende dat in Potti's klinische vervolgstudie honderden patiënten het gevaar liepen de verkeerde behandeling te krijgen. Baggerly kwam in actie, maar kreeg zowel bij Duke University als bij de subsidiegever nul op het rekest. De klinische studie ging gewoon door, totdat in 2010 bleek dat Potti in zijn curiculum vitae had geclaimd een studiebeurs te hebben gekregen die hij in werkelijkheid nooit gehad had. Dat was aanleiding voor ontslag van Potti en het stopzetten van de klinische studie.[23]

Ik vind het wrang om te concluderen dat wij het kennelijk erger vinden als een wetenschapper met zijn CV fraudeert dan als patiënten gevaar lopen als gevolg van wetenschappelijke fouten. Dit is overigens geen Amerikaanse toestand. In Nederland hadden we de affaire rond Diederik Stapel, de Tilburgse hoogleraar die jarenlang iedereen om zich heen had opgelicht met gefingeerde onderzoeksgegevens. Hierbij viel het me op dat de maatschappelijke verontwaardiging zich vooral richtte op de status die Stapel onterecht had gekregen met zijn wetenschappelijke werk, en niet op de schade die het vakgebied had geleden omdat het door zijn vals positieve resultaten de verkeerde kant op was gestuurd. En als die laatste schade er niet was, was de wetenschappelijke en maatschappelijke impact van Stapels resultaten kennelijk klein en kan de vraag dus worden opgeworpen waar de status van Stapel dan eigenlijk op gebaseerd was.[24]

De focus op status maakt dat wetenschappers voortdurend vergeleken moeten worden. Het is de taak van subsidiegevers om ranglijstjes te maken van wetenschappers en hun voorge-

stelde onderzoek op basis van kwaliteit. De inschatting van kwaliteit die de commissie moet maken heeft echter een grote onzekerheidsmarge, en dat maakt de ranglijstjes ook onzeker. Het is heel lastig om de waarde van onderzoek te bepalen dat nog niet is uitgevoerd. We weten vaak helemaal niet zo goed of het voorstel van de ene onderzoeker betere wetenschap gaat opleveren dan het voorstel van de andere. De ranglijstjes zijn onzeker.

Onzekerheid in ranglijstjes is een ander onderwerp waar ik me in mijn eigen onderzoek mee bezig houd. Samen met mijn collega's Diaa al Mohamad en Erik van Zwet kwantificeer ik de onzekerheid in ranglijsten die het gevolg is van onzekerheid in de achterliggende kwaliteitsindicatoren. Dit is wiskundig gezien een heel leuk en moeilijk probleem omdat rangen op een nogal verknoopte manier van de indicatoren afhangen. Onze motivatie om hieraan te werken komt voort uit ranglijsten van kwaliteit van ziekenhuizen, die door verzekeraars en beleidsmakers graag gemaakt worden. We brengen de onzekerheid in kaart en zeggen dan bijvoorbeeld dat best scorende ziekenhuis waarschijnlijk inderdaad bij de top 5 hoort, en dat van het slechtst scorende ziekenhuis alleen maar betrouwbaar gezegd kan worden dat het bij de slechtste 50% van de ziekenhuizen zit. We vinden meestal dat ziekenhuizen in Nederland bijna allemaal ongeveer even goed zijn, en dat de verschillen tussen ziekenhuizen heel klein zijn relatief aan de onzekerheid in de kwaliteitsmeting.[25]

Wat voor verschillen tussen ziekenhuizen geldt, geldt misschien ook voor verschillen tussen wetenschappers, temeer omdat de meting van de toekomstige maatschappelijke of wetenschappelijke waarde van een onderzoeksvoorstel nog veel moeilijker en dus onzekerder is dan de meting van de kwaliteit van ziekenhuizen nu.[26] Wat nu als de van de gehonoreerde top 10% van onderzoeksvoorstellen alleen betrouwbaar kunnen zeggen dat ze niet bij de slechtste 10% zitten? Meten is niet altijd weten: daar zit statistiek tussen. Geen ranglijst is soms beter dan een slechte ranglijst. Als de waarde van toekomstig onderzoek niet te voorspellen is, kun je jezelf de moeite van het beoordelen van onderzoeksvoorstellen besparen. De kosten van de beoordeling zijn enorm, en de tijd en energie die wetenschappers besteden aan het opstellen en beoordelen van onderzoeksvoorstellen, besteden ze niet aan onderzoek.

Krist Vaesen en Joel Katsav hebben in een recent artikel becijferd hoeveel geld iedere Nederlandse wetenschapper zou ontvangen als NWO zou worden afgeschaft en alle gelden gelijkelijk over alle wetenschappers zouden worden verdeeld.[27] Zij berekenen dat iedere wetenschapper van het niveau universitair docent of hoger, in alle vakgebieden, dan elke vijf jaar een subsidie van 390 duizend euro zou krijgen. Als we differentieren tussen vakgebieden met een hoog of laag subsidieniveau zou iedere onderzoeker in de geneeskunde zelfs elke vijf jaar 5,5 ton kunnen besteden, een kleine VIDI-subsidie. Deze cijfers zijn waarschijnlijk een factor 2 te hoog,[28] maar illustreren zelfs dan nog de kosten van het systeem. De macro-methodologische invloed van het subsidiesysteem is gigantisch. Op zeker twee punten is die invloed negatief: de tijdsinvestering van het schrijven en beoordelen van subsidies, en de beloning die het subsidiesysteem geeft aan vals positieve resultaten.[29] Per saldo ontstaat alleen een positief resultaat als subsidiegevers inderdaad heel goed in staat zijn om goed toekomstig onderzoek van slecht te onderscheiden. Ik ben daar niet van overtuigd, en ik zou NWO willen uitdagen om dat eens netjes te onderzoeken. Ik bied mijn hulp als methodoloog hierbij graag aan.

Rondom de wetenschap zijn we teveel gefocust op wetenschappelijke status. We denken in termen van goede en slechte wetenschappers in plaats van in termen van goede of slechte wetenschap. De methodologie is hierbij de grote verliezer. Veel liever kijk ik naar wetenschap als een groepsproces waarbij wetenschappers samenwerken aan het vergroten van kennis. Veel liever benadruk ik het plezier en de schoonheid van onderzoek doen. Het mooiste moment van de documentaire over Andrew Wiles is het moment waarop hij vertelt over het inzicht dat hij kreeg waardoor alles uiteindelijk op zijn plek viel,

het lichtknopje dat hij had gevonden. Bij het vertellen daarover kreeg hij opnieuw tranen in de ogen en een brok in de keel. Dat is waarom onderzoek doen zo prachtig is. Onderzoekers moeten zoeken en spelen. Het management van wetenschap moet onderzoekers de vrijheid geven om zich op de inhoud te richten zonder voortdurend te vragen verantwoording af te leggen. Methodologie moet onderzoekers beschermen tegen onjuiste conclusies zonder de vrijheid om te exploreren onnodig in te perken.

Wetenschap is een groepsproces, en dat geldt zeker ook voor het onderzoek waarin ik mijn rol heb gehad. Ik wil een aantal mensen hartelijk bedanken voor hun inspiratie, hun steun en voor het voorrecht om met hun samen te mogen werken.

Het College van Bestuur van de Universiteit Leiden en de Raad van Bestuur van het LUMC dank ik voor het in mij gestelde vertrouwen. Ik dank Theo Stijnen dat hij mij de gelegenheid heeft gegeven weer terug te keren naar het mij vertrouwde LUMC. Ik dank Ewout Steyerberg voor de voortvarendheid en energie waarmee hij de leiding van de afdeling heeft overgenomen.

Hans van Houwelingen, mijn promotor. Ik dank je voor de brede visie op het vakgebied die ik van je geleerd heb, waarin je wiskundige precisie combineert met een groot inzicht in de noden van de praktijk. Je weet als geen ander de achterliggende concepten helder te krijgen. Ik heb heel veel van je geleerd.

Hein, Saskia, Erik, Joanna en al mijn andere collega-statistici, zowel in Leiden als in Nijmegen, dank ik voor de inspirerende omgeving waarin ik mijn onderzoek mag doen. Ik waardeer de open-kritische sfeer en het enthousiasme voor het vak dat jullie uitstralen. Above all, I would like to thank all the PhD students and postdocs who have worked with me: Diaa, Jakub, Jesse, Mathijs, Monika, Nimisha, Ningning, and Rosa. It is those moments we stand in front of the whiteboard working things out together that I feel most in my element as a researcher. I am grateful to my long-time Italian collaborators, Aldo and Livio, for sharing your joy in science and your great ideas with me. Ik ben ook alle biomedisch onderzoekers dankbaar die mij betrokken hebben in de wondere wereld van het medisch onderzoek, en leuke problemen gaven om over me te denken.

Ik ben heel blij dat mijn vrienden hier in groten getale aanwezig zijn. Jullie maken voor mij Leiden de ideale plaats om te leven en onderzoek te doen. Mijn lieve moeder Christa en mijn vader Henk, die dit moment helaas niet meer heeft mogen meemaken. Ik dank jullie voor het enthousiasme voor redeneren, voor wetenschap, wiskunde en methodologie. Dat heb ik van jullie meegekregen. Zonder die achtergrond had ik hier nu niet gestaan.

Iedereen hier aanwezig: het is mij een voorrecht om hier te zijn en jullie te mogen toespreken.

Ik heb gezegd.

**Noten**

1 Wiles: "Perhaps I could best describe my experience of doing mathematics in terms of entering a dark mansion. One goes into the first room and it's dark, really dark. One stumbles around bumping into the furniture. Gradually you learn where each piece of furniture is. Finally after six months or so you find the light switch. You turn it on and it's all illuminated. You can see exactly where you were." Quoted in: BBC Horizon, Fermat's Last Theorem. Director: Simon Singh, Writer: John Lynch.

2 Het onderscheid tussen fundamenteel en toegepast onderzoek zit in het type onderzoeksvraag. Fundamenteel onderzoek vraagt: "Hoe zit het in elkaar?" terwijl toegepast onderzoek vraagt: "Wat moeten we doen?".

3 Het is een paradox dat een diep technisch begrip van statistiek een wiskundige achtergrond vereist, terwijl juist wiskundigen zelden of nooit met empirische data in aanraking komen en daardoor soms moeite hebben de methodologische component van statistiek goed te begrijpen. Ikzelf ben statistiek pas gaan waarderen doordat ik tijdens mijn geschiedenisstudie met methodologie in aanraking kwam. Ik zou er dan ook voor willen pleiten om voorafgaand aan de statistiekcolleges voor wiskundestudenten een paar colleges wetenschapsfilosofie te geven.

4 Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, Berelowitz M, Dhillon AP, Thomson MA, Harvey P, Valentine A, Davies SE and Walker-Smith JA (1998). "Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children". *The Lancet*. 351 (9103): 637-41 (Retracted).

5 RA Fisher (1935). The design of experiments. Oliver and Boyd.

6 Dit experiment past prachtig in het weddenschapsmodel voor wetenschap waarover ik in mijn oratie in Nijmegen in 2014 heb verteld. Deze oratie is in bewerkte vorm gepubliceerd als JJ Goeman (2016), "Randomness and the games of Science" in: K Landsman, E van Wolde (eds) *The challenge of chance: a multidisciplinary approach from science and the humanities.* Springer.

7 Aan Bristol werd tevoren verteld dat er vier kopjes van elk type waren.

8 Merk op dat om deze conclusie te trekken de aanname noodzakelijk is dat de onderzoeker iedere mogelijke volgorde kiest met gelijke kans 1/70. In beschrijvingen van dit experiment wordt soms gesuggereerd dat Bristol onder de nulhypothese iedere volgorde met gelijke kans 1/70 kiest. Daar mogen we echter niet vanuit gaan: de mens is geen goede generator van willekeurige sequenties.

9 Merk op dat hiermee nog niet is bewezen dat Bristol gelijk had met haar bewering dat ze het verschil altijd foutloos kan proeven. De waarheid kan ook nog ergens in het midden liggen. Waar ongeveer, dat is waar het betrouwbaarheidsinterval over gaat.

10 Aannemende dat per experiment één p-waarde wordt uitgerekend. Hier kom ik later op terug.

11 CM Bennett, AA Baird, MB Miller and GL Wolford (2011). Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: an argument for proper multiple comparisons correction. *Journal of Serendipitous and Unexpected Results* 1 (1-5).

12 JJ Goeman (2014). Toevalstreffers. Radboud Repository of the Radboud University Nijmegen. http://repository.ubn.ru.nl/bitstream/handle/2066/130405/130405.pdf.

13 RL Wasserstein and NA Lazar (2016). The ASA's statement on p-values: Context, process, and purpose. *American Statistician*, 70(2), 129-133.

14 Kahnemann (2011). *Thinking fast and slow*. Macmillan. Samengevat als: "Our brain is not a good statistician" (Hans van Houwelingen op de International Meeting on Statistical Methods in Biopharmacy, Parijs, Sept. 2017).

15 Deze uitspraak wordt toegeschreven aan Albert Einstein.

16 Het toetsen van dergelijke nulhypothesen brengt weer interessante multiple testing-problemen met zich mee, waaraan ik werk samen met Jules Ellis en Jakub Pecanka.

17 JC de Winter and D Dodou (2015). A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *PeerJ*, 3, e733. Ik denk dat de toename van deze piek in de laatste decennia te maken heeft met (1.) publicatiedruk en (2.) gebruiksgemak van statistische software.

18 R Silberzahn et al. (2017). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. Unpublished preprint.

19 DJ Benjamin et al. (2017). Redefine statistical significance. Nature Human Behaviour (in press).

20 Deze extra factor 10 is natuurlijk slechts een grove vuistregel voor de benodigde compensatie voor de vrijheid die een onderzoeker heeft om via de keuze van analysemethode de p-waarde te beïnvloeden. Liever zou ik daar ook nog preciezere methoden voor ontwikkelen.

21 Het culturele archetype dat hieronder ligt is dat van de wetenschapper als eenzame genie, zoals gepersonifieerd door Andrew Wiles of Albert Einstein. De paradox hiervan is dat waarschijnlijk Wiles noch Einstein tijdens hun meest creatieve periodes een grote kans zouden hebben gemaakt op een subsidie.

22 Het auteurschap van wetenschappelijke artikelen heeft ook geleden onder de wet van Goodhart. Omdat auteurschap van artikelen zo'n sterk beloningscriterium is geworden, is auteurschap geen goede maat meer voor de daadwerkelijke bijdrage aan het wetenschappelijke werk.

23 Ook het onderzoek van Wakefield over vaccinatie en autisme, dat ik eerder al noemde, viel niet door de mand vanwege de methodologische fouten die waren geconstateerd, maar als gevolg van het aan het licht komen van financiële belangen van Wakefield bij het onderzoek.

24 In mijn voorbeelden van vals positieve resultaten gebruik ik vooral frauderende wetenschappers, omdat er in die gevallen geen discussie is over de vraag of het resultaat wel vals positief is. De meerderheid van vals positieve resultaten komt natuurlijk van goed bedoelende wetenschappers die eenvoudigweg pech hebben gehad.

25 Een veelgehoorde mening onder statistici is dat ranglijstjes ontmoedigd moeten worden omdat ze een vertekend beeld geven. Er zijn hier goede argumenten voor. Ik denk echter dat het maken van ranglijstjes een onuitroeibare menselijke behoefte is, en dat we de wens om dat te doen als een gegeven moeten beschouwen. Ook hier pas ik liever de statistiek aan dit gegeven aan.

26 Wat beoordeeld moet worden is tenslotte niet de kwaliteit van een onderzoeksvoorstel, maar de toekomstige maatschappelijke of wetenschappelijke waarde van de uit te voeren onderzoek. "Voorspellen is moeilijk, vooral als het om de toekomst gaat" (toegeschreven aan Niels Bohr).

27 K Vaesen and J Katzav (2017). How much would each researcher receive if competitive government research funding were distributed equally among researchers?. *PloS One* 12(9); e0183967.

28 Reactie van NWO op Vaesen en Katsav: https://www.nwo.nl/actueel/nieuws/2017/reactie-stan-gielen-basisbeurs-voor-wetenschappers-betekent-achteruitgang-voor-de-wetenschap.html.

29 Als derde pervers effect van het subsidiesysteem zou ik willen noemen dat subsidies, via het besluit wetenschappers te beoordelen via artikelen en impact factors, het verouderde instituut van het wetenschappelijke tijdschrift in leven houden, en daarmee de woekerwinsten van wetenschappelijke uitgevers ten koste van belastinggeld. Zonder de bestendiging vanuit het subsidiesysteem zouden wetenschappelijke tijdschriften mijns inziens allang zijn opgevolgd door modernere en efficiëntere manieren van verspreiding van wetenschappelijke kennis.

## Prof.dr. Jelle Goeman (Leiderdorp, 1976)

| | |
|---|---|
| 2001 | Doctoraal Geschiedenis, Universiteit Leiden |
| 2001 | Doctoraal Wiskunde, Universiteit Leiden |
| 2006 | Promotie, Universiteit Leiden (Statistical methods for microarray data) |
| 2006 | Postdoc, Imperial College London |
| 2006-2007 | Postdoc, afdeling Medische Statistiek en Bioinformatica, LUMC |
| 2007-2008 | Assistant Professor, afdeling Medische Statistiek en Bioinformatica, LUMC |
| 2008-2013 | Associate Professor, afdeling Medische Statistiek en Bioinformatica, LUMC |
| 2013-2016 | Hoogleraar biostatistiek, Radboudumc |
| 2016-heden | Hoogleraar analyse van hoog-dimensionele medische data, LUMC |

Vals positieve wetenschappelijke bevindingen zijn bevindingen die in een enkel onderzoek naar boven komen maar niet in vervolgonderzoek kunnen worden herhaald. Er is in de recente jaren veel publieke aandacht voor dergelijke bevindingen en hun invloed op wetenschap en maatschappij. Statistiek biedt methoden om de frequentie van dergelijke vals positieve resultaten binnen de perken te houden. Deze methoden gaan echter over het algemeen ten koste van de vrijheid die onderzoekers om te grasduinen in hun data, en werken dus niet optimaal voor exploratief fundamenteel onderzoek. Daarom moeten nieuwe statistische methoden worden ontwikkeld die de vrijheid van wetenschappers minder inperken.

Het verminderen van vals positieve resultaten is echter niet alleen de verantwoordelijkheid van individuele onderzoekers. Ook subsidiegevers moeten hun rol in de huidige reproduceerbaarheidsproblemen van het wetenschappelijk onderzoek onder ogen zien. De huidige praktijk van het beoordelen van onderzoeksaanvragen beloont snel onderzoek boven grondig en bevordert daarmee vals positieve resultaten en niet-reproduceerbaar onderzoek.

Universiteit Leiden