JUDITH TER SCHURE OVER EEN NIEUW SOORT STATISTIEK

'DE *P*-WAARDE GEBRUIKEN IS DE MOEILIJKSTE MANIER OM DATA IN ÉÉN GETAL SAMEN TE VATTEN'

Judith ter Schure promoveert op onderzoek naar een nieuwe statistische methode aan het Centrum Wiskunde & Informatica (CWI) in Amsterdam. Met deze prille vorm van statistiek, genaamd 'safe testing', wil ze sneller conclusies trekken en efficiënter kennis verzamelen door data uit verschillende studies samen te voegen. Het liefst nog vóórdat de resultaten gepubliceerd zijn. Daarnaast pleit ze voor een betere balans tussen replicatiestudies en nieuw onderzoek. 'Ik specialiseer me eigenlijk in 'small data' en wil voorkomen dat er onnodig veel gegevens worden verzameld in wetenschappelijk onderzoek.'

'DE *P*-WAARDE GEBRUIKEN IS DE MOEILIJKSTE MANIER OM DATA IN ÉÉN GETAL SAMEN TE VATTEN'

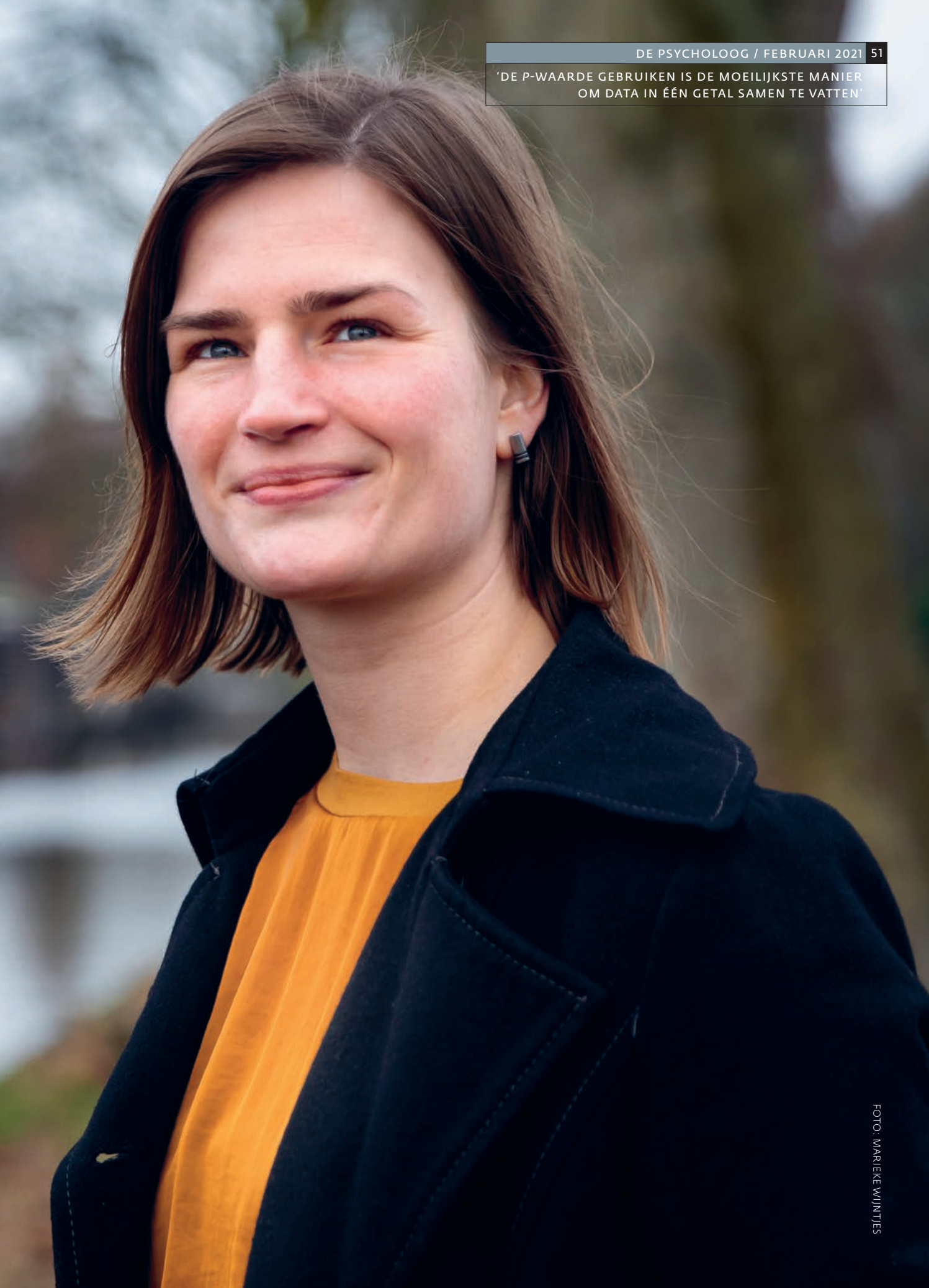


FOTO: MARIEKE WIJNTJES

'DE *P*-WAARDE GEBRUIKEN IS DE MOEILIJKSTE MANIER OM DATA IN ÉÉN GETAL SAMEN TE VATTEN'

ANOUK BERCHT

'Ik vergeef het psychologen onmiddellijk als ze niet begrijpen wat een p-waarde is'

WAT SPREEKT JE ZO AAN IN DE STATISTIEK?

'Wiskunde vond ik altijd erg leuk en op de middelbare school was ik er goed in. Maar de wiskunde leek bedoeld voor banen in de techniek of in labs, en dat leek me niets. Dus koos ik voor een bachelor kunstmatige intelligentie, een mooie mix van informatica, wiskunde, filosofie, taalkunde en psychologie. Big Data kwamen toen – heel ouderwets – nog niet aan bod. Na kunstmatige intelligentie volgde ik een master in statistiek. Statistiek is zo'n mooi vakgebied! De beroemde Amerikaanse statisticus John Tukey zei: "The best thing about being a statistician, is that you get to play in everyone's backyard." Zelf heb ik al gewerkt met seksuologen, mediadeskundigen, biologen, dierenartsen, natuurkundigen, geologen, demografen…

Sinds mijn bachelor ben ik gefascineerd door speltheoretisch onderzoek. Speltheorie is de analyse van beslissingen in een complexe context van regels, zoals in het bekende *prisoner's dilemma*. Als statisticus speel je ook een soort spel met de data, en met de kans dat de data je misleiden. Echt een heel spannend vak!'

JOUW PROMOTIEONDERZOEK IS ONDER MEER GEÏNSPIREERD OP DE REPLICATIECRISIS IN DE PSYCHOLOGIE.

'Ja. In de kankerbiologie vond trouwens eerder een replicatiecrisis plaats. Toen daar dierproeven naar chemische componenten om geneesmiddelen mee te maken op grote schaal herhaald werden, bleek zo'n vijfenzeventig tot negentig procent te mislukken.[1] Dus eigenlijk is het onterecht dat de psychologie zo geassocieerd wordt met de replicatiecrisis. Veel andere vakgebieden, zoals disciplines binnen de geneeskunde, zijn minstent zo wankel. En die lopen met hervormingen achter op de psychologie. Studies die behandelingen op patiënten testen, worden voor een publicatie in een medische tijdschrift bijvoorbeeld niet standaard van tevoren beoordeeld door collega-wetenschappers. Dat snap ik écht niet! Zo'n beoordeling vooraf door een wetenschappelijk tijdschrift heet een registered report en is een soort peer review om de kwaliteit van je onderzoeksvraag en methode te checken. Het kan immers om leven en dood gaan in de medische wetenschap, dus het lijkt me belangrijk dat wel te doen. Peer-reviewers kunnen dan nog helpen de studie beter te maken, in plaats van ze enkel achteraf te kunnen afwijzen. Voor steeds meer psychologievakbladen vindt zo'n registered report nu wel plaats.'

WAT BEN JE AL TE WETEN GEKOMEN OVER DE REPLICATIECRISIS IN DE PSYCHOLOGIE?

'Als statisticus moet ik eerst even kwijt dat statistici het er niet over eens wat nu wel en niet repliceert. Resultaten van replicaties kunnen tegenvallen in vergelijking met originele studies. Maar hoe je moet aantonen dat een studie niet repliceert, als een soort binair oordeel, weten we nog niet. Het is een interessante vraag. Maar ik vind het eigenlijk belangrijker te achterhalen of die studies samen wel of geen bewijskracht leveren tegen of voor een bepaalde theorie.'

HOE IS DE REPLICATIECRISIS TOT STAND GEKOMEN?

'In de psychologie is er tientallen jaren verschrikkelijk slordig werk gedaan. De Schotse psycholoog en wetenschapscommunicator Stuart Ritchie maakt in zijn boek *Science Fictions* onderscheid tussen vier categorieën aan knelpunten in de wetenschap die onder andere een rol spelen in replicatiecrises: bias, hype, nalatigheid en fraude. Bias is de neiging te zien wat je wilt zien. Denk aan financiële bias als je een rechtstreeks financieel belang hebt bij de uitkomsten. Of aan publicatiebias, het enkel publiceren van significante bevindingen. Bij hype worden er veel te grote claims gemaakt over de resultaten. Er ontstaat dan een groot gat tussen het abstract en de resultaten. Wetenschappers kloppen ook wel eens zelf de persberichten op. Soms is het zo erg dat er geen enkel verband meer is tussen wat de gewone burger in de krant leest en wat er in het paper stond. Dat is erg slordig. Nalatigheid gaat over statistische afrondingsfouten en zelfs typfouten waardoor de resultaten niet kloppen. En fraude spreekt voor zich.'

1 Begley, C. G. & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*(7391), 531-533.

WAAROM ZIJN WETENSCHAPPERS NALATIG?

'Er zijn te weinig prikkels en beoordelingscriteria die de nalatigheid voorkomen of afstraffen. Peer-review processen zijn bijvoorbeeld niet bedoeld om die fouten op te sporen. Misschien worden papers van eerstejaarsstudenten wel strenger beoordeeld dan artikelen die onderzoekers naar een groot vakblad opsturen. Bovendien zit het nog niet in de onderzoekscultuur om data en analysescripts te delen. Zo kunnen andere wetenschappers de fouten ook niet tegenkomen.

Te veel papers zijn inconsistent in zichzelf. Ze hebben bijvoorbeeld een gemiddelde Likert-score die niet klopt met de hoeveelheid proefpersonen.[2] Een mooie tool die een ander type fout kan opsporen is de Nederlandse uitvinding Statcheck, van Michele Nuijten en haar collega's van de Universiteit Tilburg. Die controleert bijvoorbeeld of de drie getallen van de t-toets (*t*-waarde, *p*-waarde en het aantal vrijheidsgraden) wel met elkaar kloppen. Deze wetenschappers vonden trouwens een groot percentage papers waarbij zulke inconsistenties aanwezig zijn. En bij een deel maakten die ook nog eens het verschil tussen wel en niet significant![3] Dat vind ik heel slordig.'

ONTSTAAT SLORDIGHEID DOOR ONWETENDHEID?

Lachend: 'Als wetenschapper is het je werk om te wéten wat je weet. Te weinig kennis hebben zou toch geen excuus mogen zijn om slordig werk te leveren?'

BINNEN DE WETENSCHAP IS ER EEN STROMING OP KOMST DIE AF WIL VAN DE *P*-WAARDE.

'Ja. Deze wetenschappers pleiten ervoor dat onderzoekers gewoon geen *p*-waardes meer rapporteren in vakbladen of ze niet meer met een grens vergelijken.[4] Ik vind dat gekke suggesties. Er moet eerst een alternatief komen, want *p*-waardes hebben een functie. Maar er zijn inderdaad ook problemen. De *p*-waarde is ongeveer het moeilijkste getal om data samen te vatten in één getal. Ik vergeef het psychologen ook onmiddellijk als ze niet begrijpen wat een *p*-waarde is. Tijdens mijn promotieonderzoek heb ik tientallen krantenartikelen gelezen en eens geteld hoe vaak de *p*-waarde verkeerd werd uitgelegd. Wat denk je? In al die artikelen stond het fout![5] Het is eigenlijk gek dat *p*-waardes zo'n grote rol spelen in de wetenschap terwijl bijna niemand ze begrijpt.

Een tweede probleem met de *p*-waarde is dat wetenschappers deze gebruiken als kwaliteitsstempel. Alleen onderzoeken met kleine *p*-waardes worden gepubliceerd. Daardoor worden wetenschappers verleid hun analyses een beetje te *tweaken*, *p*-hacken heet dat. Of er ontstaat publicatiebias omdat studies met een te hoge *p*-waarde in een la verdwijnen.'

'Hoe je moet aantonen dat een studie niet repliceert, als een soort binair oordeel, weten we nog niet'

WELKE ALTERNATIEVEN ZIJN ER VOOR DE P-WAARDE?

'Een optie is niet meer tegen vijf procent te toetsen, maar tegen een half procent voordat je resultaten significant mag noemen. Dat is geen alternatief, maar wel een simpele aanpassing. De statistische rechtvaardiging daarvoor is heel uitgebreid. Daarnaast voorkomt het misschien dat er te kleine steekproeven worden gebruikt, wat bijvoorbeeld vaak gebeurt in de neurowetenschap.[6]

Onderzoekers moeten eigenlijk in de ontwerpfase van hun onderzoek al anders handelen dan ze nu vaak doen. Grotere studies worden misschien beter ontworpen. Daarmee wordt de neiging ze achteraf in

2 Brown, N. J. & Heathers, J. A. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science, 8*(4), 363-369.

3 Nuijten, M. B., Hartgerink, C. H., van Assen, M. A., Epskamp, S. & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods, 48*(4), 1205-1226.

4 Amrhein, V., Greenland, S. & McShane, B. (2019). Scientists rise up against statistical significance. *Nature* 567, 305-307. doi: https://doi.org/10.1038/d41586-019-00857-9.

5 Meer over de communicatie van p-waardes: https://www.nih.gov/about-nih/what-we-do/science-health-public-trust/perspectives/science-health-public-trust/tips-communicating-statistical-significance.

6 Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience, 14*(5), 365-376.

een la te stoppen kleiner. Grotere studies hebben bovendien een robuustere bewijskracht. Toch blijft met een andere grens de focus liggen op individuele studies. Om efficiënter kennis te verzamelen, moeten we daar vanaf.'

JOUW ONDEROEK GAAT OVER EEN HEEL ANDER SOORT STATISTIEK.

'Ons team onderzoekt safe testing. Het is een type statistiek waarbij we makkelijk meerdere onderzoeken samen kunnen nemen. Maar wel op een net andere manier dan bij een meta-analyse. Als je safe testing gebruikt op de manier die ik voorstel, kijk je voordat je aan een experiment begint niet enkel naar hoeveel proefpersonen jouw onderzoek nodig heeft, maar naar hoeveel proefpersonen de wetenschap nodig heeft. Dat betekent dat je al het eerdere onderzoek met hetzelfde type vraag, de studies die je normaal gezien ook samenneemt in meta-analyses, bekijkt en je afvraagt hoeveel proefpersonen er nog nodig zijn om stevige conclusies te trekken. Niemand is immers te overtuigen op grond van één experiment, daaruit kun je geen conclusies trekken. En op deze manier kan de hele onderzoekslijn samenkomen. Onderzoekers moeten hun experimenten zo ontwerpen dat ze iets toevoegen aan de bestaande kennisbasis. Met *p*-waardes is dat moeilijk, want die zeggen niets over een hele onderzoekslijn, maar iets over maar één studie.'

HOE ZIET SAFE TESTING ER CONCREET UIT?

'Voor mijn eerste paper[7] laat ik dit zien met een theoretische tabel. In de kolommen staan alle onderzoeksvragen binnen een bepaald vakgebied. En in de rijen staan alle bijbehorende studies. Dat zijn studies die je samen in één meta-analyse zou stoppen. Dus hoe meer studies er naar eenzelfde onderzoeksvraag zijn, hoe langer de rijen onder elkaar. Idealiter moeten wetenschappers op het juiste moment stoppen met nieuwe studies in een bestaande kolom, en vervolgens aan een andere vraag beginnen. Maar in realiteit zie je dat dit niet gebeurt. In de medische wetenschappen zijn soms ruim vijftig soortgelijke studies gedaan waarbij de helft van de proefpersonen, en dat zijn dus

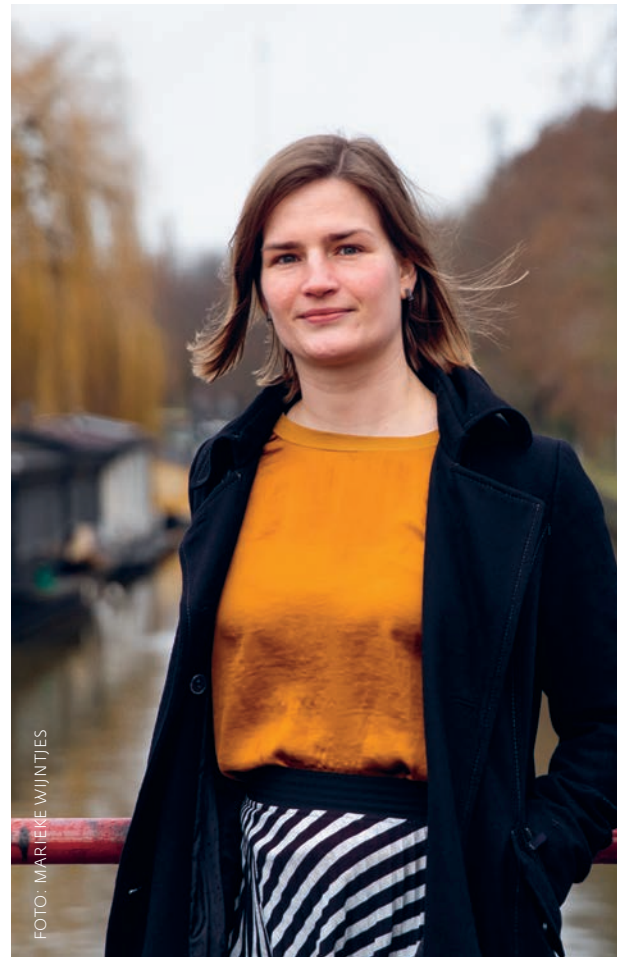


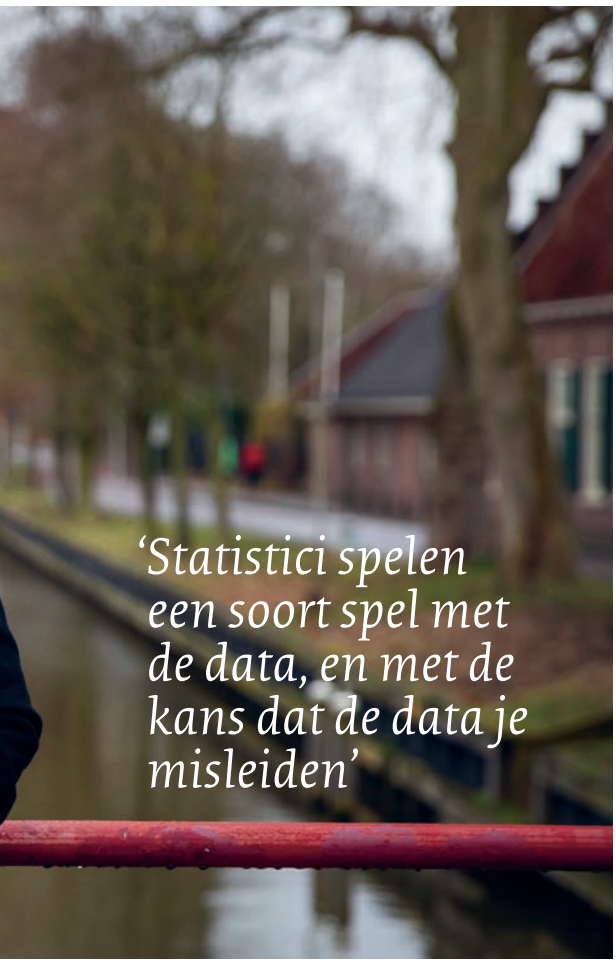
FOTO: MARIEKE WIJNTJES

soms doodzieke mensen, placebo's kregen. Terwijl er allang conclusies getrokken konden worden. Dat vind ik onethisch. In de sociale wetenschappen komen er juist veel nieuwe kolommen bij. En ook dat is hinderlijk voor het opbouwen van kennis.'

HOE ZIET DE STATISTIEK VAN SAFE TESTING ER DAN UIT?

'Wij gebruiken geen *p*-waardes maar *e*-waardes. En een e-waarde zegt, net als de *p*-waarde, iets over hoe verrassend de resultaten zijn. En dus over hoeveel bewijs je voor of tegen je onderzoeksvraag hebt. In de wiskundige betekenis staat '*e*' voor expectation, ongeveer zoals '*p*' in *p*-waarde staat voor probabillity. E-waardes leggen wij vaak uit met behulp van gokken. Als onderzoeker voer je een experiment uit en aan de hand van een soort gokstrategie verwacht je bepaalde resultaten. Stel bijvoorbeeld dat je de allereerste studie naar priming ooit uitvoert. Je wilt weten of bepaalde informatie vooraf het antwoord van je proefpersonen op een moeilijke vraag zal beïnvloeden. De nulhypothese zou zijn dat priming niet werkt, en dat al je proefpersonen een willekeurig antwoord geven op een moeilijke vraag waarvan de kans nihil is dat ze het antwoord weten. Maar de gokstrategie, dus je

7 ter Schure, J. & Grünwald, P. (2019). Accumulation Bias in meta-analysis: the need to consider time in error control. *F1000Research, 8*.



verwachting, vertelt je dat de prime werkt en dat de antwoorden niet willekeurig zijn maar afhankelijk van de prime. Je zet dan geld in op wat je denkt dat er gaat gebeuren. Als die uitkomsten zich dan inderdaad voordoen, verdien je geld. En als de proefpersonen zich niets van je prime aantrekken, verlies je geld. De e-waarde kun je vergelijken met hoeveel geld je aan het gokspel overhoudt. Het belangrijkste daaraan is dat een onderzoekersteam dat daarna met hetzelfde onderwerp aan de slag gaat jouw "winst" moet gebruiken om meer wetenschappelijke resultaten op te stapelen.

Je zet natuurlijk niet echt geld in. Maar als je een e-waarde-analyse draait op je wetenschappelijke data mag je het resultaat als gokwinst interpreteren. We hopen dat onderzoekers er met die interpretatie aan herinnerd worden dat hun resultaten op toeval gebaseerd kunnen zijn. Als je met tien euro op zak naar een casino gaat en je er met twintig uitkomt, zeg je ook niet direct dat je weet hoe je rijk moet worden in het casino. Maar als je de gokstrategie herhaalt met je verdiende geld en daarna met tweehonderd euro het casino uitloopt, en daarna met vierduizend, dan lijkt die gokstrategie toch redelijk goed te werken. Datzelfde idee willen wij introduceren in de wetenschap.'

HOE WERKT HET IN DE PRAKTIJK?

'Als alle wetenschappers met *e*-waardes zouden werken, zou het heel gemakkelijk zijn. Voorafgaand aan een experiment check je in welke kolom je werkt met een systematische zoektocht in de literatuur. In een ideale wereld heeft iedere studie in die kolom al een e-waarde en hoef je die alleen maar met elkaar te vermenigvuldigen om te kijken hoe het ervoor staat. Een gokstrategie die van één euro twintig euro maakt kan immers van tien euro uit een vorige studie tweehonderd euro maken.

Van elke onderzoekersvraag moet vooraf vastgelegd zijn hoe hoog de e-waarde moet worden om stevige conclusies te trekken. Hoe hoger die grens, hoe minder garantie op een verkeerde conclusie. Daarmee kunnen wetenschappers ook bepalen hoeveel proefpersonen nog nodig zijn. Het uitrekenen van de e-waarde zelf is ook heel flexibel. Dat kan ook tijdens het experiment. Als je de afgesproken e-waarde al haalt met de helft van je proefpersonen, mag je eerder stoppen. Voor snellere resultaten kun je ook tussentijds de resultaten met andere studies samenvoegen die tegelijkertijd lopen, zelfs als die nog niet zijn afgerond. De coronapandemie laat zien hoe belangrijk het is niet te hoeven wachten tot studies gepubliceerd zijn.'

WETENSCHAPPERS MOETEN DAN WEL LEREN OM NIET STEEDS MET NIEUWE IDEEËN TE KOMEN.

'De mindset van een onderzoeker is inderdaad vaak gericht op iets nieuws doen. Maar het hangt ook af van het veld. In de medische wetenschappen gebeurt het juist ook dat onderzoekers overal ter wereld exact dezelfde vragen stellen. Bijvoorbeeld of een bepaald middel bloedingen tijdens hartoperaties kan voorkomen. En in de sociale wetenschappen werken onderzoekers juist steeds aan nieuwe onderzoeksvragen, terwijl er nog onvoldoende studies gedaan zijn om de al bestaande onderzoeksvragen te beantwoorden.'

WORDT SAFE TESTING AL GEBRUIKT?

'Zelf zijn we bezig met een analyse van studies naar of het BCG-vaccin tegen tuberculose ook beschermt tegen het coronavirus. Dat onderzoek gebeurt gelijktijdig op verschillende plaatsen, onder andere in Denemarken en in Nederland. Ik ga tussentijds kijken

of de e-waarde die verdiend is met de tussentijdse data uit Nederland, vermeerdert als ik die inzet op de data uit Denemarken. Verder gebruikt niemand nog de e-waardes. Al wordt binnen de psychologie de Bayesiaanse statistiek steeds populairder. En dat lijkt wel op onze methode. Ik denk dat de wetenschap klaar is voor e-waardes. Maar wat wij willen kan ook botsen met de belangen van wetenschappers. Wij willen zo slim mogelijk resultaten van allerlei studies bij elkaar vegen. En niet dat iedere wetenschapper alleen bezig is met z'n eigen paper.'

JE HOUDT JE VOORAL BEZIG MET SMALL DATA. HOE DENK JE OVER BIG DATA?

'Grotere steekproeven zijn natuurlijk vaak goed. Al moet je wel goede experimenten opzetten en afwegen hoe groot je steekproef nu echt moet zijn. Het is veel werk om met veel proefpersonen van alles in de gaten te houden. Als je controles hebt, een blind onderzoek uitvoert, en de proefpersonen willekeurig indeelt in verschillende groepen, gaan de voordelen van een grote steekproef pas tellen. Als je zomaar van alles gaat meten, kunnen data misleiden. Bovendien valt het voordeel van die grote steekproef ook weg als je én bloeddruk, én hartslag, én schermtijd gaat meten, bijvoorbeeld. Heel veel variabelen. Het is wel tof dat er nu een heel veld bezig is om met nieuwe ideeën over statistiek toch conclusies te trekken uit allerlei observationele data.'

HEB JE TIPS VOOR PSYCHOLOGEN DIE ZELF ONDERZOEK DOEN?

'Het is heel menselijk om je in de luren te laten leggen door opvallende patronen in je data. Terwijl die ook kunnen zijn ontstaan door toeval. Het is nuttig je daar steeds bewust van te zijn. Als onderzoeker kun je je daar ook tegen wapenen door je te verdiepen in de replicatiecrisis. Dan zie je dat herhalingen van experimenten soms gewoon mislukken omdat een van de twee bevindingen op toeval is gebaseerd. Om geen stomme fouten te maken, kun je ook tools gebruiken die resultaten rechtstreeks in je paper plakken uit codes die je voor je data-analyse gebruikt. Dan kunnen er geen typefouten en afrondingsfouten ontstaan. De belangrijkste aanbeveling is om zoveel mogelijk open te werken en data en analysescripts te delen. Dan pas kunnen anderen je werk echt goed nalopen.'