

WETENSCHAP & ONDERWIJS



Betekent de doorbraak van deze wiskundige het einde van de p-waarde?

Wetenschap 29 APRIL 2024

De p-waarde deugt niet, roepen statistici al ruim zestig jaar. VU-statisticus Rianne de Heide en collega's presenteren een waardige opvolger om verbanden mee aan te tonen: de e-waarde.

Het is 1995 als in New York dertig proefpersonen meedoen aan een experiment dat de wetenschap later zal doen opschudden. De deelnemers krijgen te horen dat ze getest worden op hun taalvaardigheid, maar stiekem willen de onderzoekers iets heel anders meten: kun je het gedrag van mensen beïnvloeden zonder dat ze het doorhebben? De proefpersonen moeten woordpuzzels oplossen met woorden die verwijzen naar ouderen. Zodra ze klaar zijn en de testruimte verlaten, meet een onderzoeker heimelijk hoe lang ze erover doen om naar de uitgang te lopen. Wat blijkt: de proefpersonen die onbewust een beeld geschetst kregen van ouderen lopen langzamer dan de controlegroep.

Waarheid of ruis?

Bovenstaand onderzoek is inmiddels meer dan 6500 keer geciteerd en stond aan de basis van een heel nieuw onderzoeksveld. Maar hoe betrouwbaar is de uitkomst? Om die vraag te beantwoorden herhaalden Belgische wetenschappers in 2012 het onderzoek met twee keer zoveel proefpersonen. Tot hun verbazing zagen ze geen verschil tussen de test- en controlegroep. Het is een tekenend voorbeeld van wat sommige wetenschappers de 'replicatiecrisis' noemen: wetenschappelijke uitkomsten die niet na te bootsen blijken. Zo toonde een Science-publicatie uit 2015 aan dat minder dan de helft van de resultaten uit psychologisch onderzoek standhoudt. En van een selectie van 21 sociologische papers in Nature en Science waren er slechts 13 replicerbaar, de overige 8 bleken ruis.

P-waarde is failliet

“Dat wetenschappelijke resultaten geen stand houden, heeft allerlei oorzaken”, zegt statisticus Rianne de Heide. “Denk aan publication bias [alleen positieve resultaten die gepubliceerd worden, red.] of mensen zoals Diederik Stapel die fraude plegen. Maar wat volgens mij een onderbelichte oorzaak is, is het gebruik van de p-waarde. Of zoals wij statistici zeggen: p-waarde gebaseerd nulhypothese-significantietoetsing. Dat is de manier waarop vrijwel elke wetenschapper in de afgelopen honderd jaar statistisch onderzoek deed.”

Mede door de replicatieproblemen zwol de kritiek op het gebruik van deze toetsingsmethode de afgelopen tien jaar aan, al is kritiek niet alleen van nu. In 1963 verscheen er al een Amerikaans paper die beschrijft wat er allemaal mankeert aan de p-waarde. “Sindsdien zijn veel statistici het erover eens dat het een slechte manier is van hypothesetoetsing. Maar ze konden het maar niet eens worden over een alternatief.”

P-waarde

Wat is de p-waarde ook alweer? De precieze uitleg is ingewikkeld, maar simpel gezegd is het een getal waarmee je kunt bepalen of de uitkomst van een steekproef statistisch gezien significant is. De p-waarde is altijd een getal tussen 0 en 1.

Statistisch geruzie

Dat de p-waarde tot dusver nooit een waardige opvolger kreeg, komt onder meer doordat er verschillende stromingen bestaan binnen de statistiek. De bekendste zijn het frequentisme en het Bayesianisme. “Aanhangers van deze stromingen vechten elkaar weleens de tent uit”, zegt De Heide. “En als statistici het al niet met elkaar eens kunnen worden, waarom zouden toegepaste wetenschappers dan iets veranderen aan hun werkwijze?”

Na jaren van soebatten is er nu eindelijk een waardige opvolger van de p-waarde: de e-waarde. De nieuwe theorie van De Heide en collega's Peter Grünwald en Wouter Koolen lijkt een oplossing te kunnen bieden voor de problemen van de p-waarde. Het elegante van de e-waarde is namelijk dat aanhangers van beide stromingen die kunnen gebruiken. En misschien wel belangrijker: de e-waarde is eenvoudig te begrijpen én te gebruiken. “De definitie van de p-waarde is heel ingewikkeld en bijna niet uit te leggen aan mensen zonder wiskundige achtergrond. In een onderzoek uit 1996 onder Deense artsen bleek dat 70 procent van hen de p-waarde niet goed begreep. En als mensen iets niet snappen, dan gaan ze er fouten mee maken.”

Tussentijds stoppen

Een ander probleem van de p-waarde is te vinden in het zogenaamde optional stopping. De Heide legt uit wat dat inhoudt. “Optional stopping is dat je tijdens je onderzoek naar je data kijkt, misschien alvast berekent of je resultaten significant zijn en op basis daarvan besluit om nog een aantal proefpersonen aan je experiment toe te voegen. In veel onderzoeken is dat een handige werkwijze, omdat je eerst onderzoek kan doen op een kleine groep patiënten of proefpersonen. Als dat nergens toe leidt, stop je en hoef je geen tijd en geld meer te verspillen aan het onderzoek. Als het wel veelbelovend lijkt, kun je stapsgewijs proefpersonen toevoegen totdat je tevreden bent met je resultaat.”

Het grote probleem met de p-waarde is dat veel onderzoekers er optional stopping mee doen, terwijl dat volgens De Heide absoluut niet de bedoeling is. “Als je het wel doet, vergroot je de kans op een foutpositief enorm. Dat betekent dat je een verband vindt dat helemaal niet bestaat, bijvoorbeeld dat een pil beter werkt dan een placebo, terwijl er in werkelijkheid geen verschil bestaat tussen die twee. Zo'n foutpositief wil je uiteraard voorkomen. Daarom is de p-waarde zo ontworpen dat er een grens is, die vaak bij 5 procent ligt. Dat betekent dat als je dit experiment heel vaak doet – en er geen effect is – je in 5 procent van die experimenten een foutpositief resultaat krijgt. Doe je aan optional stopping, dan wordt de kans op een foutpositief vele malen groter.”

Het gemak van de e-waarde

Dat dit een groot probleem is, blijkt onder meer uit een studie uit 2012 onder 2.155 Amerikaanse psychologen. Maar liefst 55 procent van hen maakte zich schuldig aan optional stopping met p-waardes. Geen wonder dus dat een groot deel van het onderzoek in de psychologie niet na te bootsen valt. De e-waarde van De Heide en collega's biedt ook voor dit probleem een oplossing: de wiskunde erachter is ingewikkeld, maar 'tussentijds stoppen' mag met de e-waarde wél. Verder is de e-waarde een stuk eenvoudiger uit te leggen dan de p-waarde. Hoe groter de e-waarde, hoe sterker het bewijs voor je bevinding.

Voor wie iets dieper de materie in wil geeft De Heide een toelichting op de e-waarde. “Wil je, net als bij de p-waarde, een grenswaarde van vijf procent kans op foutpositieven? Dan deel je 1 door vijf procent om de grenswaarde voor de e-waarde te berekenen. Oftewel, de e-waarde moet 20 of hoger zijn.”

Niet alleen maar voordelen

Heeft de e-waarde dan alleen maar voordelen? Volgens de Heide biedt de e-waarde veel flexibiliteit en lever je daarmee wat in op de p-waarde. “Bij de p-waarde weet je van tevoren precies hoeveel proefpersonen je nodig hebt, wat handig is als je bijvoorbeeld een MRI-scanner moet affhuren. Verder blijkt uit onze simulaties dat je gemiddeld net iets minder data nodig hebt met de e-waarde, maar dat de e-waarde voor individuele gevallen soms minder gunstig is en je dan juist meer data nodig hebt.”

2,5 miljoen euro

Deze maand ontving Leids hoogleraar Peter Grünwald, co-auteur en oud-promotor van Rianne de Heide, een ERC Advanced Grant van 2,5 miljoen euro. Hiermee gaat hij de theorie rondom de e-waarde verder uitwerken en die bestendig maken voor ingewikkelde analyses.

Briljant idee

Haar onderzoek, nogal theoretisch van aard, ging niet zonder slag of stoot, vertelt De Heide. “Soms zat ik maanden vast. Dan besloot ik om iets anders te gaan doen en las ik opeens een paper waarin toevallig iets stond dat me verder kon helpen. Zo gaat het in de wiskunde. Je verzint tien dingen, negen ervan kloppen niet, maar dat ene idee blijkt opeens briljant te zijn”, lacht ze.

Collega-wetenschappers van over de hele wereld zijn enthousiast over de nieuwe theorie, zo merkte De Heide begin dit jaar. Samen met collega-auteurs Grünwald en Koolen kreeg ze in januari de kans om haar onderzoek te presenteren bij de Royal Statistical Society in London, de meest vooraanstaande statistiekvereniging ter wereld. Een eer die alleen is weggelegd voor de allerbeste statistische publicaties.

Waakhonden overtuigen

Of de wetenschap over tien jaar volop met de e-waarde werkt, is lastig te zeggen. De Heide denkt dat er meer tijd voor nodig is. “Over tien jaar zal er in ieder geval veel meer theoretische kennis over de e-waarde zijn, want het is een spannend nieuw onderzoeksveld waaraan nu al veel mensen werken. Zo ontwikkelen wij nu softwarepakketten waarmee je met een druk op de knop e-waardes kunt berekenen. Maar voordat het echt in toegepast onderzoek gebruikt gaat worden, zijn we wel een paar decennia verder. Iedereen is zo gewend om p-waardes te gebruiken en alle regels zijn daaromheen gebouwd. De waakhonden die over medisch onderzoek gaan, moeten er bijvoorbeeld van overtuigd raken dat we de p-waarde door de e-waarde moeten vervangen.”

Reactie Casper Albers

Hoogleraar Toegepaste Statistiek & Datavisualisatie bij de Rijksuniversiteit Groningen (niet betrokken bij het onderzoek)

Statistische toetsen worden meestal uitgevoerd op basis van p-waarden. Hoewel zeer nuttig, kennen p-waarden ook de nodige problemen. Het voornaamste probleem is dat te veel wetenschappers de onderliggende wiskunde te slecht begrijpen en daardoor denkfouten maken. Daarnaast zijn p-waarden ongeschikt voor zogenaamde sequentiële analyses [een vorm van optional stopping, red.].

Het gebruik van e-waarden als alternatief voor p-waarden is veelbelovend en hier is de afgelopen jaren ruimschoots over gepubliceerd in wiskundig-statistische literatuur. Met e-waarden heb je geen probleem bij sequentieel onderzoek. De onderliggende wiskunde is nóg ingewikkelder dan die van p-waarden, maar omdat het inhoudelijk beter aansluit bij hoe wetenschappers intuïtief nadenken over onderzoek zou het risico op denkfouten kleiner zijn.

Of dit daadwerkelijk zo is zal nog moeten blijken. Om succesvol te worden als vervanger van p-waarden is het essentieel dat ook bijvoorbeeld aan sociale en medische wetenschappers uitgelegd wordt wat de voordelen zijn en hoe ze gebruikt moeten worden. Dat vereist het nodige zendingswerk van Rianne de Heide en Peter Grünwald.

70 procent van de onderzochte Deense artsen begreep de p-waarde niet.

"Aanhangers van twee stromingen in de statistiek vechten elkaar weleens de tent uit"