Schets en scheef

JOOST VAN GEMEREN
Redactiemedewerker ESB
j.v.gemeren@sdu.nl

Weet u nog, de eerste keer dat uw wiskundeleraar op de middelbare school de normale verdeling op het bord probeerde te krijten, maar zich vervolgens meermaals verontschuldigde omdat hij die mooie curve weer niet symmetrisch uit de losse pols had kunnen fabriceren? Zo nu en dan wijzend naar de curve die meer weg had van de toren van Pisa dan van de beoogde klokvorm, vertelde hij dan over de p-waarde, iets wat wij wetenschappers tegenwoordig aan onze studenten uitleggen als de kans dat een empirisch ontdekte relatie ontstaat uit puur toeval. Als wordt uitgegaan van een standaard significantieniveau van 5 procent, zijn slechts 5 procent van de door ons 'empirisch bewezen' verbanden ontstaan uit puur toeval. Toch?

De psychologen Simmons *et al.* (2011) tonen aan dat dit vermeende ABC'tje niet opgaat voor onderzoeken door wetenschappers die zich bezondigen aan *p-hacking*, een vorm van fraude waardoor onder andere Dirk Smeesters zijn wetenschappelijke Waterloo vond (mede dankzij de inspanningen van Simonsohn). Onder p-hacking verstaat men het strategisch gebruiken van in principe legale onderzoeksvrijheden om een statistisch insignificant verband als significant te presenteren. Als meerdere vormen van p-hacking worden gecombineerd, zoals het selectief gebruiken van controle variabelen en het strategisch verwijderen van uitbijters, is de kans dat een verband onterecht wordt gepresenteerd als bewezen niet langer 5 procent, maar veel groter. Het onderzoek van Simmons *et al.* toont aan dat de kans op zo'n Type I fout in het extreemste geval kan oplopen tot liefst 61 procent. Als onderzoekers maar genoeg p-hacken kan het dus zijn dat een significant verklaard resultaat vaker onzin is dan zin. Bewezen is daarmee dat de kansverdeling die p-hackers hanteren bewust scheef is, in tegenstelling tot die van een goeie ouwe docent.

Dat p-hacking ook in de economische wetenschap plaatsvindt, bevestigen Brodeur *et al.* in hun onlangs verschenen paper met de memorabele naam '*Star wars: the empirics strike back*'. In een aantal economische topjournals komen p-waarden van vlak onder de 0,05 opvallend vaak voor, terwijl p-waarden tussen de 0,10 en 0,25 juist erg weinig worden gesignaleerd. Volgens de onderzoekers is dit verdacht genoeg om te concluderen dat de p-waarden tussen de 0,25 en 0,10 met het laserzwaard een kopje kleiner zijn gemaakt tot vlak onder de 0,05. Daarbij moet gezegd worden dat de galactische oorlog tegen het insignificante echter minder hevig lijkt te woeden onder economen die zich bezighouden met veld- of laboratoriumexperimenten.

Wat beweegt wetenschappers om zich te verlagen tot p-hacking? Ten eerste worden lezers niet warm of koud van een verband dat niet kon worden aangetoond. Journals die lezers geboeid willen houden publiceren dus voornamelijk resultaten met sterretjes erachter. Ruud Abma, die eind mei zijn boek over de beweegredenen van fraudemastermind Diederik Stapel publiceerde, voegt daaraan toe dat het wetenschappelijke motief informeren in het huidige academische klimaat niet altijd meer opweegt tegen de drang om te presteren.

Gelukkig wordt er door wetenschappers zelf aan de weg gebouwd om p-hackers te ontmaskeren. Onlangs publiceerden Simonsohn *et al.* (2013) een paper waarin een nieuwe methode wordt gepresenteerd om dit voor elkaar te krijgen. Een p-hacker publiceert namelijk vaker p-waarden die tussen de 0,04 en 0,05 liggen dan p-waarden die kleiner zijn dan 0,01. Voor 'schone' wetenschappers geldt dit niet. Dus: door zijn significante p-waarden te plotten in een frequentiediagram (de *p-curve*) kan men de integriteit van een wetenschapper polsen. Deze methodologie kan ook worden toegepast op afzonderlijke journals, instellingen, Duitse politici, enzovoorts.

De p-curve is echter nog onderontwikkeld. Het ziet bijvoorbeeld p-hacking bij niet-experimentele onderzoeken soms door de vingers. Ook zijn onderzoekers die slechts mild of sporadisch p-hacken volgens de p-curve nauwelijks verdacht. In de toekomst zal dit veranderen, zodra er een toets ontwikkeld is die p-hackers op een robuustere en minder tijdrovende wijze kan ontmaskeren. Tot die tijd is het hopen dat de conclusies die onderzoekers trekken niet schever zijn dan de schets van een wiskundeleraar.

LITERATUUR

Brodeur, A., M. Lé, M. Sangnier en Y. Zylberberg (2013) Star wars: the empirics strike back. *IZA Discussion Paper*, 7268.

Simmons, J.P., L.D. Nelson en U. Simonsohn (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359-1366.

Simonsohn, U., L.D. Nelson en J.P. Simmons (2013) P-curve: a key to the file drawer. *Journal of Experimental Psychology: General*, te verschijnen.