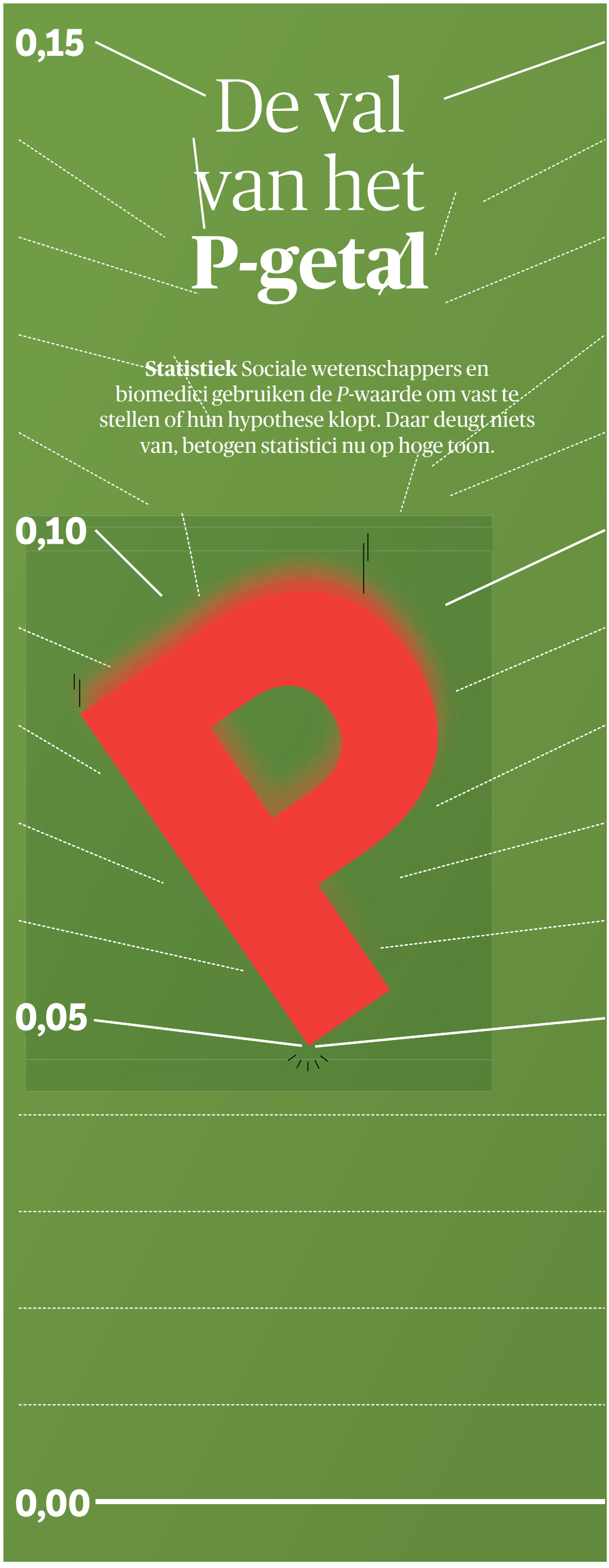
De val van het P-getal

Statistiek Sociale wetenschappers en biomedici gebruiken de *P*-waarde om vast te stellen of hun hypothese klopt. Daar deugt niets van, betogen statistici nu op hoge toon.



ILLUSTRATIE FOKKE GERRITSMA

Door **Ellen de Bruin**

Onderzoekers die willen weten of een medicijn werkt, geven willekeurig één groep patiënten dat medicijn en een andere groep een placebo, en verzamelen daarna cijfers over hoe het met die mensen gaat. Gaat het beter met de mensen die het medicijn hebben gekregen? De onderzoekers doen een statistische toets. En vervolgens zijn ze opgetogen als die toets een *P*-waarde kleiner dan 0,05 oplevert. Dan achten de onderzoekers hun hypothese dat het medicijn werkt bewezen. Maar dat is onterecht.

De medicijnonderzoekers menen bijvoorbeeld dat $p < 0,05$ betekent dat de kans dat hun hypothese onjuist is, of de kans dat hun data op toeval berusten, kleiner is dan 5 procent. Maar dat klopt allemaal niet. De *P*-waarde is nooit bedoeld geweest om dit soort gevolgtrekkingen te maken, ook al wordt hij over het algemeen wel zo gebruikt, in met name de biomedische en sociale wetenschappen. Onbegrip of verkeerd gebruik van statistiek is een belangrijke oorzaak van de huidige crisis in de wetenschap, waarbij veel onderzoeksresultaten niet gerepliceerd blijken te kunnen worden. Dat schreven statistici van de American Statistical Association (ASA) begin maart in een alarmerende publieke verklaring. Steven Goodman van het Meta-Research Innovation Center aan de Amerikaanse Stanford-universiteit herhaalt die boodschap in *Science* (3 juni).

Nulhypothese

Wat kun je op basis $p < 0,05$ dan wel concluderen? Het probleem is dat de definitie van de *P*-waarde niet meteen intuïtief duidelijk is voor niet-statistici. De *P*-waarde zegt niets over de kans dat een bepaalde hypothese of theorie juist is; *P* zegt iets over de kans op de gevonden data. Een $p < 0,05$ in de simpele medicijntest uit het voorbeeld hierboven betekent dat de kans op het gevonden verschil tussen medicijn en placebo of een nog groter effect minder dan vijf procent is, wanneer álle aannamen die aan de berekening van de *P*-waarde ten grondslag liggen zouden kloppen - inclusief het idee dat het medicijn geen effect zou hebben (dat is de zogeheten 'nulhypothese').

Als p heel klein is kún je dus gaan twijfelen aan de nulhypothese, maar ook aan alle andere aspecten van je onderzoeks- en analysemethode. Daarom is het ook zo belangrijk om onderzoek te herhalen. En om precies te rapporteren wat je hebt gedaan: alle beslissingen die je hebt genomen bij het verzamelen en analyseren van de data.

Er is dus geen magische grens op $p = 0,05$ waar een hypothese waar of onwaar wordt. Dat wordt wel algemeen gedacht: het idee dat een p kleiner dan 0,05 zó klein is, 'significant' klein, dat je hypothese juist is en dat je de nulhypothese kunt verwerpen, wordt door veel onderzoekers in de biomedische en sociale wetenschappen voor waar aangenomen. En die leren het ook aan hun studenten en aio's, wat de gewoonte om *P* verkeerd te interpreteren in stand houdt.

Maar er is nog een andere reden waarom dat fout is, schrijven de ASA-statistici. Niet alleen zegt de *P*-waarde niets over de waarheid van je hypothese - bovendien kun je, als je steekproef maar groot genoeg is, of je meetmethode nauwkeurig genoeg, altijd wel een kleine p verkrijgen, hoe minuscuul het gemeten effect ook is. Iets wat onderzoekers als 'statistisch significant' rapporteren, hoeft helemaal niet betekenisvol te zijn. Er is geen heldere grens die *P*-waarden opdeelt in betekenisvol en betekenisloos. De veelgebruikte grens van 0,05 is meestal een arbitraire, op gewoonte gebaseerde keuze.

Ronald Aylmer Fisher (1890-1962), in de jaren twintig en dertig van de vorige eeuw een van de grondleggers van de statistiek, ageerde al tegen zulke slechte gewoontes, schrijft Goodman in *Science*. „[G]een wetenschappelijk werker heeft een vast significantieniveau waarop hij, jaar na jaar en in alle omstandigheden, hypothesen verwerpt", citeert Goodman Fisher in diens handboek *Statistical Methods and Scientific Inference* uit 1956. Fisher vond een kleine *P* reden voor verder onderzoek. Dit is dus ook echt allemaal niet nieuw, benadrukken zowel Goodman als de ASA-statistici. En er is al heel vaak gewaarschuwd voor verkeerd gebruik van *P*-waarden en nulhypothesesignificantietests.

Maar hoe kunnen hele wetenschapsgebieden die gewend zijn aan het gebruik van *P*-waarden om nulhypothesen te testen hun manier van werken veranderen? Wat moet er gebeuren? Onderzoekers zouden sowieso alles helder moeten beschrijven wat ze hebben gedaan, vinden de ASA-statistici, en veel meer van hun data moeten weergeven dan alleen een P-waarde. Die *P*-waarden moeten in elk geval worden aangevuld, en misschien zelfs wel vervangen door andere, modernere soorten statistische tests, zoals Bayesiaanse analyses die meer rekening houden met de *a priori* waarschijnlijkheid van verschillende hypotheses.

Of dat nu echt gaat gebeuren is natuurlijk afwachten, maar er beweegt al wel iets. Er verschijnen bijvoorbeeld steeds vaker artikelen over verkeerde interpretaties van statistiek. Zoals dat van een internationaal team van epidemiologen en statistici in *European Journal of Epidemiology* (21 mei online), die vijfentwintig veel voorkomende verkeerde interpretaties van *P*-waarden, betrouwbaarheidsintervallen en statistische *power* opnoemen en uitleggen (ze laten elk misverstand volgen door een gealarmeerd „*No*!"). Verder besloten de hoofdredacteuren van *Basic and Applied Social Psychology* vorig jaar helemaal geen artikelen meer te accepteren die gebaseerd zijn op nulhypothese-tests met *P*-waarden.

> Er is geen heldere grens die *P*-waarden opdeelt in betekenisvol en betekenisloos

Altijd in beweging

En de Amsterdamse psycholoog Eric-Jan Wagenmakers heeft samen met collega's het gratis *open source*-programma JASP (Just Another Statistical Program) ontwikkeld, waarmee onderzoekers ook Bayesiaanse statistische methoden kunnen toepassen. Met het veelgebruikte programma SPSS (Statistical Package for the Social Sciences) kon dat niet.

Maar voor wetenschappers die middenin hun biomedische of sociaal-wetenschappelijke onderzoek zitten, is het misschien nog niet makkelijk om te accepteren dat ook de statistiek een wetenschap is met verschillende theorieën en benaderingen en, zoals het wetenschap betaamt, altijd in beweging. Die zullen hun onderzoek willen blijven doen zoals ze het altijd deden en zoals op dit moment nog de meeste redacteuren van tijdschriften het zelf doen en herkennen. Wie iets nieuws gaat doen, zal zich aanvankelijk isoleren - tot het nieuwe de nieuwe standaard wordt. Als dat ooit gebeurt.