Het probleem van Popper

De nadelen van zijn hypothetico deductieve model

Paul Lodder
Universiteit van Amsterdam, Amsterdam, Nederland

Volgens Poppers deductieve wetenschapsmodel moet uit iedere theorie een toetsbare voorspelling kunnen worden afgeleid. Ondanks dat dit model tegenwoordig de norm is binnen veel wetenschapsgebieden, kent het een aantal nadelen. De twee voornaamste problemen van de hedendaagse wetenschappelijke methode zullen worden besproken. Deze problemen zullen worden gekoppeld aan wetenschapsfilosofische theorieën.

Poppers model

Als reactie op de inductieve logica van de logisch positivisten stelde Popper een deductief wetenschapsmodel voor (Suppe, 1977). Dit model houdt in dat uit een theorie of hypothese een toetsbare en falsifieerbare voorspelling wordt afgeleid, en dat de theorie op basis van het resultaat van deze toetsing kan worden gecorroboreerd of gefalsifiseerd. Corroboratie houdt in dat het resultaat van de toetsing overeenkomt met de vooraf gestelde hypothese, en falsificatie houdt in dat het resultaat van de toetsing de hypothese tegenspreekt (Popper, 1959).

Poppers deductieve model heeft een grote rol gespeeld in de tot stand koming van de hypothetisch-deductieve methode; een manier van wetenschap uitoefenen die op dit moment dominant is binnen wetenschapsgebieden zoals de sociologie en de psychologie. Volgens deze methode wordt op basis van theorie of een hypothese een voorspelling afgeleid die vervolgens met een experiment kan worden getoetst. De resultaten worden geanalyseerd met behulp van statistische analyses en op basis van de uitkomst van de analyse wordt geconcludeerd of de vooraf aangegeven hypothese wordt gesteund of wordt verworpen.

De hypothetisch-deductieve methode (HD-methode) is de norm binnen een aantal wetenschapsgebieden en wordt soms ook wel de wetenschappelijke methode genoemd. Maar wordt deze methode in de praktijk eigenlijk wel echt toegepast? En hoe erg is deze methode vatbaar voor verkeerde interpretaties of zelfs fraude? De recente fraudegevallen binnen de wetenschap – met Diederik Stapel als toppunt – hebben wetenschappers en filosofen gestimuleerd nog eens goed te kijken naar de manier waarop wetenschappers nu eigenlijk te werk gaan. In dit essay wil ik een aantal zaken aan het licht brengen die aangeven dat de wetenschappelijke praktijk niet altijd even zuiver en objectief verloopt zoals de hypothetisch-deductieve methode ons doet geloven.

► E-mail: **p.lodder@uva.nl**

HARKing

HARKing (Hypothesizing After the Results are Known) kan worden gedefinieerd als het in de introductie van een onderzoekrapport presenteren van een post hoc hypothese (hypothese die is bedacht op basis van de resultaten van het onderzoek) en vervolgens doen alsof deze hypothese a priori is (is voorgesteld voordat de data is verzameld). Als onderzoekers gebruik maken van HARKing, dan nemen ze de plausibiliteit van post hoc hypotheses mee bij het beslissen over welke hypotheses in de introductie van het onderzoek worden gerapporteerd (Kerr, 1998).

Er zijn verschillende vormen van HARKing, waarvan de drie bekendste nu zullen worden besproken. In de eerste plaats is er een pure vorm van HARKing, waarbij na het zien van de resultaten van een onderzoek een hypothese wordt gekozen die het meest overeenkomt met de gevonden resultaten. Bij deze vorm van HARKing kan het goed voorkomen dat de post hoc hypothese totaal geen rol heeft gespeeld in het ontwerpen van het onderzoek, omdat de hypothese in eerste instantie als implausibel werd beschouwd, of zelfs niet eens was voorgesteld. Bij een tweede vorm van HARKing, namelijk empirische inspiratie, is de initiële onderzoekshypothese door de data verworpen en wordt er vervolgens gezocht naar theorieën die wel bij de data passen en op basis waarvan de initiële hypothese verfijnd kan worden. Een derde vorm van HARKing is het verwijderen van verworpen hypotheses uit het onderzoeksverslag en daarmee dus alleen de bevestigde hypotheses rapporteren (Kerr, 1998).

HARKing in de praktijk?

Er is onderzocht hoe vaak HARKing ongeveer wordt toegepast binnen de gedragswetenschappen. Aan 156 wetenschappers, afkomstig uit de disciplines Sociale psychologie, Klinische psychologie, en Sociologie, werd gevraagd om aan te geven hoe vaak ze de HD-methode en de hierboven beschreven drie vormen van HARKing in hun vakgebied hebben opgemerkt, en wat hun verwachting is van hoe vaak deze methodes daadwerkelijk

voorkomen. Ten slotte werd gevraagd hoe vaak iedere methode idealiter zou moeten voorkomen.

Het bleek dat de HD-methode even vaak door de deelnemers werd geobserveerd als de pure vorm van HARKing en de vorm empirische inspiratie. Daarnaast werd door de deelnemers verwacht dat deze twee vormen van HARKing in de werkelijkheid zelfs vaker zouden voorkomen dan de HD-methode. Ten slotte bleek dat de HD-methode en de empirische inspiratie vorm van HARKing volgens de deelnemers idealiter zouden moeten worden toegepast in de wetenschap en dat de pure vorm van HARKing en de vorm waarbij verworpen hypotheses niet worden vermeld, niet door de beugel kunnen bij goede wetenschap.

Er zijn een aantal redenen die wetenschappers kunnen hebben om een vorm van HARKing toe te passen. In de eerste plaats bespaart HARKing enorm veel geld. Als na het zien van de resultaten een nieuwe hypothese wordt bedacht dan is het veel goedkoper om te doen alsof deze hypothese al was voorgesteld voordat het onderzoek is uitgevoerd, dan dat een compleet nieuw onderzoek wordt ontworpen waarin deze nieuwe hypothese wordt getoetst. Ten tweede verwachten wetenschappers dat er een kleinere kans bestaat om een artikel gepubliceerd te krijgen wanneer eerlijk wordt beschreven dat een bevestigde hypothese post hoc is toegevoegd. Het als a priori presenteren van deze post hoc hypothese zou de kans op publicatie dus vergroten (Kerr, 1998). Verder is onderzocht dat bevestiging van een hypothese door zowel wetenschappers als nonwetenschappers als informatiever wordt ervaren dan verwerping van een hypothese (Mahoney, 1976; aangehaald in Kerr ,1998). Uit hetzelfde onderzoek bleek dat wetenschappers aangeven dat ze vaak geen energie willen steken in het uitwerken van een artikel over een onderzoek met negatieve resultaten, omdat ze vermoeden dat dergelijke artikelen een kleinere kans hebben om gepubliceerd te worden dan artikelen waar een bepaalde theorie juist wordt bevestigd in plaats van verworpen.

Popper over HARKing

De hierboven beschreven manier van het uitvoeren van wetenschap staat in sterk contrast met Poppers visie op goede wetenschap. Popper beweerde namelijk dat disconfirmatie van een theorie informatiever is dan confirmatie. Een bevestiging heeft namelijk slechts tot gevolg dat de theorie gecorroboreerd wordt en dus maar tijdelijk wordt gesteund. Er zal immers altijd een kans zijn dat de theorie in de toekomst nog zal worden verworpen. Disconfirmatie van een theorie is volgens Popper informatiever omdat het tot gevolg kan hebben dat de theorie in zijn geheel wordt verworpen en op die manier wordt dus het aantal mogelijke verklaringen voor een bepaald fenomeen verminderd (Popper, 1959).

Vanuit Poppers filosofie kan kritiek worden geleverd op HARKing. Een hypothese moet volgens Popper falsifieerbaar zijn. Als een hypothese niet kan worden verworpen dan kan deze nooit als een goede wetenschappelijke verklaring gelden. Wetenschappers die gebruik maken van HARKing kunnen voorkomen dat een hypothese wordt gefalsificeerd. Ze kunnen na het zien van de resultaten de hypothese aanpassen zodat deze consistent is met de data, of zelfs met een compleet nieuwe passende hypothese komen die voor het uitvoeren van het onderzoek nog niet was bedacht. Daarnaast zal er door het verwijderen van een verworpen hypothese uit de onderzoeksrapportage nooit inzicht ontstaan in welke theorieën eigenlijk verworpen zouden moeten worden. Er worden immers alleen bevestigde hypotheses gerapporteerd.

P-waarde

Binnen veel wetenschapsgebieden worden hypotheses getoetst met behulp van statistische analyses. De meest dominante vorm van statistiek is de frequentische, waarbij een nulhypothese wordt getoetst tegen een significantieniveau van bijvoorbeeld 5%. Een nulhypothese is een hypothese die een basissituatie of uitgangspositie weergeeft waarin er geen sprake is van een bepaald effect. Een significantieniveau van 5% houdt in dat wanneer de nulhypothese wordt verworpen, er nog steeds een kans van 5% is dat deze eigenlijk juist is. Een kleiner significantieniveau geeft meer statistische zekerheid in het verwerpen van een nulhypothese. Het significantieniveau wordt idealiter voorafgaand aan het onderzoek door de wetenschapper bepaald. De uitkomst van de meeste frequentische statistische analyses is een p-waarde. Een p-waarde kan worden gedefinieerd als de kans op de geobserveerde data, gegeven dat de nulhypothese waar is (Gigerenzer, 2004; Wetzels et al., 2011). Als de uit een analyse gekomen p-waarde kleiner is dan het vooraf gestelde significantieniveau, dan wordt de nulhypothese verworpen en de eventuele alternatieve hypothese geaccepteerd.

Dit laatste punt is vreemd, omdat het niet zo is dat er bewijs voor de alternatieve hypothese is gevonden. Het is alleen zo dat de nulhypothese wordt verworpen en dat zegt niets over de plausibiliteit van de alternatieve hypothese (Rouder & Morey, 2011). Een tweede nadeel van het gebruik van de p-waarde is dat niet duidelijk wordt hoe groot een effect is. Een klein effect in een onderzoek met een grote steekproef kan dezelfde p-waarde hebben als een onderzoek met een groot effect en een kleine steekproef. De p-waarde op zich is dus vaak niet geschikt om conclusies te trekken over het gevonden effect, terwijl in veel onderzoeken de nadruk juist ligt op rapportage van de p-waarde, en niet de effectgrootte (Greenwald, Gonzalez, Harris & Guthrie, 2007).

Een derde probleem van de p-waarde is dat deze het bewijs voor veel analyses groter uitdrukt dan het werkelijk is (Goodman, 1999). Goodman vergeleek de p-waarde met de Minimum Bayes factor – een Bayesiaans alternatief voor hypothese testen – zodat inzicht verkregen kan worden in de verschillen tussen deze twee

statistieken. De Minimum Bayes factor geeft de kleinste hoeveelheid bewijs weer dat op basis van de data gevonden als steun voor de nulhypothese gegeven kan worden. Het bleek dat voor een p-waarde van 0.05 de Minimum Bayes factor 0.15 is, wat betekent dat de nulhypothese een relatieve steun krijgt van 0.15 ten opzichte van de beste alternatieve hypothese. Dus de p-waarde 0.05 die in veel onderzoeken wordt gebruikt als de grens van een significant resultaat, is volgens Bayesiaanse statistiek eigenlijk maar een middelmatig resultaat. Goodman concludeert dat de sterkte van het bewijs tegen de nulhypothese bij lange na niet zo sterk is als de waarde van een bepaalde p-waarde doet vermoeden.

Ten slotte is een vierde nadeel van de p-waarde dat deze vrij makkelijk verkeerd kan worden geïnterpreteerd. Uit een onderzoek met statistiek-docenten bleek dat 80% van de docenten uit zes interpretaties van de p-waarde aangaf dat minimaal 1 van deze interpretaties juist was, terwijl er niet één juist was (Gigerenzer, 2004). Het bleek dat zelfs statistiekdocenten moeite hebben met de interpretatie van p = 0.01. Ze kregen zes mogelijke interpretaties van de p-waarde te zien en het bleek dat 80% van hen aangaf dat minimaal één van deze interpretaties juist was, terwijl er niet één juist was. Een voorbeeld van een verkeerde interpretatie van p = 0.01 is: de kans dat de nulhypothese waar is, is 1%. Volgens Gigerenzer kan er op basis van een significantietoets niets worden gezegd over de kans dat een bepaalde hypothese optreedt. Hij concludeert dat de p-waarde vaak informatiever wordt ingeschat dan deze werkelijk is. Een correcte interpretatie van p = 0.01 zou zijn dat de kans op het verkrijgen van een test statistiek (zoals de t waarde) die minstens zo extreem is als de test statistiek die geobserveerd wordt in het experiment 1% is, gegeven dat de nulhypothese waar is en de steekproef getrokken is volgens een vooraf vastgestelde procedure, zoals een vaste steekproefgrootte. Een simplere maar ook correcte interpretatie zou zijn: de kans op de geobserveerde data, gegeven dat de nulhypothese waar is, is 1%. Deze interpretaties gaan over een voorwaardelijke of conditionele kans. Het is dus niet zo dat je met behulp van een p-waarde conclusies kunt trekken over de kans dat een bepaalde hypothese waar is.

Filosofen over de p-waarde

Gegeven de hierboven besproken nadelen kan de vraag worden gesteld hoe objectief wetenschappelijk onderzoek eigenlijk nog is. Wetenschappers worden grotendeels beoordeeld op basis van hun publicaties. Tegelijkertijd denken ze dat de kans groter is om te publiceren als een hypothese is bevestigd, dan wanneer een hypothese wordt gefalisificeerd (Kerr, 1998). Dit stimuleert het gebruik van HARKing in plaats van de HD-methode. Verder geeft de p-waarde die gebruikt wordt om te onderbouwen of de hypothese wordt bevestigd een te optimistische schatting. Dit impliceert dat er soms hypotheses geaccepteerd worden, terwijl daar helemaal geen goede reden voor is.

Bovenstaande beschrijving van het wetenschappelijk proces sluit goed aan bij de wetenschapsfilosofie van Bourdieu. Volgens Bourdieu kan de wetenschap worden gezien als een sociaal veld dat net als andere sociale velden ook haar machtsverhoudingen en conflicten kent. Een veld definieert Bourdieu als een autonoom netwerk van relaties die dwang uitoefenen op de actoren die bepaalde posities binnen het veld bekleden (Bourdieu, 1989). Het gegeven dat confirmatieve wetenschappelijke studies meer kans op publicatie hebben dan disconfirmatieve studies oefent dwang uit op de manier waarop wetenschappers te werk gaan. Bovendien is het kapitaal waar de competitie binnen het wetenschappelijke veld om plaats vindt het verkrijgen van een monoplie op de waarheid. Gedrag dat de kans op publicaties verhoogt zou volgens Bourdieu ook nog eens kunnen worden gevoed door onderlinge strijd om een monopolie op de waarheid. Een oplossing voor deze problematiek zou volgens Bourdieu zijn dat wetenschappers meer reflexiviteit tonen. Ze moeten rekening houden met de machtsstructuren die van invloed zijn binnen het vakgebied.

Een andere wetenschapsfilosoof die van mening is dat de wetenschap geen speciale status heeft is Rorty. Hij staat positief tegenover de wetenschap, maar vindt dat we ons niet blind moeten richten op objectiviteit. Volgens Rorty bestaat er een vooringenomenheid over de wetenschap. Ze wordt gezien als een instrument dat objectieve en onbetwijfelbare kennis over de realiteit levert en subjectiviteit daarmee aan de kant zet (Rorty, 1979). De hierboven beschreven nadelen van het gebruik van een p-waarde steunen Rorty in zijn stelling dat wetenschap niet op een magische manier tot universele waarheden komt. Wat voor waar wordt gehouden binnen veel takken van wetenschap is afhankelijk van afspraken die zijn gemaakt tussen wetenschappers met betrekking tot statistische analyses en methoden van onderzoek.

Een mogelijke oplossing voor de problemen met de p-waarde is om niet meer vast te blijven houden aan de kern van dit onderzoeksprogramma (in Lakatosiaanse zin) en in te zien dat de Bayesiaanse statistiek een nieuw onderzoeksprogramma is waarbij de genoemde problemen niet optreden (Wagenmakers et al., 2011). Een oplossing om HARKing tegen te gaan is om wetenschappers voordat het onderzoek wordt uitgevoerd de hypotheses, onderzoeksmethoden, en het te gebruiken significantieniveau te laten rapporteren in een online database, zodat er meer transparantie ontstaat en op die manier meer inzicht in hoe wetenschappers te werk zijn gegaan tijdens het onderzoek.

Referenties

Bourdieu, P. (1989). *Opstellen over smaak, habitus en het veldbegrip*. Amsterdam: Van Gennep.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Goodman, S. N. (1999). Toward evidence-based medical statistics, 2: the bayes factor. *Annuals of Internal Medicine*, *130*, 1005–1013.

Greenwald, A. G., Gonzalez, R. G., Harris, R. J. & Guthrie, D. (2007). Effect sizes and p-values: What should be reported and what should be replicated? *Psychophysiology*, *33*, 175–183.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

Popper, K. (1959). *The logic of scientific discovery*. London: Routledge & Kegan Paul.

Rorty, R. (1979). *Objectivity, relativism, and truth*. Cambridge: Campridge University Press.

Rouder, J. N. & Morey, R. D. (2011). An assessment of the evidence for feeling the future with a discussion of bayes factor and significance testing. *Manuscript submitted for publication*.

Suppe, F. (1977). *The structure of scientific theories*. Urbana: University of Illinois Press.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J. & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298.