Hoe betrouwbaar is 95 procent betrouwbaar?

Hoekstra, Rink

Published in:
Skepter

Hoe betrouwbaar is 95 procent betrouwbaar?

Als een uitkomst 'significant' is, betekent dat niet dat de kans op toeval klein is, en een '95%-betrouwbaarheidsinterval' geeft niet aan dat een uitkomst met 95 procent zekerheid binnen dat interval ligt. Over deze sleutelbegrippen in de statistiek heersen vooral misverstanden en onbegrip — hoe significant en betrouwbaar kan de wetenschap dan zijn?

DOOR RINK HOEKSTRA

In de sociale wetenschappen is het gebruikelijk om op basis van uitkomsten van een relatief kleine groep uitspraken te doen over een veel grotere groep, de groep waarover je eigenlijk iets zou willen weten. Voor dat generaliseren is er de statistiek.

De meest gebruikte statistische techniek is de significantietoets, zeker binnen de sociale wetenschappen. Zelfs iemand die nooit een cursus statistiek heeft gevolgd komt het regelmatig tegen: in krantenartikelen wordt geschreven dat een resultaat 'significant' was, en de journalist gaat er kennelijk vanuit dat de lezer weet wat dit betekent.

Dat is natuurlijk de vraag. Het is zelfs de vraag of de journalist zelf weet wat 'significantie' precies inhoudt, en het is, eerlijk gezegd, zelfs de vraag of de onderzoeker het begrip helemaal begrijpt. Zeker iemand die het Engels machtig is, zou kunnen denken dat het een wat ingewikkelde manier is om aan te geven dat het gevonden effect 'veelbetekenend' of 'belangrijk' is. Immers, dat is de vertaling van het Engelse woord *significant*. Helaas zal zelfs de trouwste supporter van de significantietoets moeten toegeven dat dit niet klopt: een inhoudelijk triviaal resultaat kan statistisch significant zijn, en een inhoudelijk groot effect kan in sommige gevallen niet significant zijn. Dit maakt significantie een lastig te hanteren begrip. Wat betekent het dan wel? En als er blijkbaar voorstanders van het gebruik van significantietoetsing zijn: zijn er dan ook tegenstanders van deze toch bijna universeel toegepaste techniek?

De significantietoets

Ik zal proberen significantietoetsing uit te leggen aan de hand van een voorbeeld. Stel dat je onderzoek wilt doen naar de vraag of mensen meer 'out of the box' gaan denken als ze naast een grote kartonnen doos zitten of erin (ik verzin dit niet: het resultaat werd begin 2012 door Angela Leung gepubliceerd in het vakblad *Psychological Science*). De significantietoets houdt in dat je als onderzoeker bewijs probeert te verzamelen tegen het idee dat er *geen* effect is (dus dat de doos geen effect heeft op creativiteit). Deze

voorlopige aanname van geen effect heet de 'nulhypothese'. Vervolgens kun je berekenen hoe groot de kans is dat, als die nulhypothese waar zou zijn, je een uitkomst vindt die minstens zo extreem is als de uitkomst die in het onderzoek gevonden is. Als je bijvoorbeeld ziet dat personen die naast de doos zaten bij een taakje gemiddeld 1,65 punten hoger scoren dan degenen die in de doos zaten, kun je uitrekenen hoe bijzonder deze uitkomst is als de waarde onder alle mensen (ook degenen die je niet hebt onderzocht) gemiddeld precies 0 punten zou zijn. Als die kans kleiner is dan een vooraf gestelde grenswaarde (meestal 5 procent, wiskundig genoteerd als $p < 0{,}05$), heet het gevonden effect 'significant'.

Vaak wordt vervolgens geconcludeerd dat, bij een significant effect, de nulhypothese wel niet waar zal zijn. De achterliggende redenering klinkt logisch: als de nulhypothese waar zou zijn en de doos *geen* invloed heeft op creativiteit, is de kans klein om zo'n sterke toename van creativiteit (of een nog sterkere) te vinden. Toch is er die toename gevonden, en dus zal de nulhypothese wel niet waar zijn. Leung en haar collega's concludeerden dan ook dat het in de praktijk brengen van 'metaforen voor creativiteit' leiden tot meer creativiteit.

Bijna blind

Toch is er, zeker ook binnen de statistiek, veel kritiek op dit soort gevolgtrekkingen, en op de significantietoets in het algemeen. Een eerste punt, waarop Eric-Jan Wagenmakers verderop in dit nummer ingaat, is dat eigenlijk geen rekening wordt gehouden met alle mogelijke alternatieve hypotheses en verklaringen. Maar er is meer.

Een fundamenteel kritiekpunt is dat de hele redenering in feite onlogisch is. Die redenering was:

- Als de nulhypothese waar is, is het onwaarschijnlijk dat we dit effect vinden.
- Wij hebben zo'n effect gevonden.
- Dus zal de nulhypothese wel onwaarschijnlijk zijn.

Dit lijkt misschien nog niet eens zo vreemd. Maar kijk eens naar de volgende redenering:

- Als we een vrouw zien, is het onwaarschijnlijk dat die zwanger is.
- Wij zien iemand die zichtbaar zwanger is.
- Dus zal de aanname dat die persoon een vrouw is wel onwaarschijnlijk zijn.

De voorbeelden zijn qua inhoud verschillend, maar de gedachtegang is gelijk. En die gedachtegang is, helaas, volstrekt niet logisch, wat voor een zo vaak gebruikte techniek toch wel problematisch is.

Andere kritiek op de significantietoetsing richt zich op de manier waarop in de wetenschap met de techniek wordt omgegaan. Omdat tijdschriften graag significante resultaten publiceren, zijn onderzoekers erop gebrand significante uitkomsten te krijgen. Als een uitkomst na de eerste analyse van de data niet significant blijkt kan een onderzoeker besluiten om dan nog maar een paar extra proefpersonen te onderzoeken, of een bepaalde groep uit het onderzoek te laten, of nog wat meer effecten te bekijken, in de hoop dat er alsnog een resultaat uitkomt dat wel significant is. Hoewel dit misschien niet eens heel onredelijk klinkt, maakt dit de interpretatie van de uitkomsten van de significantietoets zo goed als onmogelijk.

Daarnaast blijkt dat de bijna dwangmatige focus op significantie

De hele redenering achter de significantietoets is in feite onlogisch

De focus op significantie maakt onderzoekers bijna blind voor de vraag hoe groot een gevonden effect eigenlijk is

onderzoekers bijna blind maakt voor de vraag hoe groot een gevonden effect nu eigenlijk is. Als je bijvoorbeeld het effect van een nieuwe lesmethode op de schoolprestaties wilt onderzoeken aan de hand van CITO-scores, is het wel degelijk van belang of die uiteindelijk tot een gemiddelde stijging van 2 of van 20 punten heeft geleid — ook al geven beide lesmethodes significante resultaten.

Betrouwbaarheidsinterval

Een veelgenoemd alternatief voor de significantietest is het 'betrouwbaarheidsinterval'. Dit is een marge rond het gevonden effect, met als eerste voordeel dat je als lezer direct zicht krijgt op de grootte van het effect. De laatste jaren wordt deze techniek binnen de sociale wetenschappen behoorlijk gepropageerd, zeker ook door critici van significantietoetsing. Dus wat is nu dit betrouwbaarheidsinterval? En vooral: zijn onderzoekers, journalisten en krantenlezers beter in staat dit interval te interpreteren dan die vermaledijde significantietoets?

De onderzoekers naar het effect van buiten-de-doos-denken vonden een verschil van 1,65 punten, en hadden dus een interval rond die 1,65 punten kunnen geven (bijvoorbeeld een interval dat loopt van 1,33 tot 1,97). De redenering achter betrouwbaarheidsintervallen is: als je het onderzoek heel vaak herhaalt, en telkens zo'n betrouwbaarheidsinterval rond de gevonden waarde construeert, dan zal een bepaald percentage van die intervallen de werkelijke waarde — dus de waarde waarnaar je op zoek bent — bevatten. Gewoonlijk is dat percentage 95: 95 van de 100 95%-betrouwbaarheidsintervallen zullen de gezochte, maar onbekende waarde bevatten. Dus als je heel veel 95%-betrouwbaarheidsintervallen rond het effect van homeopathie hebt, zal in 95 procent van die intervallen de waarde 0 liggen. Als het gemiddelde verschil in lengte tussen mannen en vrouwen 10 centimeter is, zal die 10 centimeter in 95 procent van de 95%-betrouwbaarheidsintervallen van de steekproeven liggen.

Zelfde traditie

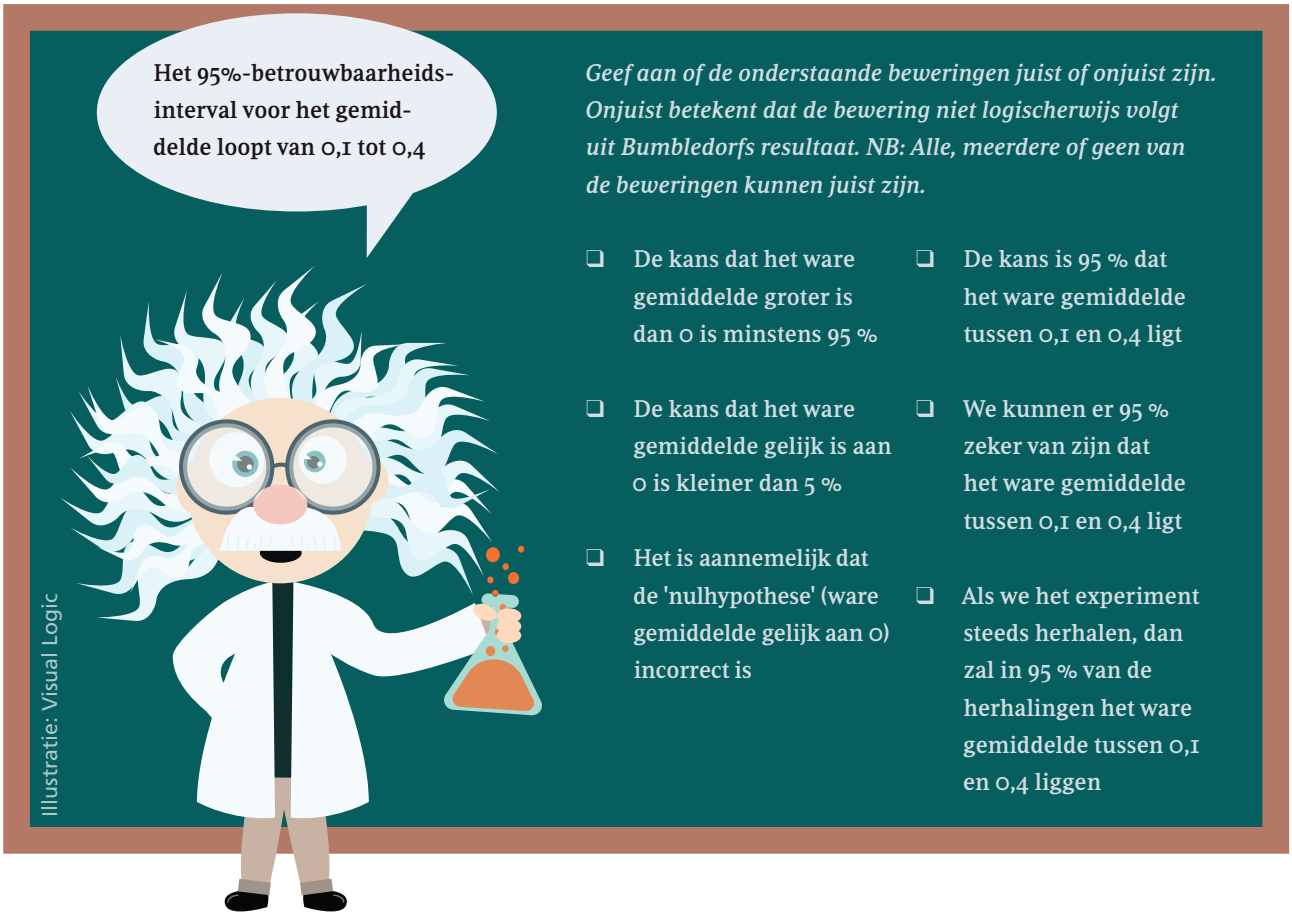
Betrouwbaarheidsintervallen komen voort uit dezelfde statistische traditie als significantietoetsen, namelijk dat het begrip 'kans' is gebaseerd op het heel vaak herhalen van een bepaald experiment. We weten pas echt wat de kans op een 5 bij een willekeurige dobbelsteen is, als we er heel vaak mee hebben gegooid.

Dit is meteen het lastige aan een betrouwbaarheidsinterval: in de praktijk doen wetenschappers één onderzoek, en beschrijven ze de resultaten voor alleen dat onderzoek. Dit levert voor die uitkomst dus ook één betrouwbaarheidsinterval op. Maar hoe interpreteren we vervolgens zo'n enkel interval? Kunnen we een uitspraak doen over het interval rond de waarde die wij gevonden hebben?

Een op het eerste gezicht logische gedachte is, dat als 95 procent van de intervallen de populatiewaarde bevat, de kans bij ieder interval afzonderlijk ook 95 procent is. Maar helaas zit het toch iets ingewikkelder in elkaar. Het probleem zit hem in het feit dat die '95 procent' niet hoort bij het interval, maar bij het hele proces — theoretisch moet er elk (bij het experiment horend) willekeurig interval uit kunnen komen. Maar als je zo'n interval krijgt voorgeschoteld, kijk je naar dat ene interval en kun je, bijvoorbeeld op grond van voorkennis, een inschatting maken over de waarschijnlijkheid van dat interval. Hierdoor kun je strikt genomen niet meer van 95 procent zekerheid spreken, omdat die zekerheid direct al is beïnvloed door het interval

Het experiment van professor Bumbledorf

Professor Bumbledorf voert een experiment uit, analyseert de data en rapporteert:



Geef aan of de onderstaande beweringen juist of onjuist zijn. Onjuist betekent dat de bewering niet logischerwijs volgt uit Bumbledorfs resultaat. NB: Alle, meerdere of geen van de beweringen kunnen juist zijn.

- ❑ De kans dat het ware gemiddelde groter is dan 0 is minstens 95 %
- ❑ De kans dat het ware gemiddelde gelijk is aan 0 is kleiner dan 5 %
- ❑ Het is aannemelijk dat de 'nulhypothese' (ware gemiddelde gelijk aan 0) incorrect is
- ❑ De kans is 95 % dat het ware gemiddelde tussen 0,1 en 0,4 ligt
- ❑ We kunnen er 95 % zeker van zijn dat het ware gemiddelde tussen 0,1 en 0,4 ligt
- ❑ Als we het experiment steeds herhalen, dan zal in 95 % van de herhalingen het ware gemiddelde tussen 0,1 en 0,4 liggen

zelf. Ik zal proberen dit te illustreren met het voorbeeld over homeopathie.

Effect homeopathie

Stel dat een onderzoeker een effect voor homeopathie vindt van 0,25 (wat dat 'effect' ook moge zijn), en dat het bijbehorende 95%-betrouwbaarheids-interval loopt van 0,1 tot 0,4. Is er daarmee 95 procent zekerheid dat er een effect van homeopathie is dat ligt tussen de 0,1 en 0,4? Je zou mogen hopen dat een beetje kritisch persoon zich niet zo eenvoudig laat overtuigen. Maar wat is er nu fout aan deze op zich logisch lijkende redenering?

De crux zit hem zoals gezegd in het feit dat het interval zelf al extra informatie geeft. Het is immers een interval dat de *afwezigheid* van een effect van homeopathie — de waarde 0 — niet bevat, wat toch behoorlijk verbazingwekkend is. Hierdoor lijkt er in dit geval toch wel een grote kans dat dit net een van die 5 procent van de intervallen is die de juiste waarde niet bevatten. Als immers ieder interval een gelijke kans zou hebben om de juiste waarde te omsluiten, zou je moeten concluderen dat er 95 procent kans is dat homeopathie werkt.

Ander voorbeeld. Er is een app die met een cirkel op de kaart aangeeft waar je vrienden zijn. Uiteraard is de app niet feilloos, maar de maker beweert dat in 95 procent van de gevallen de gezochte persoon zich werkelijk binnen de cirkel bevindt. Stel nu dat je op een kille herfstavond je app opent, en je ziet dat het cirkeltje voor een vriend zich volledig boven een vijver bevindt. Natuurlijk kan het zijn dat hij toevallig even aan het zwemmen is of door andere oorzaak in de vijver beland is, maar erg waarschijnlijk is dat toch niet. Met andere woorden: niet ieder interval heeft evenveel kans de juiste waarde te omsluiten.

Om een lang en ingewikkeld verhaal samen te vatten: de eis voor het construeren van een 95%-betrouwbaarheidsinterval is dat je *vooraf* weet dat 95% van de op deze manier geconstrueerde intervallen de ware waarde bevat. Maar bij een gegeven interval *achteraf* is het niet te zeggen hoe groot die kans is, omdat het interval zelf vaak al informatie weggeeft, informatie die niet per se is meegenomen in het interval. Natuurlijk weet je niet altijd, zoals bij het homeopathie- of app-voorbeeld hoe onwaarschijnlijk het betreffende interval is — soms heb

je amper een idee. En mensen kunnen erover van mening verschillen. Maar helemaal blanco is bijna niemand: je weet bijvoorbeeld dat gigantische effecten hoe dan ook buitengewoon zeldzaam zijn.

Dus hoe moet een betrouwbaarheidsinterval nu wel geïnterpreteerd worden? Eigenlijk zou je iets moeten zeggen als: 'Dit interval is gebaseerd op een proces dat in 95 procent van de gevallen de gezochte waarde bevat. Hopelijk is dat nu ook zo, maar dat weten we niet. Wat denkt u zelf?' Erg leesbaar en handzaam is het inderdaad niet.

Geen benul

Toch zou je mogen verwachten dat onderzoekers, die in recente statistiekboeken en in richtlijnen van tijdschriften en overkoepelende organisaties worden aangemoedigd om toch vooral betrouwbaarheidsintervallen rond hun uitkomsten te geven, van deze beperkingen op de hoogte zijn. Helaas blijkt dit niet het geval.

Een paar jaar geleden legden we aan zowel onderzoekers als studenten een zestal stellingen voor over de correcte interpretatie van een betrouwbaarheidsinterval. De deelnemers moesten aangeven of de stellingen logisch volgden uit de gegeven informatie — de stellingen staan bij dit artikel, dus wie de eigen vaardigheden wil testen, moet nu eerst de quiz doen en dan pas verder lezen.

Wij vonden de uitkomsten schokkend: onderzoekers presteerden niet of nauwelijks beter dan eerstejaars, en gaven gemiddeld aan dat ruim drie (in plaats van nul) stellingen logisch volgden uit de gegeven informatie. Ook bleken er bij de deelnemende onderzoekers nauwelijks patronen in de antwoorden te zitten, wat volgens ons aangeeft dat er niet zo zeer sprake lijkt van bepaalde wijdverbreide misverstanden, maar eerder van een totaal onbegrip over hoe je zo'n interval moet interpreteren.

Hoewel sommige statistici onze conclusies te streng vinden, lijkt het duidelijk dat er onder onderzoekers geen duidelijke consensus is over de interpretatie, laat staan dat deze overeenkomt met de formeel logische interpretatie.

Statistiekboeken

Als lezer zou je nu kunnen denken dat onderzoekers beter moeten opletten in statistieklessen en hun huiswerk beter moeten doen. Dit is wel waar, maar niet de enige oplossing van het probleem. Uit onderzoek dat we op dit moment afronden, blijkt namelijk dat het in leerboeken statistiek wemelt van interpretaties van betrouwbaarheidsintervallen die je op zijn best als kort door de bocht, maar eigenlijk als fout zou moeten bestempelen. Sommige auteurs geven hun worsteling zelfs expliciet aan — 'eigenlijk klopt dit niet helemaal, maar doe het toch maar zo...'

Natuurlijk mogen we van onderzoekers verwachten dat zij goed op de hoogte zijn van de technieken waarvan vaak gezegd wordt dat ze die zouden moeten gebruiken, maar als het in leerboeken al niet goed wordt uitgelegd is het niet zo verrassend dat dit niet helemaal goed gaat.

Hoe nu verder?

Als zowel significantietoetsen als betrouwbaarheidsintervallen zo'n mijnenveld zijn, hoe moet het dan verder?

Dat hangt een beetje af van wat je verwacht van de wetenschap. Wil je een wetenschap waarin de technieken die je gebruikt op de lange termijn vaak het juiste antwoord geven, maar waarbij je bij een individueel artikel niet in staat bent aan te geven hoe groot de kans in dit specifieke geval is? Of wil je expliciet kunnen uitdrukken per onderzoek wat je op basis van deze data hebt geleerd?

In het eerste geval kun je de huidige technieken wel blijven gebruiken, met de kanttekening dat je veel voorzichtiger zult moeten zijn bij de interpretatie dan in de meeste artikelen gebeurt.

In het tweede geval zou je gebruik kunnen maken van een andere filosofische stroming binnen de statistiek, die bayesiaans genoemd wordt, en hierop gaat Eric-Jan Wagenmakers in het volgende stuk wat uitgebreider in. Deze stroming is gericht op het expliciet kwantificeren van hoe een rationeel persoon zijn of haar mening zou moeten bijstellen op basis van de gevonden data. Het ingewikkelde daarbij is wel dat je je mening vooraf in getallen zult moeten omzetten, wat nog niet zo eenvoudig is.

Het is voor onderzoekers, die niet altijd dankzij een voorliefde voor statistiek in de wetenschap zijn geraakt, niet eenvoudig statistiek op een verantwoorde manier te gebruiken en te interpreteren. Je moet kennis hebben van vele technieken, en misschien zelfs van verschillende filosofische stromingen binnen de statistiek. De verleiding om een bochtje af te snijden is dan erg groot, zeker als je je kunt verschuilen achter het feit dat ongeveer al je collega's voor je dit ook al deden. En toch zou je juist van een wetenschapper mogen verwachten dat die niet alleen kritisch op anderen, maar juist ook op zichzelf is. Ook wat betreft het gebruik van statistische methoden.

Rink Hoekstra is universitair docent aan de Rijksuniversiteit Groningen, afdeling GION onderwijs/onderzoek.