ESSAY

Effecten op schaal en betrouwbaarheidsintervallen als alternatief voor p < 0,05

A.I. WIERDSMA, J. VAN OS

ACHTERGROND *Door onderzoekers en referenten wordt bij statistische toetsen vaak de conventie p < 0,05 aangehouden. De interpretatie van p-waarden is echter in veel gevallen onjuist.*
DOEL *Beschrijven waar de 5%-norm vandaan komt en welke interpretatieproblemen voorkomen en aangeven van alternatieven.*
METHODE *Op basis van literatuur schetsen wij de betekenis en de herkomst van de p < 0,05-norm. In oorspronkelijke artikelen en korte bijdragen in het Tijdschrift voor Psychiatrie vanaf het jubileumjaar 2008 werden voorbeelden van methodologische problemen gezocht die zijn gerelateerd aan het routinematige gebruik van p-waarden.*
RESULTATEN *Diverse voorbeelden werden gevonden van het toetsen van a priori onwaarschijnlijke of zelfs onmogelijke hypothesen, het rapporteren van kleine effecten, mogelijk verkeerde berekeningen, en verkeerde interpretaties van statistische parameters en p-waarden.*
CONCLUSIE *Net als in andere disciplines wordt in psychiatrisch onderzoek te veel nadruk gelegd op p-waarden. In de aanwijzingen voor auteurs dient expliciet aandacht te worden gevraagd voor effect size, betrouwbaarheidsintervallen en de schaal waarop de uitkomsten zijn gemeten.*
[TIJDSCHRIFT VOOR PSYCHIATRIE 55(2013)7, 471-480]

TREFWOORDEN alfa, hypothesen, power, significantie, type I-fout

Vaak volgen onderzoekers het wetenschappelijke ritueel van statistische toetsing van een hypothese met alfa = 5% (p < 0,05). De interpretatie van p-waarden is echter in veel gevallen onjuist. Daarom zijn er al decennia lang protesten tegen significantietesten en de p < 0,05-regel: in de sociologie (Morrison & Henkel 1970), de psychologie (Cohen 1994) en de economie (Ziliak & McCloskey 2008), maar ook in de epidemiologie (Rothman 2002) en de medische wereld (Goodman 1999). De wetenschapsjournalistiek heeft de kritiek op significantietesten bij een breder publiek onder de aandacht gebracht (Matthews 1998; Senn 2010). Ook worden er proefschriften aan gewijd (Hoekstra 2009), maar toch wordt in de regel de p < 0,05-norm aangehouden.

Over p-waarden bestaan enkele hardnekkige misverstanden. Een p-waarde wordt ten onrechte vaak opgevat als maat voor de omvang van het gevonden verschil of voor de praktische relevantie. In een grote studie worden echter alle verschillen significant en ook kleine, niet statistisch significante verschillen kunnen relevant zijn. Bij p < 0,05 is er ook niet een 95%-kans dat een replicatiestudie eenzelfde uitkomst oplevert. Die kans is namelijk berekend op grond van de aanname dat de hypothese correct is en bij een kleine p-waarde wordt juist die hypothese verworpen. Daarom is het ook onjuist om niet-significante uitkomsten te interpreteren als bewijs voor de hypothese dat er geen verschil of geen effect is.

Een aantal tijdschriften benadrukt alternatieven voor p-waarden – in Nederland bijvoorbeeld het *Tijdschrift voor Gezondheidswetenschappen*, maar het *Tijdschrift voor Psychiatrie* doet dat nog niet. Van Os (2000) en Van Harten (2008) wezen eerder in dit tijdschrift op het belang van het *number needed to treat* dat als maat voor de effectiviteit van een interventie concreter is dan *effect size* (bijvoorbeeld een maat voor de sterkte van de samenhang of verschil in gemiddelden). Maar tot een richtlijn voor auteurs kwam het niet. Dergelijke aanwijzingen worden alsmaar relevanter. Tegenwoordig zijn de meest complexe analysetechnieken voor iedereen beschikbaar en volgens Goodman (1999) verplicht dat onderzoekers en referenten tot meer statistische discipline.

In deze bijdrage gaan we daarop in. Eerst komen de verschillen tussen significantie en hypothesetoetsen aan de orde en de vraag waar de 5%-norm vandaan komt. Vervolgens geven we enkele voorbeelden uit de praktijk hoe het met p-waarden fout kan gaan. Die voorbeelden haalden we uit oorspronkelijke artikelen en korte bijdragen in het *Tijdschrift voor Psychiatrie*, overwegend vanaf het jubileumjaar 2008. De aandacht was vooral gericht op de variatie in de problemen rond p-waarden, zodat het aantal referenties in deze bijdrage slechts een ondergrens aangeeft (ruim 20% van in totaal 69 beoordeelde rapportages van kwantitatief onderzoek). Tot slot dagen we de redactie uit als alternatief voor $p < 0{,}05$ in de aanwijzingen voor auteurs ook meer aandacht te vragen voor effectmaten, betrouwbaarheidsintervallen, grafische presentaties, en de schaal waarop de uitkomstmaten zijn gemeten.

SIGNIFICANTIE- EN HYPOTHESETOETSEN

In het begin van de vorige eeuw introduceerde Sir Ronald Aylmer Fisher (1890-1962; figuur 1) een reeks statistische begrippen die nu in wetenschappelijk onderzoek gemeengoed zijn: randomisatie, variantie en variantieanalyse, interactie, nulhypothese en significantietest. David (1995) stelde een lijst samen met de namen van statistici die bekende termen hebben gemunt. Wij telden dat Fisher 53 keer voorkomt, twee keer meer dan Pearson, maar dan gaat het om vader Karl (van Pearsons correlatiecoëfficiënt) en zoon Egon (van de *likelihood ratio test*) samen. Fisher wordt dan ook algemeen gezien als de vader van de moderne statistiek (Hald 2004; Lehmenn 2011). In Fishers significantietest zijn de data een selectie uit een theoretisch oneindige populatie met een bepaalde kansverdeling. P is de kans (Pr) op de testwaarde (Data) of een extremere waarde wanneer de nulhypothese (H_0) waar is – in korte notatie: $Pr(Data|H_0)$. Voor Fisher was een p-waarde een index die de mate van vertrouwen in de hypothese uitdrukt. *'If P … is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05 …'* (Fisher 1990/1925).

FIGUUR 1 *Portret van Sir Ronald Aylmer Fisher (foto uit 1941, Special Collections Research Center at NCSU Libraries, itemnr. 0006973 Rare and Unique Materials, http://d.lib.ncsu.edu/collections/catalog/0006973)*

De gedachte dat we een hypothese alleen verwerpen omdat er een andere hypothese is die de gevonden testwaarde aannemelijker maakt, lag aan de basis van de Engelse en Russisch/Poolse samen-

werking van Egon Sharpe Pearson (1895-1960) en Jerzy Neyman (1894-1981). Aan de tandem Neyman-Pearson danken we onder andere de alternatieve hypothese en de alfafout en het onderscheidend vermogen van een test (*power*). De combinatie van de nulhypothese en de alternatieve hypothese betekent dat er ook twee soorten fouten zijn. Figuur 2 vat de verschillende mogelijkheden samen. Bij een type I-fout verwerpen we de nulhypothese, maar is deze correct; bij een type II-fout wordt de alternatieve hypothese verworpen, maar is deze juist. In plaats van de p-waarde van één test kijken we of de testwaarde in het kritische gebied ligt (alfa = 5%) en alleen door herhaalde testen kunnen op de lange duur type I- en II-fouten worden ingeperkt. '*These two sources of error can rarely be eliminated completely; in some cases it will be more important to avoid the first, in others the second ... The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator*' (Neyman & Pearson 1928).

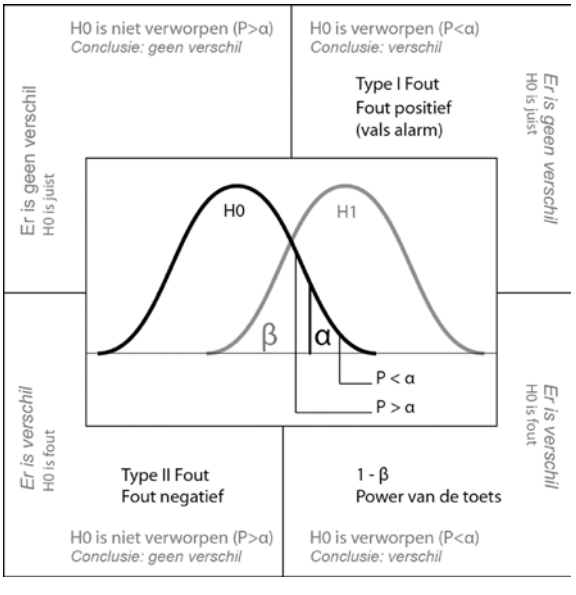
Vaak wordt in onderzoek wel de terminologie van Neyman-Pearson van nulhypothese en alternatieve hypothese en alfafout gebruikt, maar volstaat men met het rapporteren van Fishers p-waarden (Goodman 1999). De power van een test komt meestal alleen aan de orde bij de opzet van een studie (vaak alleen voor de benodigde steekproefomvang), niet bij de interpretatie van de uitkomsten in het licht van concurrerende hypothesen en replicatiestudies. Maar ook al zijn er fundamentele verschillen in benadering, in de praktijk zijn statistische significantie- en hypothesetoetsen moeilijk te onderscheiden (Senn 2001). De kritiek richt zich dan ook vooral op het routinematige gebruik van de 5%-norm en de verkeerde interpretatie van p-waarden.

DE 5%-NORM

Fisher beschouwde de p-waarde als een index en volgens Neyman-Pearson gaat het om een afweging van de type I- en type II-fout die alleen door de onderzoeker in een bepaalde context gemaakt kan worden. Waar komt dan de 5%-norm vandaan? Deels was dat toeval. Toen Fisher werkte aan zijn in 1925 verschenen *Statistical methods for research workers*, kreeg hij vanwege auteursrechten geen toestemming van Karl Pearson, de eindredacteur van *Biometrika*, om de in dat tijdschrift gepubliceerde tabellen van Pearsons χ^2- en Students t-toets over te nemen. Fisher maakte zijn eigen versies, maar beperkte zich tot waarden voor bepaalde p-niveaus in plaats van omgekeerd de *probabilities* van de verschillende toetsingsgrootheden te tabuleren. Daarbij introduceerde hij grenswaarden p = 0,05 en p = 0,01 die in andere statistische handboeken werden gepopulariseerd (Lehmenn 2011). Neyman en Pearson adopteerden Fishers 5%-grenswaarde voor hun type I-fout.

In de praktijk bleek de 5%-norm ook voor onderzoekers een bruikbare grens: de objectieve p-waarde gaf toch voldoende publiceerbare resultaten. De 5%-norm was echter niet meer dan een praktische handreiking. In Fishers formulering: '*... no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each par-*

FIGUUR 2 *Type I- en type II-fouten in statistische toetsen; de H_0-curve is de verdeling van de steekproefuitkomst onder de 'nulhypothese' dat er in de populatie geen verschil is (bijvoorbeeld tussen mannen en vrouwen in ernst van psychiatrische symptomen); H_1 is de verdeling wanneer de 'alternatieve hypothese' juist is (de populaties verschillen)*



ticular case in the light of his evidence and his ideas' (Fisher 1956).

PROBLEMEN MET P-WAARDEN

Elke statistische toets is een variatie op het basisthema: de signaal-ruisratio. De toets koppelt een bepaalde uitkomst, bijvoorbeeld een verschil in gemiddelden, percentages of tellingen aan de spreiding van de populatiewaarde. Voor deze toetsingsgrootheid wordt een p-waarde berekend. De ingrediënten voor een statistische toets zijn: (1) de hypothese, (2) een toetsingsgrootheid, en (3) en de kansverdeling van de toetsingsgrootheid, vaak de normaalverdeling of de binomiale of poissonverdeling. Per ingrediënt, en in verschillende combinaties, kunnen p-waarden problemen opleveren.

Wij beschrijven een aantal voorbeelden uit de praktijk: de onwaarschijnlijke nulhypothese, de rapportage van kleine verschillen, niet voldoen aan de testvoorwaarden, de interpretatie van geen verschil en de interpretatie van complexe modellen.

De nulhypothese

Ten eerste worden vaak nulhypothesen getoetst die onwaarschijnlijk zijn of zelfs onmogelijk. Dat er geen verschillen zijn, is niet erg aannemelijk als het bijvoorbeeld gaat om civiel- en strafrechtelijke plaatsingen (Hamerlynck e.a. 2009), agressieve en niet-agressieve patiënten (Penterman & Nijman 2009), of patiënten die in zorg bleven versus degenen die de zorg verlieten na een opname met een rechterlijke machtiging (Stobbe e.a. 2009). Met een significante p-waarde bij een 'geen verschil'-hypothese kunnen we concluderen in welke richting het effect gaat. Dat kan de primaire vraagstelling zijn, bijvoorbeeld in medicatiestudies, maar alleen de richting bepalen is niet voldoende om verklaringen te vinden.

In de praktijk is het effect nooit exact nul zodat de 'geen verschil'-hypothese altijd onwaar is. Dan kan geen type I-fout gemaakt worden, maar de kans op een type II-fout is groot – correctie voor multipele testen verlaagt de power nog verder (Cohen 1994). Wanneer Penterman e.a. (2011) een bonferroni-correctie uitvoeren, verdwijnen onder andere 'doelmatigheid' en 'ordelijkheid' uit beeld als persoonskenmerken van crisisdienstmedewerkers. Een dergelijke selectie is echter niet nodig wanneer een volledig overzicht wordt gegeven van de verschillende onderzoeksvragen (Mayo & Cox 2006). Die correctie hadden Bisseling en Braam (2009) wel moeten toepassen bij de analyse van verschillen in de totale tijdsduur van de crisisinterventie naast de samenstellende delen van die tijdsduur: aanrijtijd en duur beoordeling.

Ook bij Penterman en Nijman (2009) is er sprake van multipele testen omdat per item van hun risicochecklist, bijvoorbeeld de melder van de crisissituatie, alle categorieën apart zijn getoetst. De 6 checklistonderwerpen zijn onderzocht met 38 χ^2-toetsen, waarvan 20 toetsen statistisch significant waren bij $p < 0{,}05$ en 12 bij $p < 0{,}001$. Dergelijke toetsen zouden niet moeten worden uitgevoerd. Wanneer de nulhypothese onjuist is en de power van de test gering, dan is elk resultaat onduidelijk en de kans op een foutieve conclusie groter (Murphy & Myors 2004).

In sommige gevallen is de nulhypothese zelfs logischerwijs onmogelijk. Van der Post e.a. (2009) toetsten het verschil tussen aantallen vrijwillige en gedwongen opnamen in de perioden 1983 en 2004-2005 in Amsterdam. De besluitvormingsprocessen zijn echter moeilijk te vergelijken en onduidelijk is waarom er geen verschil tussen vrijwillige en gedwongen opnamen zou zijn of in de tijd een vaste verdeling zou bestaan. Bovendien bleek uit de tijdreeks over enkele decennia meer dan een verdubbeling van het aantal acute gedwongen opnamen per 100.000 inwoners. Maar in de χ^2-toets is het verschil in absolute aantallen gedwongen opnamen tussen 1983 en 2004-2005 niet groot: 77 versus 87. Vermoedelijk is in de samenstelling van de cohorten een selectiebias opgetreden en heeft de berekende p-waarde geen betekenis.

Grote aantallen en kleine verschillen

Een tweede fout betreft de interpretatie van kleine verschillen. De kans op deze fout neemt toe wanneer de dataset groot genoeg is en elk resultaat statistisch significant wordt. Tijdink e.a. (2008) concludeerden op basis van 2297 vragenlijsten ingevuld door studenten geneeskunde en basisartsen dat de belangstelling voor de psychiatrie verschilt tussen de twee betrokken universiteiten. De opvallende verschillen varieerden van slechts 2,9 tot 5,6%, terwijl de respons maar liefst 28% verschilde.

Grote aantallen en kleine verschillen zien we ook in het artikel van Brook (1999) over heropnames van thuislozen (n = 893) in vergelijking tot de overige opgenomen patiënten (n = 114.308). Tabel 1 van zijn bijdrage toont van 13 patiënt- en opnamekenmerken alleen de p-waarden (p = 0,001) – in de tekst zijn de verschillen in aantallen en percentages vermeld. Deze statistisch significante verschillen zijn verder niet gecorrigeerd voor bijvoorbeeld geslacht en diagnose. Geconcludeerd werd dat er tussen beide groepen een significant verschil is in het aantal (her)opnames in het algemeen psychiatrisch ziekenhuis: 72% van de thuislozen is eenmaal opgenomen, tegenover 74% van de referentiegroep. Ook dergelijke kleine effecten worden in de slotparagraaf vaak als substantiële verschillen benoemd.

Testvoorwaarden

Een derde fout die gemaakt kan worden, is dat de berekening niet voldoet aan de assumpties van de test. Bijvoorbeeld: een verschil in percentages volgt een binomiale verdeling, die met de normale verdeling in een gewone t-toets goed benaderd wordt wanneer de percentages rond de 50% liggen, maar niet bij extreem hoge of lage percentages. Een t-toets of variantieanalyse (ANOVA) gaat uit van normaal verdeelde uitkomsten en past niet bij een vergelijking van bijvoorbeeld aantal ziekenhuisopnamen of aantal psychiatrische stoornissen. Wanneer de gebruikte toetsingsgrootheid of de kansverdeling niet goed is gekozen, dan zijn de p-waarden zonder betekenis. Damhuis e.a. (2011) gebruikten een t-toets voor een vergelijking van twee kleine groepen met niet normaal verdeelde gemiddelde aantallen van 1,45 en 2 DSM-IV-classificaties per patiënt. Ten Have e.a. (2010) schatten met regressieanalyses verschillen in gemiddelden tussen landen in attitudes tegenover hulpzoekgedrag. De verschillen varieerden van 0,02 tot 0,60, maar werden gebaseerd op een vierpuntsschaal met scores 1, 2, 3 of 4.

Mulder en Kortrijk (2012) meldden significante verschillen in gemiddelde aantallen niet-vervulde zorgbehoeften in relatie tot de duur van de behandeling. In groepen van 88 tot 364 patiënten varieerden de gemiddelden van ongeveer 4 tot iets meer dan 5 zorgbehoeften op de 22 items van de '*Camberwell Assessment of Need Short Appraisal Schedule*'. Echter, de gebruikte variantieanalyse is niet robuust voor een discrete en scheef verdeelde afhankelijke variabele met ongelijke patiëntgroepen (Glass e.a. 1972).

Soms wordt logtransformatie toegepast om de scheve verdeling te corrigeren, bijvoorbeeld aantallen opnamedagen en ambulante contacten (Bak e.a. 2008), maar dat werkt zeker bij kleine aantallen bias in de hand en is geen alternatief voor de moderne (*multilevel*)poisson- of negatieve binomiale regressieanalyse (O'Hara & Kotze 2010).

Ook aan de voorwaarden van een Pearsons χ^2-toets voldoen sommige studies niet. Deze toets veronderstelt een grote steekproef om zogenoemde structurele nullen (per definitie lege cellen) en lage verwachte celfrequenties te voorkomen. Hamerlynck e.a. (2009) toetsten of er verschil was in het soort delict tussen civiel- of strafrechtelijk geplaatste meisjes in justitiële jeugdinrichtingen. Bij doodslag en moord kwam geen civielrechtelijke plaatsing voor en verschillen bij enkele andere delicten werden getoetst met niet meer dan één of twee casussen in totaal.

Complexe modellen

De ene p-waarde is de andere niet. In regressiemodellen gaat het zowel om de *goodness-of-fit* van het model als om de bijdragen van de afzonderlijke variabelen. Daarbij worden de significantieniveaus van de coëfficiënten berekend gegeven het model. Maindonald en Braun (2010) laten in een simulatiestudie zien dat de p-waarden van de geselecteerde variabelen worden overschat. In 40 variabelen met elk 100 willekeurige getallen werden de 3 beste voorspellers geselecteerd: steeds was ten minste één coëfficiënt statistisch significant op 5%- of zelfs 1%-niveau.

Toch worden regressiecoëfficiënten vaak niet gepresenteerd of los van informatie over modelselectie en goodness-of-fit. Den Held e.a. (2011) beschreven een regressieanalyse zonder vermelding van coëfficiënten, betrouwbaarheidsintervallen, p-waarden of verklaarde variantie. Swolfs e.a. (2011) vermeldden één p-waarde van een niet benoemde toets, De Wachter e.a. (2008) vermeldden alleen de p-waarden van een regressieanalyse en De Jonge e.a. (2009) alleen de globale test- en p-waarden van 34 *mixed model*(multilevel)-analyses.

Met meer modelinformatie worden soms problemen in de analyses duidelijk. Een voorbeeld: Bruffaerts e.a. (2010) relateerden onderwijsniveaus aan vroege psychische stoornissen, gecorrigeerd voor geslacht, leeftijd, opleidingsniveau ouders en traumatische ervaring in de kindertijd. Samenvattende informatie over de verschillende modellen ontbreekt, maar dat de uitkomsten onzeker zijn, valt op te maken uit de kleine aantallen en de betrouwbaarheidsintervallen van significante oddsratio's die oplopen tot ver boven 200.

Ook Winter-Van Rossum e.a. (2010) geven geen modelinformatie van longitudinale analyses van het effect van cannabisgebruik op het ziektebeloop van bipolaire stoornissen (BS). Gebaseerd op multilevelregressieanalyses werden gestandaardiseerde coëfficiënten en betrouwbaarheidsintervallen geschat, onder andere voor het effect van cannabisgebruik op algemeen ziektebeeld en manie (B = 0,15 en 0,13 respectievelijk). De auteurs concludeerden dat de uitkomsten duiden op een ongunstige relatie op de lange termijn tussen cannabisgebruik en BS-symptomen. Echter, de factor tijd en een interactieterm van cannabisgebruik en tijd komen in het model niet voor. Uit de tabel van de ruwe klinische behandeluitkomsten blijkt dat zowel voor cannabisgebruikers als niet-gebruikers de score in 12 maanden tijd verbeterde van matig/ernstig naar minimaal ziek (gescoord als '*Clinical Global Impressions*').

Penterman en Nijman (2009) rapporteerden de resultaten van een multipele logistische-regressieanalyse van voorspellers van agressie-incidenten. In een *forward stepwise procedure* zijn twee predictoren geselecteerd: de globale inschatting van het agressierisico en de aanwezigheid van agressieve personen in de omgeving van de patiënt. Geen regressiecoëfficiënten en betrouwbaarheidsintervallen, maar alleen de p-waarden werden vermeld. Het voorspellend vermogen van het predictiemodel bleek beperkt. Van de agressie-incidenten werd in verschillende modellen 67 tot 74% correct voorspeld en 80 tot 83% van de checklistregistraties in totaal, terwijl 90% van de patiënten in de crisisdienst niet agressief was – in het algemeen mag de crisisdienst dus een niet-agressief contact verwachten. De auteurs concludeerden dat het vooraf invullen van een checklist zal leiden tot meer routine en standaardisatie in het overwegen van mogelijke risicofactoren. Maar voor die conclusie zijn geen statistische toetsen nodig (zie ook Wierdsma 2013).

Interpretatie van geen effect

Tot slot: soms worden conclusies geformuleerd terwijl er geen significante verschillen zijn gevonden. Volgens Penterman en Nijman (2009) kan ook het gebrek aan samenhang informatief zijn, bijvoorbeeld dat agressie-incidenten niet zijn gerelateerd aan eerder contact met de patiënt. Schadé e.a. (2011) verwachtten 20% minder terugval na behandeling van comorbide angststoornissen bij alcoholafhankelijke patiënten. Zij toetsten ech-

ter een 'geen-verschil'-hypothese en concludeerden op basis van de niet significante p-waarde dat behandeling geen invloed heeft. In een minder expliciete formulering wordt bijvoorbeeld gerapporteerd dat de ernst van de problemen niet lijkt toe te nemen met leeftijd (Broersma & Sytema 2008).

De interpretatie van 'geen effect' is echter een omkering van zaken. De toets betreft $Pr(Data|H_0)$ en dat is een heel andere kans dan $Pr(H_0|Data)$: de kans dat iemand Dood is na verHanging is groot: $Pr(DH) > 95\%$, maar de kans is klein dat we kunnen concluderen dat iemand zich heeft verHangen als hij Dood is: $Pr(HD) < 1\%$. In een sterke onderzoeksopzet is een niet verworpen hypothese misschien met enig vertrouwen te interpreteren als geen verschil of geen effect (Mayo & Cox 2006). Dus toch $Pr(H_0|Data)$ onder bepaalde omstandigheden, maar in de meeste gevallen zegt een niet verworpen 'geen verschil'-hypothese alleen dat de richting van het verband nog onzeker is (Cohen 1994).

ALTERNATIEVEN

Uitsluitend p-waarden beoordelen heeft alleen zin in explorerend onderzoek waarbij het significantieniveau kan worden gebruikt als *'a sort of crude surprise index'* (Senn 2001). Een significante p-waarde vermengt echter in één maat zowel informatie over de omvang van de verschillen als de omvang van de studie. De significantietest leidt af van de hoofdvragen: hoe groot is het verschil of hoe sterk is het verband en hoe nauwkeurig is de schatting van het effect?

Effectmaten en betrouwbaarheidsintervallen geven een beter beeld van de sterkte van de onderzochte verbanden en de precisie van de schattingen. Gestandaardiseerde effectmaten, zoals Cohens d of correlaties, kunnen worden samengevoegd in meta-analyses die betere 'overall'-schattingen opleveren of de variatie in uitkomsten relateren aan kenmerken van de studies. Daarmee verschuift de aandacht van enkelvoudige, grootschalige onderzoeksprojecten naar replicaties waarbij ook plaats is voor kleinschalige studies in verschillende contexten. Betrouwbaarheidsintervallen geven het bereik aan waarbinnen de testwaarde zal liggen. Wanneer de waarde van de nulhypothese buiten het interval ligt, is de uitkomst statistisch significant. Dit is dezelfde informatie als de p-waarde, maar betrouwbaarheidsintervallen benadrukken het effect en de variatie van de schatting van het effect.

Net als p-waarden vragen ook effectmaten en betrouwbaarheidsintervallen om een niet-routinematige aanpak. Soms worden effect sizes wel in de tabel opgenomen, maar blijven deze uitkomsten marginaal in de discussiesectie (Van Houdenhove e.a. 2010; Wilson e.a. 2010). Cohen (1994) benoemde effect sizes als '*small*', '*medium*', en '*large*', maar benadrukte dat er geen algemene grenswaarden zijn. De betekenis van het effect moet steeds worden bezien in vergelijking tot andere studies en in zinvolle eenheden.

Ook een betrouwbaarheidsinterval heeft beperkingen, vergelijkbaar met de significantietest. Wanneer de nul buiten het schattingsinterval ligt, is de uitkomst statistisch significant, maar de 'geen verschil'-hypothese is weinig informatief en bij een groot aantal worden alle verschillen significant. Exploratieve analyses met behulp van grafische technieken, zoals staafgrafieken met *error-bars*, kunnen bijdragen aan een beter begrip van de data, maar net als p-waarden worden de figuren vaak verkeerd geïnterpreteerd (Belia e.a. 2005). Bijvoorbeeld Mulder en Kortrijk (2012) beschrijven een verandering in zorgbehoeften op basis van staafdiagrammen met 95%-betrouwbaarheidsinterval. Maar de meeste betrouwbaarheidsintervallen in de figuur overlappen met de helft of meer en dan is $p > 0{,}05$ zodat statistisch geen beloop is te zien (Cumming & Finch 2005).

Andere alternatieven voor significantietesten, zoals het rapporteren van conditionele p-waarden (Berger 2003) of bayesiaanse methoden (Wijeysundera e.a. 2009), zullen moeilijker ingang vinden. Maar meer nadruk op effectmaten en betrouwbaarheidsintervallen in plaats van p-waarden is met de beschikbare hardware en statistische software al lang geen probleem meer.

Aanwijzingen voor auteurs

Alleen alternatieven geven voor p-waarden helpt niet om problemen bij de interpretatie van de uitkomsten te voorkomen. Effectmaten en betrouwbaarheidsintervallen geven een indicatie van de sterkte van de samenhang en de precisie van de schattingen. De praktische betekenis van de onderzoeksresultaten wordt daarmee nog niet duidelijk. Daarom zou de uiteindelijke conclusie van het hoofdeffect van de studie moeten worden samengevat in eenheden van de schaal waarop de uitkomstmaat is gemeten. In de woorden van de richtlijn van de American Psychological Association: '*Always present effect sizes for primary outcomes. If the units of measurement are meaningful on a practical level..., then we usually prefer an unstandardized measure (regression coefficient or mean difference) to a standardized measure (r or d)*' (Wilkinson & APA Task Force on Statistical Inference 1999). Ook wanneer complexe statistische modellen zijn gebruikt, heeft een interpretatie van de 'ruwe' coëfficiënten de voorkeur. Effectmaten en betrouwbaarheidsintervallen in eenheden van de afhankelijke variabele helpen om problemen bij de interpretatie van de resultaten te voorkomen. Deze aanwijzing voor auteurs zal een deel van de fouten die nu nog worden gemaakt wegvangen, omdat het onderzoekers dwingt heel concreet de betekenis en relevantie van de uitkomsten te duiden.

LITERATUUR

Bak M, Drukker M, de Bie A, à Campo J, Poddighe G, van Os J, e.a. Een observationele trial naar 'assertive outreach' met remissie als uitkomstmaat. Tijdschr Psychiatr 2008; 50: 253-62.

Belia S, Fidler F, Williams J, Cumming G. Researchers misunderstand confidence intervals and standard error bars. Psychol Methods 2005; 10: 389-96.

Berger J. Could Fisher, Jeffreys and Neyman have agreed on testing? Statistical Science 2003; 18: 1-31.

Bisseling EM, Braam AW. Tijdslimieten in de samenwerking tussen politie en crisisdienst: een praktijkevaluatie in Utrecht. Tijdschr Psychiatr 2009; 51: 687-92.

Broersma TW, Sytema S. Implementatie van het meetinstrument HoNOS65+. Onderzoek op een afdeling voor ouderenpsychiatrie. Tijdschr Psychiatr 2008; 50: 77-82.

Brook OH. Heropnames van thuislozen in het algemeen psychiatrisch ziekenhuis. Tijdschr Psychiatr 1999; 41: 567-74.

Bruffaerts R, Bonnewyn A, Demyttenaere K. Effect van vroege psychische stoornissen op opleidingsniveau in België; een bevolkingsstudie. Tijdschr Psychiatr 2010; 52: 133-42.

Cohen J. The earth is round (p < .05). Am Psychol 1994; 49: 997-1003.

Cumming G, Finch S. Inference by eye. Conficence intervals and how to read pictures of data. Am Psychol 2005; 60: 170-80.

Damhuis N, van Megen HJGM, Peeters CFW, Vollema MG. De MiniScan als psychiatrische interventie; pilotonderzoek naar de toegevoegde waarde van een gecomputeriseerd classificatiesysteem. Tijdschr Psychiatr 2011; 53: 175-80.

David HA. First (?) occurrence of common terms in mathematical statistics. The American Statistician 1995; 49: 121-33.

Held OM den, Hegge IRHJ, van Schaik DJF, van Balkom AJLM. Medisch studenten en hun houding ten opzichte van het vak psychiatrie. Tijdschr Psychiatr 2011; 53: 519-30.

Fisher RA. Statistical methods and scientific inference. Edinburgh: Oliver & Boyd; 1956.

Fisher RA. Statistical methods for research workers. In: Bennett JH, red. Statistical methods, experimental design, and scientific inference: A re-issue of statistical methods for research workers, the design of experiments, and statistical methods and scientific inference. Oxford: Oxford University Press; 1990/1925.

Glass GV, Peckham PD, Sanders JR. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. Review of Educational Research 1972; 42: 237-88.

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999; 130: 995-1004.

Hald A. A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713 to 1935. Copenhagen: Department of Applied Mathematics and Statistics, University of Copenhagen; 2004.

Hamerlynck SMJJ, Jansen LMC, Doreleijers TAH, Vermeiren RRJM, Cohen-Kettenis PT. Civiel- en strafrechtelijk geplaatste meisjes in justitiële jeugdinrichtingen; psychiatrische stoornissen, traumatisering en psychosociale problemen. Tijdschr Psychiatr 2009; 51: 87-96.

Harten PN van. Number needed to treat: een nuttige maat voor de effectiviteit van een behandeling. Tijdschr Psychiatr 2008; 50: 337-43.

Have M ten, de Graaf R, Ormel J, Vilagut G, Kovess V, Alonso J. Attituden aangaande zoeken van professionele hulp voor psychische problemen en werkelijk hulpzoekgedrag: verschillen in Europa. Tijdschr Psychiatr 2010; 52: 205-17.

Hoekstra R. The use and usability of inferential techniques. Vol PhD. Groningen: Rijksuniversiteit Groningen; 2009.

Houdenhove L Van, Buyse B, Gabriëls L, Van Diest I, Van den Bergh O. Cognitieve gedragstherapie voor primaire insomnia: effectiviteit in een klinische setting. Tijdschr Psychiatr 2010; 52: 79-88.

Jonge E de, Nijman HLI, Lammers SMM. Gedragsveranderingen tijdens tbs-behandeling: een multicenteronderzoek. Tijdschr Psychiatr 2009; 51: 205-15.

Lehmenn EL. Fisher, Neyman, and the Creation of Classical Statistics. New York: Springer; 2011.

Maindonald J, Braun WJ. Data Analysis and Graphics Using R (3de dr). Cambridge: University Press; 2010.

Matthews R. The great health hoax. Sunday Telegraph 13 september, 1998.

Mayo DG, Cox DR. Frequentist statistics as a theory of inductive inference. IMS Lecture Notes–Monograph Series; 2nd Lehmann Symposium – Optimality 2006; 49: 77-97.

Morrison DE, Henkel RE, red. The Significance Test Controversy - A Reader. London: Butterworths; 1970.

Mulder CL, Kortrijk HE. De invloed van de duur van behandeling op het interpreteren van ROM-metingen bij ACT. Tijdschr Psychiatr 2012; 54: 191-6.

Murphy KR, Myors B. Statistical Power Analysis, (2de druk). Mahwah: Lawrence Erlbaum; 2004.

Neyman J, Pearson ES. On the use and interpretation of certain test criteria. Biometrika 1928; 20A: 175-240, 263-95.

O'Hara RB, Kotze DJ. Do not log-transform count data. Methods in Ecology and Evolution 2010; 1: 118-122.

Os J van. Cochrane-speak: een introductie. Tijdschr Psychiatr 2000; 42: 423-9.

Penterman EJM, Nijman HLI. Het inschatten van agressie bij patiënten van de ggz-crisisdienst. Tijdschr Psychiatr 2009; 51: 355-64.

Penterman EJM, Smeets JML, van der Staak CPF, Nijman HLI. Persoonlijkheidskenmerken van crisisdienstmedewerkers in de ggz. Tijdschr Psychiatr 2011; 53: 145-51.

Post LFM van der, Dekker JJM, Jonkers JFJ, Beekman ATF, Mulder CL, de Haan L, e.a. Veranderingen in crisisinterventie en acute psychiatrie; Amsterdamse consulten in 1983 en 2005. Tijdschr Psychiatr 2009; 51: 139-50.

Rothman KJ. Epidemiology: an introduction. New York: Oxford University Press; 2002.

Schadé A, Marquenie LA, van Balkom AJLM, Koeter MMJ, van den Brink W, van Dyck R. Effectiviteit van een toegevoegde behandeling voor angstklachten bij alcoholafhankelijke patiënten met een fobische stoornis. Tijdschr Psychiatr 2011; 50: 137-48.

Senn S. Two cheers for p-values? J Epidemiol Biostat 2001; 6: 193-204.

Stobbe J, Wierdsma AI, van Beest RHP, Mulder CL. Drop-out na gedwongen opname - hoe groot is het probleem? Tijdschr Psychiatr 2009; 51: 801-12.

Swolfs SN, Boerkoel RA, Rijnders CAT. De meerwaarde van een somatische screening op een polikliniek psychiatrie. Tijdschr Psychiatr 2011; 53: 201-10.

Tijdink JK, Soethout MBM, Koerselman GF, ten Cate TJ. De belangstelling voor het beroep van psychiater bij studenten geneeskunde en basisartsen. Tijdschr Psychiatr 2008; 50: 9-17.

Wachter D De, Neven A, Vandewalle S, Vanderlinden J, Lange A. Dissociatieve verschijnselen: verband met stress uit heden én verleden. Tijdschr Psychiatr 2008; 50: 83-8.

Wierdsma AI. Reactie op 'Inschatten van agressie tijdens contacten met de ggz-crisisdienst met een checklist: een replicatiestudie'. Tijdschr Psychiatr 2013; 55: 312-4.

Wijeysundera DN, Austina PC, Hux JE, Beattie WS, Laupacis A. Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials. J Clin Epidemiol 2009; 62: 13-21.

Wilkinson L, APA Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. American Psychologist 1999; 54: 594-604.

Wilson S, van Loo S, Geuens T, Claes SJ. Persoonlijkheidskenmerken bij patiënten die volledig hersteld zijn van een depressieve stoornis. Tijdschr Psychiatr 2010; 52: 9-16.

Winter-van Rossum I, Boomsma MM, Tenback DE, Reed C, van Os J. De invloed van cannabis op het ziektebeloop van de bipolaire stoornis; een longitudinale analyse. Tijdschr Psychiatr 2010; 52: 287-98.

Ziliak ST, McCloskey DN. The cult of statistical significance: how the standard error costs us jobs, justice, and lives: Ann Arbor: University of Michigan Press; 2008.

AUTEURS

ANDRÉ WIERDSMA, socioloog/methodoloog en universitair docent, afd. Psychiatrie, Erasmus MC, Rotterdam.
JIM VAN OS, voorzitter vakgroep Psychiatrie en Psychologie, Maastricht Universitair Medisch Centrum, en Visiting Professor Psychiatric Epidemiology, King's College London, King's Health Partners, Department of Psychosis Studies, Institute of Psychiatry, Londen.

Correspondentieadres: dr. André Wierdsma, Erasmus MC, afd. Psychiatrie, Postbus 2040, 3000 CA Rotterdam.
E-mail: a.wierdsma@erasmusmc.nl

Geen strijdige belangen meegedeeld.

Het artikel werd voor publicatie geaccepteerd op 11-2-2013.

SUMMARY

Effects on scale and confidence intervals as alternatives to $p < 0.05$ – A.I. Wierdsma, J. van Os –

BACKGROUND *Researchers and reviewers often use the conventional $p < 0.05$ as threshold in statistical tests. In many cases, however, the interpretation of p-values is incorrect.*

AIM *To explain where the 5% norm originates, identify the interpretation problems that often arise and suggest some alternatives.*

METHOD *On the basis of recent literature we examine the meaning and origin of the $p < 0.05$ norm. We looked closely at entire articles and short reports in the Tijdschrift voor Psychiatrie, starting with the Jubilee issue of 2008, in order to find examples of methodological problems relating to the routine use of p-values.*

RESULTS *We found several examples of the problematic use of p-values; these included the testing of a priori unlikely, or even impossible null hypotheses, the reporting of small effects calculations based on erroneous assumptions, and incorrect interpretations of statistical parameters and p-values.*

CONCLUSION *Research in psychiatry, like research in other disciplines, attaches too much weight to p-values. Guidelines for authors should advise authors to focus explicitly on effect sizes, confidence intervals and the scale on which the results are presented.*

[TIJDSCHRIFT VOOR PSYCHIATRIE 55(2013)7, 471-480]

KEY WORDS hypothesis, power, significance, type 1 error