Podosophia
https://doi.org/10.1007/s12481-019-00220-x

Zijn statistisch significante resultaten ook klinisch relevant?

Sil Aarts

Samenvatting Hypothesetests worden vaak gebruikt voor het leveren van 'evidence' met als doel medische onderzoeksresultaten te genereren. Door hypothesen te toetsen, probeert de onderzoeker conclusies te trekken over een gehele populatie op basis van informatie die is verkregen uit een relatief kleine steekproef. Echter, de interpretatie van het testen van hypothesen en de bijbehorende *p*-waarde biedt vele uitdagingen. In artikelen die in wetenschappelijke tijdschriften worden gepubliceerd staat bijvoorbeeld 'is statistisch significant' of 'een statistisch significante *p*-waarde van 0,031 werd gevonden'. Helaas vergemakkelijkt deze focus op de *p*-waarde beslissingen over de relevantie van de verkregen resultaten niet. Want wat zeggen deze uitspraken eigenlijk? Het huidige artikel beschrijft de interpretatie van *p*-waarden en betrouwbaarheidsintervallen en het klinisch belang van deze maten.

Trefwoorden statistische significantie · *p*-waarden · klinische relevantie

Inleiding

Het doel van het nemen van een willekeurige steekproef is het schatten van een bepaalde paramater in de populatie. Dit doen we door een zogenoemde steekproef te nemen. Een steekproef is een kleine selectie van de populatie. Dergelijke steekproeven worden verzameld om onderzoeksvragen te beantwoorden. Die vragen worden veelal beantwoord door middel van statistische toetsen; het toetsen van *hypothesen*. Op basis van eerdere kennis- en praktijkervaring wordt een zogenaamde *nulhypothese* (H_0) opgesteld. Dat wil zeggen, de veronderstelling dat er *geen* verschil (tussen bijvoorbeeld groepen) of samenhang/relatie (tussen variabelen) bestaat. Ook wordt er een *alternatieve hypothese* geformuleerd (H_a). Dat wil zeggen: de veronderstelling dat er *wel* een verschil of samenhang/relatie bestaat. Het doel van een statistische toets is dan om na te gaan welke hypothese door de data wordt ondersteund: de nulhypothese of de alternatieve hypothese. Om dit te bepalen, maakt men gebruik van een zogenaamde *p*-waarde. De *p*-waarde is de kans of waarschijnlijkheid (*p* = *probability*) dat een gevonden resultaat berust op toeval. De *p*-waarde is dus een berekening die aangeeft of een bepaald resultaat dat gevonden wordt in een steekproef, bijvoorbeeld een verschil tussen mannen en vrouwen in hun tevredenheid over podotherapeutische behandelingen, door toeval is ontstaan of niet.

In de wetenschap wordt meestal een *p*-waarde lager dan 0,05 beschouwd als 'statistisch significant'. Dit betekent dat wetenschappers hebben 'afgesproken' dat, als de kans op toeval 5 % of kleiner is (0,05 op een schaal van 0–1 is 5 %), we dit in de wetenschap een acceptabele kans op toeval vinden. Bij een $p < 0{,}05$ wordt de nulhypothese verworpen; de alternatieve hypothese die meestal luidt dat er een verschil of samenhang/relatie is, wordt dan aangenomen. Dit is zo, omdat de kans op een toevalsbevinding dan 5 % of kleiner is. Bij een $p \geq 0{,}05$ wordt de nulhypothese niet verworpen. De kans op toeval is dan groter dan 5 %, wat wetenschappers als een te grote kans beschouwen. Uit bovenstaande informatie is dan ook meteen te concluderen, dan de *cut-off* van 5 % nogal arbitrair is.

In deze rubriek dragen de auteurs een steentje bij aan het vergroten van de kennis over wetenschappelijk onderzoek en de toepasbaarheid ervan in de podotherapeutische praktijk.

dr. S. Aarts (✉)
Health Services Research, Universiteit Maastricht, Maastricht, Nederland
s.aarts@maastrichtuniversity.nl

Published online: 11 February 2019

Voorbeeld CI

Stel we nemen een steekproef om de gemiddelde leeftijd van een populatie (bijv. de Nederlandse bevolking) te schatten. De vraag is hoe goed deze steekproef (bijvoorbeeld 2.000 willekeurig geselecteerde Nederlanders) de onderliggende waarde, in dit geval de leeftijd van de Nederlandse bevolking, schat. Een CI geeft dan ook, in tegenstelling tot bijvoorbeeld één gemiddelde, een heel interval van betrouwbare waarden/schattingen van dit gemiddelde. Zo geeft een breed CI (een breed interval met veel tussenliggende waarden) aan dat het schatten de gemiddelde leeftijd in de populatie onzeker is; de werkelijke gemiddelde leeftijd van de populatie kan veel mogelijke waarden aannemen. Een klein CI (een klein interval) kan de gemiddelde leeftijd van de populatie met meer zekerheid schatten.

Betrouwbaarheidsintervallen

De meeste medische onderzoeken maken gebruik van betrouwbaarheidsintervallen, in het Engels *confidence intervallen* (CI) genoemd. Een CI van 95 % wordt het vaakst gehanteerd. Dit houdt in dat, als 100 willekeurige steekproeven uit dezelfde populatie worden getrokken, 95 % van de betrouwbaarheidsintervallen die in deze steekproeven zijn verkregen, de werkelijke waarde die in de populatie geldt, zal omvatten (zie tab. 1) [1]. Bij grote steekproeven, dat wil zeggen een groot aantal deelnemers (een grote *n*), zal een CI met meer precisie de betreffende parameter, bijvoorbeeld een gemiddelde, kunnen schatten dan met een kleine steekproef. Met andere woorden, CI's van grote steekproeven bevatten een kleiner interval (i. e. minder mogelijke waarden die de parameter in de populatie kan aannemen), dan CI's die worden verworven bij kleine steekproeven. CI's gaan ook over statistische significantie; een associatie of effect is statistisch significant wanneer het 95 %CI niet de waarde bevat die als nulhypothese wordt gesteld (zie tab. 1; voorbeeld III). Echter, CI's verschaffen – door de intervallen – iets meer informatie dan bijvoorbeeld alleen een gemiddelde of een *p*-waarde.

Statistische significantie

Het testen van hypothesen wordt al decennialang gebruikt om overtuigingen te kwantificeren tegen een bepaalde (nul)hypothese of ten gunste van een andere, alternatieve hypothese. Deze analyses steunen op een willekeurige verdeling van 'statistisch significant' of 'niet statistisch significant', gebaseerd op een verkregen *p*-waarde. Deze *p*-waarden hangen sterk af van de steekproefgrootte: een grotere steekproefomvang zal de zogenoemde '*power*' van een studie vergroten, ofwel het vermogen van de studie om een statistisch significant verschil of statistisch significant(e) samenhang/relatie te detecteren. Het is dus mogelijk dat klinisch irrelevante effecten of associaties statistisch significant zijn in onderzoeken met een grote steekproefomvang. Vice versa is het ook mogelijk dat klinisch belangrijke verschillen die in kleine studies zijn waargenomen, worden genegeerd vanwege hun niet-statistische significantie. Met klinisch belangrijk verschillen bedoelen we verschillen die relevant zijn voor het handelen in de (para)medische praktijk. *P*-waarden (en inherent ook het CI) zijn alleen een maat voor het bewijs voor of tegen de nulhypothese en zijn daarom geen indicatie voor een klinisch of maatschappelijk belang.

Laten we drie hypothetische studies overwegen. In deze studies wordt een vragenlijst gebruikt om depressieve klachten in kaart te brengen. Dit wordt gedaan door middel van de GDS-15 (*Geriatric Depression Scale*); een vragenlijst met 15 items die beantwoord kunnen worden met 'ja' of 'nee' [2]. De minimumscore van de GDS-15 is 0 (geen klachten; alle vragen zijn beantwoord met 'nee') en de maximumscore is 15 (maximale klachten: alle vragen zijn beantwoord met 'ja') [2]. De onderzoekers willen nagaan of er verschillen zijn tussen mannen en vrouwen in het ervaren van depressieve klachten, gemeten met de GDS-15. Het volgende geldt:

- H_0 *(de nulhypothese)*: er is geen verschil in de depressieve klachten gemeten met de GDS-15 tussen mannen en vrouwen.
- H_a *(de alternatieve hypothese)*: er is een verschil in de depressieve klachten gemeten met de GDS-15 tussen mannen en vrouwen.

We voeren dit onderzoek drie keer uit, steeds met een ander aantal, fictieve, deelnemers. De resultaten van deze drie fictieve studies zijn te vinden in tab. 1. Aan de eerste fictieve studie namen 2.000 personen deel. Het verschil tussen mannen en vrouwen is hier 0,15. Met andere woorden, mannen vertonen gemiddeld 0,15 punt meer depressieve klachten op de GDS-15 dan vrouwen. De berekening is dan: gemiddelde score GDS-15 mannen minus gemiddelde score vrouwen GDS-15 (mannen – vrouwen = 3,15 – 3,00 = 0,15). De *p*-waarde, die je verkrijgt middels een statistische analyse, is hier 0,001 (0,1 % kans dat dit resultaat op toeval berust) en dus veel kleiner dan 0,05, wat duidt op een statistisch significant resultaat. Het 95 %CI laat een klein interval zien.

Hoewel dit voorbeeld statistisch significant is, is het onwaarschijnlijk dat dit resultaat enige klinische relevantie heeft, omdat het verschil tussen vrouwen en mannen klein is, en mannen gemiddeld 5 % hoger 'scoren' op de GDS-15 dan vrouwen ($3{,}15/3 \times 100 = 105$ %).

De tweede fictieve studie toont ook dat mannen meer depressieve klachten rapporteren dan vrouwen; gemiddeld 'scoren' zij 2,10 punt hoger op de GDS-15 (mannen – vrouwen = 4 – 1,90 = 2,10). Dit verschil is statistisch significant bij een *p*-waarde van 0,005.

Tabel 1 De resultaten van drie fictieve studies middels de GSD-15

Fictieve studie	Verschil[a]	*P*-waarde	95 %CI [b]	*N*
I	0,15	0,001	0,05–0,25	2.000
II	2,10	0,005	1,25–2,95	1.000
III	1,30	0,089	–2,20–3,70	400

GDS-15: een hogere score duidt op meer klachten (min. 0; max. 15)
[a] Verschil berekening: mannen – vrouwen
[b] 95 %CI rondom het verschil

Dit *statistisch* significante verschil tussen mannen en vrouwen in score op de GDS-15 kan ook duiden op een *klinisch* relevant verschil, aangezien het absolute verschil tussen mannen en vrouwen meer dan twee punten betreft; mannen score 210 % hoger dan vrouwen. Bovendien is het 95 %-CI 'vrij smal', wat aangeeft dat de steekproefomvang groot genoeg is ($n = 1.000$) om een goede schatter van het verschil te geven.

Het derde en laatste voorbeeld betreft een studie met 400 deelnemers en laat een niet-statistisch significant resultaat zien ($p = 0{,}089$) bij een gemiddeld verschil van 1,30 punten op de GDS-15. Dit verschil tussen mannen en vrouwen is niet statistisch significant, wat ook te zien is aan het CI: de waarde 0 ligt binnen het CI. Met andere woorden, de berekening van het gemiddelde verschil (mannen minus vrouwen) kan dus ook 0 zijn, wat erop duidt dat mannen dezelfde gemiddelde score hebben op de GDS-15 als vrouwen. Daarnaast is het betrouwbaarheidsinterval in dit voorbeeld 'erg breed' (bijna 6 punten: van –2,20 tot +3,70), waardoor het moeilijk is om adequate conclusies te trekken. Omdat het betrouwbaarheidsinterval in dit voorbeeld zowel negatieve als positieve waarden bevat, is het nog niet duidelijk óf er een verschil is tussen mannen en vrouwen en óf mannen meer depressieve klachten rapporteren dan vrouwen (een positief verschil) of dat het juist andersom is (een negatief verschil). Dit onderzoek zal dan ook moeten worden herhaald met een grotere steekproefomvang, waardoor de breedte van het betrouwbaarheidsinterval afneemt en er dus een meer adequate conclusie kan worden getrokken. Let wel: het absolute verschil in dit voorbeeld (1,30) is groter dan in het eerste voorbeeld (0,15), terwijl in het eerste voorbeeld dit verschil wel statistisch significant was.

Klinische relevantie

Dikwijls wordt statistische significantie nog beschouwd als gelijkwaardig aan klinische relevantie. Aangezien een *p*-waarde een eenvoudige en dichotome maatstaf is voor het bekijken van toeval, biedt een *p*-waarde geen informatie over klinisch belang. De (para)medische wetenschap zal dan ook niet verbeteren als onderzoekers resultaten eenvoudigweg interpreteren met behulp van het arbitraire verschil tussen statistische significantie of niet-significantie [3]. Het is daarom van zeer groot belang om een *p*-waarde niet als gouden standaard te beschouwen. Bovendien moet de interpretatie van de resultaten worden beoordeeld in het licht van andere beschikbare statistische methoden, zoals de hoeveelheid verklaarde variantie en regressiecoëfficiënten. Deze statistische maten zijn beschreven in artikelen die eerder binnen deze reeks verschenen:

- Wouters EMJ en Aarts S. Regressie-analyse. Podosophia nr. 2, juni 2018.
- Aarts S en Wouters EMJ. T-toets en ANOVA. Podosophia nr. 1, maart 2018.

Take Home Message

Medisch onderzoekers en de lezers van wetenschappelijke literatuur, zouden meer geïnteresseerd moeten zijn in de grootte van het waargenomen resultaat dan of het resultaat statistisch significant is. Aangezien de conclusies die in medische onderzoeken worden getrokken input vormen voor verder medisch onderzoek en ook tot medische beslissingen en richtlijnen kunnen leiden, is het zaak dat onderzoekers hun medische kennis en klinische expertise gebruiken om de sterkte van de geobserveerde resultaten te evalueren, ongeacht of deze statistisch significant zijn of niet. Gezien het feit dat een dergelijke kritische beoordeling een subjectieve aangelegenheid kan zijn, moeten de waargenomen resultaten worden geïnterpreteerd in termen van context en type onderzoek en worden vergeleken met de beschikbare kennis en literatuur.

Dit artikel is een bewerking van: S. Aarts, B. Winkens en M. van den Akker. The insignificance of statistical significance. Eur J Gen Pract. 2012;18:50–2. Dit is het laatste artikel in deze reeks.

Literatuur

1. Fetheny J. Statistical and clinical significance, and how to use confidence intervals to help interpret both. Austr Crit Care. 2010;23:93–7.
2. Sheikh JL, Yesavage JA. Geriatric depression scale (GDS): recent evidence and development of a shorter version. In: Brink TL, redactie. Clinical gerontology: a guide to assessment and intervention. New York: Haworth Press; 1986:165–73.
3. Sterne JA, Davey Smith G. Shifting the evidence-what's wrong with significance tests? Br Med J. 2001;322:226–31.

dr. Sil Aarts, universitair docent