Podosophia (2017) 25:72–76
DOI 10.1007/s12481-017-0153-x

Wat is toetsende statistiek?

Eveline Wouters

Published online: 15 May 2017
© Bohn Stafleu van Loghum 2017

Samenvatting Om te weten of er een verschil is (bijvoorbeeld tussen twee groepen), of juist een samenhang (bijvoorbeeld tussen twee kenmerken), wordt toetsende statistiek gebruikt. Toetsen betekent daarbij dat er, met behulp van statistische berekeningen, wordt nagegaan of een verschil of relatie reëel is, dan wel mogelijk op toeval berust. Er wordt daarbij een hypothese getoetst. In dit artikel wordt besproken hoe dat in de praktijk gaat, en hoe een keuze gemaakt kan worden voor een passende toets.

Trefwoorden toetsende statistiek · hypothese · type-I-fout · type-II-fout · *p*-waarde · betrouwbaarheidsinterval

Inleiding

In het vorige artikel is uitgelegd dat er grofweg drie soorten variabelen bestaan, die verschillen wat betreft hun meetniveau en centrummaten: de nominale variabelen (voorbeeld: geslacht), waaraan geen getalsmatige waarde gegeven kan worden, de ordinale variabelen (voorbeeld: opleidingsniveau), waarin wel een ordening aan te brengen is (van lager naar hoger), maar waarmee niet echt gerekend kan worden en – op het hoogste, derde meetniveau – de intervalvariabelen (voorbeeld: *body mass index*, BMI). Met deze laatste soort variabelen kunnen berekeningen gemaakt worden, bijvoorbeeld om gemiddelden uit te rekenen. Deze kennis is nodig voor het uitvoeren van toetsende statistiek.

Bij toetsende statistiek wordt, met behulp van een statistische berekening, een hypothese getoetst. Er wordt bijvoorbeeld bepaald of het verschil dat we zien tussen twee groepen op toeval berust of niet. Of er wordt bepaald of de samenhang die we zien tussen twee variabelen toevallig is of niet.

Dit artikel begint met een uitleg over welke stappen er genomen worden bij het uitvoeren van een statistische toets in kwantitatief onderzoek. Daarna komt aan de orde welke statistische toetsen grofweg geschikt zijn voor welke soort onderzoeksvraag en de daarbij behorende variabelen. In latere artikelen zullen dan per toets de details volgen.

Hypothese toetsen

Bij statistische toetsen wordt een onderzoeksvraag beantwoord door een *hypothese* te toetsen. Op basis van eerdere kennis- en praktijkervaring wordt eerst een *nulhypothese* (H_0) opgesteld. Dat wil zeggen, de veronderstelling dat er *geen* verschil (tussen de groepen) of samenhang (tussen de variabelen) bestaat. Het doel van de statistische toets is dan het omgekeerde aantonen of de nulhypothese verwerpen. De tegenhanger van de nulhypothese is de *alternatieve hypothese* (H_1). Ofwel, er bestaat wél een samenhang of verschil. Het lijkt ingewikkeld om het op deze twee verschillende manieren te benoemen. Dat dit hier toch kort aangestipt wordt, is omdat statistische (computer)programma's, zoals het bekende statistiekprogramma SPSS (*Statistical Package for the Social Sciences*), werken met de nulhypothese, en niet met de alternatieve hypothese.

Nadat de (nul)hypothese is geformuleerd, wordt deze hypothese getoetst. Laten we nog eens kijken naar het voorbeeld uit het artikel over beschrijvende

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.

E. Wouters, PhD MD (✉)
Lectoraat Health Innovations & Technology, Fontys Paramedische Hogeschool, Eindhoven, Nederland
e.wouters@fontys.nl

Tabel 1 Fictief onderzoek naar drie behandelingen voor voetproblematiek.

	type behandeling		
variabele	podotherapie ($n = 22$)	huismiddel A ($n = 20$)	huismiddel B ($n = 21$)
leeftijd			
mediaan	53	55	45
min.–max.	23–69	23–74	25–68
IQR	24,5	26,75	19
geslacht (% vrouw)	85	86	82
BMI			
mediaan	28	27	28
min.–max.	19–41	20–36	22–41
IQR	9,75	4,75	6,25
gemiddelde duur behandeling	5 dagen	8 dagen	15 dagen
kans op terugkeer voetproblematiek na behandeling binnen een half jaar	10 %	25 %	15 %

BMI body mass index, *IQR* interkwartielrange (*inter quartile range*)

Tabel 2 Kansen op het nemen van de juiste of onjuiste beslissing naar aanleiding van een statistische toets.

beslissing naar aanleiding van de toets	de werkelijkheid	
	H_0 is waar	H_1 is waar
H_0 geaccepteerd	$(1 - \alpha)$	β
H_0 verworpen	onbetrouwbaarheid α[a]	power $(1 - \beta)$**[b]

[a]Voor de onbetrouwbaarheid alfa wordt vaak 5 % (0,05) genomen
[b]Voor de power $1 - \beta$ wordt vaak 80 % (0,80) genomen, dus: $\beta = 20$ % (0,20)

statistiek in het vorige nummer van dit tijdschrift [1]. In dat artikel werd het resultaat van podotherapeutische behandeling versus dat van twee andere behandelingen (huismiddelen) – behandeling A en B genoemd – met elkaar vergeleken. Nu willen we ook weten hoe per behandelconditie de gemiddelde duur van de behandeling is en wat de kans is dat de klachten binnen een half jaar terugkeren.

In tab. 1 is te zien dat er een verschil is in de gemiddelde duur van de behandeling tussen de drie groepen (5, 8 en 15 dagen), en dat er een verschil is in de kans dat de klachten terugkeren (10 %, 25 % en 15 %). De nulhypothese zou zijn: 'er is geen verschil in gemiddelde duur van de behandeling tussen de podotherapeutische behandeling en huismiddel A, en huismiddel B'. De tweede nulhypothese is: 'er is geen verschil in terugkeer van de klachten tussen podotherapeutische behandeling, en behandeling met huismiddel A, en behandeling met huismiddel B'. De alternatieve hypothese 1 luidt: 'er is wel een verschil tussen de gemiddelde duur van de drie behandelingen', en de alternatieve hypothese 2: 'er is wel verschil tussen terugkeer van de voetproblematiek bij de drie behandelingen'. De vraag is echter of, als er een verschil gevonden wordt, dit verschil dan een echt verschil is, ofwel een significant verschil. Of is het slechts toeval dat er een verschil wordt gevonden? Anders gezegd, is het gevonden verschil een betrouwbaar gegeven, geldt het in principe voor iedereen met deze klachten?

Om die vraag te kunnen beantwoorden, moet de betrouwbaarheid van de uitkomst worden vastgesteld. Om te beginnen: niet *alle* mensen met deze voetklachten werden in de studie onderzocht. Er is slechts een steekproef gedaan uit de hele populatie van mensen met deze voetklachten. Onderzoekers hebben hun best gedaan zo representatief mogelijke groepen samen te stellen, maar toch, het is een steekproef. Als er een verschil of verband wordt aangetoond dat er in de werkelijkheid niet is, wordt dit een *type-I-fout* genoemd, of ook wel de onbetrouwbaarheid *alfa*. Voor de praktijk wordt meestal gezegd: een kans van minder dan 5 % onbetrouwbaarheid (alfa) is acceptabel. Ofwel, we vinden het voldoende als er een kans is van 95 % dat het gevonden verschil of het verband wel bestaat. Soms (afhankelijk van wat onderzocht wordt, de omvang van het onderzoek en daarmee het aantal data) vinden we die 95 % te laag en wordt een (veel) lagere onbetrouwbaarheid acceptabel gevonden (bijv. 0,01 (1 %) of 0,001 (0,1 %)).

Het omgekeerde bestaat ook: er wordt ten onrechte geen verschil of verband aangetoond, terwijl dat er in werkelijkheid wel is. Dit heet *type-II-fout*, ook wel *bèta* geheten. Dit laatste komt veel voor als de onderzoeksgroepen (zoals in het voorbeeld) (te) klein zijn of het verschil of de samenhang klein of zwak is. Men spreekt in dat geval over een onderzoek met te weinig *power*. Type-I- en type-II-fouten zijn, met andere woorden, de kansen dat er naar aanleiding van een toets een verkeerde beslissing wordt genomen. In tab. 2 worden de begrippen H_0, H_1, alfa, bèta, onbetrouwbaarheid en power, samengevat.

Tabel 3 Keuze van de statistische toets bij samenhangvragen.

intervalvariabele(n)	ordinale variabele(n)	nominale variabele(n)
– Pearsons correlatie (twee variabelen, beide intervalvariabelen) – regressie (meer dan twee intervalvariabelen)	– Spearmans correlatie (een of meer ordinale variabelen)	– Chi-kwadraat

Tabel 4 Keuze van de statistische toets bij verschilvragen.

bij intervalvariabelen		bij ordinale variabelen of intervalvariabelen in een kleine steekproef		bij nominale variabelen
2 groepen	*>2 groepen*	*2 groepen*	*>2 groepen*	– Chi-kwadraattoets
– niet gepaard: *t*-toets – gepaard: gepaarde *t*-toets	– niet gepaard: ANOVA – gepaard: ANOVA gepaarde metingen	– niet gepaard: Mann-Whitney-U-test – gepaard: Wilcoxon signed rank toets	– niet gepaard: Kruskal-Wallis-toets – gepaard: Friedman-toets	
ANOVA analysis of the variance (variantieanalyse)				

Statistisch significant?

In de praktijk worden de getallen die beschreven staan in tab. 1, onderworpen aan een geschikte statistische toets (daarover later meer). De toets, die in een computerprogramma zoals SPSS wordt uitgevoerd, geeft altijd een zogenaamde *p*-waarde aan, en/of het betrouwbaarheidsinterval (BI, in het Engels CI, wat staat voor *confidence interval*).

De *p*-waarde is de kans of waarschijnlijkheid (*p* = *probability*) dat het gevonden verschil tussen groepen of de gevonden associatie, berust op toeval. Dus dat lijkt heel erg op het begrip alfa. Het verschil is subtiel: de alfa wordt van tevoren door de onderzoekers vastgesteld (wat vinden zij een redelijke waarschijnlijkheid dat een gevonden resultaat mag berusten op toeval). De *p*-waarde berekent de echte waarschijnlijkheid. Als van tevoren is vastgesteld dat de alfa onder 0,01 moet liggen, en er wordt een *p*-waarde van 0,04 (4 %) gevonden, dan kan de nulhypothese *niet* verworpen worden. Anderzijds, als alfa van tevoren op 0,05 (5 %) wordt gesteld, en er wordt 0,04 gevonden, dan mag de hypothese *wel* verworpen worden. Wat een subtiel verschil lijkt, heeft dus wel grote gevolgen voor de conclusie die getrokken wordt uit het onderzoek.

Naast de *p*-waarde wordt ook het 95 %-betrouwbaarheidsinterval (BI, CI) steeds vaker weergegeven. Het 95 %-BI van een bepaalde uitkomst (in het voorbeeld bijvoorbeeld de gemiddelde BMI) geeft aan dat indien deze meting 100 keer wordt herhaald, de gemeten waarde 95 keer binnen het interval zal liggen dat wordt aangegeven. In de overige vijf keer zal het percentage buiten het interval liggen. Met andere woorden, hoe kleiner dit 95 %-BI, des te preciezer de schatting is van (in het voorbeeld) de BMI. Het BI kan ook worden gebruikt om te bepalen of verschillen tussen groepen op toeval berusten of dat er een daadwerkelijk verschil is. Hoe minder de BI's overlappen, des te groter de kans dat er tussen die groepen een werkelijk (statistisch significant) verschil is. Het BI zegt echter niet alleen iets over de statistische significantie, zoals de *p*-waarde, het geeft zoals gezegd ook een indruk van de precisie van de uitkomsten.

Soorten toetsen

Als de data zijn verzameld en ingevoerd in een statistisch computerprogramma, kan de hypothese getoetst worden. Er zijn verschillende soorten toetsen, afhankelijk van de onderzoeksvraag (gaat het om een verschil tussen groepen of om een verband tussen variabelen?), het meetniveau van de variabelen (nominaal, ordinaal, interval), en het aantal en de omvang van de onderzoeksgroepen die vergeleken worden.

Om de juiste statistische toets te kiezen, is de eerste vraag: gaat het bij de onderzoeksvraag om een *verschil* tussen groepen of om een *samenhang* (verband) tussen variabelen? In het voorbeeld van de verschillende voetbehandelingen zou een verschilvraag kunnen zijn: is er een verschil tussen huismiddel A en huismiddel B wat betreft het terugkeren van de klachten? Een samenhangvraag zou kunnen zijn: is er een verband tussen de gekozen behandeling en de duur van de behandeling? Of, is er een verband tussen leeftijd en het terugkeren van de klachten?

Voor het bepalen van de juiste toets is het ook van belang om vast te stellen op welk meetniveau er wordt gemeten: gaat het om nominale (bijvoorbeeld geslacht), ordinale (bijvoorbeeld opleidingsniveau) of om intervalvariabelen (bijvoorbeeld BMI)? De grootte van de steekproef is van belang (zie kader 'Steekproefomvang en verdeling van de data').

Als er groepen met elkaar worden vergeleken, speelt bij het kiezen van de toets ook het aantal groepen (twee of meer) een rol. Ten slotte is van belang of er sprake is van zogenaamde *gepaarde* of *ongepaarde* data. Een voorbeeld van gepaarde data is als er twee keer wordt gemeten in dezelfde groep (bijv. meten voor en na behandeling). Een voorbeeld van ongepaarde data is als twee groepen die een verschillende behandeling krijgen met elkaar worden vergeleken (in het genoemde voorbeeld: de drie verschillende behandelingsmogelijkheden voor de voetklachten).

Steekproefomvang en verdeling van de data

Niet in alle gevallen is het mogelijk om de toets te gebruiken die op het eerste gezicht past bij de variabelen. Dit heeft te maken met de grootte van de steekproef en de verdeling van de data.

Bij intervalvariabelen (het hoogste meetniveau) is het voor het bepalen van de juiste toets van belang of de data *normaal verdeeld* zijn. Dat wil zeggen dat de getallen rond een bepaald gemiddelde scoren. Met andere woorden, er zijn weinig 'uitschieters' (omlaag of omhoog). Zijn die uitschieters er wel, dan gelden de toetsen voor intervalvariabelen niet meer, maar is een toets voor ordinale variabelen nodig.

Hoe groter de steekproef, hoe groter de kans is dat er sprake is van een normale verdeling van de variabelen. Om die reden wordt bij kleine steekproeven doorgaans gekozen voor de toetsen die passen bij ordinale variabelen.

Verder lezen

- Baarda B, Goede M D, Dijkum C van. Basisboek Statistiek met SPSS. 4e druk. Houten/Groningen: Noordhoff uitgevers; 2011.
- Bouter M, Dongen M van, Zielhuis G. Epidemiologisch onderzoek. Opzet en interpretatie. 5e druk. Houten: Bohn Stafleu van Loghum; 2005.
- Field, A. Discovering Statistics Using IBM SPSS Statistics. 4e druk. Londen: Sage publications; 2013.
- Wouters E, Zaalen Y van, Bruijning J. Praktijkgericht onderzoek in de (para)medische zorg. Bussum: Coutinho; 2015.

We zullen de toetsen voor samenhangvragen en verschilvragen kort de revue laten passeren.

Samenhangvragen

Bij een samenhangonderzoeksvraag is het meetniveau bepalend voor de keuze van de statistische toets (de drie kolommen van tab. 3). Zou bijvoorbeeld het verband tussen BMI en leeftijd worden berekend voor alle mensen met de genoemde voetklachten uit ons voorbeeld (alle 63), dan zou gekozen worden voor de Pearsons correlatie. Zou de samenhang tussen opleidingsniveau en BMI worden berekend, de Spearmans correlatie, en bij het verband tussen geslacht en terugkeren van de klachten binnen een half jaar, de chi-kwadraat.

Samenhang berekenen tussen meer dan twee variabelen wordt gedaan met regressie. Zo zou bijvoorbeeld bepaald kunnen worden wat de voorspellende samenhang is tussen leeftijd en BMI op de duur van de behandeling van de voetklachten.

Verschilvraag

Bij een verschilvraag zijn, naast het meetniveau, het aantal groepen en het al of niet gepaard zijn van de metingen van belang. In tab. 4 wordt hiervan een overzicht gegeven. Een voorbeeld: als we het verschil tussen de drie behandelgroepen willen weten voor de variabele BMI, dan gaat het om:

1. meer dan twee groepen die vergeleken worden,
2. een intervalvariabele, die niet gepaard is en
3. vrij kleine groepen (ongeveer 20 personen per groep).

Uit tab. 4 valt af te lezen dat dan gekozen wordt voor de Kruskal-Wallis. Een ander voorbeeld. Stel dat er sprake was van veel grotere steekproeven (bijvoorbeeld: 200 per groep), die normaal verdeeld zijn, waarbij de volgende vraag wordt gesteld: 'Wat is het verschil tussen huismiddel A en B wat betreft het terugkeren van de klachten (% binnen een half jaar)?' Welke keuze zou u maken?[1]

In dit artikel is in vogelvlucht een overzicht gegeven van de verschillende toetsen die gebruikt worden bij het analyseren van onderzoeksdata. In de volgende artikelen zal nader worden ingegaan op een aantal veelvoorkomende analyses.

Literatuur

1. Wouters W, Aarts S. Beschrijvende statistiek. Podosophia. 2016;24:122–7.

Eveline Wouters, lector health innovations and technology

[1] Het gaat in dit geval om een nominale variabele (% terugkeer van de klachten binnen een half jaar). In alle gevallen wordt dan de chi-kwadraattoets gebruikt.

Hier staat een advertentie.



Bohn
Stafleu
van Loghum

Springer Media

Houten 2017