Groepsverschillen analyseren in kwantitatief onderzoek

Grip op wetenschappelijk onderzoek (4)

'Uit onderzoek blijkt ...', lees je vaak. Maar hoe kun je nu zelf nagaan wat er dan blijkt? En hoe de onderzoekers tot deze conclusie zijn gekomen? Dit vierde deel van de serie 'Grip op wetenschappelijk onderzoek' helpt je om de technische en statistische vaktaal in kwantitatief onderzoek te begrijpen; het artikel bevat tevens een overzicht met veelvoorkomende begrippen en toetsen.

Anja Visser

Dit is deel 4 van een vijfdelige serie over empirisch onderzoek in de geestelijke verzorging. Doel van deze serie is om je meer bekend te maken met verschillende vormen van onderzoek, zodat je inzichten uit onderzoek op waarde kunt schatten en kunt toepassen in je eigen praktijk. En misschien wil je zelf wel onderzoek doen? Dan helpt deze serie om je te oriënteren op welke onderzoeksbenadering zou kunnen passen bij wat jij wilt onderzoeken en geeft ze je een aantal aandachtspunten mee bij de planning van je onderzoek.

In het vorige deel (zie nummer 2021/103) besprak ik hoe en wanneer je vragenlijsten kunt gebruiken in onderzoek naar geestelijke verzorging. Maar hoe worden de gegevens uit deze vragenlijsten (of uit andere vormen van kwantitatief onderzoek) geanalyseerd? In dit artikel staat de analyse van groepsverschillen centraal. Voor geestelijke verzorging kan het onderzoeken van groepsverschillen belangrijk zijn, omdat dit bijvoorbeeld inzicht kan geven in welke mensen meer of minder baat hebben bij begeleiding door een geestelijk verzorger of welke vorm van begeleiding de grootste invloed heeft op (aspecten van) welzijn of spirituele, levensbeschouwelijke of persoonlijke ontwikkeling. Ook kan onderzoek naar groepsverschillen inzicht geven in de vraag of begeleiding door geestelijk verzorgers een grotere of andere bijdrage levert aan het (spiritueel) welzijn van mensen in vergelijking met begeleiding door bijvoorbeeld huisartsen of verpleegkundigen.

Figuur 1. Scan deze QR-code om meteen naar het artikel van Van der Geer e.a. (2017) te gaan.

Al deze onderzoeken helpen om inzichtelijk te maken wat geestelijke verzorging betekent en te bepalen of de gehanteerde werkwijzen ook de meest effectieve zijn. Onderzoek naar groepsverschillen maakt het ook mogelijk om inzicht te krijgen in verschillen tussen geestelijk verzorgers zelf. Zijn er wezenlijke verschillen in de verhouding van denominaties, genders of leeftijdscategorieën tussen geestelijk verzorgers? Verschillen typen geestelijk verzorgers misschien in hun houding ten opzichte van bijvoorbeeld het belang van ambtelijke zending of werkzaamheid in geestelijke verzorging in de thuissituatie?

Of er sprake is van een verschil tussen groepen wordt in kwantitatief onderzoek vastgesteld met behulp van verschillende soorten statistische toetsen. Zo'n toets vergelijkt de informatie die bij de deelnemers (de steekproef) verzameld is met een veronderstelling (de nulhypothese). Bij toetsen voor groepsverschillen is de nulhypothese meestal dat er geen verschil is tussen de groepen. Deze hypothese wordt verworpen als de verzamelde gegevens te ver afwijken van de getallen die je op basis van de veronderstelling zou verwachten. Het verschil tussen de groepen is dan statistisch significant.

Hierna zal ik de meest gebruikte toetsen voor groepsverschillen toelichten: de *t*-toets, de variantieanalyse (ANOVA) en de Chi-kwadraattoets. Ik sta ook stil bij alternatieven voor de *t*-toets en ANOVA, namelijk de Wilcoxon signed-rank-toets of Mann-Whitney-U-toets en de Kruskall-Wallis-toets. Deze alternatieven worden vaak gebruikt in onderzoek naar geestelijke verzorging. Ik zal daarbij niet ingaan op de formules die onderlig-

Handige websites en boeken over statistische toetsen

Versnellingsplan Onderwijsinnovatie met ICT, *Statistisch handboek studiedata* (https://sh-studiedata.nl).
Wikipedia, *Statistische toets* (https://nl.wikipedia.org/wiki/Categorie:Statistische_toets).
Allan Field, *Discovering statistics using IBM SPSS* (2017).
Deborah Rumsey, *Statistiek voor dummies* (2014).
Nel Verhoeven, *Statistiek in stappen* (2021).

gend zijn aan de toetsen, want dat is voor het doel van dit artikel niet nodig en meestal gebruiken onderzoekers software voor het uitvoeren van de toetsen. Als je daar meer over wilt weten, verwijs ik je graag naar de boeken en websites in het kader. Ik zal hier vooral ingaan op wat er met deze toetsen getoetst wordt, wanneer deze toetsen wel en niet toegepast mogen worden, en hoe je de getallen die hierover in artikelen gerapporteerd worden kunt interpreteren. Zo ben je goed toegerust om artikelen over dergelijke onderzoeken te begrijpen en kritisch te evalueren.

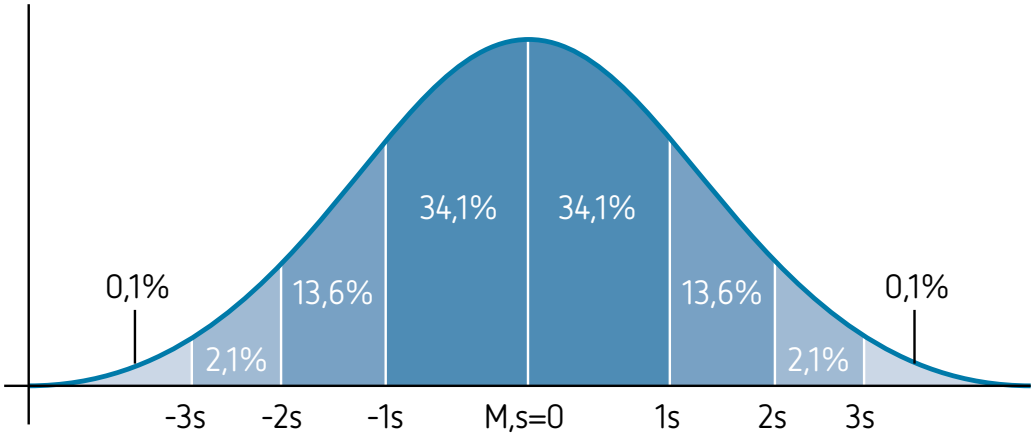
Het is belangrijk dat je als lezer kritisch naar statistische toetsen kunt kijken, want veel onderzoekers – vooral in de geesteswetenschappen, maar ook in de sociale wetenschappen – kregen zelf weinig training in het gebruik van statistiek en hebben niet altijd een statisticus tot hun beschikking die kan helpen met het kiezen van de juiste toets en het correct toepassen, interpreteren en rapporteren ervan. Bovendien zijn typefouten snel gemaakt, zoals we eerder in deze serie ook hebben gezien, wat soms vergaande consequenties kan hebben voor de betekenis van de resultaten. Als lezer van een wetenschappelijk artikel moet je dus altijd op je hoede zijn en in ieder geval controleren of de gekozen toets passend is en of de resultaten die in de tabellen staan passen bij de conclusie (zie ook nummer 2021/101).

Ik gebruik het artikel 'Training hospital staff on spiritual care in palliative care influences patient-reported outcomes: results of a quasi-experimental study' van Van de Geer e.a. (2017) als voorbeeld.[1] In dit onderzoek hebben geestelijk verzorgers in acht ziekenhuizen in Nederland trainingen in spirituele zorg gegeven aan verschillende zorgverleners. In het artikel vergelijken de onderzoekers de fysieke gezondheid, spirituele noden, spirituele attitudes en ontvangen spirituele zorg van twee groepen met elkaar: *a)* een groep patiënten die in het ziekenhuis lag nadat de zorgverleners van die afdelingen getraind waren in spirituele zorg (de interventiegroep) en *b)* patiënten die ofwel op diezelfde afdelingen lagen voordat de zorgverleners getraind waren ofwel patiënten die op afdelingen lagen waar de zorgverleners niet getraind waren in spirituele zorg (de controlegroep).

De *t*-toets

Een *t*-toets wordt ingezet als je wilt bepalen of twee groepen van elkaar verschillen op één eigenschap. Voor elke groep wordt de gemiddelde score op de eigenschap berekend en vervolgens wordt bepaald of deze gemiddelden statistisch significant van elkaar verschillen. Er zijn verschillende soorten *t*-toetsen voor verschillende vergelijkingen. Een 'independent samples'-*t*-toets wordt gebruikt om twee onafhankelijke groepen met elkaar te vergelijken, bijvoorbeeld of een groep mensen die geestelijke verzorging hebben ontvangen (een interventiegroep) minder angstig is dan een onafhankelijke maar vergelijkbare groep mensen die geen geestelijke verzorging hebben ontvangen (een controlegroep).

Een 'paired samples'-*t*-toets wordt gebruikt om verandering over de tijd mee te kunnen vaststellen, bijvoorbeeld om te bepalen of een groep



Figuur 2. Normaalverdeling. M = de gemiddelde score in de groep, s = de spreiding van de scores in de groep (standaarddeviatie, ook wel SD). Bij een normaalverdeling heeft ongeveer 68 procent van de deelnemers een score dicht bij het gemiddelde en hebben steeds minder mensen een score daar verder vanaf.[3]

mensen voor deelname aan een retraite minder spiritueel welbevinden ervoer dan na afloop van de retraite. Een 'paired samples'-*t*-toets kan ook gebruikt worden om verschillen te bepalen tussen twee groepen mensen die veel met elkaar te maken hebben (zoals stellen of gezinsleden).

Een 'one sample'-*t*-toets, ten slotte, wordt gebruikt om te bepalen of een groep mensen verschilt van een vooraf bekende waarde, bijvoorbeeld of een groep mensen die geestelijke verzorging hebben ontvangen ouder zijn dan de gemiddelde leeftijd van mensen in Nederland. Deze toets wordt overigens niet vaak gebruikt.
Wil een *t*-toets een betrouwbaar resultaat opleveren, dan moeten de data waar de berekeningen op uitgevoerd worden aan een aantal voorwaarden voldoen. Bij het lezen van wetenschappelijke artikelen waarin een *t*-toets wordt gebruikt, wil je controleren of aan deze voorwaarden is voldaan. Ten eerste moet de eigenschap waar je groepsverschillen op wilt bepalen een eigenschap zijn waarop mensen hoger of lager kunnen scoren en waarvan de verschillende getallen ook een duidelijk verschil in hoeveelheid aangeven (een schaalvariabele).[2]

Laat ik opleiding als voorbeeld nemen. Je kunt opleiding op verschillende manieren vaststellen, bijvoorbeeld door te vragen naar het aantal jaren opleiding dat iemand heeft gehad, maar ook door naar het opleidingsniveau te vragen. Het aantal jaren opleiding is geschikt voor een *t*-toets, want mensen kunnen meer of minder jaren opleiding hebben gehad en vijf jaar opleiding is duidelijk meer dan drie jaar opleiding. Opleidingsniveau is niet geschikt voor een *t*-toets, want een hoger opleidingsniveau is niet automatisch meer jaren opleiding; het is een ander soort opleiding. Het geven van een getal aan opleidingsniveau is daarom enigszins willekeurig.

Naast dat de data dus op een duidelijke kwantitatieve schaal gemeten moeten zijn, moeten de scores in de groep ook op een bepaalde manier verdeeld zijn. De data moet een zogenaamde normaalverdeling volgen (zie figuur 2). In een normaalverdeling hebben de meeste mensen een score die dicht bij het gemiddelde ligt en zijn er steeds minder mensen die scores hebben aan de uiteinden van de schaal. Deze eis wordt minder belangrijk naarmate er meer mensen met het onderzoek meegedaan hebben – de meningen verschillen of dit al vanaf vijftig deelnemers is of pas bij honderd – maar grote afwijkingen waarbij de meeste mensen aan een van de uiteinden van de schaal scoren, zijn altijd problematisch. Dan is het beter om een Wilcoxon signed-rank toets of de Mann-Whitney U toets toe te passen (zie hierna).

In de paragraaf Resultaten van wetenschappelijke artikelen wordt er een aantal getallen gerapporteerd als er een *t*-toets is uitgevoerd. Allereerst het gemiddelde (*M*) en de standaarddeviatie (*SD* of *s*) van de groepen of meetmomenten die met elkaar vergeleken worden. De standaarddeviatie is een maat die aangeeft hoe groot de spreiding van de scores in de groep is. In tabellen zie je vaak ook een getal aangegeven met *N* (of *n*), wat de groepsgrootte(s) weergeeft.

Onderzoek helpt inzichtelijk te maken wat geestelijke verzorging betekent en of werkwijzen effectief zijn

Vervolgens zie je de resultaten van de toets, bijvoorbeeld $t_{45} = 1{,}53$, $p < ,001$. Het getal na het isgelijkteken is de *t*-waarde. Je kunt op zichzelf niet zoveel aflezen aan dit getal, behalve dat als het negatief is (er staat een minteken voor) het gemiddelde van de tweede groep of op het tweede meetmoment groter is dan het gemiddelde van de eerste groep of het eerste meetmoment en andersom als het getal positief is (er staat geen minteken voor).

Het getal achter *p* ('probability') geeft de statistische significantie van de toets weer. De precieze betekenis van dit getal is behoorlijk complex (zie de begrippenlijst), maar men heeft afgesproken dat als de *p*-waarde kleiner is dan 0,05 het resultaat statistisch significant is. In het geval van een *t*-toets zou je kunnen zeggen dat dit betekent dat het waarschijnlijk is dat de twee groepsgemiddelden echt van elkaar verschillen. Tot een waarde van 0,001 wordt in de tekst van wetenschappelijke artikelen meestal het precieze getal weergegeven tot op drie decimalen. Als *p* kleiner is dan 0,001, wordt dat weergegeven als $p < ,001$.[4]

Wilcoxon signed-rank toets en Mann-Whitney U toets

In veel sociaalwetenschappelijk onderzoek volgen de scores van de deelnemers op de gemeten eigenschap niet de normaalverdeling. Zoals gezegd, is dat bij kleine afwijkingen niet zo erg. Maar als de afwijking groter wordt, dan worden traditionele toetsen (zoals de *t*-toets) onnauwkeurig en moet er een alternatief gebruikt worden. Dit worden non-parametrische toetsen genoemd. Zulke toetsen stellen nauwelijks eisen aan de manier waarop de scores verdeeld zijn.

Non-parametrische toetsen kunnen ook toegepast worden als de eigenschap waarop de groepen vergeleken worden bestaat uit categorieën, in plaats van dat het op een schaal gemeten is (een categorische variabele). Deze toetsen kun je dus ook gebruiken om groepsverschillen op bijvoorbeeld opleidingsniveau te onderzoeken. De eigenschap moet echter wel meer dan twee waarden kunnen aannemen en de waarden moeten geordend zijn in grootte, waarbij een hogere score dus betekent dat mensen meer of beter scoren op de eigenschap.[5]

Een non-parametrisch alternatief voor de 'paired samples'-*t*-toets is de Wilcoxon signed-rank toets, terwijl de Mann-Whitney U toets een alternatief vormt voor de 'independent samples'-*t*-toets. In plaats van de gemiddelden van de twee groepen of meetmomenten te vergelijken, zoals bij de *t*-toets, vergelijken deze toetsen de rangordening van de scores. In de Wilcoxon signed-rank toets worden de verschilscores tussen de eerste en de tweede meting op volgorde van klein naar groot geplaatst en wordt er gekeken of er een verschil is in het aantal negatieve verschillen (de score bij de tweede meting is hoger dan bij de eerste) en positieve verschillen (de score bij de tweede meting is lager dan bij de eerste). In de Mann-Whitney U toets worden alle scores van beide groepen samen op volgorde gezet, van klein naar groot, en wordt er vervolgens gekeken of de scores van de ene groep hoger of lager in de rangorde staan dan de scores van de andere groep. Beide toetsen gebruiken hiervoor rangnummers: het laagste getal krijgt rangnummer 1, het volgende getal krijgt 2, het daaropvolgende getal krijgt 3, enzovoort. Op deze rangnummers worden de berekeningen uitgevoerd.

In de paragraaf Resultaten van wetenschappelijke artikelen vind je de volgende getallen voor

een Wilcoxon signed-rank toets: de mediaan van de scores van beide meetmomenten of groepen, eventueel het aantal en de som van de negatieve en van de positieve rangnummers, een getal *Z*, de *p*-waarde en eventueel een getal aangeduid met *r* (dit getal wordt een effectgrootte genoemd; bij een *t*-toets is het getal *t* de effectgrootte). Een *r* van rond 0,1 geeft aan dat het verschil klein is, van rond 0,3 dat het matig is en van rond 0,5 dat groot is (Cohen, 1992).
Voor de Mann-Whitney U toets vind je de volgende getallen bij de resultaten: de mediaan van de scores van beide groepen, eventueel het gemiddelde of de mediaan van de rangnummers van beide groepen, het getal *U*, de *p*-waarde en eventueel de effectgrootte *r*. Net als bij de *t*-toets wil je kijken naar de *p*-waarde, die aangeeft of er een statistisch significant verschil is, en wil je *r* bekijken (als die gegeven wordt) om te zien hoe groot of belangrijk het verschil is. Om meer inzicht te krijgen in de aard van het verschil tussen de groepen, bekijk je de medianen en/of rangnummers.

In het onderzoek van Van de Geer e.a. (2017) zijn onder andere Mann-Whitney U toetsen gebruikt.

Met deze toetsen hebben de onderzoekers onderzocht of een groep patiënten die in het ziekenhuis lag nadat de zorgverleners van die afdelingen getraind waren in spirituele zorg (de interventiegroep) verschilde van patiënten die op diezelfde afdelingen lagen voordat de zorgverleners getraind waren en van patiënten die op afdelingen lagen waar de zorgverleners niet getraind waren in spirituele zorg (de controlegroep). De eigenschappen die getoetst zijn, zijn de scores op verschillende lichamelijke klachten, spirituele noden en de evaluatie van aandacht voor levensvragen door de zorgverlener.

Verschillen tussen groepen worden in kwantitatief onderzoek vastgesteld met behulp van statistische toetsen

De onderzoekers leggen niet uit waarom ze hier de Mann-Whitney U toets gebruiken. Het meest aannemelijk is dat de scores niet een normaalverdeling volgden, maar dan is het niet passend dat in tabel 5 gemiddelde scores en standaarddeviaties worden gerapporteerd en niet de medianen. Gemiddelde scores en standaarddeviaties zijn namelijk alleen betekenisvol als de scores ongeveer een normaalverdeling volgen, vandaar ook dat *t*-toetsen niet toegepast worden in zo'n geval; *t*-toetsen maken immers gebruik van het gemiddelde en de standaarddeviatie.

Wat betreft het resultaat van de toetsen geven de onderzoekers weinig informatie. Ze rapporteren in tabel 5 enkel de *p*-waarden en in de tekst wordt geen toelichting gegeven bij toetsen die niet statistisch significant waren. Van de toetsen over verschillen in lichamelijke klachten was alleen voor slaapkwaliteit de *p*-waarde kleiner dan 0,05, namelijk $p = ,020$. Bij de toetsen over de evaluatie van aandacht voor levensvragen was de *p*-waarde voor de mate aandacht van zorgverleners voor levensvragen 0,008. Er is een kans dat de groepen ook verschilden in het belang dat ze hechten aan aandacht van zorgverleners voor levensvragen, want de *p*-waarde voor die toets was $p = ,056$. Er leek geen verschil tussen de groepen in spirituele noden. Op basis van de gerapporteerde gemiddelden zou dit betekenen dat de interventiegroep, vergeleken met de controlegroep, een betere slaapkwaliteit rapporteerde, meer aandacht van zorgverleners voor levensvragen ervoer en misschien meer belang hechtte aan aandacht van zorgverleners voor levensvragen.

ANOVA

Waar je met de *t*-toets, Wilcoxon signed-rank toets en Mann-Whitney U toets maar twee groepen of meetmomenten met elkaar kunt vergelijken, kun je met een Analaysis of Variance (ANOVA) meerdere groepen of meetmomenten vergelijken. Bij meerdere groepen gebruik je een 'one way'-ANOVA, terwijl je bij meerdere meetmomenten een 'repeated measures'-ANOVA gebruikt. Misschien wil je bijvoorbeeld niet alleen mensen die wel en geen geestelijke verzorging hebben gekregen met elkaar vergelijken, maar wil je mensen die geestelijke verzorging hebben gekregen vergelijken met mensen die begeleiding hebben gehad van een psycholoog, verpleegkundige of maatschappelijk werker of die geen begeleiding hebben gekregen. Van de Geer e.a. (2017) hebben een ANOVA gebruikt om te controleren of de vier groepen in hun onderzoek verschilden in gemiddelde leeftijd. Dit bleek niet het geval ($p = ,632$).

Net als bij een *t*-toets kan een ANOVA alleen gebruikt worden om scores op een schaalvariabele te vergelijken en zijn de resultaten van een ANOVA het meest betrouwbaar als de scores ongeveer een normaalverdeling volgen. Daarnaast moet bij een ANOVA de spreiding van de scores in alle groepen vergelijkbaar zijn. In de paragraaf Resultaten in wetenschappelijke artikelen vind je ook ongeveer dezelfde informatie terug als bij de *t*-toets, namelijk de gemiddelden en standaarddeviaties, de toetswaarde, de *p*-waarde en eventueel een effectgrootte. De effectgrootte wordt weergegeven met eta-kwadraat (η^2). Als deze rond 0,01 is, betekent dit dat de verschillen klein zijn. Is ze rond de 0,06, dan zijn de verschillen matig. En is ze rond de 0,14, dan zijn de verschillen groot (Cohen, 1988).

Als een *F*-waarde statistisch significant is (dus $p <$,05), dan betekent dit dat er een verschil is tussen de groepen of meetmomenten, maar het is dan nog niet duidelijk welke groepen dan van elkaar verschillen. Daarom worden er bij een ANOVA vaak zogenaamde post-hoctoetsen gedaan. Bij een 'one way'-ANOVA wordt hiervoor meestal de Tukey Honestly Significant Difference-toets (TukeyHSD) gebruikt.[6] In deze toets worden alle mogelijke paren van groepen met elkaar vergeleken in iets als een serie van *t*-toetsen (dus geestelijke verzorging met psycholoog, met verpleegkundige, met maatschappelijk werk, met niets; psycholoog met geestelijke verzorging, met verpleegkundige, met maatschappelijk werk, met niets, enzovoort). Daarbij wordt er statistisch rekening mee gehouden dat er heel vaak dezelfde soort toets wordt uitgevoerd, zodat je minder 'kanskapitalisatie' krijgt. Dat wil zeggen dat de kans steeds groter wordt dat je per toeval een statistisch significant groepsverschil vindt, in plaats van een 'werkelijk' groepsverschil. Kanskapitalisatie is een belangrijke valkuil in onderzoek en dus ook een om alert op te zijn bij het lezen van wetenschappelijke artikelen. Het betekent namelijk dat de resultaten van het onderzoek minder betrouwbaar zijn en je ze dus niet zomaar kunt toepassen in de praktijk.

> Het is belangrijk dat je als lezer kritisch naar statistische toetsen kunt kijken

Bij een post-hoctoets wordt per vergeleken paar groepen of meetmomenten het gemiddelde verschil in scores gerapporteerd ('mean difference' ofwel *MD*) en de *p*-waarde. Bij een negatieve *MD* (met minteken) is de score van de tweede groep gemiddeld hoger dan de score van de eerste groep. Bij een positieve *MD* (zonder minteken) is de score van de tweede groep gemiddeld lager dan de score van de eerste groep. Bij een *p*-waarde kleiner dan 0,05 is dit een betekenisvol verschil. Zo kun je dus achterhalen tussen welke groepen het verschil aanwezig is.

Kruskall-Wallis toets

De Kruskall-Wallis toets is het non-parametrische alternatief voor de 'one way'-ANOVA.[7] Voor post-hoctoetsen wordt de Mann-Whitney-U-toets gebruikt. Net als de Wilcoxon signed-rank toets en Mann-Whitney U toets, maakt de Kruskall-Wallis toets gebruik van de rangorde van de scores. Ook hier worden dus de mediane scores van de groepen op de gemeten eigenschap gerapporteerd, samen met de groepsgrootte, de *H*-waarde, de *p*-waarde en eventueel de effectgrootte η^2.
Bij de post-hoctoets worden de gemiddelde rangnummers en de *p*-waarden gerapporteerd. Net als bij de andere toetsen wil je hier controleren of de *p*-waarde lager is dan 0,05, hoe groot de effectgrootte is en wat de medianen zijn, zodat je de aard van het verschil kunt controleren. Van de Geer e.a. hebben de Kruskall-Wallis toets gebruikt om te onderzoeken of de vier groepen in hun onderzoek verschilden in hun spirituele attitudes. Er werden geen statistisch significante verschillen gevonden.

Chi²-toets

Een laatste toets die veel gebruikt wordt om groepsverschillen mee te onderzoeken, is de Chi-kwadraattoets (Chi² of χ^2). De 'Chi²-toets for goodness of fit' wordt gebruikt voor het vergelijken van een groep met een referentiegetal (zoals bij de 'one sample'-*t*-toets) en de 'Chi²-toets voor onafhankelijkheid' wordt gebruikt voor het vergelijken van twee of meer onafhankelijke groepen (zoals bij de 'independent samples'-*t*-toets en bij de 'one way'-ANOVA). De Chi²-toets wordt gebruikt om groepen te vergelijken op eigenschappen die in twee of meer categorieën verdeeld zijn die niet logisch geordend kunnen worden, bijvoorbeeld geslacht of denominatie (categorische variabelen).[8] Het toeschrijven van een getal of score aan deze categorieën om ze in een statistische toets te kunnen gebruiken, is volledig willekeurig. Overigens kan de Chi²-toets ook worden gebruikt als de antwoordcategorieën wel geordend kunnen worden, zoals bij opleidingsniveau of bij mate van tevredenheid, maar vaak worden dan de genoemde non-parametrische toetsen gebruikt.

Bij de Chi²-toets for goodness of fit wordt de verdeling van de groep over de gemeten eigen-

Toets	Wanneer toepassen	Voorwaarden	Minimaal gerapporteerde getallen
***t*-toets**	Vergelijking van twee groepen of twee meetmomenten van één groep	De te vergelijken eigenschap is gemeten op een schaal De scores volgen ongeveer een normaalverdeling	*M* = te vergelijken gemiddelden *SD* = spreiding van de scores t_{df} = effectgrootte *p* = statistische significantie
Wilcoxon signed-rank toets	Vergelijking van twee afhankelijke groepen of gemiddelden		Mediane scores *Z* = toetswaarde *p* = statistische significantie
Mann-Whitney U toets	Vergelijking van twee onafhankelijke groepen		Mediane scores *U* = toetswaarde *p* = statistische significantie
ANOVA	Vergelijking van meerdere groepen of meetmomenten	De te vergelijken eigenschap is gemeten op een schaal De scores volgen ongeveer een normaalverdeling De spreiding van de scores in de groepen is vergelijkbaar	*M* = te vergelijken gemiddelden *SD* = spreiding van de scores $F_{df,df}$ = effectgrootte *p* = statistische significantie Bij post-hoc: *MD* = gemiddeld verschil *p* = statistische significantie
Kruskall-Wallis toets	Vergelijking van meerdere onafhankelijke groepen		Mediane scores *H(df)* = toetswaarde *p* = statistische significantie Bij-post hoc: Gemiddelde rangnummers *p* = statistische significantie
Chi²-toets	Vergelijking van twee of meer onafhankelijke groepen	De te vergelijken eigenschappen zijn gemeten in categorieën Niet meer dan 20 procent van de cellen heeft minder dan 5 verwacht aantal deelnemers	Kruistabellen χ^2_{df} = toetswaarde *p* = statistische significantie

Tabel 1. Overzicht van de besproken toetsen voor groepsvergelijking met hun eigenschappen.

schap vergeleken met een bekende verdeling. Als we bijvoorbeeld weten dat ongeveer 60 procent van de Nederlandse bevolking zichzelf niet tot een religieus genootschap rekent en ongeveer 40 procent wel, dan kunnen we toetsen of deze verdeling in een groep geestelijk verzorgers vergelijkbaar is. Om dit te bepalen, wordt er een tabel gemaakt met de geobserveerde aantallen mensen in elk van de categorieën (dus het aantal geestelijk verzorgers dat zich wel of niet tot een religieus genootschap rekent) en de verwachte aantallen op basis van de verdeling in de Nederlandse populatie. Bij een *p*-waarde kleiner dan 0,05 wijkt de verdeling binnen de groep geestelijk verzorgers statistisch significant af van de verdeling in de Nederlandse populatie. Bij de Chi²-toets for goodness of fit worden naast de tabel met geobserveerde en verwachte aantallen, alleen de toetswaarde χ^2 en de *p*-waarde gerapporteerd.

Bij de Chi²-toets voor onafhankelijkheid worden kruistabellen gemaakt, met de categorieën van de groepen in de rijen en de categorieën van de eigenschap in de kolommen. In de cellen in de tabel staat hoeveel deelnemers aan beide categorieën voldoen. Vervolgens wordt bepaald of het aantal mensen in de cellen overeenkomt met

het verwachte aantal, als er geen relatie zou zijn tussen de twee eigenschappen. Zo kunnen we bijvoorbeeld onderzoeken of mensen die geestelijke verzorging ontvangen vaker een hoger opleidingsniveau hebben dan mensen die geen geestelijke verzorging ontvangen. Is de *p*-waarde kleiner dan 0,05 dan betekent dit dat er een relatie bestaat tussen het ontvangen van geestelijke verzorging en opleidingsniveau. Het is dan nog niet duidelijk wat die relatie precies is: hebben mensen die geestelijke verzorging ontvangen een hoger of lager opleidingsniveau dan mensen die het niet hebben ontvangen? Dat moeten de onderzoeker en lezer zelf uit de tabellen opmaken. Die moeten dus ook altijd in een wetenschappelijk artikel worden opgenomen. Naast de tabellen worden bij de Chi²-toets voor onafhankelijkheid gerapporteerd: de toetswaarde χ^2, de *p*-waarde en eventueel een effectgrootte *w*.

Een voorwaarde voor beide soorten Chi²-toetsen is dat in niet meer dan 20 procent van de cellen in de kruistabel minder dan vijf deelnemers verwacht worden. Als er niet aan die voorwaarde voldaan wordt, zijn er alternatieven beschikbaar; deze zal ik hier niet verder toelichten, omdat ze niet vaak gebruikt worden.

Begrippenlijst

Categorische variabele. Een gemeten eigenschap die opgedeeld is in categorieën, waarbij de categorieën voorzien zijn van een cijfer om ze te identificeren (bijvoorbeeld: geslacht).

Controlegroep. Een groep met dezelfde eigenschappen als de interventiegroep, maar waarbij niet of op een andere manier ingegrepen wordt op de eigenschap van interesse.

Effectgrootte. Een statistische grootheid die informatie geeft over het belang van een groepsverschil of effect. In de meeste gevallen geeft een effectgrootte van rond 0,1 een klein effect aan, van rond 0,3 een matig effect en van rond 0,5 een groot effect.

Interventiegroep. Een groep waarbij op een specifieke manier ingegrepen wordt op de eigenschap van interesse om een verandering teweeg te brengen.

Kanskapitalisatie. Een statistisch fenomeen, waarbij de kans steeds groter wordt dat een toets per toeval statistisch significant is naarmate de toets vaker herhaald wordt op dezelfde data.

***N* of *n*.** De grootte van de gehele populatie (*N*) of de steekproef (*n*) van de groep waar onderzoek naar gedaan wordt. Meestal is de populatiegrootte echter onbekend en wordt *N* gebruikt voor de hele steekproef en *n* voor subgroepen in de steekproef.

Non-parametrische toets. Een statistische toets waarvoor scores geen normaalverdeling hoeven te volgen.

Normaalverdeling. Een verdeling van scores waarbij de meeste scores rond het gemiddelde liggen en steeds minder scores daar verder vanaf liggen.

Nulhypothese. Een aanname over het verband dat getoetst wordt (bijvoorbeeld: er is geen groepsverschil of er is geen relatie tussen eigenschappen), waartegen de gemeten getallen worden afgezet in een statistische toets.

Schaalvariabele. Een eigenschap die gemeten kan worden op een oplopende kwantitatieve schaal (bijvoorbeeld: leeftijd).

Standaarddeviatie. Een maat voor de spreiding van de scores in een steekproef.

Statistische significantie. De waarschijnlijkheid dat de nulhypothese verworpen wordt, onder de aanname dat de nulhypothese waar is. Dit wordt uitgedrukt met *p*. Bij $p < 0{,}05$ wordt aangenomen dat de nulhypothese 'veilig' verworpen kan worden en er waarschijnlijk sprake is van een groepsverschil of een relatie tussen eigenschappen.

Statistische toets. Een berekening waarmee wordt onderzocht in welke mate de verzamelde gegevens afwijken van de nulhypothese. Er zijn twee grote families van toetsen: toetsen voor groepsverschillen en toetsen voor relaties tussen eigenschappen (correlaties).

Steekproef. De deelnemers aan het onderzoek die een grotere groep (de populatie) met dezelfde eigenschappen vertegenwoordigen.

Tot besluit

In tabel 1 staat een overzicht van de besproken toetsen met hun belangrijkste eigenschappen. Als lezer van een wetenschappelijk artikel is het belangrijk om te controleren of de gekozen toets past bij de vraag die de onderzoekers stellen en of de resultaten die in de tabellen staan passen bij de conclusie. Dit artikel kan een handige referentie vormen voor een aantal van deze toetsen. Ook is er in dit artikel enige terminologie geïntroduceerd die veel gebruikt wordt in kwantitatief onderzoek. Deze is terug te vinden in het onderstaande kader. Dat helpt ook om sneller te begrijpen welke keuzes onderzoekers gemaakt hebben.

Naast de hier besproken toetsen voor groepsvergelijkingen is er echter nog een familie toetsen die veel gebruikt wordt: de correlationele toetsen. Die staan centraal in het vijfde en laatste deel van deze serie.

Dr. A. Visser is universitair docent geestelijke verzorging aan de Rijksuniversiteit Groningen. E-mail: a.visser-nieraeth@rug.nl.

Literatuur

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2e dr.). New York: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112* (1), 155-159.

Geer, J. van de, Groot, M., Andela, R., Leget, C., Prins, J., Vissers, K. & Zock, H. (2017). Training hospital staff on spiritual care in palliative care influences patient-reported outcomes: results of a quasi-experimental study. *Palliative Medicine, 31* (8), 743-753.

Noten

1. Dit artikel is (gratis) te downloaden via https://journals.sagepub.com/doi/pdf/10.1177/0269216316676648?casa_token=UxMZB2NSVtYAAAAA:RK63yRWbXPO3cIqB2bCvCtA6x9WWGcEniIJtoSSpHwvvJ9CUmtwbKVYNkr5FXv-LjsC402CFasA

2. Dit wordt een interval- of ratiomeetniveau genoemd. Bij eigenschappen op een ratiomeetniveau is er sprake van een absoluut nulpunt (bijvoorbeeld: een lengte van 0 is de afwezigheid van lengte). Bij eigenschappen op een intervalmeetniveau is er geen absoluut nulpunt (bijvoorbeeld: een temperatuur van 0 °C is niet de afwezigheid van temperatuur).

3. Afbeelding door MADe, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1788643

4. Omdat een *p*-waarde altijd kleiner is dan 1, wordt de 0 meestal weggelaten in notitie. Er wordt dan dus *p* = ,013 geschreven, in plaats van *p* = 0,013.

5. Dit wordt een ordinaal meetniveau genoemd. Een ander voorbeeld hiervan is tevredenheid gecodeerd als 1 = erg ontevreden, 2 = enigszins ontevreden, 3 = niet tevreden/niet ontevreden, 4 = enigszins tevreden, 5 = erg tevreden. Het is bij dit soort variabelen niet helemaal duidelijk hoeveel meer er is van een eigenschap als iemand hoger scoort en er hadden ook andere getallen gekozen kunnen worden om de antwoorden mee te coderen, maar er is wel sprake van een logische volgorde in de scores. Er is, met andere woorden, een rangvolgorde.

6. Bij een 'repeated measures'-ANOVA moet de onderzoeker zelf een serie 'paired sample'-*t*-toetsen uitvoeren en een statistische correctie doen op de *p*-waarde tegen kanskapitalisatie.

7. Een non-parametrisch alternatief voor de 'repeated measures'-ANOVA is de Friedmans ANOVA. Deze wordt echter nauwelijks gebruikt en zal ik hier dus niet toelichten.

8. Dit wordt een nominaal meetniveau genoemd.