Podosophia (2018) 26:28–33
https://doi.org/10.1007/s12481-018-0187-8

De *t*-toets en de analysis of variance, ANOVA

Sil Aarts · Eveline Wouters

Published online: 9 February 2018

Samenvatting Bij toetsende statistiek wordt met behulp van een statistische berekening een hypothese getoetst. Bijvoorbeeld om te bepalen of het verschil tussen groepen op toeval berust, of niet. De keuze welke statistische toets van toepassing is op de gestelde onderzoeksvraag, is afhankelijk van verscheidene factoren, zoals het meetniveau van de variabelen, de (on)afhankelijkheid van de data (onafhankelijke groepen zoals man *vs.* vrouw, of afhankelijke data, zoals voormeting *vs.* nameting bij dezelfde patiënten) en de normaliteit van de data. Dit artikel zal dieper ingaan op een van de meest gebruikte toetsen in de statistiek: de *t*-toets en de ANOVA (*analysis of variance*). Het basisidee van een *t*-toets en de ANOVA is na te gaan of gemiddelden van groepen gelijk zijn of van elkaar verschillen. Om een *t*-toets of ANOVA te kunnen uitvoeren, dient de afhankelijke variabele, de uitkomstmaat, van ratio- of intervalniveau te zijn.

Trefwoorden *t*-toest · ANOVA · variantieanalyse

Inleiding

In Podosophia nr. 3, 2017 stond de Chi-squaretoets centraal [1]. Deze toets kan alleen worden gebruikt indien de variabelen van nominaal, dichotoom of ordinaal meetniveau zijn (zie kader 'Meetniveau van variabelen'). Wanneer de variabelen uit een onderzoek echter van interval- en rationiveau zijn, gebruiken we *t*-toetsen of ANOVA.

Verschillende *t*-toetsen

Er zijn verschillende soorten *t*-toetsen, waarvan de ongepaarde *t*-toets en de gepaarde *t*-toets (Engels: *t*-test) de meest gebruikte zijn. Hierna lichten we de drie *t*-toetsen verder toe.

T-toets voor één steekproef

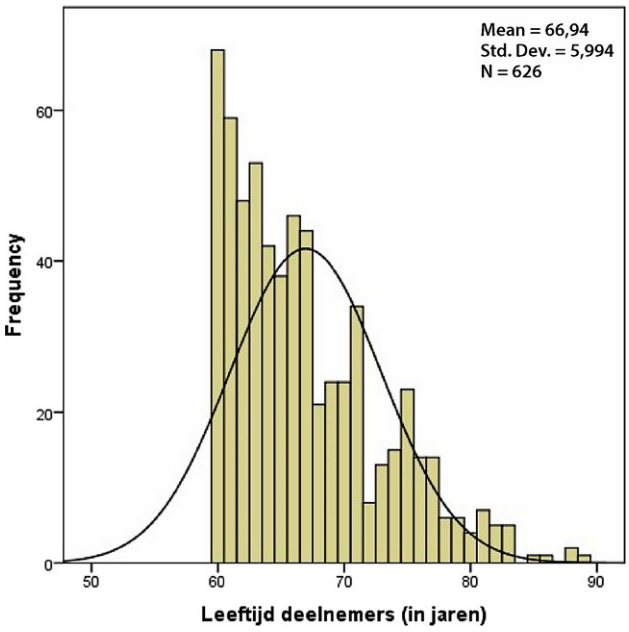
Deze toets, ook wel de *one sample t-toets* genoemd, is bedoeld voor één groep waarbij wordt getoetst of het populatiegemiddelde afwijkt van een bepaalde waarde.

Meetniveau van variabelen

- Nominale variabelen zijn variabelen die slechts een naam hebben, een 'benoeming' (bijv. land van herkomst: Europa/Azië/Amerika).
- Dichotome variabelen zijn variabelen met een naam, een 'benoeming', maar met slechts twee waarden (bijv. ja *vs.* nee).
- Ordinale variabelen beschrijven variabelen met een naam, een benoeming, maar mét een natuurlijke rangorde (bijv. opleidingsniveau: mavo/havo/vwo).
- Interval- en ratiovariabelen zijn variabelen met numerieke waarden (bijv. graden Celsius, leeftijd in jaren, lengte in cm, score van 0 tot 100 op een vragenlijst).

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.
Alle informatie is te vinden in diverse handboeken, zoals die van Field [3] en Howell [4].

S. Aarts, PhD (✉) · E. Wouters, PhD MD
Health Innovations & Technology, Fontys Paramedische Hogeschool, Eindhoven, Nederland
s.aarts@fontys.nl



Figuur 1 Histogram voor de verdeling van leeftijd van 626 deelnemers. Gemiddelde = 66,94, Standaarddeviatie = 5,994

Voorbeeld: Een tevredenheidsvragenlijst wordt in diverse podotherapeutische praktijken afgenomen en het gemiddelde van al die praktijken is een 7. Een podotherapeut wil onderzoeken of patiënten hoger of lager scoren dan het algemeen gemiddelde.

H_0 Het gemiddelde van de steekproef is 7 (er is geen verschil tussen het populatiegemiddelde en de steekproef).

H_1 Het gemiddelde van de steekproef is niet 7 (er is wel een verschil tussen het populatiegemiddelde en de steekproef).

Deze toets wordt in de praktijk slechts sporadisch gebruikt, omdat onderzoekers meestal geen idee hebben welke populatiegemiddelde ze moeten gebruiken om de H_0 te formuleren.

T-toets voor twee steekproeven

Ongepaarde t-toets

De ongepaarde *t*-toets, ook wel onafhankelijke *t*-toets (Engels: *independent samples-test*), is bedoeld voor twee groepen waarbij wordt getoetst of het populatiegemiddelde van de ene groep afwijkt van het populatiegemiddelde van de andere groep.

Voorbeeld: Er zijn twee podotherapeutische praktijken (bijv. patiënten van praktijk A versus patiënten van praktijk B) en men wil weten of, en zo ja, in hoeverre, de tevredenheid in deze praktijken van elkaar verschilt.

H_0 Er is geen verschil tussen de praktijken.

H_1 Er is wel een verschil tussen de praktijken.

Gepaarde t-toets

De gepaarde *t*-toets (Engels: *dependent samples t-test*) is bedoeld voor één groep; één groep wordt vergeleken op twee verschillende momenten.

Voorbeeld: Er wordt een interventie voor patiënten aangeboden. Men wil weten of de patiënten na de therapie minder pijn ervaren dan vóór de toepassing van de interventie. Op twee momenten vullen de patiënten een pijnvragenlijst in; een vóór de start van de interventie (meting 1) en een ná de interventie (meting 2).

H_0 Er is geen verschil tussen meting 1 en meting 2.

H_1 Er is wel een verschil tussen meting 1 en meting 2.

Voorwaarden

Voordat een *t*-toets kan worden uitgevoerd, dient te worden nagegaan of er aan bepaalde belangrijke voorwaarden wordt voldaan.

1. Aselecte steekproef uit een normale verdeling

Ten eerste dient de getrokken streekproef een aselecte steekproef te zijn uit een normale verdeling. Dat betekent dat alle deelnemers die tot een bepaalde populatie behoren een even grote kans hebben om in de steekproef terecht te komen. De steekproef dient een natuurgetrouwe afspiegeling van de hele populatie te zijn. In statistische termen wordt dit ook wel 'normaal verdeeld' genoemd; de steekproefwaarden dienen bij benadering normaal verdeeld te zijn. Om na te gaan of aan die voorwaarde is voldaan, wordt meestal een histogram met bijbehorende normaalcurve opgevraagd in SPSS (zie fig. 1; [2]). Mochten de data niet normaal verdeeld zijn, dan wordt voor een andere toets gekozen. De *Wilcoxon signed rank toets* kan dan gebruikt worden als alternatief voor de gepaarde *t*-toets; de *Mann-Whitney-U-toets* wordt vaak gebruikt als alternatief voor de ongepaarde *t*-toets. Beide toetsen veronderstellen geen normaal verdeelde data.

2. Homogeniteit van varianties

Ten tweede geldt de voorwaarde van homogeniteit van varianties. In het geval van twee onafhankelijke steekproeven (bijvoorbeeld twee onafhankelijke groepen: mannen en vrouwen) dienen bij toepassing van de standaard *t*-toets de beide steekproeven, bij benadering, dezelfde variantie te hebben (zie het kader 'Spreiding, standaarddeviatie en variantie'). Met andere woorden, de spreiding van de getallen in de twee groepen moeten bij benadering even groot zijn. Een vuistregel hierbij is dat de grootste standaarddeviatie niet groter mag zijn dan tweemaal de kleinste standaarddeviatie. In de SPSS-output (bijvoorbeeld tab. 1 en 2) zijn dan altijd twee rijen te zien: *Equal variances*

Tabel 1 Onafhankelijke *t*-toets voor de relatie tussen geslacht en tevredenheid: beschrijvende statistiek (aangepaste SPSS-tabel)

	Group statistics				
	Gender	*N*	Mean	Std. deviation	Std. error mean
Tevredenheid	Man	108	5,74	0,586	0,056
	Vrouw	105	5,70	0,798	0,078

N aantal deelnemers, *Mean* gemiddelde van de tevredenheid van de deelnemers, *Std. devation* standaarddeviatie, *Std. error mean* standaardfout van het gemiddelde

Tabel 2 Onafhankelijke *t*-toets voor de relatie tussen geslacht en tevredenheid (aangepaste SPSS-tabel)

		Levene's test for equality of variances		*t-test for Equality of Means*						
		F	Sig.	t	df	Sig. (2-tailed)	MD	Std. error difference	95%-CI	
									Lower	Upper
Tevredenheid	Equal variances assumed	1,788	0,183	0,475	211	0,635	0,046	0,096	–0,143	0,234
	Equal variances not assumed			0,473	1900,671	0,637	0,046	0,096	–0,144	0,235

CI Confidence interval of the difference, *MD* Mean difference, *Std. error mean* standaardfout van het gemiddelde

Tabel 3 ANOVA van de relatie tussen provincie en leeftijd van patiënten in de podotherapeutische praktijk (aangepaste SPSS-tabel)

ANOVA					
Leeftijd patiënten					
	Sum of Squares	df	Mean Square	F	Sig
Between Groups	1600,000	2	800,000	25,097	0,000
Within Groups	6693,953	210	31,876		
Total	8293,953	212			

assumed en *Equal variances not assumed.* In de kolom *Levene's Test for Equality of Variances* wordt getoetst of de standaarddeviaties inderdaad als even groot kunnen worden benaderd. Bij $p > 0{,}05$ worden gelijke varianties aangenomen (*equal variances*); bij $p \leq 0{,}05$ worden ongelijke varianties aangenomen (*unequal variances*).

Wat analyseert een *t*-toets?

In tab. 1 en 2 is de SPSS-output van een ongepaarde sample *t*-test weergegeven (verkregen met behulp van het statistiekprogramma SPSS [2]). De onderzoeksvraag 'Verschillen de scores tussen de twee groepen?' wordt in statistische termen vertaald en uitgevoerd: de onafhankelijke variabele 'groep' (een nominale variabele) en de afhankelijke variabele 'score vragenlijst' (een numerieke variabele). De vraag die met de *t*-toets beantwoord wordt is: 'Hoeveel van de verschillen in de scores op de afhankelijke variabele worden verklaard doordat de scores uit twee verschillende groepen afkomstig zijn?' Er wordt dus berekend hoeveel procent van de totale variantie kan worden verklaard door de variabele 'groep'. Dit noemt men ook wel de *verklaarde variantie.*

Wat betekenen de uitkomsten van deze toets?

Voordat we ons richten op de interpretatie van de resultaten van deze output, eerst de belangrijkste begrippen uit deze tabel. *T-test for Equality of Means* duidt op de ongepaarde *t*-toets: de toets om na te gaan of gemiddelden gelijk zijn. *Sig. (2-tailed)* duidt op de *p*-waarde. De *p*-waarde is, in de betreffende SPSS-output, de kans dat het gevonden verschil tussen mannen en vrouwen in hun tevredenheid berust op toeval. *Mean difference* rapporteert het verschil tussen mannen en vrouwen in hun tevredenheid: 5,74–5,70 = 0,046 (afgerond). De *St. error difference* laat de standaardfout van dit het verschil zien. Het *95% confidence interval* (hierna afgekort tot CI) *of the difference* staat voor het betrouwbaarheidsinterval van het verschil (ofwel de *mean difference*). Bij het 100 keer herhalen van dit onderzoek, verwachten we dat in 95% van de gevallen (dus in 95 van deze 100 onderzoeken) de *mean difference* binnen het CI zal vallen (dus tussen de getallen die te vinden zijn in '*Lower*' en '*Upper*').

In tab. 3 en 4, ook een SPSS-output, geeft de *Levene's Test for Equality of Variances* een *p*-waarde van 0,183 ($p = 0{,}183$). Aangezien de *p*-waarde groter is dan 0,05 kunnen we ervan uitgaan dat de twee varianties bij benadering gelijk zijn. De bovenstaande rij van

Juiste steekproefgrootte

Voordat men een statistische analyse uitvoert, dient men vast te stellen hoe groot de steekproef moet zijn. Veelal is de gedachte: hoe meer deelnemers, hoe beter. Dit is onjuist. Als er te veel deelnemers worden geworven, brengt dit niet alleen extra (onnodige) onderzoekskosten met zich mee, maar zijn mogelijk alle relaties die worden geanalyseerd statistisch significant (NB Statistische significantie is veelal afhankelijk van de *n* van een steekproef, dit is het aantal deelnemers. Meer hierover in een volgend artikel). Tevens is het niet ethisch om meer deelnemers te vragen dan nodig (waarom meer mensen deel laten nemen dan voor het onderzoek noodzakelijk is?). Te weinig deelnemers is echter ook onwenselijk; dit geeft geen enkele mogelijkheid om de juiste conclusies te trekken. Voordat deelnemers worden geworven, is dan ook een steekproefberekening nodig. Er zijn veel websites die, veelal gratis, deze berekening, aanbieden; een veel gebruikte website is (http://www.gpower.hhu.de/).

Spreiding, standaarddeviatie en variantie

De standaarddeviatie, ook wel standaardafwijking genoemd, is een getal dat de spreiding van getallen (waarden) aangeeft rondom het gemiddelde. Hoe groter de spreiding, hoe groter de standaarddeviatie, hoe meer de verkregen waarden (van de deelnemers) onderling verschillen. De variantie is ook een maat voor de spreiding van getallen. Een grote variantie duidt op veel verschillen in gekregen waarden, terwijl een kleine variantie duidt op meer eenheid in de gekregen waarden. Als er helemaal geen variatie in een getallenreeks zit, dan is de variantie nul; alle getallen zijn dan hetzelfde en komen dus overeen met het gemiddelde. Daar de variantie is gedefinieerd als het kwadraat van de standaarddeviatie, zijn de twee termen sterk aan elkaar gerelateerd.

Voorbeeld

We bekijken de volgende waarden: 4, 5, 7, 3 en 6. De minimumscore is 3, de maximumscore is 7. Het gemiddelde is $(4+5+7+3+6)\ /\ 5=5$. De variantie is dan $((4-5)^2+(5-5)^2+(7-5)^2+(3-5)^2+(6-5)^2)\ /\ 5=(1+0+4+4+1)\ /\ 5=2$. Als we de wortel van dit gemiddelde pakken, krijgen we de standaarddeviatie $=\sqrt{2}=1{,}41$.*

*Soms wordt de variantie en de standaarddeviatie gedeeld door *n*–1, en niet door *n*. Dit is een zuiverdere schatting dan delen door *n*. Meer informatie is te vinden in statistiekhandboeken.

de SPSS output ('equal variances assumed') kan dus worden gebruikt voor het interpreteren van de toets. Uitgaande van deze '*equal variances assumed*' zien we een *p*-waarde van 0,635 ($p=0{,}635$). De uitslag die hier wordt gevonden, is vrij groot. Als de kans zo groot is, dan is er weinig 'evidence' voor de H_1. Dit betekent dat H_1 moet worden verworpen en H_0 wordt aangenomen: er is geen *evidence* voor een verschil tussen mannen en vrouwen. De conclusie luidt dan ook dat er geen statistisch significant verschil is in de tevredenheid tussen mannen en vrouwen bij een significantieniveau van 5 %.

ANOVA ofwel variantieanalyse

De ongepaarde *t*-toets kan slechts twee groepen vergelijken. Stel, men wil nagaan of er een verschil is in de gemiddelde leeftijd van patiënten van podotherapeutische praktijken in de provincies Limburg, Zuid-Holland en Drenthe. De ongepaarde *t*-toets volstaat nu niet; er zijn immers drie provincies (dus drie groepen) te vergelijken. Een ANOVA biedt dan uitkomst. ANOVA staat voor *analysis of variance* (analyse van variantie) en is, in tegenstelling tot de *t*-test, in staat om twee of meer groepen met elkaar te vergelijken (NB Een ANOVA is gelijk aan een ongepaarde *t*-toets als slechts twee groepen worden vergeleken). De toets bevat hetzelfde principe als de onafhankelijke (ongepaarde) *t*-toets; gemiddelden worden vergeleken om hypotheses te toetsen.

Bij de ANOVA worden de volgende hypotheses geformuleerd:

H_0 Er is geen verschil tussen de provincies.
H_1 Er is wel een verschil tussen de provincies.

Er wordt een ANOVA uitgevoerd waarvan de SPSS-output te vinden is in tab. 3 en 4 (voor de leesbaarheid en begrijpelijkheid van dit artikel, zullen slechts de belangrijke begrippen uit deze tabel worden benoemd). De tabel is opgedeeld in *between-groups* (variantie tussen de groepen) en *within groups* (variantie binnen de groepen). Als de verhouding van de variantie tussen de groepen groot genoeg is ten opzichte van de variantie binnen de groepen zal een *p*-waarde kleiner dan 0,05 worden getoond. Met andere woorden, een *sum of squares* of *between groups* die groot is in verhouding tot de *sum of squares* of *within groups*, duidt op een statistisch significante relatie (*Sig.*).

Dit wordt ook duidelijk uit de *F* (F-toets) die deze verhouding berekent als *mean square between groups*/*mean square within groups*: 800 / 31,876 = 25,097. In tab. 3 en 4 geldt bij deze *F* een $p<0{,}001$ (een *p*-waarde die bijna nul is, wordt genoteerd als <0,001, aangezien een *p*-waarde nooit nul kan zijn). H_0 wordt dus verworpen; er is een statistisch significant verschil tussen de provincies.

Bonferonnie-correctie

Zoals te zien is in de geformuleerde hypothese van een ANOVA, geeft een ANOVA alleen aan of er wel of geen verschil is tussen de groepen; welke groep

Histogram met normaalcurve

Voordat een onderzoeker data gaat analyseren met statistische toetsen, wordt er veelal een histogram aangevraagd van de data. Een onderzoeker krijgt zo inzicht in hoe de data 'eruitzien'. Een algemeen patroon is de klokvormige lijn, die bekend staat als de normale verdeling. Bij een normale verdeling zijn de getallen aan de ene kant van het gemiddelde hetzelfde verdeeld als aan de andere kant van het gemiddelde. Een histogram kan dus de vorm van een normale verdeling hebben (de mooie klokvorm volgen), maar dat hoeft niet. De vorm van een histogram kan bijvoorbeeld ook scheef zijn. Dat is het geval in fig. 1. Je ziet daaraan dat in dit onderzoek de meeste deelnemers rond de 60 jaar zijn (bij die leeftijd zijn de staven van het histogram heel lang), terwijl slechts een heel klein aantal deelnemers 80 jaar is. De verdeling volgt dus niet mooi de klokvorm. Ook kan een verdeling tweetoppig zijn. Dit zou betekenen dat er twee pieken in het histogram zijn, bijvoorbeeld een piek bij 60 jaar en dan weer een piek bij 80 jaar. Ook dan volgt de verdeling geen klokvorm. Veelal is het moeilijk om te beoordelen of data normaal verdeeld zijn. Daar is ervaring voor nodig: hoe meer ervaring, hoe makkelijker het wordt. Er zijn nog meer mogelijkheden om normaliteit na te gaan, maar die vallen buiten de beschouwing van dit artikel.

Daarnaast is er de centrale limietstelling ('*central limit theorem*'). Deze limiet houdt in dat op het moment dat de steekproef groot is, we van normaliteit mogen uitgaan. Hierbij geldt als vuistregel: een steekproef dient uit minstens 30 participanten (of observaties) te bestaan. Het bekijken van normaliteit is dus met name van belang in kleine samples.

verschilt ten opzichte van de andere groepen (of verschillen ze allemaal?) wordt daaruit nog niet duidelijk. Door een zogenoemde post-hoctest uit te voeren, kan worden nagegaan welke groepen van elkaar verschillen (zie tab. 3 en 4). De post-hoctest die gekozen is in tab. 3 en 4 is de zogenoemde Bonferonni. In deze tabel worden de drie provincies met elkaar vergeleken. We bespreken de bovenste regel uit de tabel met de Bonferonni-correctie; de relatie tussen Noord-Holland (I) en Drenthe (J).

Links staat de provincie (I) die wordt vergeleken met een andere provincie (J). Het verschil tussen deze twee provincies in de gemiddelde leeftijd van de patiënten is te vinden onder *Mean Difference* (I – J). Het verschil in gemiddelde leeftijd tussen Noord-Holland en Drenthe is –2,819; het verschil tussen Drenthe en Noord-Holland is dan ook 2,819. *Std. error* staat voor standaardfout en toont hoe groot de standaarddeviatie van het steekproefgemiddelde is (hoe meer deelnemers een steekproef heeft, hoe kleiner de standaardfout is). *Sig* duidt op statistisch significantie (i.e. de *p*-waarde). Het *95% Confidence Interval* staat voor het betrouwbaarheidsinterval. Bij het 100 keer herhalen van dit onderzoek, verwachten we dat in 95% van de gevallen (dus in 95 van deze 100 onderzoeken) het gemiddelde (*mean difference*) in het CI zal vallen (dus tussen *lower bound* en *upper bound*).

Als we de gehele ANOVA-analyse bekijken, kunnen we dus concluderen dat er statistisch significante verschillen zijn tussen de drie provincies: Noord-Holland verschilt van Limburg, en Drenthe verschilt ook van Limburg. Drenthe en Noord-Holland verschillen daarentegen niet statistisch significant van elkaar.

Kanttekening bij deze toetsen

De *t*-toets en de ANOVA houden geen rekening met eventuele andere factoren die van invloed kunnen zijn op de uitkomstmaat. Als we bijvoorbeeld kijken naar de relatie tussen geslacht en tevredenheid, wordt er geen rekening gehouden met factoren zoals leeftijd, type aandoening van de patiënt, aantal behandelingen die de patiënten al hebben gehad en grootte van de podotherapeutische praktijk. Deze factoren kunnen echter ook een invloed hebben op de tevredenheid van patiënten van een podotherapeutische praktijk. Daarmee doen de *t*-toets en de ANOVA onvoldoende recht aan de werkelijkheid en di ent men op

Tabel 4 post-hocanalyse van de relatie tussen de provincie en leeftijd van patiënten in de podotherapeutische praktijk (aangepaste SPSS-tabel)

Multiple Comparisons

Dependent Variable: Leeftijd patiënten

Bonferroni (I) Provincie	(J) Provincie	Mean Difference (I – J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Noord-Holland	Drenthe	–2,819	1,188	0,056	–5,69	0,05
	Limburg	–12,245[a]	1,794	0,000	–16,57	–7,92
Drenthe	Noord-Holland	2,819	1,188	0,056	–0,05	5,69
	Limburg	–9,426[a]	1,476	0,000	–12,99	–5,86
Limburg	Noord-Holland	12,245[a]	1,794	0,000	7,92	16,57
	Drenthe	9,426[a]	1,476	0,000	5,86	12,99

[a]The mean difference is significant at the 0,05 level

te passen om, enkel op basis van deze testen, een conclusie te trekken. In het volgende artikel gaan we uitvoerig in op een analyse die meer recht doet aan de werkelijkheid, namelijk de regressieanalyse.

Take home message

- Er zijn drie verschillende *t*-toetsen.
- Een *t*-toets is bedoeld voor het analyseren van twee groepen (ofwel de ongepaarde *t*-toets), terwijl een ANOVA twee of meer groepen kan analyseren.
- Een Bonferonni-correctie kan uitkomst bieden om na te gaan welk van de groepen bij een ANOVA verschillen.

Literatuur

1. Aarts A, Wouters E. De chi-squaretoets. Podosophia. 2017;3:124–7.
2. IBM. SPSS Statistics for Windows. Version 21.0. IBM corp: Armonk, N.Y., USA.
3. Field A. Discovering statistics using SPSS. 3e druk. band 9. Londen: SAGE; 2009.
4. Howell DC. Statistical methods for psychology, international edition. 8e druk. Boston: Cengage Learning; 2012.

Sil Aarts, docent/onderzoeker

Eveline Wouters, lector health innovations and technology