Associatiematen: tot slot 9

In de vorige hoofdstukken zijn associatiematen behandeld naar meetniveau. Een onderzoeker moet een keuze maken uit de associatiematen, waarbij het meetniveau van de variabelen een belangrijk, maar niet het enige criterium is. In dit hoofdstuk bespreken we de overwegingen die een rol kunnen spelen bij het kiezen van een associatiemaat. In de regel geldt dat als twee variabelen een verschillend meetniveau hebben, je alleen de associatiematen kunt gebruiken die horen bij het laagste meetniveau. Daarop bestaat een uitzondering. In dit hoofdstuk bespreken we *eta,* een associatiemaat die je gebruikt als de afhankelijke variabele een interval- of rationiveau heeft en de onafhankelijke variabele nominaal is. In deze situatie hoeven we ons niet te beperken tot associatiematen op nominaal niveau.

9.1 Eta en eta-kwadraat

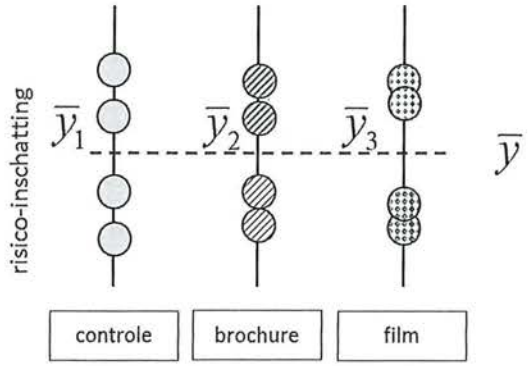
Eta is een associatiemaat die geschikt is voor een asymmetrische relatie, waarbij de onafhankelijke variabele een nominaal meetniveau heeft en de afhankelijke variabele een interval of ratio meetniveau. Eta hoort bij de *variantieanalyse.*[1] Bij het uitvoeren van een variantieanalyse worden de gemiddelden van verschillende groepen met elkaar vergeleken door te kijken naar de spreiding binnen en tussen de groepen.

9.1.1 Interpretatie

Als we spreken van een variantieanalyse, wordt gekeken naar de *gemiddelden* van verschillende groepen en naar de spreiding van die gemiddelden. Stel je voor dat je een experiment uitvoert om te kijken of de manier van voorlichten over de risico's van drugsgebruik invloed heeft op de inschatting van die risico's. De proefpersonen worden aselect (dat wil zeggen, op toevalsbasis) toegewezen aan drie groepen, waarvan de eerste groep een brochure te lezen krijgt over de risico's, de tweede groep een voorlichtingsfilm te zien krijgt en de derde groep dient als controlegroep die geen enkele voorlichting krijgt. Vervolgens wordt bij alle groepen dezelfde vragenlijst afgenomen, waaruit onder andere moet blijken in welke mate zij zich bewust zijn van de risico's die drugs met zich meebrengen. Deze variabele (die we hier 'risico-inschatting' noemen) is een gemiddelde intervalschaal (variërend van 0 tot 10) waarbij hoe hoger wordt

gescoord, hoe meer de proefpersonen zich bewust zijn van de mogelijke gevolgen van drugs. In hoofdstuk 10 (schaalconstructie) zullen we dieper ingaan op het maken van een dergelijke schaal. We hebben hier twee variabelen, waarvan de onafhankelijke variabele (de groep waarin de proefpersonen zijn ingedeeld, de conditie) nominaal is, en de afhankelijke variabele (de risico-inschatting) interval is. We kunnen nu per groep kijken hoe hoog de gemiddelde risico-inschatting is, en hoe goed de variantie in de conditie de variantie in de risico-inschatting kan verklaren.

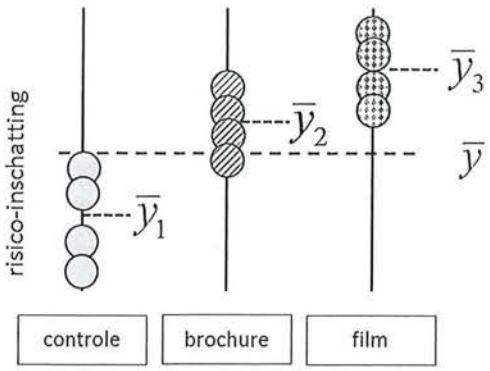
We schetsen hieronder twee mogelijke scenario's. In figuur 9.1 zie je de mogelijkheid dat de proefpersonen onderling veel van elkaar verschillen, maar dat tussen de groepen niet veel verschil is in de gemiddelde risico-inschatting. Zowel in de controlegroep, als in de groep die de brochure heeft gelezen, als in de groep die de film heeft gezien, zijn er proefpersonen (elk bolletje is een proefpersoon) die of hoog of laag scoren. We zouden aan de hand van dit figuur waarschijnlijk concluderen dat het niet uitmaakt of, en op welke manier, er voorlichting wordt gegeven over de risico's van drugsgebruik, omdat de gemiddelden in de drie groepen niet van elkaar verschillen en het totale gemiddelde hetzelfde is als de afzonderlijke groepsgemiddelden. We kunnen dus stellen dat er geen spreiding (geen variantie) is *tussen* de verschillende groepen. We zeggen ook wel: de *tussenvariantie* is nul. In het Engels wordt dit aangeduid met variantie *between groups*.



Figuur 9.1 Geen variantie tussen de groepen: gemiddelden gelijk

Een tweede mogelijkheid is dat de groep waarin een proefpersoon is ingedeeld wél effect heeft op de risico-inschatting, en dat de gemiddelden van de groepen van elkaar verschillen, zoals in figuur 9.2.

In dit figuur zie je dat de groepsgemiddelden wel van elkaar verschillen. De controlegroep scoort gemiddeld lager dan de experimentele groepen, en van de experimentele groepen scoort de groep die de film heeft gezien gemiddeld het hoogst. We kunnen ook zeggen dat er spreiding is tussen de groepsgemiddelden, en dat we aan de hand daarvan kunnen concluderen dat voorlichting geven een effect heeft op de risico-inschatting.



Figuur 9.2 Wel variantie tussen de groepen: verschillende gemiddelden

In de figuren 9.1 en 9.2 is te zien dat hoe hoger de variantie is tussen de groepen, hoe meer de gemiddelden van de verschillende groepen van elkaar verschillen. Ook is te zien dat in figuur 9.2, waar de gemiddelden inderdaad van elkaar verschillen, de variantie binnen de groepen kleiner is. Waar in figuur 9.1 binnen de groepen de scores ver van elkaar verwijderd waren, liggen deze scores binnen de groepen in figuur 9.2 dichter bij elkaar. De spreiding *binnen* de groepen wordt in het Engels aangeduid met de term *Within groups*, en in het Nederlands met *binnenvariantie* of de *onverklaarde variantie*. We kunnen nu beredeneren dat wanneer de tussenvariantie groot is en de binnenvariantie (de onverklaarde variantie) klein is, er inderdaad sprake is van verschillen tussen de groepsgemiddelden.

Dit komt overeen met wat we al eerder hebben gezien bij een regressieanalyse. Bij een regressieanalyse spraken we niet over *between* en *within*, maar werden de termen *regression* en *residual* gebruikt. We hebben ook gezien dat we voor de berekening van R^2 de totale variantie en de onverklaarde variantie nodig hebben. De totale variantie is bij een regressieanalyse het aantal voorspellingsfouten dat je maakt als je uitgaat van het gemiddelde als voorspeller, E_1. Bij een regressieanalyse hadden we deze informatie, samen met de E_2, nodig om de proportie verklaarde variantie (R^2) te kunnen berekenen. Dat kunnen we ook doen bij een variantieanalyse. Bij variantieanalyse noemen we de proportie verklaarde variantie *eta²*, wat ook wel wordt aangeduid met de Griekse letter η^2. De sterkte van de samenhang drukten we bij een regressieanalyse uit met β (die bij een enkelvoudige regressie gelijk was aan | r |). Ook bij een variantieanalyse krijgen we de sterkte van de samenhang: η, ofwel: *eta*, wanneer we de wortel nemen uit η^2.

De formule van eta² is dan ook niet anders dan de formule van R^2:

$$\eta^2 = \frac{E_1 - E_2}{E_1}$$

Formule voor eta²

Ook bij een variantieanalyse noemen we de totale variantie de E_1. Bij het berekenen van de totale variantie kijk je naar de mate waarin de gemiddelden afwijken van het totale gemiddelde. Onze afhankelijke variabele is minimaal interval, en we kunnen dus de informatie van het gemiddelde gebruiken om uit te rekenen hoeveel voorspellingsfouten we maken. Daarbij houden we nog geen rekening met de onafhankelijke variabele. Net als bij het berekenen van de E_1 bij een regressieanalyse, moeten we de verschillen kwadrateren voordat we ze bij elkaar optellen, omdat anders de som altijd op nul uitkomt.

Als we dat in formulevorm schrijven, krijgen we, net als bij een regressieanalyse:

$$E_1 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Formule voor E_1 bij eta²

Wanneer we wel rekening houden met de informatie van de onafhankelijke variabele, berekenen we E_2. Bij de regressievergelijking keken we dan naar de individuele waarnemingen ten opzichte van de voorspelde regressielijn. Bij de variantieanalyse kijken we per groep naar de individuele waarnemingen ten opzichte van dat groepsgemiddelde. Je kunt de waarde van y voorspellen door voor elke waarde van x apart het gemiddelde van y te berekenen en dat gemiddelde als voorspeller te gebruiken voor die speciale waarde van x. Voor die waarde van x is het aantal gemaakte fouten dan weer de kwadratensom van alle afwijkingen van dat gemiddelde. Als je dit voor elke waarde van x herhaalt (1 t/m k) en vervolgens deze kwadratensommen bij elkaar optelt, krijg je E_2. Dit is de onverklaarde variantie.

$$E_2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

Formule voor E_2 bij eta²

Wanneer je deze twee onderdelen samenvoegt, krijg je de uiteindelijke formule voor eta²:

$$\eta^2 = \frac{E_1 - E_2}{E_1} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Formule voor eta²

In paragraaf 9.2 zullen we laten zien dat dit ingewikkelder klinkt dan het in werkelijkheid is. Voordat we de berekening met de hand uitvoeren, kijken we eerst naar hoe dit er in SPSS uitziet. Wanneer we deze informatie uit SPSS halen, krijgen we een tabel waarin per groep de gemiddelden en de standaarddeviaties worden weergegeven, zoals in tabel 9.1, die grofweg overeenkomt met figuur 9.1:

Tabel 9.1 Vergelijken van groepsgemiddelden per conditie op de afhankelijke variabele risico-inschatting (SPSS-output)

Report

risico-inschatting

conditie	Mean	N	Std. Deviation
1,00 controle	5,0000	4	2,94392
2,00 brochure	5,0000	4	2,38048
3,00 film	5,0000	4	2,61406
Total	5,0000	12	2,40265

We zien dat er geen verschil is tussen de gemiddelde risico-inschatting van drugsgebruik per groep. In alle groepen is de gemiddelde inschatting 5,00, en de totale gemiddelde risico-inschatting (ongeacht de groep) is daarmee ook 5,00.

Net als bij een regressieanalyse wordt in SPSS een tabel uitgedraaid waar ANOVA boven staat. Letterlijk betekent dit 'Analysis of Variance' (wat niet vreemd is omdat de proportie verklaarde variantie wordt berekend). Hierin wordt de informatie van E_1 (de totale variantie, *Total*) en E_2 (de variantie binnen de groepen, *within groups*) gegeven. In tabel 9.2 is te zien dat de variantie tussen de groepen nul is, en de onverklaarde variantie (de variantie binnen de groepen, *within groups*) gelijk is aan de totale variantie:

Tabel 9.2 ANOVA-tabel gebaseerd op figuur 9.1 (SPSS-output)

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
risico-inschatting * conditie	Between Groups	(Combined)	,000	2	,000	,000	1,000
	Within Groups		63,500	9	7,056		
	Total		63,500	11			

Zou je nu eta of eta² berekenen, dan komt deze dus uit op 0:

$$\eta^2 = \frac{E_1 - E_2}{E_1} = \frac{63{,}5 - 63{,}5}{63{,}5} = 0$$

SPSS bevestigt dit uiteraard:

Tabel 9.3 Associatiemaat eta en eta² (SPSS-output)

Measures of Association

	Eta	Eta Squared
risico-inschatting * conditie	,000	,000

In welke groep de proefpersonen zijn ingedeeld, verklaart voor 0% de variantie in de gemiddelde risico-inschatting.

We laten ook nog zien hoe de analyse eruit zou zien als we de informatie van figuur 9.2 (grofweg) overnemen:

Tabel 9.4 Vergelijking van gemiddelden met associatiematen (SPSS-output)

Report

risico-inschatting

conditie	Mean	N	Std. Deviation
1,00 controle	3,0000	4	1,29099
2,00 brochure	5,5000	4	1,29099
3,00 film	6,5000	4	1,29099
Total	5,0000	12	1,93061

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
risico-inschatting * conditie	Between Groups	(Combined)	26,000	2	13,000	7,800	,011
	Within Groups		15,000	9	1,667		
	Total		41,000	11			

Measures of Association

	Eta	Eta Squared
risicoinschatting * conditie	,796	,634

We zien in tabel 9.4 dat de gemiddelden per groep nu wel van elkaar verschillen. Het totale gemiddelde is nog steeds 5,00: als we geen rekening zouden houden met in welke groep de proefpersonen zijn ingedeeld, is de gemiddelde risico-inschatting 5,00 (SD = 1,93). In je conclusie noem je dan ook de gemiddelden per groep, en de standaarddeviaties per groep. In de ANOVA-tabel daaronder zien we dat de variantie tussen de groepen (*between groups*) groter is dan de variantie binnen de groepen (*within groups*), wat wijst op een verschil tussen de groepen op de afhankelijke variabele. Wanneer we nu de formule voor eta² zouden

invullen, krijgen we, zoals ook in de tabel *Measures of Association* te zien is, een proportie verklaarde variantie van 0,634:

$$\eta^2 = \frac{E_1 - E_2}{E_1} = \frac{41-15}{41} = 0{,}634$$

Trekken we hier de wortel uit, dan krijgen we eta, de sterkte van de samenhang:

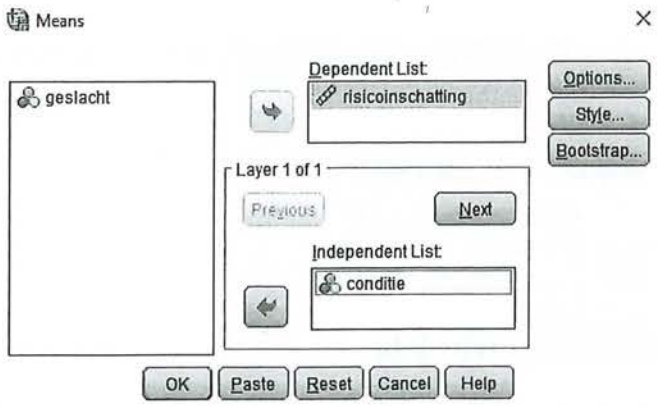
$$\eta = \sqrt{0{,}634} = 0{,}796$$

Onze conclusie aan de hand van bovenstaande output is:

> *Uit een experiment waarin is gekeken naar het effect van de manier van voorlichten over de risico's van drugs op de mate waarin proefpersonen zich bewust zijn van die risico's, blijkt dat er een zeer sterk verband is tussen de manier van voorlichten en de risico-inschatting (η = 0,80, n = 12). Op een schaal van 0 tot 10 schatten proefpersonen die geen voorlichting hebben gekregen het risico gemiddeld het laagst in (M = 3,00, SD = 1,29). Proefpersonen die een voorlichtingsfilm te zien krijgen, zijn zich het meest bewust van de risico's (M = 6,50, SD = 1,29) en proefpersonen die een brochure te lezen krijgen, kunnen de risico's redelijk inschatten (M = 5,50, SD = 1,29). Het soort voorlichting verklaart voor 63,4% de variantie in de risico-inschatting, wat het tot een sterk verklaringsmodel maakt.*

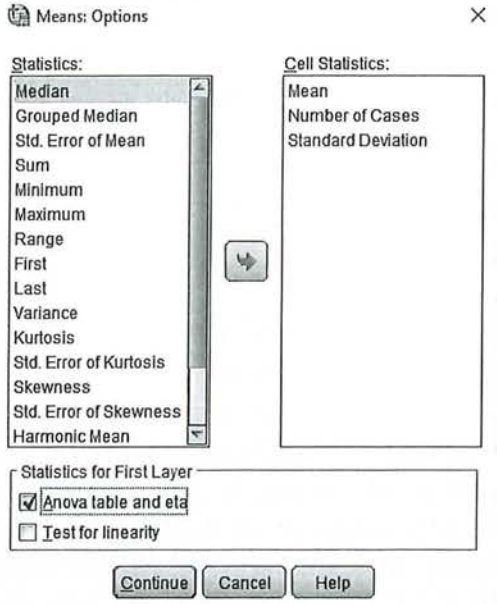
SPSS Vergelijken van gemiddelden tussen afzonderlijke groepen

Voor het vergelijken van gemiddelden tussen groepen maak je in SPSS gebruik van het commando *Means*. Dat commando vind je via *Analyze → Compare Means → Means*.
SPSS vraagt je dan in het *Means*-venster aan te geven wat de onafhankelijke en wat de afhankelijke variabele is (figuur A).



Figuur A Means-venster

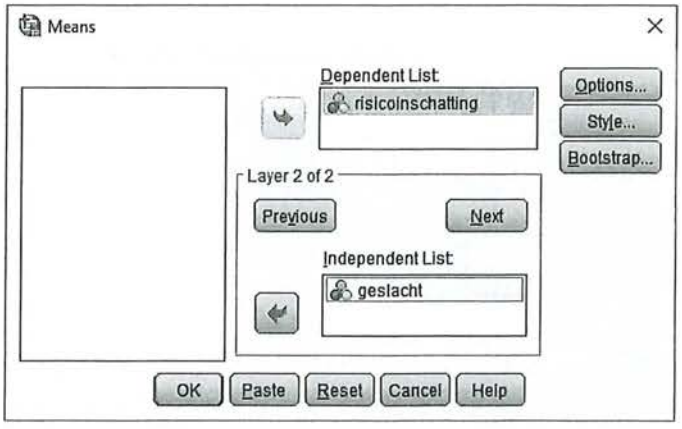
Via *Options* kun je aangeven welke informatie SPSS moet vermelden. Standaard staan hier het gemiddelde (*Mean*), het totaal aantal onderzoekseenheden (*Number of Cases*) en de standaarddeviatie (*Standard Deviation*) weergegeven. Bij *Statistiscs for First Layer* kan worden aangegeven dat eta (en daarmee ook eta^2) berekend moet worden (figuur B).



Figuur B Options-venster

NB: Eta kun je ook via het *Statistics*-venster van *Crosstabs* berekenen, analoog aan de associatiematen voor nominale en ordinale variabelen. Maar dan ontbreekt de ANOVA-tabel, en de informatie van de gemiddelden en standaarddeviaties.

Wanneer je een derde variabele zou willen toevoegen (zie paragraaf 9.1.3), kan dat door in het *Means*-venster op *Next* te klikken bij *Layer 1 of 1*. SPSS voegt dan een nieuwe 'laag' toe (figuur C) waar een derde variabele toegevoegd kan worden als onafhankelijke variabele. Wanneer je ook een uitspraak wilt doen over de sterkte van de interactie-effecten, zul je een andere manier moeten gebruiken, zie daarvoor kader 9.2.



Figuur C Means-venster Layer 2 of 2

9.1.2 Berekening

Op basis van een zeer beperkt aantal onderzoekseenheden rekenen we een voorbeeld met de hand uit. Je wilt weten of er verschil is tussen mannen en vrouwen en de gemiddelde boekenleestijd per week. Geslacht is een nominale variabele, leestijd is een ratiovariabele. Je start met de datamatrix van de twee variabelen: de onafhankelijke variabele sekse (1 = vrouw, 0 = man)[2] en de afhankelijke variabele leestijd (aantal uur per week dat respondenten een boek lezen) (zie tabel 9.5).

Tabel 9.5 Datamatrix van sekse en leestijd

sekse (x)	leestijd (y)
1	8
1	7
1	6
1	8
0	6
0	3
0	4

We hebben zeven onderzoekseenheden die gemiddeld (ongeacht of ze man of vrouw zijn) 6 uur per week een boek lezen ($\overline{y} = 6$). Dit is het totale gemiddelde. De vier vrouwen ($x = 1$) hebben een hogere gemiddelde score op y ($\overline{y}_{\text{vrouw}} = 7{,}25$) (want: $\frac{8+7+6+8}{4} = 7{,}25$) dan de drie mannen ($x = 0$), die gemiddeld maar 4,333 uur per week lezen (want: $\frac{6+3+4}{3} = 4{,}333$).

Je gaat nu eerst de totale variantie uitrekenen (E_1). Voor elke onderzoekseenheid verminder je de waarde van y met het gemiddelde (6). Dit verschil kwadrateer je. Vervolgens tel je deze kwadraten bij elkaar op voor E_1, die hier dus 22 is (zie tabel 9.6).

Tabel 9.6 Waarde van E_1 berekenen

y	$(y-\bar{y})$	$(y-\bar{y})^2$
8	8 – 6 = 2	$2^2 = 4$
7	7 – 6 = 1	$1^2 = 1$
6	6 – 6 = 0	$0^2 = 0$
8	8 – 6 = 2	$2^2 = 4$
6	6 – 6 = 0	$0^2 = 0$
3	3 – 6 = –3	$-3^2 = 9$
4	4 – 6 = –2	$-2^2 = 4$
E_1=	Σ	2^2

Om E_2 te bepalen moet je hetzelfde doen voor elke waarde van x afzonderlijk, dus voor elke groep apart (elke j van k). Daarbij heb je voor elke waarde van x een ander gemiddelde voor y (respectievelijk 7,25 voor vrouwen en 4,333 voor mannen). Als $x = 1$, is de som van de kwadraten 2,752 en als $x = 0$, is dit 4,667. E_2, het aantal voorspellingsfouten dat overblijft als je de informatie over x gebruikt (de onverklaarde variatie), is 2,752 + 4,667 = 7,419 (zie tabel 9.7).

Tabel 9.7 Waarde van E_2 berekenen

		voor $x = 1$ (vrouw)		voor $x = 0$ (man)	
x	y	y –7,25	$(y-\bar{y})^2$	y – 4,33	$(y-\bar{y})^2$
1	8	8 –7,25 = 0,75	$0{,}75^2 = 0{,}563$		
1	7	7 –7,25 = –0,25	$-0{,}25^2 = 0{,}063$		
1	6	6 –7,25 = –1,25	$-1{,}25^2 = 1{,}563$		
1	8	8 –7,25 = 0,75	$0{,}75^2 = 0{,}563$		
0	6			6–4,33 = 1,67	$1{,}67^2 = 2{,}789$
0	3			3–4,33 = –1,33	$-1{,}33^2 = 1{,}769$
0	4			4–4,33 = –0,33	$-0{,}33^2 = 0{,}109$
	E_2=	Σ	2,752	+	4,667

Nu kun je eta² berekenen en daarna eta:

$$\eta^2 = \frac{E_1 - E_2}{E_1} = \frac{22 - 7{,}419}{22} = 0{,}663$$

en

$$\eta = \sqrt{0{,}663} = 0{,}814$$

Bij de interpretatie van eta en/of eta² vermeld je naast de associatiemaat ook altijd het gemiddelde per groep en de standaarddeviatie per groep. De standaarddeviatie hebben we nog niet berekend in voorgaande tabellen, maar met al het voorwerk voor het berekenen van eta en eta² hebben we bijna alle informatie al. We hebben immers de variatie al berekend voor vrouwen (2,752) en voor mannen (4,667). We hoeven dus alleen nog maar te delen door $n - 1$ en de wortel uit dat product te trekken. Voor vrouwen komt de standaarddeviatie daarmee op 0,958 en voor mannen op 1,528. We concluderen:

> *Uit een vergelijking tussen gemiddelden blijkt dat vrouwen gemiddeld vaker een boek lezen ($M = 7{,}25$, $SD = 0{,}96$) dan mannen ($M = 4{,}33$, $SD = 1{,}53$). Dit is een zeer sterk verband ($\eta = 0{,}81$, $n = 7$). De variantie in geslacht verklaart 66,3% de variantie in leestijd.*

Als je deze berekening door SPSS laat maken, vind je dezelfde resultaten (tabel 9.8).

Tabel 9.8 Vergelijking van het gemiddelde tussen twee groepen (SPSS-output)

Measures of Association

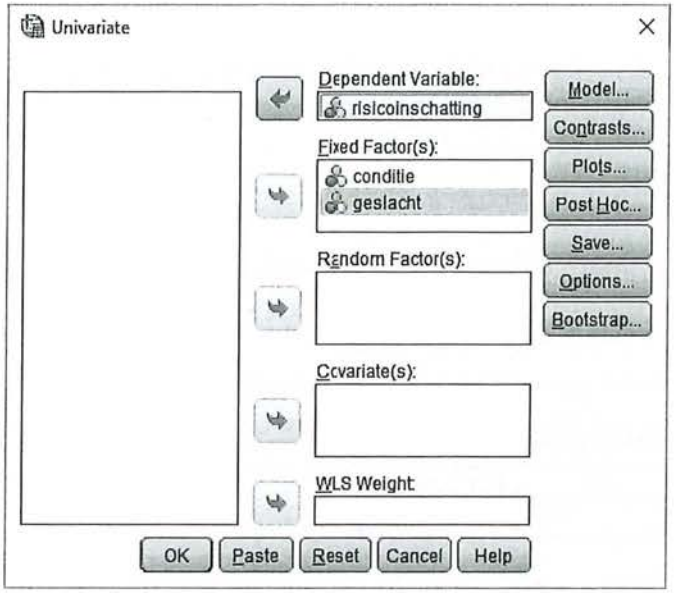
	Eta	Eta Squared
leestijd * geslacht	,814	,663

9.1.3 *Interactie-effecten bij variantieanalyse*

Net als bij kruistabellen en bij regressieanalyse kunnen we bij een variantieanalyse een variabele toevoegen waar we op willen controleren. Het handmatig berekenen van dergelijke interactie-effecten zullen we in dit boek achterwege laten, wel willen we stilstaan bij de (beschrijvende) interpretatie van een interactie-effect bij een variantieanalyse. Bij een variantieanalyse kijken we naar het gezamenlijke effect van twee (of meer) categorische (nominaal of ordinaal) onafhankelijke variabelen op een numerieke (interval of ratio) afhankelijke variabele. We onderscheiden daarbij twee soorten effecten: de hoofdeffecten, en interactie-effecten. Wanneer we twee onafhankelijke variabelen hebben, bijvoorbeeld opleiding en sekse, en één afhankelijke variabele, bijvoorbeeld aantal online aankopen, hebben we twee hoofdeffecten en mogelijk één interactie-effect.

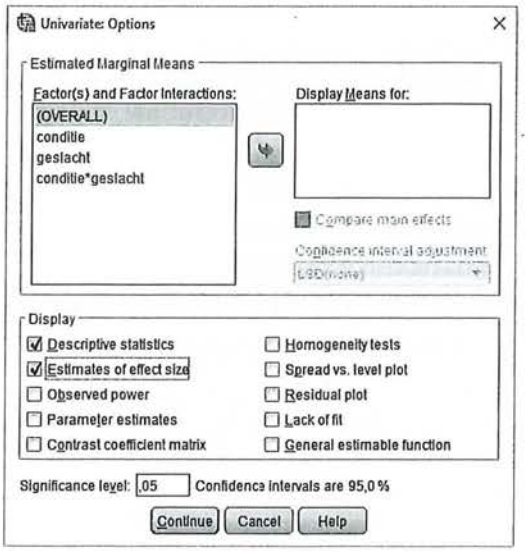
SPSS Berekenen van interactie-effecten via GLM

Het berekenen van een eta² van een interactie-effect en het opvragen van een grafische weergave gaat op een andere manier dan het bereken van eta en eta² zoals besproken in kader 9.1. Ga via *Analyze* naar *General Linear Model → Univariate*. Hier kan de afhankelijke variabele worden ingevoerd onder *Dependent Variable* en de onafhankelijke variabelen in *Fixed Factors* (figuur A).



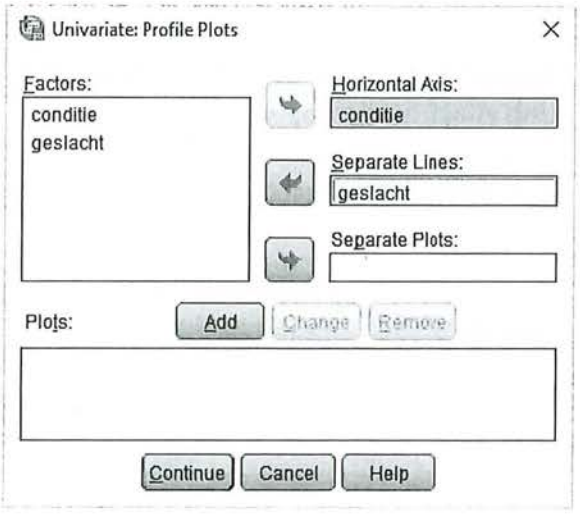
Figuur A Univariate-venster

Onder de knop *Options* moeten onder *Display* de opties *Descriptive Statistics* (om de beschrijvende tabel met gemiddelden en standaarddeviaties te krijgen) en *Estimates of effect size* (om de eta² op te vragen van de hoofdeffecten en het interactie-effect) aangeklikt worden (figuur B).



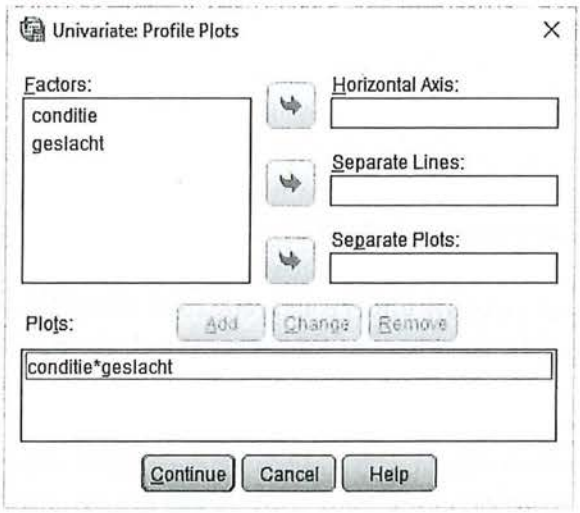
Figuur B Options-venster

Tot slot kan via de knop *Plots* een grafische weergave van de hoofdeffecten en het mogelijke interactie-effect worden gemaakt (figuur C). Zet indien mogelijk de variabele met de minste waarden (hier: geslacht) in de *Separate Lines* en de variabele met meer waarden op de *Horizontal Axis*.



Figuur C Profile Plots-venster

Om de plot ook uit te draaien moet op *Add* worden geklikt, zodat in het venster onder *Plots* de variabelen met daartussen een * verschijnen (figuur D)



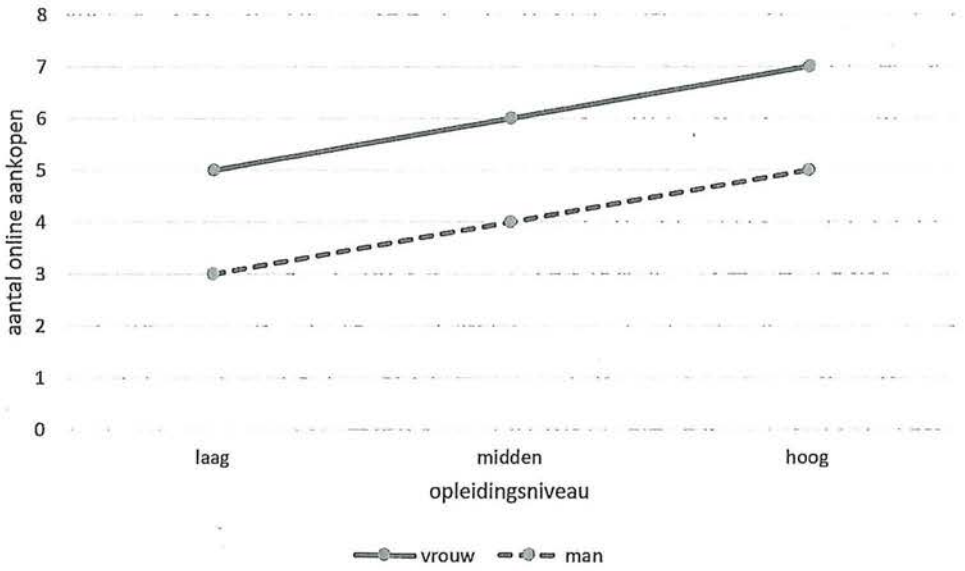
Figuur D Profile Plots-venster na Add

Kader 9.2

We hebben ten eerste een hoofdeffect van opleiding op aantal online aankopen (je verwacht bijvoorbeeld dat hoogopgeleiden meer online aankopen doen dan laag- en middelhoog opgeleiden). We hebben een tweede hoofdeffect van geslacht op aantal online aankopen (je verwacht bijvoorbeeld dat vrouwen vaker online aankopen doen dan mannen). We zouden ook een interactie-effect kunnen hebben, als je verwacht dat er een gezamenlijk effect van opleiding en

geslacht op aantal online aankopen is (je verwacht bijvoorbeeld dat het effect van opleiding op aantal online aankopen voor mannen anders is dan voor vrouwen).

Ook nu zijn er weer twee mogelijke scenario's: er gebeurt niets door toevoeging van deze derde variabele (figuur 9.3), of het blijkt dat de twee onafhankelijke variabelen een gezamenlijk effect hebben op de afhankelijke variabele (figuur 9.4).

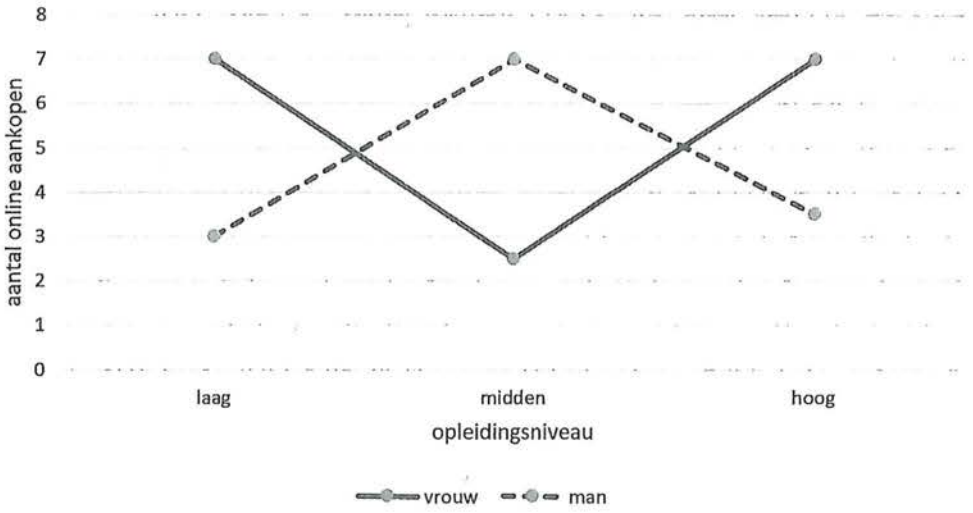


Figuur 9.3 Geen interactie-effect van opleiding en geslacht op aantal online aankopen

In figuur 9.3 is het fictieve onderzoek naar het effect van opleiding en geslacht op het aantal online aankopen in een lijndiagram getekend. In het diagram zie je de gemiddelde scores van zes groepen onderzoekseenheden (namelijk laag opgeleide mannen, laag opgeleide vrouwen, middelhoog opgeleide mannen, middelhoog opgeleide vrouwen, hoog opgeleide mannen en hoog opgeleide vrouwen). Opleidingsniveau (ordinaal) en geslacht (nominaal) zijn de onafhankelijke variabelen, aantal online aankopen (ratio) is de afhankelijke variabele. We zien ten eerste dat er inderdaad een hoofdeffect is van opleidingsniveau op aantal online aankopen. Onder constanthouding van het geslacht doen lager opgeleiden gemiddeld de minste online aankopen en hoogopgeleiden gemiddeld het meest. Middelhoog opgeleiden zitten daar precies tussenin. Er is ook een hoofdeffect van geslacht op aantal online aankopen. Onder constanthouding van het opleidingsniveau kopen vrouwen vaker online dan mannen. Er is echter geen interactie-effect. Het effect van opleiding op aankopen is voor mannen hetzelfde als voor vrouwen.

In figuur 9.4 is een situatie te zien waarin er wel een hoofdeffect is van geslacht op online aankopen (vrouwen scoren anders dan mannen) maar geen hoofdeffect van opleidingsniveau. Wanneer je de lijnen van mannen en vrouwen

samen zou nemen (je zou dus de variabele geslacht constant houden), krijg je een rechte streep in het midden. Er is wel sprake van een interactie-effect. Laag en hoog opgeleide mannen scoren namelijk laag op aantal online aankopen, terwijl laag en hoog opgeleide vrouwen juist hoog op online aankopen scoren. Bij de middelhoog opgeleiden is het precies andersom: daar doen de mannen juist vaak online aankopen en vrouwen weinig online aankopen. Het effect van opleiding op aantal online aankopen is dus voor mannen anders dan voor vrouwen.



Figuur 9.4 Interactie-effect van opleiding en geslacht op aantal online aankopen

We laten het toevoegen van een derde variabele nogmaals zien aan de hand van ons voorbeeld van het experiment van voorlichting op risico-inschatting, en nemen nu de controlevariabele 'geslacht' mee in de analyse. Wanneer we nu in SPSS de informatie opvragen met als derde variabele geslacht, dan wordt in de overzichtstabel met gemiddelden een opdeling gemaakt naar mannen en vrouwen per groep (zie tabel 9.9).

We zien in tabel 9.9 dat de twaalf proefpersonen per conditie ingedeeld zijn naar geslacht. In de rijen met *Total* zie je per conditie de gemiddelde risico-inschatting ongeacht het geslacht. Deze waarden komen overeen met de waarden zoals we ze al hadden gezien in tabel 9.4: de controlegroep scoort gemiddeld 3,00 (SD = 1,29), de groep met een brochure gemiddeld 5,50 (SD = 1,29) en de groep die de film ziet gemiddeld 6,50 (SD = 1,29). In de onderste rij zien we de gemiddelden van geslacht, ongeacht de conditie. Vrouwen schatten gemiddeld de risico's over drugsgebruik hoger in (M = 5,33, SD = 2,68) dan mannen (M = 4,67, SD = 0,88), waarbij het niet uitmaakt of ze nu wel of niet, en zo ja welke voorlichting hebben gekregen.

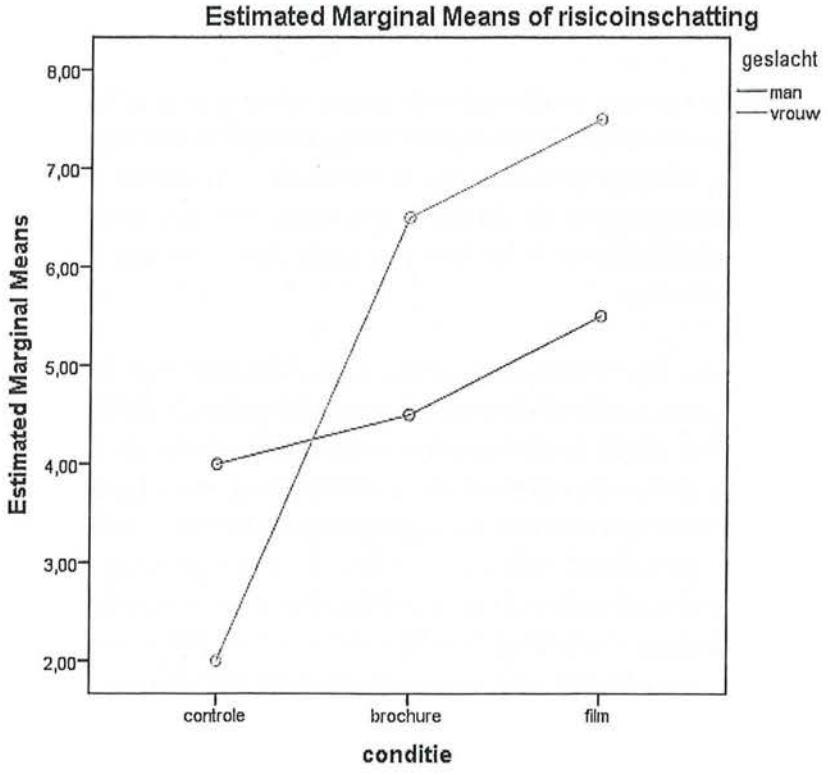
Tabel 9.9 Gemiddelden naar conditie en geslacht voor risico-inschatting (SPSS-output)

Report

risico-inschatting

conditie	geslacht	Mean	N	Std. Deviation
1,00 controle	0 man	4,0000	2	,70711
	1 vrouw	2,0000	2	,70711
	Total	3,0000	4	1,29099
2,00 brochure	0 man	4,5000	2	,70711
	1 vrouw	6,5000	2	,70711
	Total	5,5000	4	1,29099
3,00 film	0 man	5,5000	2	,70711
	1 vrouw	7,5000	2	,70711
	Total	6,5000	4	1,29099
Total	0 man	4,6667	6	,87560
	1 vrouw	5,3333	6	2,67706
	Total	5,0000	12	1,93061

We zien dat er sprake is van een interactie-effect: mannen in de controlegroep hebben een hogere risico-inschatting ($M = 4{,}00$, $SD = 0{,}71$) dan vrouwen in de controlegroep ($M = 2{,}00$, $SD = 0{,}71$), terwijl mannen in de experimentele groepen juist lager scoren dan de vrouwen. Dit interactie-effect kan in SPSS ook grafisch worden weergegeven, zoals te zien is in figuur 9.5.



Figuur 9.5 Grafische weergave in SPSS van interactie-effect (SPSS-output)

Het berekenen van eta² wanneer je een derde variabele gebruikt, gaat niet via *Compare Means*, maar door middel van een *General Linear Model*. In dit boek zullen wij uit deze analyse en de tabel die daarbij hoort alleen eta² bespreken, zoals die te zien is in tabel 9.10.

Tabel 9.10 Eta-kwadraten van twee hoofdeffecten en een interactie-effect (SPSS-output)

Tests of Between-Subjects Effects

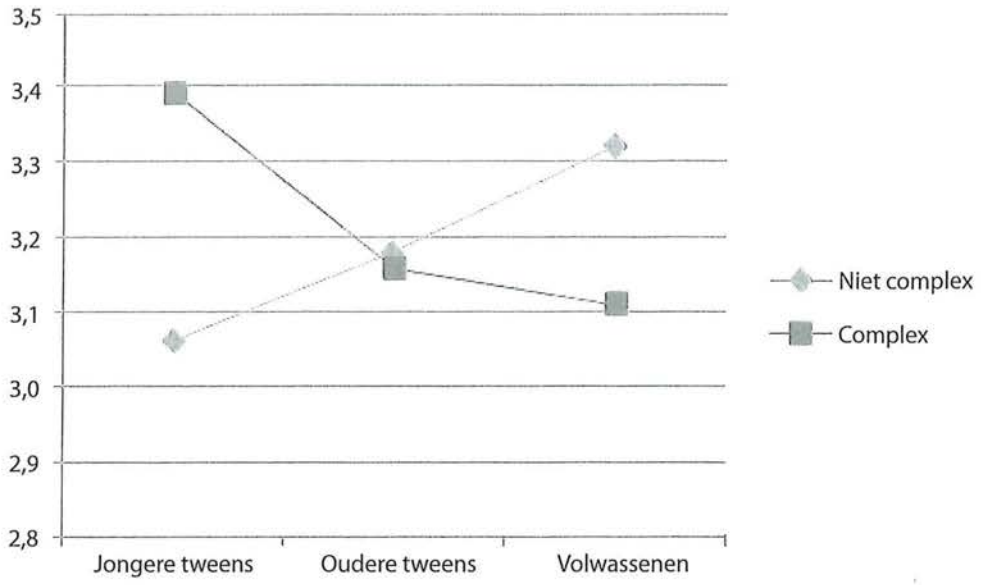
Dependent Variable: risico-inschatting

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	38,000[a]	5	7,600	15,200	,002	,927
Intercept	300,000	1	300,000	600,000	,000	,990
conditie	26,000	2	13,00	26,000	,001	,897
geslacht	1,333	1	1,333	2,667	,154	,308
conditie * geslacht	10,667	2	5,333	10,667	,011	,780
Error	3,000	6	,500			
Total	341,000	12				
Corrected Total	41,000	11				

a. R Squared = ,927 (Adjusted R Squared = ,866)

Toen we alleen naar het effect van conditie keken op de risico-inschatting, vonden we een eta² van 0,63 (zie tabel 9.8). We zien dat het model verbetert als we ook rekening houding met geslacht. Het model verklaart nu 92,7% van de variantie in risico-inschatting (eta² = 0,927). Er blijkt inderdaad een interactie-effect te zijn tussen conditie en geslacht. Dit interactie-effect is verantwoordelijk voor 78,0% van de variantie in de risico-inschatting. Als er een interactie-effect is, zijn de verklaarde varianties van de twee hoofdeffecten niet meer goed te interpreteren, het effect van conditie blijkt immers afhankelijk te zijn van het effect van geslacht.

Interactie-effecten worden veel beschreven in onderzoeken waarin experimenten worden gebruikt, omdat hier twee of meer groepen (controlegroep en één of meer experimentele groepen) met elkaar worden vergeleken op een afhankelijke variabele, waarbij gecontroleerd wordt voor een derde mogelijke verklarende factor. Dat is ook te zien in het experiment *Evaluating books by their covers*[3], dat we ook al ter illustratie gebruikten in hoofdstuk 3, waarbij de waardering van kinderboekomslagen werd onderzocht. In figuur 9.6 zien we dat er geen hoofdeffect is van de leeftijdscategorieën op de waardering van de omslag van een kinderboek waar een detective op staat afgebeeld. Er is wel een hoofdeffect van de conditie, namelijk of de proefpersonen een complex of een niet complex omslag hebben gezien, en er is een duidelijk interactie-effect te zien.



Figuur 9.6 Interactie-effect van leeftijd proefpersonen en complexiteit omslag op waardering van omslag met detective, in artikel van Hartman et al. (2014)

Uit de figuur valt af te lezen dat jongere *tweens* die een niet complexe omslag hebben gezien een lage waardering hebben, terwijl jongere *tweens* die een complexe omslag hebben gezien een hoge waardering hebben. Bij oudere *tweens* is er geen verschil in de gemiddelde waardering tussen het complexe en het niet complexe omslag, maar bij volwassenen zien we weer een duidelijk verschil: volwassenen die een complex omslag hebben gezien, hebben minder waardering dan volwassenen die een niet complex omslag hebben gezien.
De figuur zoals hierboven gepresenteerd, gaat in de tekst van het artikel nog wel altijd gepaard met het noemen van belangrijke gemiddelden en standaarddeviaties!

9.2 Het kiezen van een associatiemaat

In de volgende paragraaf zullen we geen nieuwe associatiematen meer bespreken, maar stilstaan bij de keuze voor een associatiemaat bij bepaalde verwachtingen. Het meetniveau van de variabelen bepaalt welke associatiematen mogelijk en onmogelijk zijn. Maar dan nog blijven er vaak meerdere maten over waaruit je kunt kiezen. Bij de uiteindelijke keuze spelen naast het meetniveau ook de volgende twee zaken mee: de uitspraken die je op basis van je onderzoek zou willen doen en de toepasbaarheid van de specifieke kenmerken van de associatiemaat die je gebruikt.

9.2.1 *Formulering van uitspraken op basis van je onderzoek*

De uitspraken die je op basis van je onderzoek wilt of kunt doen, hangen samen met de associatiemaat die je gebruikt. Dat kan een reden zijn om soms een andere associatiemaat te gebruiken dan je op grond van het meetniveau en de symmetrie of asymmetrie in eerste instantie zou kiezen. We zullen dit uitleggen aan de hand van een voorbeeld.

Tabel 9.11 geeft een kruistabel van de variabelen televisiekijken (1 = weinig, 2 = matig, 3 = veel) en leeftijdsgroepen (1 = 10-25 jaar, 2 = 26-60 jaar, 3 = 61 jaar en ouder). Op basis van deze tabel hebben we de indruk dat deze variabelen met elkaar samenhangen. We hebben twee ordinale variabelen en een asymmetrische relatie. Op grond daarvan zou Somers' d een geschikte associatiemaat zijn. Als je echter in deze kruistabel de concordante en discordante paren gaat tellen, dan blijkt dat er evenveel concordante als discordante paren zijn. Daardoor komt de teller van de formule van Somers' d ($Nc - Nd$) op 0 uit. Somers' d is dus 0. Dit geldt uiteraard ook voor gamma en Kendalls tau-b. Op grond daarvan zou je kunnen concluderen dat er geen verband is tussen de twee variabelen, maar dat klopt niet met onze eerste indruk.

Tabel 9.11 Kruistabel televisiekijken naar leeftijdsgroepen

		Leeftijdsgroepen							
		(1) 10-25 jaar		(2) 26-60 jaar		(3) 61+ jaar		Totaal	
Televisie-kijken	(3) veel	10	(33,3%)	0	(0%)	20	(40%)	30	(30%)
	(2) matig	10	(33,3%)	30	(100%)	0	(0%)	40	(40%)
	(1) weinig	10	(33,3%)	0	(0%)	20	(40%)	30	(30%)
	Totaal	30	(100 %)	30	(100%)	40	(100%)	100	(100%)

Bij associatiematen voor ordinale variabelen maken we gebruik van de ordening in de waarden van de variabelen. Bij een positief verband zou een oudere leeftijdsgroep vaker televisiekijken dan een jongere leeftijdsgroep. Maar deze veronderstelling wordt door onze resultaten weerlegd ($d_{yx} = 0$, $n = 100$).
Je kunt ervoor kiezen geen gebruik te maken van de ordening in de waarden van de variabelen en een associatiemaat kiezen die geschikt is voor nominale variabelen. Rekening houdend met de asymmetrie zou je dan Goodman en Kruskals tau of lambda kunnen gebruiken.
Als je geen rekening houdt met de asymmetrie kun je Cramers V uitrekenen. Dan blijkt dat er wel een verband is tussen de twee variabelen ($\tau = 0{,}39$ en $V = 0{,}60$). Op basis van de waarden van deze associatiematen kun je nu *wel* concluderen dat 'er verschillen zijn tussen de leeftijdsgroepen in de mate waarin deze groepen naar de televisie kijken'. Die uitspraak wordt door onze data niet weerlegd.

Dit voorbeeld maakt duidelijk dat de associatiemaat die je kiest niet alleen afhankelijk is van meetniveau en (a)symmetrie, maar ook van wat je met je onderzoek wilt aantonen. Als je op basis van je onderzoek iets wilt zeggen over verschillen tussen leeftijdsgroepen en de ordening in die leeftijdsgroepen niet van belang is, kies je voor een associatiemaat die daarop de nadruk legt. Om die keuze goed te kunnen maken is het nodig te weten hoe je een associatiemaat precies berekent. Je moet daarom goed weten waarop de formule is gebaseerd.

9.2.2 Kenmerken van de associatiematen

De wijze waarop je een associatiemaat berekent, geeft die maat zijn specifieke kenmerken. Deze kenmerken kunnen een rol spelen bij het maken van een keuze uit de associatiematen. Als je een interval- of ratiovariabele hebt, ligt als centrummaat het rekenkundig gemiddelde het meest voor de hand. Maar het rekenkundig gemiddelde is alleen de meest geschikte centrummaat als de verdeling over de waarden van de variabelen niet al te scheef is. Als de verdeling erg scheef is of als er extreme waarden onder je onderzoekseenheden zijn, kan de mediaan een beter inzicht geven in de verdeling van de onderzoekseenheden over de betreffende variabele. Dit is bijvoorbeeld het geval bij de in tabel 9.12 gegeven frequentieverdeling.
In deze frequentieverdeling is de modus 20, de mediaan 21 en het gemiddelde 23,16. Dit rekenkundig gemiddelde is sterk beïnvloed door de vier personen die respectievelijk 82 en 92 jaar oud zijn. In dit voorbeeld geeft de mediaan een beter inzicht in de leeftijdsverdeling van de 120 personen. Het rekenkundig gemiddelde is minder geschikt. Dit heeft ook consequenties voor de keuze van een associatiemaat als je wilt nagaan of en hoe in dit onderzoek leeftijd samenhangt met andere variabelen. Associatiematen die gebaseerd zijn op het gemiddelde, zoals r en β, zijn nu misschien niet de beste keuze.
Spearmans rho is gebaseerd op de rangordening van waarden. Deze maat heeft geen last van de extreme waarden en is in dit geval een van de mogelijkheden. Maar Spearmans rho is in deze specifieke situatie weer minder geschikt omdat er in de leeftijdsverdeling veel onderzoekseenheden zijn met dezelfde leeftijd. Hierdoor zijn er erg veel knopen (*ties*) en zullen vele onderzoekseenheden een rangordepositie moeten delen. Spearmans rho komt het best tot zijn recht als er juist veel verschillende waarden zijn, waardoor onderzoekseenheden een uniek rangordenummer kunnen krijgen.

Tabel 9.12 Frequentieverdeling van de variabele leeftijd (n = 120)

Leeftijd	Frequentie	%
19	5	4,2
20	50	41,7
21	35	29,2
22	15	12,5
23	5	4,2
24	4	3,3
25	2	1,7
82	1	0,8
92	3	2,5
	120	100

Een goede keuze is in dit geval Kendalls tau-b (mits uiteraard de andere variabele ook minstens een ordinaal niveau heeft). Kendalls tau-b is bijna altijd een betere keuze dan gamma. Anders dan bij gamma wordt in de formule voor Kendalls tau-b ook rekening gehouden met geknoopte paren, waardoor tau-b over het algemeen lagere waarden heeft dan gamma. Kendalls tau-b is daardoor preciezer dan gamma. Gamma is eigenlijk in vergelijking met tau-b een heel grove maat.

Tabel 9.13 Kruistabellen met ordinale variabelen *x* en *y* (*n* = 100)

y \ x	1	2	
2	20	36	56
1	44	0	44
	64	36	100

γ = 1

tau-b = .67

y \ x	1	2	
2	20	36	56
1	24	20	44
	44	56	100

γ = .37

tau-b = .19

In tabel 9.13 is dit met een voorbeeld geïllustreerd. Dat in de linkertabel het verband sterker is dan in de rechtertabel klopt wel, maar het verband in de linkertabel is zeker niet 'perfect', zoals gamma ons wil doen geloven.
Een vergelijkbaar onderscheid is te maken tussen lambda en Goodman en Kruskals tau. Voor de berekening van lambda gebruik je minder informatie dan voor de berekening van Goodman en Kruskals tau. Goodman en Kruskals tau is daardoor preciezer, minder grof dan lambda.

In de vorige paragraaf is beschreven dat een onderzoeker soms toch voor een associatiemaat op nominaal niveau kiest, ook al zijn de variabelen ordinaal of zelfs van een hoger niveau. Dit is niet altijd mogelijk, want Cramers V, die gebaseerd is op de geobserveerde en verwachte waarden in de cellen van een kruistabel, kent ook beperkingen. Als in een kruistabel veel lege of bijna lege cellen staan, krijgt Cramers V een veel te hoge waarde. De kans op lege of bijna lege cellen wordt groter als het aantal waarden van een variabele groot is (in verhouding tot het aantal onderzoekseenheden). De onderzoekseenheden zijn dan over veel cellen van een kruistabel verdeeld (zie tabel 9.14). Er zal dan ook niet aan de voorwaarde worden voldaan dat er geen verwachte waarden lager dan 1 in de tabel aanwezig zijn.

Een oplossing in deze situatie is het samenvoegen van de waarden van de variabelen in een beperkter aantal groepen. Als je beide variabelen hercodeert in weinig (1 tot en met 5) en veel (6 tot en met 9) internetgebruik en televisiekijktijd is Cramers V veel lager (0,27). Er zijn nu geen lege cellen meer bij de geobserveerde frequenties (zie tabel 9.15), en de verwachte frequenties zijn allemaal hoger dan 1.

Tabel 9.14 Kruistabel van internetgebruik en televisiekijktijd (n = 100)

TV / Internet	Zelden							Vaak		
	1	2	3	4	5	6	7	8	9	
1 zelden	1	0	0	2	0	0	0	0	0	3
2	0	0	0	0	0	0	0	14	0	14
3	4	4	0	0	0	0	0	0	0	8
4	1	5	5	0	0	1	16	5	0	33
5	0	0	3	0	0	3	0	0	0	6
6	0	0	0	1	0	0	0	0	0	1
7	0	0	0	5	0	0	0	0	0	5
8	0	0	0	7	1	0	0	0	0	8
9 vaak	0	0	0	0	10	10	0	0	2	22
	6	9	8	15	11	14	16	19	2	100

Cramers V = 0,62

Tabel 9.15 Kruistabel van internetgebruik en televisiekijktijd (n = 100)

Internet \ TV	Weinig 1	Veel 2	Totaal
1 weinig	25	39	64
2 veel	24	12	36
	49	51	100

Cramers V=0,27

9.3 Samenvatting

Wanneer in een kruistabel één van de twee variabelen nominaal is, kies je bijna altijd voor een associatiemaat op nominaal niveau, behalve als de onafhankelijke variabele x op nominaal of ordinaal niveau is gemeten en de afhankelijke variabele y op interval- of rationiveau. In dat geval kun je heel goed eta en eta^2 gebruiken. Deze associatiematen maken onderdeel uit van een variantieanalyse, omdat door middel van de spreiding (variantie) binnen en tussen de groepen naar de verschillen in gemiddelden wordt gekeken. Eta^2 is een maat die qua interpretatie analoog is aan R^2: het is de mate waarin de varia(n)tie in de afhankelijke variabele verklaard wordt door de varia(n)tie in de onafhankelijke variabele.
Naast een ander meetniveau van de onafhankelijke variabele is een tweede verschil tussen eta en een regressieanalyse dat eta uitspraak doet over gemiddelde verschillen tussen groepen, en regressieanalyse over de voorspelling van de afhankelijke variabele op basis van de onafhankelijke variabele. Ook bij eta kan een derde variabelen worden toegevoegd, waardoor je een mogelijk interactie-effect kunt vaststellen.

Welke associatiemaat je kiest, is voor een belangrijk deel afhankelijk van het meetniveau van je variabelen. Daarnaast spelen bij deze keuze de probleemformulering en de aard en kenmerken van de maat een rol.

Ga naar de website om de opdrachten bij dit hoofdstuk te maken.

Noten

1 Een variantieanalyse kan op verschillende manieren worden uitgevoerd, zoals door een ANOVA of een GLM. Omdat we in dit boek niet ingaan op de inferentiële statistieken, kiezen wij er hier voor om de analyses uit te voeren op de meest simpele manier door het vergelijken van gemiddelden (zie kader 9.1).

2 Hoewel de variabele geslacht hier de waarden nul en 1 heeft, blijft het een nominale variabele. Een enkelvoudige regressieanalyse is hier niet geschikt omdat wij bij een enkelvoudige regressieanalyse voor de onafhankelijke variabele altijd minimaal intervalniveau hanteren.

3 Hartman, L., Okken, V. & Rompay, T. van (2014). 'Evaluating books by their covers; de invloed van realisme en complexiteit in fotografiegebruik op de waardering van tweens'. *Tijdschrift voor Communicatiewetenschap, 42*(2), pp. 221-243.