Associatiematen op nominaal niveau 5

In hoofdstuk 1 zijn de meetniveaus behandeld. Het meetniveau is belangrijk bij het bepalen van de analyses die mogelijk zijn. Dit zagen we al bij de centrum- en spreidingsmaten (hoofdstuk 2 en hoofdstuk 3), en dit geldt ook voor de associatiematen. In dit hoofdstuk bespreken we vier bivariate associatiematen op nominaal niveau: Cramers V, phi, Goodman en Kruskals tau en lambda. Bij elk van deze associatiematen wordt steeds een interpretatie gegeven aan de hand van een kruistabel en een SPSS-uitdraai, en wordt de handmatige berekening uitgelegd.

5.1 Wat zijn associatiematen?

Associatiematen, of, samenhangmaten, geven aan of er een verband is tussen twee variabelen. Als bepaalde combinaties van waarden vaak voorkomen, is er een verband tussen de variabelen. Als PvdA'ers bijvoorbeeld relatief vaak Radio 2 als meest favoriete radiostation noemen en CDA'ers Radio 4, is er een verband tussen partijkeuze en radiozenderkeuze. Bij nominale variabelen geeft de associatiemaat de *sterkte* van dat verband aan. De sterkte van het verband druk je uit in een numerieke waarde die bij nominale associatiematen ligt tussen 0 (er is helemaal geen verband) en 1 (er is een perfect verband).

5.1.1 Meetniveau van de variabelen

De eerste stap bij het kiezen van een juiste associatiemaat is vaststellen wat het meetniveau van de twee variabelen is. Is ten minste één van de variabelen nominaal, dan kies je voor een associatiemaat op nominaal niveau. Alleen wanneer beide variabelen op (minimaal) ordinaal niveau zijn gemeten, is een ordinale associatiemaat geoorloofd. Ordinale associatiematen komen aan bod in hoofdstuk 6. Wanneer beide variabelen op minimaal intervalniveau zijn gemeten, zijn interval of ratio associatiematen geoorloofd, die in hoofdstuk 8 worden besproken.

5.1.2 Symmetrische en asymmetrische relaties

Nadat je hebt vastgesteld wat het meetniveau van de variabelen is (in dit hoofdstuk is dat steeds nominaal), bepaal je wat de veronderstelde relatie tussen de twee variabelen is. Wanneer je bijvoorbeeld veronderstelt dat partijkeuze invloed heeft op het favoriete radiostation, maak je onderscheid tussen een onafhankelijke variabele (partijkeuze) en een afhankelijke variabele (favoriete radiostation). Wanneer dit onderscheid aanwezig is, spreken we van een *asymmetrische relatie.*
Is dit onderscheid niet duidelijk, dan spreken we van een *symmetrische relatie.* Als je je bijvoorbeeld afvraagt of er een verband is tussen de partij waarop iemand stemt en de krant die hij leest, en je niet weet wat door wat wordt beïnvloed, dan is de relatie symmetrisch. De partijkeuze zou de krantenkeuze kunnen beïnvloeden, maar de krantenkeuze kan net zo goed de partijkeuze bepalen. In dit geval is er geen onafhankelijke en afhankelijke variabele. De variabelen zijn gelijkwaardig; er is een symmetrische relatie. In paragraaf 1.4 hebben we al aan de hand van voorbeelden gezien wat het verschil is tussen een afhankelijke en onafhankelijke variabele. Bij deze voorbeelden kunnen we nu bepalen of het om een symmetrische of asymmetrische relatie gaat.

- In welke mate heeft woonplaats invloed op het inkomen dat iemand verdient? – *asymmetrisch*
- In hoeverre wordt de krant die iemand leest bepaald door zijn inkomen? – *asymmetrisch*
- Is er een verband tussen iemands favoriete televisieserie en zijn favoriete boekgenre? – *symmetrisch*

Als niet duidelijk is wat de afhankelijke en wat de onafhankelijke variabele is, kun je geen asymmetrische associatiemaat uitrekenen. Een associatiemaat voor een symmetrische relatie kun je wel uitrekenen als de relatie asymmetrisch is. Je maakt dan alleen niet gebruik van de asymmetrie.
Of een verband symmetrisch of asymmetrisch is, kan duidelijk worden door de vraagstelling en door de hypothesen die zijn geformuleerd. Het kan ook zijn dat een van de twee variabelen onbeïnvloedbaar is, bijvoorbeeld geslacht of leeftijd. In dat geval kan die variabele als de onafhankelijke variabele worden behandeld, ook als daar verder geen aanwijzingen voor zijn in de theorie, vraagstelling of hypothesen.

5.1.3 Samenhang in kruistabellen

Aan de hand van de percentages in een kruistabel kun je al een eerste (voorzichtige) conclusie trekken over de samenhang tussen twee variabelen. Stel dat je onderzoek doet naar het verband tussen de favoriete televisieserie van jongvolwassenen en hun favoriete boekgenre. Van deze variabelen maak je vervolgens een kruistabel, waarbij je percenteert op de kolommen.

Tabel 5.1 Kruistabel van favoriete serie en boekgenre (SPSS-output), sterk verband

Boekgenre * Serie Crosstabulation

			Serie			Total
			1 True Detective	2 Game of Thrones	3 Dr. Who	
Boekgenre	1 thrillers	Count	20	0	0	20
		% within Serie	90,9%	0,0%	0,0%	30,8%
	2 avontuur	Count	1	19	1	21
		% within Serie	4,5%	95,0%	4,3%	32,3%
	3 fantasy/SF	Count	1	1	22	24
		% within Serie	4,5%	5,0%	95,7%	36,9%
Total		Count	22	20	23	65
		% within Serie	100,0%	100,0%	100,0%	100,0%

Je ziet aan deze kruistabel dat er bijna een perfecte samenhang is tussen de twee variabelen. Per kolom (per waarde van 'favoriete serie') is er één cel met bijna 100%. Bijna alle (namelijk 90,9% van de) jongvolwassenen die *True Detective* als favoriete televisieserie hebben, hebben *thrillers* als favoriete boekgenre. Hetzelfde zien we voor de *Game of Thrones*-fans: 95,0% heeft *avontuur* als favoriete genre, en de *Dr Who*-fans, waarvan 95,7% een voorkeur voor *fantasy/SF* heeft. We zien dus een zeer sterk, bijna perfect verband (bijna, want er staat niet drie keer het percentage 100 in de cellen).
In tabel 5.2 zien we daarentegen een voorbeeld van een kruistabel waarbij er zo goed als geen verband tussen de twee variabelen is.

Tabel 5.2 Kruistabel van favoriete serie en boekgenre (SPSS-output), geen verband

Boekgenre * Serie Crosstabulation

			Serie			Total
			1 True Detective	2 Game of Thrones	3 Dr. Who	
Boekgenre	1 thrillers	Count	8	7	7	22
		% within Serie	36,4%	35,0%	30,4%	33,8%
	2 avontuur	Count	8	7	8	23
		% within Serie	36,4%	35,0%	34,8%	35,4%
	3 fantasy/SF	Count	6	6	8	20
		% within Serie	27,3%	30,0%	34,8%	30,8%
Total		Count	22	20	23	65
		% within Serie	100,0%	100,0%	100,0%	100,0%

In deze kruistabel zien we dat de percentages in de kolommen niet veel afwijken van de totale kolompercentages. We zien dat in totaal 33,8% van de jongvolwassenen *thrillers* als favoriete boekgenre heeft, en dat deze 33,8% vrij regelmatig over de rij verdeeld is. Het maakt met andere woorden dus niet veel uit wat je

favoriete televisieserie is voor je favoriete boekgenre. Hetzelfde zien we in de rijen daaronder. 35,4% van de respondenten heeft *avontuur* als favoriete genre, en deze percentages liggen dicht bij de percentages per favoriete televisieserie. Hier is dus sprake van zo goed als geen verband. We kunnen niet zeggen: er is helemaal geen verband, want dan zouden de cellen per rij helemaal gelijk zijn aan de totale percentages. Hier zullen we verder op ingaan in paragraaf 5.2.2.

We kunnen dus aan de hand van de kolompercentages in een kruistabel al een inschatting maken van de sterkte van het verband. Wijken de percentages veel van elkaar af, dan zullen we een sterker verband hebben, liggen de percentages dicht bij elkaar, dan zal er een minder sterk verband zijn. We hoeven ons echter niet te beperken tot dit natte vingerwerk. Met associatiematen kunnen we laten zien hoe sterk het verband daadwerkelijk is. In dit hoofdstuk staan associatiematen centraal die gebruikt worden wanneer minimaal een van de variabelen nominaal is. We maken daarnaast nog een onderscheid tussen nominale associatiematen die het meest geschikt zijn bij symmetrische relaties, en nominale associatiematen die alleen geschikt zijn bij asymmetrische relaties.

5.2 Cramers V

Cramers V is een associatiemaat die je gebruikt als minimaal een van de variabelen nominaal is, en waarbij je geen onderscheid maakt tussen een onafhankelijke en een afhankelijke variabele. Het is dus een maat die het meest geschikt is voor symmetrische relaties.

5.2.1 Interpretatie

We gaan verder met het voorbeeld van het mogelijke verband tussen de favoriete televisieserie van jongvolwassenen en hun favoriete boekgenre. Beide variabelen zijn hier nominaal; er zit geen rangordening in favoriete serie of boekgenre. Omdat je onderzoekt of er een verband is (er is geen duidelijke afhankelijke variabele), heb je dus te maken met een symmetrische relatie waarvan minimaal één variabele nominaal is, en daarom is Cramers V hier de meest geschikte maat. Eerst maak je een kruistabel, waarbij je voor de berekening van de percentages deze over de kolommen tot 100% laat optellen (zie tabel 5.3).[1]

Aan de hand van deze kruistabel kun je een uitspraak doen over het verband tussen de favoriete serie en het favoriete boekgenre. Aan de hand van de percentages en absolute waarden kunnen we al zien dat er geen perfect verband is. Er zou een perfect verband zijn tussen de twee variabelen wanneer bijvoorbeeld alle *True Detective*-fans als favoriete boekgenre *thrillers* zouden hebben, alle *Game of Throne*-fans als genre *avontuur* en alle *Dr. Who*-liefhebbers *fantasy/SF*. Dit is niet het geval. Wel komen de combinaties van de waarden (1,1) (*True*

Tabel 5.3 Kruistabel van favoriete televisieserie en boekgenre (SPSS-output)

Boekgenre * Serie Crosstabulation

			Serie			Total
			1 True Detective	2 Game of Thrones	3 Dr. Who	
Boekgenre	1 thrillers	Count	17	3	3	23
		% within Serie	77,3%	15,0%	13,0%	35,4%
	2 avontuur	Count	2	15	3	20
		% within Serie	9,1%	75,0%	13,0%	30,8%
	3 fantasy/SF	Count	3	2	17	22
		% within Serie	13,6%	10,0%	73,9%	33,8%
Total		Count	22	20	23	65
		% within Serie	100,0%	100,0%	100,0%	100,0%

Detective, thriller), (2,2) (*Game of Thrones, avontuur*), en (3,3) (*Dr Who, fantasy/SF*) relatief vaak voor. Op grond van deze percentages kun je dus vaststellen dat er wel een verband is, en gezien de hoge percentages in sommige cellen verwachten we ook een redelijk sterk verband. Hoe sterk dat verband precies is, kunnen we zien aan de waarde van Cramers V.

Tabel 5.4 Cramers V bij de kruistabel van favoriete televisieserie en boekgenre (SPSS-output)

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,893	,000
	Cramer's V	,632	,000
N of Valid Cases		65	

Tabel 5.4 toont de output van SPSS. Daaruit blijkt dat de waarde van Cramers V 0,632 is. Wanneer je bedenkt dat bij een perfect verband Cramers V de waarde 1 heeft en de waarde 0 betekent dat er geen verband is, dan is de waarde 0,632 best wel hoog.

Voor de interpretatie van de nominale associatiematen kun je de volgende grove richtlijnen hanteren:

0 – 0,10: zeer zwak/geen verband;
0,11 – 0,30: zwak verband;
0,31 – 0,50: redelijk verband;
0,51 – 0,80: sterk verband;
0,81 – 0,99: zeer sterk verband;
1: perfect verband.

Let wel, dit zijn slechts richtlijnen! Als een onderzoeker het in een publicatie of onderzoeksverslag heeft over een sterk of een zwak verband, is het verstandig om naar de waarde van de associatiemaat te kijken om te weten hoe sterk het verband echt is. Bij nominale variabelen is het ook niet voldoende om alleen maar te vertellen in welke mate er een verband is. Dan weet je namelijk nog niet hoe het verband precies in elkaar zit en welke combinaties van waarden nu vaak voorkomen.

Bij het interpreteren van een nominale associatiemaat zoals Cramers V noemen we in de conclusie dan ook altijd de waarde van de associatiemaat (afgerond op twee decimalen), de sterkte van het verband (volgens bovenstaande richtlijnen), het aantal onderzoekseenheden (n = ...) en de variabelen waar het verband over gaat. Indien bekend worden ook de onderzoekseenheden genoemd, en worden minimaal twee percentages uit de kruistabel genoemd om het verband toe te lichten. De keuze van de percentages is afhankelijk van wat je als onderzoeker precies wilt vaststellen en wat voor jouw onderzoek van belang is. Zo zou het hoogste percentage met het laagste percentage vergeleken kunnen worden, of zouden meerdere hoge percentages toegelicht kunnen worden.

Onze conclusie zou bij bovenstaand voorbeeld dan zijn:

> *Er is een sterk verband tussen de favoriete televisieserie en het favoriete boekgenre (V = 0,63, n = 65). Zo zien we dat 73,9% van de jongvolwassenen die als favoriete televisieserie Dr. Who heeft, fantasy/SF als favoriete boekgenre heeft, dat 75,0% van de Game of Thrones-fans avontuur als favoriete genre heeft en dat 77,3% van de fans van True Detective het meest van thrillers houdt.*

In een artikel in het *Tijdschrift voor Communicatiewetenschap*[2] lezen we in het artikel *Vlaamse krantenverslaggeving over cyberpesten* bijvoorbeeld:

> *(...) Bijna een derde (30,8%) van de 182 berichten had een lokale (bijv. gemeente/provincie) focus. Daarnaast had 55,5% een nationale (Vlaanderen/België) focus en berichtte 13,7% over een ander land (meestal Nederland, de Verenigde Staten, Groot-Brittannië of Duitsland) of supranationaal nieuws over Europese initiatieven.*
> *Opgedeeld naar krant, zien we dat respectievelijk 38%/5% van de berichten uit de populaire kranten/kwaliteitskranten lokaal nieuws bracht. Er is dus een verband (V = 0.31) tussen de twee variabelen (χ^2 = 17.86) [...]. Uit de percentages valt op te maken dat populaire dagbladen meer lokaal nieuws brengen dan kwaliteitskranten.*

5.2.2 Berekening

Cramers V is gebaseerd op de χ^2 (Chi-kwadraat). De χ^2 geeft een indicatie van de sterkte van het verband tussen variabelen, maar is op zichzelf geen bruikbare associatiemaat omdat hij niet naar boven toe is begrensd en niet direct is te interpreteren. De grootte van χ^2 is namelijk niet alleen afhankelijk van de sterkte van het verband, maar ook van de grootte van *n*, en van de hoeveelheid rijen en kolommen. Door de feitelijk gevonden waarde van χ^2 te relateren aan de maximale waarde die χ^2 kan aannemen in een specifieke kruistabel, krijg je een waarde tussen 0 en 1, die als associatiemaat wel goed te interpreteren is. Dit is Cramers V.
De formule voor Cramers V luidt:

$$V = \sqrt{\frac{\chi^2}{\chi^2 \max}} = \sqrt{\frac{\chi^2}{n[(\min r,k)-1]}}$$

Formule voor Cramers V

En de formule voor χ^2 is:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Formule voor Chi-kwadraat

Centraal bij de berekening van χ^2 (en dus bij Cramers V) is het verschil tussen de feitelijk in de tabel gevonden celfrequenties, die in de formule worden aangeduid met f_o (*frequencies observed*, de geobserveerde frequenties) en de celfrequenties die je zou verwachten als er *geen* verband is tussen beide variabelen, aangeduid in de formule met f_e (*frequencies expected*). Dit verschil wordt uitgedrukt als $f_o - f_e$. Wanneer voor alle cellen geldt dat $f_o = f_e$, is er geen verband ($\chi^2 = 0$).
De f_o's zijn de celfrequenties in een kruistabel. Het zijn de absolute aantallen die in de cellen van de kruistabel staan. Het getal 17 in tabel 5.3 bijvoorbeeld is de geobserveerde frequentie in cel (1,1). Er zijn 17 mensen die *True Detective* als favoriete serie hebben en *thrillers* als favoriete boekgenre. De f_e's moeten nog berekend worden; het zijn de aantallen die je zou verwachten als je uitgaat van de randtotalen en er geen verband tussen de twee variabelen is.

Laten we dit eens toepassen in ons voorbeeld van tabel 5.3. Wanneer er tussen de variabelen geen verband zou bestaan, zou dat betekenen dat de totaalpercentages in de meest rechtse kolom van tabel 5.3 identiek zouden zijn aan de percentages in de voorgaande kolommen. Dan zou 33,8% van de respondenten die *True Detective* als favoriete serie hebben, 33,8% van de *Game of Thrones*-liefhebbers en 33,8% van de *Dr. Who*-fans, als favoriete boekgenre *thrillers* hebben. Bij alle favoriete televisieseries zou 30,8% het meest van *avonturen*boeken houden, en heeft 35,4% van alle fans *fantasy/SF* als favoriete boekgenre, ongeacht

de serie. In dat geval is er dus geen verschil in favoriete boekgenre tussen *True Detective, Game of Thrones* en *Dr. Who*-fans, en is er dus geen verband tussen de twee variabelen.

Uit de celpercentages, die worden berekend aan de hand van de kolomtotalen, blijkt dat dit niet het geval is. Van alle *True Detective*-liefhebbers heeft 77,3% *thrillers* als favoriete boekgenre, en niet 33,8%, en het percentage *Dr. Who*-liefhebbers dat veel van *fantasy/SF* houdt is 73,9, en niet 35,4. We weten dus aan de hand van deze vergelijking tussen de celpercentages en de totaalpercentages in de rechterkolom, dat deze van elkaar afwijken en dat er dus wél een verband zal bestaan.

Deze informatie hebben we nodig bij het uitrekenen van de Chi-kwadraat, en daarmee het uitrekenen van Cramers V. In deze formule worden immers de f_o's (de geobserveerde frequenties) vergeleken met de f_e's (de verwachte frequenties als er geen samenhang zou zijn). Hoe meer deze van elkaar verschillen, hoe hoger de samenhang zal zijn. De geobserveerde frequenties zijn bekend, dat zijn de waarden die je krijgt als je een kruistabel uitdraait. De verwachte waarden moeten echter nog berekend worden. Hoewel we hierboven al de percentages hebben gegeven wanneer in de kruistabel geen sprake is van samenhang, moeten we deze nog omzetten in een absolute frequenties die we kunnen gebruiken in de formule.

De randtotalen (de meest rechtse kolom en de onderste rij in de kruistabel) spelen een belangrijke rol in het berekenen van de f_e's. Om de verwachte frequenties te berekenen gebruiken we die randtotalen om de favoriete televisieserie zodanig over de verschillende boekgenres te verdelen dat de kolompercentages in alle kolommen hetzelfde zijn. In de eerste cel (*True Detective, thriller*) is de verwachte frequentie 33,8% van het totaal aantal *True Detective*-fans (23). De f_e voor deze eerste cel is dus 33,8% van 23 = 7,774. Op deze manier kun je voor elke cel de verwachte frequentie (f_e) uitrekenen.

Nog een voorbeeld. Cel (2,3) bestaat uit de onderzoekseenheden die *Game of Thrones* als favoriete serie hebben en als favoriete boekgenre *fantasy/SF*. Dit zijn er 3 (f_o). Om de verwachte frequentie uit te rekenen, vermenigvuldig je het percentage van *fantasy/SF* van het totaal aantal respondenten (35,4%) met alle respondenten die *Game of Thrones* als favoriete serie hebben (20): 0,354 × 20 = 7,08. Dit is dan de verwachte frequentie (f_e) in cel 2,3 als er geen verband is tussen de twee variabelen. Op deze manier kun je de hele tabel invullen. Merk op dat de randtotalen hetzelfde zijn voor de f_o's en de f_e's (tabel 5.5 en 5.6). Als je nu in tabel 5.6 de kolompercentages zou uitrekenen, zouden dezelfde percentages in elke kolom terugkomen, namelijk 33,8%, 30,8% en 35,4%.

Tabel 5.5 Geobserveerde frequenties

	(1)	(2)	(3)	Totaal
(1)	17	2	3	22
(2)	3	15	2	20
(3)	3	3	17	23
Totaal	23	20	22	65

Tabel 5.6 Verwachte frequenties[3]

	(1)	(2)	(3)	Totaal
(1)	7,774	6,76	7,436	22
(2)	7,084	6,16	6,776	20
(3)	8,142	7,08	7,788	23
Totaal	23	20	22	65

Nu de geobserveerde en verwachte frequenties bekend zijn, kun je de formule voor χ^2 invullen. Die formule was:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

χ^2 bereken je door per cel het verschil tussen f_o en f_e te kwadrateren en vervolgens te delen door de f_e. Tot slot tel je de uitkomsten van alle cellen bij elkaar op (zie tabel 5.7).

Tabel 5.7 Het berekenen van χ^2

Cel	$(f_o - f_e)^2$	$(f_o - f_e)^2 \div f_e$
(1,1)	(17 – 7,774)2 = 85,119	85,119 ÷ 7,774 = 10,949
(1,2)	(3 – 7,084)2 = 16,679	16,679 ÷ 7,084 = 2,354
(1,3)	(3 – 8,142)2 = 26,440	26,440 ÷ 8,142 = 3,247
(2,1)	(2 – 6,76)2 = 22,658	22,658 ÷ 6,76 = 3,352
(2,2)	(15 – 6,16)2 = 78,146	78,146 ÷ 6,16 = 12,686
(2,3)	(3 – 7,08)2 = 16,646	16,646 ÷ 7,08 = 2,351
(3,1)	(3 – 7,436)2 = 19,678	19,678 ÷ 7,436 = 2,646
(3,2)	(2 – 6,776)2 = 22,810	22,810 ÷ 6,776 = 3,366
(3,3)	(17 – 7,788)2 = 84,861	84,861 ÷ 7,788 = 10,89
Totaal		51,847

χ^2 is 51,847. Wil je Cramers V berekenen, dan moet je dit getal eerst nog delen door de maximale waarde die χ^2 kan aannemen. Deze maximale waarde is: n * [min (r,k) – 1]. 'Min (r,k)' is het kleinste getal (*minimum)* van het aantal rijen of kolommen. Hier is een 3 × 3-tabel gebruikt, en is het kleinste getal (minimum) dus 3. Wanneer we bijvoorbeeld een 5 × 4-tabel als voorbeeld hadden genomen, was het minimum 4 geweest. Van dit getal (hier: 3) wordt 1 afgetrokken, en dit verschil wordt vermenigvuldigd met n, het totale aantal onderzoekseenheden,

in dit voorbeeld dus 65. Deel nu de eerder berekende χ^2 (51,847) door χ^2 maximaal (= 65 * (3-1) = 130). Cramers V is de wortel uit dit getal. Hierna zijn de stappen van deze berekening in de formule ingevuld.

$$V = \sqrt{\frac{\chi^2}{n[(\min r,k)-1]}} = \sqrt{\frac{51{,}847}{65(3-1)}} = \sqrt{\frac{51{,}847}{130}} = \sqrt{0{,}399} = 0{,}632$$

Je ziet dat de uitkomst exact overeenkomt met de berekening van SPSS (tabel 5.4). Er bestaat een sterk verband tussen de favoriete televisieserie en het favoriete boekgenre van jongvolwassenen.

5.3 Phi

Net als Cramers V is phi (φ) een symmetrische associatiemaat die je gebruikt bij variabelen op nominaal niveau. Het verschil is dat je phi (spreek uit: fi) alleen gebruikt bij 2 × 2-tabellen. Bij 2 × 2-tabellen heeft phi dezelfde waarde als Cramers V.

5.3.1 Interpretatie

We gebruiken hetzelfde voorbeeld als voorheen. We kijken nu slechts naar twee televisieseries (*Game of Thrones* en *Dr. Who*) en twee boekgenres (*avontuur* en *fantasy/SF*).

Tabel 5.8 Kruistabel tussen favoriete televisieserie en boekgenre (SPSS-output)

Boekgenre * Serie Crosstabulation

			Serie		
			2 Game of Thrones	3 Dr. Who	Total
Boekgenre	2 avontuur	Count	15	3	18
		% within Serie	88,2%	15,0%	48,6%
	3 fantasy/SF	Count	2	17	19
		% within Serie	11,8%	85,0%	51,4%
Total		Count	17	20	37
		% within Serie	100,0%	100,0%	100,0%

Aan de hand van de percentages kunnen we weer een eerste verwachting uitspreken. Er is wel een verband tussen de serie en het boekgenre: relatief meer *Game of Thrones*-fans hebben *avontuur* als favoriet genre en relatief meer *Dr. Who*-fans *fantasy/SF*. Er is echter geen perfect verband; de waarden van de percentages zijn hoog, maar geen 100%. We verwachten dus een sterk, maar niet perfect verband. Dit blijkt ook uit de waarde van phi (zie tabel 5.9).

Tabel 5.9 Phi bij de kruistabel van favoriete televisieserie en boekgenre (SPSS-output)

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	,730	,000
	Cramer's V	,730	,000
N of Valid Cases		37	

5.3.2 *Berekening*

Ook bij de berekening van phi staat χ^2 centraal. De formule luidt:

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

De χ^2 bereken je uiteraard op dezelfde manier als bij Cramers V. In tabel 5.10 wordt per cel steeds eerst de f_o en daarna de f_e gegeven. In cel (2,2) is 15 de geobserveerde frequentie (f_o) en 8,262 de verwachte frequentie (f_e), want: 0,486 * 17 = 8,262.

Tabel 5.10 Berekenen van de f_o's en f_e's

	(2) Game of Thrones		(3) Dr. Who		Totaal	
	f_o	f_e	f_o	f_e	f_o	f_e
(2) avontuur	15	8,262	3	9,720	18	48,6%
(3) fantasy/SF	2	8,738	17	10,280	19	51,4%
Totaal	17		20		37	

Aan de f_e's is ook al te zien dat het verband tussen de twee variabelen sterk is. De verwachte frequenties liggen namelijk ver van de geobserveerde waardes (f_o's). Net als bij Cramers V geldt dat als alle f_o's gelijk zijn aan de f_e's, er geen verband is en de waarde van de associatiemaat op 0 uitkomt.

Tabel 5.11 Het berekenen van χ^2

Cel	$(f_o - f_e)^2$	$(f_o - f_e)^2 \div f_e$
(2,2)	$(15 - 8{,}262)^2 = 45{,}401$	$45{,}401 \div 8{,}262 = 5{,}495$
(2,3)	$(2 - 8{,}738)^2 = 45{,}401$	$45{,}401 \div 8{,}738 = 5{,}196$
(3,2)	$(3 - 9{,}720)^2 = 45{,}158$	$45{,}158 \div 9{,}720 = 4{,}646$
(3,3)	$(17 - 10{,}280)^2 = 45{,}158$	$45{,}158 \div 10{,}280 = 4{,}383$
Totaal		19,730

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{19{,}730}{37}} = \sqrt{0{,}533} = 0{,}730$$

Net als bij Cramers V worden in de conclusie de maat, de sterkte van de maat, de variabelen, de onderzoekseenheden en het aantal onderzoekseenheden, en minimaal twee percentages uit de kruistabel genoemd. Onze conclusie zou hier zijn:

> *Er is onder jongvolwassenen een sterk verband tussen hun favoriete televisieserie en hun favoriete boekgenre (φ = 0,73, n = 37). Uit de kruistabel blijkt dat als Game of Thrones wordt opgegeven als favoriete serie, in 88,2% van de gevallen het boekgenre avontuur als favoriet wordt gezien. Van de Dr. Who-fans heeft 85% fantasy/SF als favoriete boekgenre.*

5.4 Goodman en Kruskals tau

We hebben in de voorgaande paragrafen associatiematen gezien die het meest geschikt zijn wanneer er minimaal één nominale variabele is en wordt uitgegaan van een symmetrisch verband. Cramers V en phi maken geen onderscheid tussen een afhankelijke en een onafhankelijke variabele. Als we bij de vorige berekeningen het favoriete boekgenre in de kolommen zouden zetten en de favoriete televisieserie in de rijen, zou de berekening van de associatiematen dezelfde uitkomst hebben.
In de komende paragrafen kijken we naar associatiematen die wel gebruikmaken van het onderscheid tussen een afhankelijke en onafhankelijke variabele, namelijk lambda en Goodman en Kruskals tau. Bij deze maten moet dus duidelijk zijn wat de afhankelijke variabele is. Als basis voor deze maat gebruik je niet de geobserveerde en verwachte frequenties, zoals bij de voorgaande maten. Bij deze maten gaat het om de *voorspelbaarheid van de afhankelijke variabele.* We zullen eerst *Goodman en Kruskals tau* behandelen.

5.4.1 Berekening

Tau geeft de proportie voorspellingsverbetering van y aan wanneer rekening wordt gehouden met x. Een voorspellingsverbetering is een voorspelling 'met minder fouten'. We voorspellen de score van de afhankelijke variabele, door wel en niet gebruik te maken van de scores op de onafhankelijke variabele. Als de voorspelling van de afhankelijke variabele met gebruikmaking van de informatie over de onafhankelijke variabele veel beter is dan zonder die informatie, is de voorspellingsverbetering groot, en is er dus een sterk verband tussen de twee variabelen. De formule van tau is er een die we nog vaker tegen zullen komen. De manier waarop binnen de formule de onderdelen worden berekend, zal wel steeds verschillend zijn.

$$\tau = \frac{E_1 - E_2}{E_1}$$

Formule voor Goodman en Kruskals tau

In deze formule is E_1 het aantal voorspellingsfouten zonder gebruikmaking van de onafhankelijke variabele x en E_2 het aantal voorspellingsfouten met gebruikmaking van de onafhankelijke variabele.

De formule voor E_1 is

$$E_1 = \sum_i \left(\frac{n - R_i}{n} R_i \right)$$

Formule voor E_1 bij Goodman en Kruskals tau

Hierbij staat n voor het totaal aantal waarnemingen, en R_i voor het totaal van rij i.

De formule voor E_2 is

$$E_2 = \sum_j E_{2j}, \text{ waarbij } E_{2j} = \sum_j \left(\frac{C_j - O_{ij}}{C_j} O_{ij} \right)$$

Formule voor E_2 bij Goodman en Kruskals tau

Hierbij staat C_j voor het totaal van kolom j, en O_{ij} voor het totaal aantal waarnemingen in rij i en kolom j.

Stel, je wilt kijken of sekse (man/vrouw; de onafhankelijke variabele) invloed heeft op het favoriete boekgenre (thriller, avontuur, fantasy/SF; de afhankelijke variabele). Je wilt dan op grond van iemands geslacht het favoriete boekgenre voorspellen. Hoe beter je dat kunt voorspellen, hoe kleiner de voorspellingsfout is. Als het gebruik van de informatie over x, het geslacht, leidt tot een perfecte voorspelling van het favoriete boekgenre (geen voorspellingsfouten; $E_2 = 0$),

heeft Goodman en Kruskals tau de waarde 1. Wanneer de voorspelling helemaal niet verbeterd kan worden en er met en zonder gebruik van x evenveel voorspellingsfouten zijn, ($E_2 = E_1$), heeft Goodman en Kruskals tau de waarde 0.

Een voorbeeld ter verduidelijking. Als je wilt weten of er verband is tussen sekse en iemands favoriete boekgenre, is er sprake van een asymmetrische relatie. Je kunt je wel voorstellen dat sekse invloed heeft op het genre, maar andersom is dit niet mogelijk. Met SPSS maak je de in tabel 5.12 gepresenteerde kruistabel, waarbij je de onafhankelijke variabele (sekse) in de kolommen zet, en de afhankelijke variabele (boekgenre) in de rijen. Je percenteert zoals altijd over de onafhankelijke variabele, dus over de kolommen.

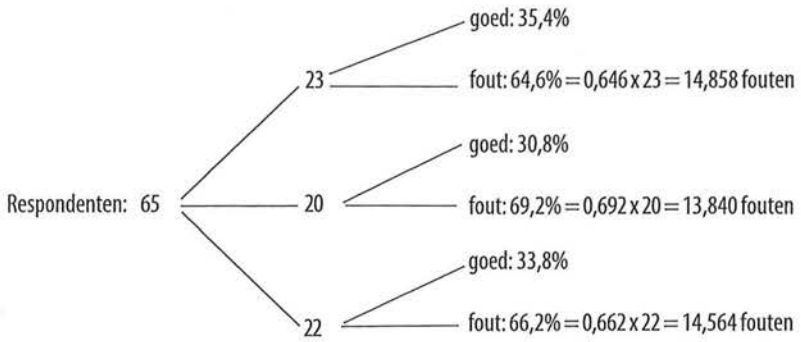
Tabel 5.12 Kruistabel van favoriete boekgenre naar sekse (SPSS-output)

Boekgenre * Sekse Crosstabulation

			Sekse		Total
			1 vrouw	2 man	
Boekgenre	1 thrillers	Count	17	6	23
		% within Sekse	44,7%	22,2%	35,4%
	2 avontuur	Count	17	3	20
		% within Sekse	44,7%	11,1%	30,8%
	3 fantasy/SF	Count	4	18	22
		% within Sekse	10,5%	66,7%	33,8%
Total		Count	38	27	65
		% within Sekse	100,0%	100,0%	100,0%

Bij Goodman en Kruskals tau wordt gekeken hoe goed je aan de hand van de onafhankelijke variabele (hier: sekse), de afhankelijke variabele (hier: iemands favoriete boekgenre) kunt voorspellen. Wanneer bijvoorbeeld in deze kruistabel in cel (1,1) 100% had gestaan en in cel (2,3) ook 100%, zou je perfect aan de hand van iemands geslacht kunnen voorspellen wat diens favoriete boekgenre is. Dan zou elke willekeurige vrouw in je databestand een voorkeur hebben voor *thrillers*, en zou elke willekeurige man een voorkeur hebben voor *fantasy/SF*. Je maakt dan dus geen voorspellingsfouten, en de waarde van tau zou in dat geval 1 zijn. Je kunt ook al zien dat er wel *een* verband is. Zoals we ook al hadden gezien in paragraaf 5.1.3 en 5.2.2 zouden bij geen verband de kolompercentages gelijk zijn aan de kolompercentages over het totaal. In bovenstaand geval zou dan dus 35,4% van de vrouwen, én 35,4% van de manen als favoriete boekgenre *thrillers* hebben. Aan de hand van de kruistabel kunnen we dus alvast de voorzichtige conclusie trekken dat er wel een verband is, maar dat dit geen perfect verband is. We kunnen ook al zien dat er geen sterk verband zal zijn. Hoewel mannen relatief het meest van *fantasy/SF* houden (66,7%), houdt 44,7% van de vrouwen het meest van *thrillers* én heeft 44,7% van de vrouwen *avontuur* als favoriete boekgenre. Hoe sterk het verband precies is, gaan we bekijken aan de hand van Goodman en Kruskals tau.

We beginnen met het berekenen van de E_1, het aantal voorspellingsfouten dat je maakt wanneer je informatie over de onafhankelijke variabele niet meeneemt. We kijken dus voor het berekenen van dit deel van de formule alleen naar de informatie over de afhankelijke variabele, in dit geval het favoriete boekgenre. In totaal hebben 65 jongvolwassenen (de onderzoekseenheden) onze enquête ingevuld, waarvan er 23 hebben aangegeven dat *thrillers* hun favoriete boekgenre is. Wanneer ik een willekeurige respondent uit deze data zou indelen in de categorie 'thrillers als favoriete boekgenre', zou ik dat in 35,4% van de gevallen dus goed doen. Dat betekent automatisch dat ik dat in 64,6% van de gevallen niet goed doe. Ik maak dus in 64,6% van de gevallen een foute voorspelling als het gaat over de categorie 'thriller als favoriete boekgenre'.
Aangezien de formule van E_1 niet vraagt om het *percentage* foute voorspellingen maar om het *aantal* voorspellingsfouten, moeten we dit percentage nog vermenigvuldigen met het totaal aantal onderzoekseenheden dat in die categorie valt. Hier is dat dus: 64,6% * 23 = 0,646 * 23 = 14,858 voorspellingsfouten voor het favoriete boekgenre *thrillers*. Dit doen we vervolgens voor alle categorieën van de afhankelijke variabele 'favoriete boekgenre'.



Figuur 5.1 Berekening E_1 (afhankelijke variabele = boekgenre)

Bij elkaar opgeteld zijn er 14,858 + 13,840 + 14,564 = 43,262 foute voorspellingen als we alleen uitgaan van de afhankelijke variabele (= E_1).

Je hoeft niet per se de hele tijd kansbomen te tekenen om de E_1 uit te rekenen, je kunt ook gebruikmaken van de formule, die uiteraard volgens hetzelfde principe werkt als de kansboom[4]:

$$E_1 = \sum_i \left(\frac{n - R_i}{n} R_i \right)$$

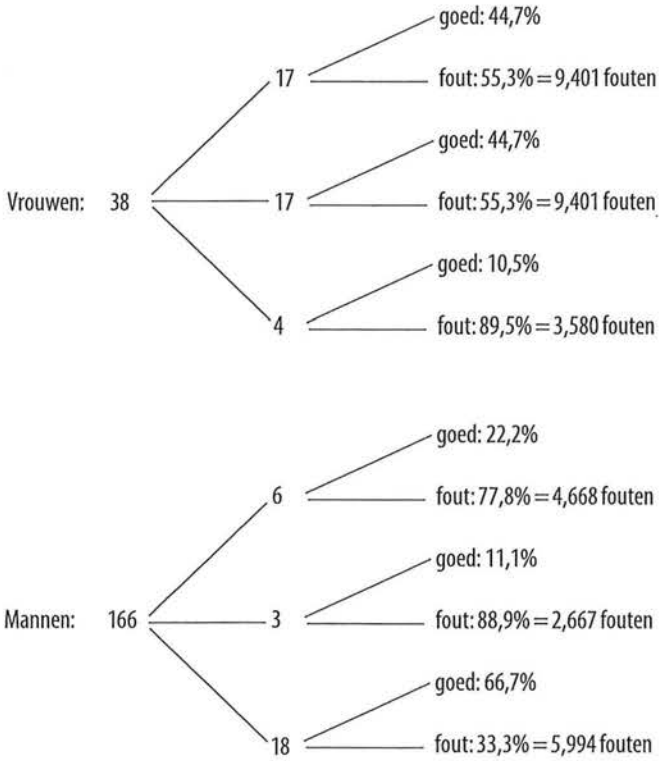
Formule voor E_1 bij tau.

Daarin staat n voor het totaal aantal waarnemingen, en R_i voor het totaal van rij i. We gaan dus per rij deze formule invullen:

$$(y = 1)\left(\frac{65-23}{65} * 23\right) = 14{,}862$$
$$(y = 2)\left(\frac{65-20}{65} * 20\right) = 13{,}846$$
$$(y = 3)\left(\frac{65-22}{65} * 22\right) = 14{,}554$$

$E_1 = 14{,}862 + 13{,}846 + 14{,}554 = 43{,}262$

Voor het berekenen van E_2 houd je wel rekening met de onafhankelijke variabele, in dit geval met sekse. Je moet dan een onderverdeling maken naar de favoriete boekgenres van mannen en de favoriete boekgenres van vrouwen (zie figuur 5.2).



Figuur 5.2 Berekening E_2 (onafhankelijke variabele = sekse)

Er zijn in totaal 38 vrouwen, en hiervan hebben 17 als favoriete boekgenre *thrillers*. Als ik in dit databestand voor een willekeurige vrouw ga kijken of zij het liefst *thrillers* leest, is dat in 44,7% het geval. In 55,3% van de gevallen maak ik dus een voorspellingsfout, wat overeenkomt met 0,553 * 17 = 9,401 fouten.

Deze berekening voer ik eerst apart voor elke categorie van favoriete boekgenre uit voor de vrouwen, en daarna nog een keer voor de mannen. Alle fouten bij elkaar opgeteld maakt 9,401 + 9,401 + 3,580 + 4,668 + 2,667 + 5,994 = 35,711 foute voorspellingen. Dat zijn de fouten die we maken als we rekening houden met de onafhankelijke variabele (= E_2).

Ook voor het berekenen van de E_2 kun je een formule gebruiken en hoef je niet per se een kansboom te tekenen.

De formule is $E_2 = \sum_j E_{2j}$

Formule voor E_2 bij tau.

Daar zit weer een formule in, namelijk $E_{2j} = \sum_j \left(\frac{C_j - O_{ij}}{C_j} O_{ij} \right)$

Hier staat dat je voor iedere groep (aangeduid met de letter *j*, zoals we dat ook zagen bij de formule voor het rekenkundig gemiddelde), het totaal aantal waarnemingen per cel (O_{ij}) van het kolomtotaal (C_j) moet aftrekken en moet delen door het kolomtotaal (je krijgt dan het percentage voorspellingsfouten zoals je dat ook in de kansboom zou berekenen). Deze uitkomst vermenigvuldig je weer met het totaal aantal waarnemingen van die cel.[5]

Vrouwen

(1,1) $\left(\frac{38-17}{38} * 17 \right) = 9{,}395$

(1,2) $\left(\frac{38-17}{38} * 17 \right) = 9{,}395$

(1,3) $\left(\frac{38-4}{38} * 4 \right) = 3{,}579$

Mannen

(2,1) $\left(\frac{27-6}{27} * 6 \right) = 4{,}667$

(2,2) $\left(\frac{27-3}{27} * 3 \right) = 2{,}667$

(2,3) $\left(\frac{27-18}{27} * 18 \right) = 6{,}000$

Nu hebben we alle benodigdheden om de E_2 te berekenen:

$$E_2 = \sum_j E_{2j}$$

Dit betekent letterlijk: neem de som van alle E_{2j}'s die je zojuist berekend hebt.
Dus: E_2 = 9,395 + 9,395 + 3,579 + 4,667 + 2,667 + 6,000 = 35,703.

Nu kunnen we de formule voor tau invullen:

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{43{,}262 - 35{,}703}{43{,}262} = 0{,}175$$

Goodman en Kruskals tau is dus gebaseerd op het aantal voorspellingsfouten, waarbij je die voorspellingsfouten berekent op basis van de frequentieverdeling van de afhankelijke variabele (voor de berekening van E_1) en de frequentieverdeling van de afhankelijke variabele voor elke waarde van de onafhankelijke variabele (voor de berekening van E_2).

5.4.2 *Interpretatie*

De waarde van tau kunnen we op dezelfde manier interpreteren als de waarde van Cramers V en phi zoals we dat eerder in dit hoofdstuk hebben gezien. In het voorbeeld van sekse en favoriete boekgenre zagen we een tau van 0,175. Dit betekent dus dat er een zwak verband is tussen deze twee variabelen. We kunnen met andere woorden geen goede voorspelling doen over iemands favoriete boekgenre als we weten of iemand een man of een vrouw is. We kunnen ook zeggen: we hebben (slechts) een verbetering van 17,5% wanneer we aan de hand van iemands sekse een voorspelling willen doen over het favoriete boekgenre.

Wanneer we in SPSS de output bekijken van Goodman en Kruskals tau (de informatie van lambda krijg je er automatisch bij, en zullen we in de volgende paragraaf behandelen) blijkt dat SPSS twee waarden voor tau geeft: één voor 'boekgenre dependent' en één voor 'sekse dependent'. SPSS weet immers niet wat wij als onafhankelijke variabele hebben gebruikt. Kijk dus altijd goed in de tabel of je bij de juiste waarde kijkt.

Tabel 5.13 Goodman en Kruskals tau van boekgenre naar sekse (SPSS-output)

Directional Measures

			Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
Nominal by Nominal	Lambda	Symmetric	,377	,104	3,073	,002
		Boekgenre favoriete boekgenre Dependent	,286	,099	2,571	,010
		Sekse sekse Dependent	,519	,121	3,213	,001
	Goodman and Kruskal tau	Boekgenre favoriete boekgenre Dependent	,175	,063		,000
		Sekse sekse Dependent	,350	,115		,000

Favoriete boekgenre is onze afhankelijke variabele, en we kijken dus achter de waarde 'boekgenre dependent'. Zoals we ook met de hand hadden berekend, zien we hier de waarde 0,175 staan. We kunnen nu dus de volgende conclusie trekken:

> *Er is een zwak verband tussen sekse en het favoriete boekgenre ($\tau = 0{,}18$). Sekse is dus geen goede voorspeller voor het favoriete boekgenre bij jongvolwassenen. Hoewel relatief veel mannen voornamelijk van fantasy/SF houden (66,7%), zijn vrouwen minder uitgesproken. Van de vrouwen houdt 44,7% voornamelijk van het genre thriller, maar ook 44,7% heeft avontuur als favoriete boekgenre.*

We kijken naar nog een voorbeeld. Aan de hand van theorie verwachten we dat onder bejaarden het opleidingsniveau van invloed is op de voorkeur voor de publieke of commerciële omroep bij het televisiekijken. In een klein onderzoekje onder tien bejaarden wordt gevraagd naar deze twee variabelen, en kan aan de hand van de antwoorden de volgende kruistabel gemaakt worden:

Tabel 5.14 Kruistabel van zendervoorkeur naar opleiding (SPSS-output)

zendervoorkeur * opleiding Crosstabulation

			opleiding			
			1 laag	2 midden	3 hoog	Total
zendervoorkeur	1 publieke omroep	Count	0	4	2	6
		% within opleiding	0,0%	80,0%	66,7%	60,0%
	2 commerciële omroep	Count	2	1	1	4
		% within opleiding	100,0%	20,0%	33,3%	40,0%
Total		Count	2	5	3	10
		% within opleiding	100,0%	100,0%	100,0%	100,0%

De onafhankelijke variabele (opleidingsniveau) is ordinaal, de afhankelijke variabele (zendervoorkeur) is nominaal. Er is sprake van een asymmetrisch verband (opleiding beïnvloedt zendervoorkeur), dus Goodman en Kruskals tau is de meest geschikte associatiemaat. Omdat we bij het berekenen en interpreteren van een associatiemaat altijd uitgaan van het laagste meetniveau (hier: nominaal), maken we hier dus niet gebruik van de rangordening van de ordinale variabele opleidingsniveau. We beschouwen de verschillende niveaus hier als afzonderlijke categorieën.
Als we kijken naar de kruistabel, zien we ten eerste dat er in ieder geval een samenhang is (de totale kolompercentages zien we niet terug bij de afzonderlijke categorieën), dat deze samenhang niet perfect is, en dat de samenhang redelijk zal zijn, aangezien de percentages wel van elkaar verschillen maar niet zeer sterk.

We beginnen met het berekenen van E_1, waarbij alleen de voorspellingsfouten worden berekend voor de afhankelijke variabele, en waar nog geen rekening wordt gehouden met het opleidingsniveau. Er zijn zes mensen met een voorkeur voor de publieke omroep, en vier met een voorkeur voor de commerciële omroep. Wat is dan van deze tien mensen de kans dat je goed voorspelt wat hun zendervoorkeur is? In 60% van de gevallen voorspel je goed dat iemand het liefst naar de publieke omroep kijkt, en dus in 40% van de gevallen fout. In 40% van de gevallen voorspel je bovendien goed dat iemand het liefst naar de commerciële omroep kijkt, en dus in 60% van de gevallen fout. Deze percentages moeten nog omgezet worden in werkelijke fouten, waarvoor we de formule van E_1 kunnen gebruiken:

$$E_1 = \sum_i \left(\frac{n - R_i}{n} R_i \right)$$

Dat doen we per rij, per categorie van de afhankelijke variabele:

$$(y = 1) \left(\frac{10-6}{10} * 6 \right) = 2{,}4$$
$$(y = 2) \left(\frac{10-4}{10} * 4 \right) = 2{,}4$$
$$E_1 = 2{,}4 + 2{,}4 = 4{,}8$$

Voor het berekenen van E_2 houden we wel rekening met de informatie die we hebben over de onafhankelijke variabele. Zo zien we bijvoorbeeld dat als we weten dat een respondent het opleidingsniveau 'laag' heeft, we geen voorspellingsfouten maken. Alle laagopgeleiden hebben namelijk in deze kruistabel een voorkeur voor de commerciële omroep.

Ook hier berekenen we het aantal voorspellingsfouten aan de hand van de formules, waarbij we eerst E_2 per categorie van de onafhankelijke variabele berekenen, en deze uitkomsten vervolgens bij elkaar optellen:

Laag opgeleiden

$$(1,1) \left(\frac{2-0}{2} * 0 \right) = 0$$
$$(1,2) \left(\frac{2-2}{2} * 2 \right) = 0$$

Midden opgeleiden

$$(2,1) \left(\frac{5-4}{5} * 4 \right) = 0{,}8$$
$$(2,2) \left(\frac{5-1}{5} * 1 \right) = 0{,}8$$

Hoog opgeleiden

$$(3,1)\ \left(\frac{3-2}{3}*2\right) = 0{,}667$$

$$(3,2)\ \left(\frac{3-1}{3}*1\right) = 0{,}667$$

$$E_2 = 0 + 0 + 0{,}8 + 0{,}8 + 0{,}667 + 0{,}667 = 2{,}934$$

Tot slot vullen we de formule voor tau in:

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{4{,}8 - 2{,}934}{4{,}8} = 0{,}389$$

SPSS bevestigt deze uitkomst:

Tabel 5.15 Goodman en Kruskals tau van opleiding en zendervoorkeur (SPSS-output)

Directional Measures

			Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
Nominal by Nominal	Lambda	Symmetric	,333	,272	1,054	,292
		zendervoorkeur Dependent	,500	,250	1,581	,114
		opleiding Dependent	,200	,310	,587	,557
	Goodman and Kruskal tau	zendervoorkeur Dependent	,389	,192		,174
		opleiding Dependent	,167	,153		,223

Weer worden twee waarden voor tau genoemd; we kijken naar de waarde die staat achter 'zendervoorkeur dependent' omdat dit onze afhankelijke variabele is.

Onze conclusie zou zijn:

> *Er is een redelijk verband tussen opleidingsniveau en zendervoorkeur (τ = 0,39, n = 10). Bejaarden met een lage opleiding hebben allemaal een voorkeur voor de commerciële omroep, 20% van de gemiddeld opgeleiden heeft hier een voorkeur voor, en 33,3% van de hoger opgeleiden kijkt het liefst naar de commerciële zenders.*[6]

5.5 Lambda

Een tweede nominale associatiemaat die je kunt gebruiken wanneer er een afhankelijke variabele is, is lambda (λ). Deze associatiemaat is net als Goodman en Kruskals tau een maat voor de voorspellingsverbetering, maar de

voorspellingsfouten bereken je nu niet door gebruik te maken van de frequentieverdeling. Lambda gebruikt de modus om een zo goed mogelijke voorspelling te doen.
Lambda is gebaseerd op dezelfde formule als Goodman en Kruskals tau, en heeft ook dezelfde conclusie als tau. Lambda is echter een grovere maat, waarbij met minder informatie rekening wordt gehouden.

$$\lambda = \frac{E_1 - E_2}{E_1}$$

Formule voor lambda

E_1 is nog steeds het aantal voorspellingfouten als je alleen de informatie over de afhankelijke variabele gebruikt. De waarde van de modus is een goede voorspeller van de waarde van de variabele voor alle onderzoekseenheden. Deze komt immers het vaakst voor. De frequentie waarmee de modus van y ($fMo(y)$) voorkomt, is het aantal onderzoekseenheden dat je goed voorspelt door die modus te gebruiken. Alleen voor de onderzoekseenheden die niet de waarde van de modus hebben, is de modus een foute voorspeller. Dit wordt in formulevorm als volgt geschreven:

$$E_1 = n - fMo(y)$$

Formule voor E_1 bij lambda

E_1 is dus het aantal voorspellingsfouten als we uitgaan van de modus van y, en niet zoals bij Goodman en Kruskals tau het percentage waar vervolgens het aantal fouten van wordt berekend.
E_2 is ook hier het aantal voorspellingsfouten als we informatie over de onafhankelijke variabele x wél bij de voorspelling betrekken. Voor elke waarde van x gebruiken we de vaakst voorkomende waarde als voorspeller ($Mo(y)kx$). We tellen vervolgens al de keren dat de vaakst voorkomende waarde een goede voorspelling is bij elkaar op. In formulevorm schrijven we:

$$E_2 = n - \sum fMo(y)_{kx}$$

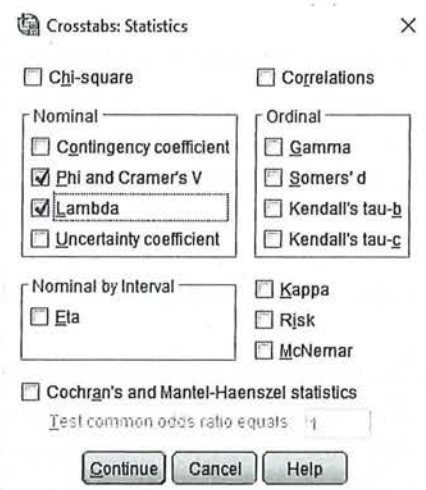
Formule voor E_2 bij lambda

Onderzoekseenheden die niet de waarde hebben die het vaakst voorkomt, hebben we fout voorspeld. Uit de berekening van lambda in paragraaf 5.5.1 zal blijken dat dit in de praktijk eenvoudiger is dan het door deze formules lijkt.

SPSS Berekenen van nominale associatiematen

In hoofdstuk 1 (kader 1.3) hebben we al gezien hoe in SPSS kruistabellen moeten worden gemaakt en hoe gepercenteerd kan worden over de rijen of de kolommen. Om een kruistabel te maken in SPSS ga je via *Analyze* → *Descriptive Statistics* naar *Crosstabs*. Hier geef je aan welke variabele je in de rijen en welke variabele je in de kolommen wilt, en je geeft bij *Cells* aan dat je wilt percenteren op de kolommen. Via de knop *Statistics* kun je aangeven welke associatiemaat je bij deze kruistabel wilt laten uitdraaien.

Bij het uitdraaien van Cramers V wordt ook de waarde van phi gegeven. Bij het uitdraaien van lambda wordt ook de waarde van Goodman en Kruskals tau gegeven. Je kunt via dit venster ook Chi-kwadraat laten berekenen.



Figuur A Statistics-venster: nominale associatiematen.

Kader 5.1

5.5.1 *Berekening*

We nemen als voorbeeld het onderzoek naar de favoriete televisieserie van jongvolwassenen, en verwachten dat de keuze voor een favoriete serie bepaalt hoe vaak er over de serie gepraat wordt. We hebben dus twee variabelen, waarvan 'favoriete serie' de onafhankelijke variabele is (deze is nominaal) en 'hoeveelheid praten over serie' de afhankelijke variabele is (deze is ordinaal). Eerst kijken we weer naar de kruistabel voor een eerste indruk van het verband.

Tabel 5.16 Kruistabel van hoeveelheid praten over favoriete televisieserie (SPSS-output)

praten * Serie Crosstabulation

			Serie			
			1 True Detective	2 Game of Thrones	3 Dr. Who	Total
praten	1 nooit	Count	2	0	9	11
		% within Serie	9,1%	0,0%	39,1%	16,9%
	2 soms	Count	4	11	6	21
		% within Serie	18,2%	55,0%	26,1%	32,3%
	3 regelmatig	Count	16	9	8	33
		% within Serie	72,7%	45,0%	34,8%	50,8%
Total		Count	22	20	23	65
		% within Serie	100,0%	100,0%	100,0%	100,0%

De eerste indruk is dat er wel een verband is, maar dat dit verband niet sterk zal zijn, voornamelijk doordat *Dr. Who*-fans vrij gelijkmatig over de kolom zijn verdeeld.

Hoe goed kunnen we aan de hand van de favoriete televisieserie voorspellen hoeveel er over de serie wordt gepraat?
Eerst rekenen we E_1 uit, met de formule $fMo(y)$. Dit is de hoogste randfrequentie van de afhankelijke variabele. De afhankelijke variabele is hier het praten over de serie, en de hoogste randfrequentie is 33 (de meeste jongvolwassenen praten regelmatig over een serie) $Mo = 3$ en $fMo(y) = 33$. Die 33 voorspellen we goed als we voor de voorspelling de modus (3) gebruiken. Bij alle overige onderzoekseenheden is onze voorspellingsfout ($E_1 = n - fMo(y)$):

$E_1 = 65 - 33 = 32.$

De E_2 berekenen we door de formule $\Sigma fMo(y)_{kx}$, en wordt gevormd door voor elk van de categorieën van de onafhankelijke variabele de aantallen in de cel met de hoogste frequenties bij elkaar op te tellen. Hier is favoriete televisieserie de onafhankelijke variabele. De hoogste kolomfrequentie voor *True Detective* is 16 (cel (1,3)), voor *Game of Thrones* 11 (cel (2,2)) en voor *Dr. Who* is dat 9 (cel (3,1)). $\Sigma fMo(y)_{kx}$ is dus 16 + 11 + 9 = 36. Voor die 36 onderzoekseenheden voorspellen we met behulp van de informatie over x de juiste waarde voor y. Bij alle overige onderzoekseenheden doen we het fout ($E_2 = n - \Sigma fMo(y)_{kx}$).

$E_2 = 65 - 36 = 29.$

We hebben nu alle informatie om de formule van lambda verder in te vullen en lambda te berekenen:

$$\lambda = \frac{E_1 - E_2}{E_1} = \frac{32-29}{32} = 0{,}094$$

Er is dus een zeer zwak verband tussen favoriete televisieserie en hoeveel er over de serie gepraat wordt.

5.5.2 *Interpretatie*

Net als bij Goodman en Kruskals tau geeft SPSS in de output voor lambda twee mogelijkheden. Dat wij televisieserie als onafhankelijke variabele zien, kan SPSS immers niet weten. We kijken in dit geval bij 'praten dependent', omdat we deze als afhankelijke variabele hebben gebruikt. Lambda is 0,094, wat overeenkomt met onze berekening.

Tabel 5.17 Lambda bij kruistabel van praten over serie naar favoriete serie (SPSS-output)

Directional Measures

			Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
Nominal by Nominal	Lambda	Symmetric	,216	,134	1,511	,131
		praten praten over serie Dependent	,094	,181	,494	,621
		Serie favoriete tvserie Dependent	,310	,127	2,098	,036
	Goodman and Kruskal tau	praten praten over serie Dependent	,129	,061		,002
		Serie favoriete tvserie Dependent	,147	,058		,001

Onze conclusie aan de hand van bovenstaande output is:

> *Er is onder jongvolwassenen een zeer zwakke samenhang tussen hun favoriete televisieserie en hoeveel zij over de serie praten ($\lambda = 0{,}09$, $n = 65$). Zo zien we dat van alle Dr. Who-liefhebbers 39,1% nooit over de serie praat, 26,1% soms en 34,8% regelmatig.*

De SPSS-output in tabel 5.17 laat ook het belang van de onafhankelijke variabele zien. Wij hebben (op grond van een bepaalde theoretische verwachting) de favoriete serie als onafhankelijke variabele bestempeld. Maar was onze verwachting geweest dat het praten over de serie juist invloed zou hebben op welke serie het liefst gekeken wordt, dan zouden we een hele andere conclusie

getrokken hebben. Dan zou namelijk de favoriete televisieserie de afhankelijke variabele zijn, en zou lambda 0,31 zijn: een redelijk sterke samenhang. Het is dus belangrijk om in de SPSS-output naar de juiste waarde te kijken die aansluit bij jouw verwachting.

5.6 Voorwaarden bij het maken van een kruistabel

Een kruistabel maken heeft niet veel zin als er te veel waarden zijn en/of als er veel lege cellen zijn. Wanneer je niet naar drie favoriete televisieseries hebt gevraagd maar naar twintig, zul je ten eerste een erg grote kruistabel krijgen (met twintig rijen of kolommen, afhankelijk of je deze variabele als afhankelijke of onafhankelijke kiest), maar zul je ook veel lege of bijna lege cellen krijgen omdat sommige series maar door één of twee personen, of zelfs door niemand, gekozen zijn, zoals te zien is in tabel 5.18.
Bij het berekenen van de hierboven genoemde associatiematen bij variabelen op nominaal niveau is daarom een voorwaarde dat geen enkele cel in de kruistabel een verwachte waarde (dus een f_e) heeft van minder dan 1, en dat minimaal 80% van de cellen een verwachte waarde heeft van minimaal 5 (we kunnen ook zeggen: maximaal 20% van de cellen mag een verwachte waarde lager hebben dan 5).

Tabel 5.18 Kruistabel met geobserveerde en verwachte waarden van boekgenre naar favoriete Nederlandse televisieprogramma (SPSS-output)

Boekgenre * tvprog Crosstabulation

			tvprog								
			1 Baantjer	2 Flikken	3 Smeris	4 All Stars	5 Costa!	6 Zeg eens Aaa	7 2 voor 12	8 Per secon-de wijzer	Total
Boekgenre	1 thrillers	Count	8	14	12	2	4	4	3	2	49
		Expected Count	5,7	16,8	6,3	,6	1,9	11,4	5,1	1,3	49,0
	2 avontuur	Count	10	39	8	0	2	0	3	0	62
		Expected Count	7,2	21,2	8,0	,8	2,4	14,4	6,4	1,6	62,0
	3 fantasy/ SF	Count	0	0	0	0	0	32	10	2	44
		Expected Count	5,1	15,0	5,7	,6	1,7	10,2	4,5	1,1	44,0
Total		Count	18	53	20	2	6	36	16	4	155
		Expected Count	18,0	53,0	20,0	2,0	6,0	36,0	16,0	4,0	155,0

We hebben door SPSS de verwachte waarden (*Expected Count*) laten berekenen (deze kunnen worden berekend onder *Cells* bij het maken van de kruistabel) en zien dat er niet aan de voorwaarden voor het berekenen van associatiematen op nominaal niveau wordt voldaan. Zo hebben meerdere cellen een verwachte

waarde lager dan 1 (namelijk de cellen (4,1), (4,2), (4,3)), en zijn er daarnaast meerdere cellen die een verwachte waarde lager dan 5 hebben. In totaal hebben in deze kruistabel tien cellen een verwachte waarde van lager dan 5.
Aangezien we een 8 x 3-tabel hebben (= 24 cellen) heeft dus 41,7% van de cellen een te lage verwachte waarde, daarmee wordt niet aan de voorwaarde van voldoende gevulde cellen van de kruistabel voldaan.
Een oplossing zou zijn om sommige waarden niet mee te nemen bij de analyses of om verschillende waarden samen te voegen. Dat moet je dan wel goed verantwoorden in het onderzoek. In dit geval zou je bijvoorbeeld kunnen beargumenteren dat de eerste drie categorieën 'politie- en detectiveseries' meten, de tweede drie categorieën 'comedy' en de laatste twee categorieën 'quiz en spel'. Door middel van *Recode* (zie paragraaf 4.4) kunnen nieuwe categorieën worden gevormd, en wanneer vervolgens een kruistabel wordt uitgedraaid, is te zien dat deze én overzichtelijker is, én dat aan de voorwaarden voor het berekenen van een associatiemaat bij een kruistabel wordt voldaan:

Tabel 5.19 Kruistabel met geobserveerde en verwachte waarden van boekgenre naar televisiegenre (gehercodeerd) (SPSS-output)

boekgenre * tvprogHER Crosstabulation

			tvprogHER			
			1 politie	2 comedy	3 quiz	Total
boekgenre	1 thriller	Count	34	10	5	49
		Expected Count	28,8	13,9	6,3	49,0
	2 avontuur	Count	57	2	3	62
		Expected Count	36,4	17,6	8,0	62,0
	3 fantasy/SF	Count	0	32	12	44
		Expected Count	25,8	12,5	5,7	44,0
Total		Count	91	44	20	155
		Expected Count	91,0	44,0	20,0	155,0

Er zijn geen verwachte waarden meer die lager zijn dan 5. Wel hebben we op deze manier wat informatieverlies, we hebben immers categorieën samengevoegd en kunnen daardoor minder genuanceerde uitspraken over alle gevraagde televisieseries doen.

5.7 Samenvatting

Een associatiemaat gebruik je om een verband tussen twee variabelen aan te duiden. Wanneer minimaal een van deze variabelen nominaal is, kies je voor een associatiemaat op nominaal niveau. Hiervoor maken we een kruistabel, waarbij aan de voorwaarden moet worden voldaan dat geen enkele verwachte waarde lager is dan 1, en dat minimaal 80% van de cellen een verwachte waarde van minimaal 5 heeft.

Naast meetniveau speelt de veronderstelde relatie een rol: bij symmetrische relaties is er geen duidelijke (on)afhankelijke variabele, bij asymmetrische wel. Cramers V en phi zijn beide associatiematen die het best gebruikt kunnen worden bij symmetrische relaties. Phi wordt echter alleen gebruikt wanneer de kruistabel 2 x 2 is. Je kunt Goodman en Kruskals tau en lambda niet berekenen als je niet weet wat de afhankelijke en wat de onafhankelijke variabele is.
Lambda en Goodman en Kruskals tau kun je in dezelfde situaties toepassen (minimaal één nominale variabele en een asymmetrische relatie). Uit de beschrijvingen blijkt dat voor het berekenen van Goodman en Kruskals tau meer informatie wordt gebruikt (de frequentieverdelingen) dan voor het berekenen van lambda (de frequentie van de vaakst voorkomende waarde). Daardoor zal de waarde van tau doorgaans iets lager uitvallen, iets conservatiever zijn. Of anders gezegd, lambda is een grovere maat dan Goodman en Kruskals tau.
Goodman en Kruskals tau en lambda mag je alleen berekenen bij een asymmetrische relatie, Cramers V mag je zowel bij een symmetrische of asymmetrische relatie berekenen, maar is het meest geschikt bij een symmetrisch verband.

Tabel 5.20 Nominale associatiematen naar relatie

	Nominale associatiemaat
Symmetrisch	Cramers V phi (φ)
Asymmetrisch	Goodman en Kruskals tau (τ) lambda (λ)

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

Noten

1 We percenteren over de kolommen, ook als er geen onafhankelijke variabele is. Op grond van deze percentages is vaak al duidelijk te zien of er verband is tussen de twee variabelen.
2 Vermeulen, A. & Vandebosch, H. (2014). Vlaamse krantenverslaggeving over cyberpesten. *Tijdschrift voor Communicatiewetenschap , 42(3)*, 286-304.
3 Omdat we deze informatie nog nodig hebben in de formule, ronden we in deze tabel af op drie decimalen.
4 Vanwege afrondingsverschillen wijkt het aantal voorspellingsfouten per categorie iets af, dit maakt niet uit voor het eindresultaat van tau.
5 Afwijkingen van de waarden in de decimalen komt door de verschillende manieren van afronden en hebben geen effect op de uiteindelijke waarde van tau.
6 Overigens mogen in het verslag over dit onderzoek ook andere percentages uit de kruistabel worden genoemd; het gaat erom dat de conclusie wordt ondersteund door minimaal twee percentages.