Associatiematen op ordinaal niveau 6

In het vorige hoofdstuk zijn associatiematen voor variabelen op nominaal niveau besproken. In dit hoofdstuk kijken we naar associatiematen voor variabelen op minimaal ordinaal niveau: gamma, Somers' d, Kendalls tau-b en Spearmans rho. Een groot verschil tussen maten op ordinaal niveau en maten op nominaal niveau is dat deze laatste alleen de *sterkte* van het verband kunnen aanduiden, terwijl ordinale maten ook iets zeggen over de *richting* van het verband. Bij ordinale variabelen is sprake van rangordening. Het is dus mogelijk om te stellen dat iets 'naar boven gaat of groter is' of 'naar beneden gaat of kleiner is'.

Voorbeelden van vraagstellingen of hypothesen die beantwoord kunnen worden met een associatiemaat voor ordinale variabelen zijn:
- In hoeverre is er een verband tussen hoe vaak adolescenten een boek lezen en hoe vaak zij televisiekijken? – *symmetrisch*
- In hoeverre is er een verband tussen de leeskans van de *Linda* en de leeskans van de *Happinez*? – *symmetrisch*
- Hoe meer educatieve spelletjes peuters op een tablet spelen, hoe groter hun woordenschat zal zijn. – *asymmetrisch*

Aan deze vraagstellingen en hypothese is overigens niet meteen te zien dat het hier om ordinale variabelen gaat, dat moet duidelijk worden in de operationalisatie van de gebruikte begrippen in je onderzoek.

6.1 Samenhang in kruistabellen met ordinaal meetniveau

Net als bij associatiematen die gebruikt worden wanneer minimaal één van de variabelen nominaal is, kan ook bij associatiematen waarbij de variabelen ordinaal zijn al een indicatie worden gegeven over de sterkte van het verband aan de hand van de kruistabel. Daarbij kan ook iets gezegd worden over de richting van het verband. Omdat we bij ordinale variabelen gebruik kunnen maken van de rangordening, kan bijvoorbeeld gesteld worden dat hoe hoger iemand is opgeleid, hoe hoger diens inkomen zal zijn. Dit noemen we een positieve samenhang; er zijn dan veel combinaties met hoog-hoog (hoog opgeleid, hoog inkomen) en met laag-laag (laag opgeleid, laag inkomen). Of: hoe vaker iemand televisiekijkt, hoe minder vaak iemand een boek leest. Dat noemen we een negatieve samenhang; er zullen dan veel combinaties met hoog-laag (veel televisiekijken, weinig lezen) en laag-hoog (weinig televisiekijken, veel lezen) voorkomen.

De richting van de samenhang kan worden afgelezen aan het al dan niet aanwezig zijn van een minteken. De waarde van een ordinale associatiemaat varieert tussen -1 (perfecte negatieve samenhang) en 1 (perfecte positieve samenhang). Een associatiemaat bij nominale variabelen kan nooit negatief zijn, omdat er niet gesproken kan worden van 'meer' of 'minder'. De interpretatie van de *sterkte* van de samenhang is bij nominale en ordinale associatiematen hetzelfde. De waarde nul (0) betekent dat er geen samenhang is, en naarmate de maat meer richting 1 gaat (of -1) is de samenhang sterker.

In tabel 6.1 zie je twee ordinale variabelen in een kruistabel, namelijk hoe vaak iemand televisiekijkt, en hoe vaak iemand een boek leest. Er is een rangordening in beide variabelen, en we zien dat de combinaties (1,3 – nooit, vaak), (2,2 – soms, soms) en (3,1 – vaak, nooit) relatief vaak voorkomen (namelijk respectievelijk 72,5%, 63,2% en 57,1%). Hoe vaker iemand televisiekijkt, hoe minder vaak diegene een boek leest, en hoe minder vaak iemand televisiekijkt, hoe vaker iemand een boek leest. We verwachten aan de hand van deze percentages dus een sterk verband tussen de twee variabelen.

Tabel 6.1 Kruistabel van hoe vaak tv-kijken en hoe vaak boeken lezen, sterk negatief verband (SPSS-output)

hoe vaak boeken lezen * hoe vaak tvkijken Crosstabulation

			hoe vaak tvkijken			
			1 nooit	2 soms	3 vaak	Total
hoe vaak boeken lezen	3 vaak	Count	37	2	4	43
		% within hoe vaak tvkijken	72,5%	10,5%	7,1%	34,1%
	2 soms	Count	8	12	20	40
		% within hoe vaak tvkijken	15,7%	63,2%	35,7%	31,7%
	1 nooit	Count	6	5	32	43
		% within hoe vaak tvkijken	11,8%	26,3%	57,1%	34,1%
Total		Count	51	19	56	126
		% within hoe vaak tvkijken	100,0%	100,0%	100,0%	100,0%

Overigens hebben we in deze tabel (zoals we in alle tabellen zullen doen in dit hoofdstuk) de waarden van de rijen af laten lopen (3 vaak – 2 soms – 1 nooit). De reden hiervoor is dat wij het op deze manier gemakkelijker vinden om te laten zien dat er een positieve (of negatieve) samenhang is. Bij een positieve samenhang zullen de percentages in de diagonaal van linksonder naar rechtsboven hoger zijn dan in de andere cellen, bij een negatieve samenhang zullen de percentages in de diagonaal van linksboven naar rechtsonder hoger zijn dan in de andere cellen. Hoe je dit zelf kunt doen in SPSS, staat in kader 1.3. Dit laatste is te zien in tabel 6.1. In tabel 6.2 hebben we dezelfde variabelen gebruikt met andere fictieve data en hier zien we een positieve samenhang.

Tabel 6.2 Kruistabel van hoe vaak tv-kijken en hoe vaak boeken lezen, sterk positief verband, (SPSS-output)

hoe vaak boeken lezen * hoe vaak tvkijken Crosstabulation

			hoe vaak tvkijken			
			1 nooit	2 soms	3 vaak	Total
hoe vaak boeken lezen	3 vaak	Count	6	5	32	43
		% within hoe vaak tvkijken	11,8%	26,3%	57,1%	34,1%
	2 soms	Count	8	12	20	40
		% within hoe vaak tvkijken	15,7%	63,2%	35,7%	31,7%
	1 nooit	Count	37	2	4	43
		% within hoe vaak tvkijken	72,5%	10,5%	7,1%	34,1%
Total		Count	51	19	56	126
		% within hoe vaak tvkijken	100,0%	100,0%	100,0%	100,0%

Tot slot kunnen we ook bij ordinale variabelen aan de hand van de percentages zien wanneer er geen verband is, zoals in tabel 6.3. De percentages in de cellen van de drie rijen verschillen niet sterk van de totale percentages in de rechterkolom (kolompercentages over het totaal aantal respondenten).

Tabel 6.3 Kruistabel van hoe vaak tv-kijken en hoe vaak boek lezen, geen samenhang (SPSS-output)

hoe vaak boeken lezen * hoe vaak tvkijken Crosstabulation

			hoe vaak tvkijken			
			1 nooit	2 soms	3 vaak	Total
hoe vaak boeken lezen	3 vaak	Count	13	12	21	46
		% within hoe vaak tvkijken	37,1%	30,0%	41,2%	36,5%
	2 soms	Count	11	14	14	39
		% within hoe vaak tvkijken	31,4%	35,0%	27,5%	31,0%
	1 nooit	Count	11	14	16	41
		% within hoe vaak tvkijken	31,4%	35,0%	31,4%	32,5%
Total		Count	35	40	51	126
		% within hoe vaak tvkijken	100,0%	100,0%	100,0%	100,0%

Ook bij de associatiematen voor ordinale variabelen is er een specifieke maat die rekening houdt met afhankelijke en onafhankelijke variabelen en die dus alleen geschikt is als er een asymmetrische relatie tussen de variabelen is. We beginnen met het bespreken van een associatiemaat die geschikt is voor ordinale variabelen en een symmetrische relatie.

6.2 Gamma

Gamma, aangeduid met de Griekse letter γ, is een associatiemaat voor ordinale variabelen waarbij je geen rekening houdt met een mogelijke afhankelijke of onafhankelijke variabele; het is dus een associatiemaat voor symmetrische

relaties. Centraal in de formule voor gamma staan *concordante paren* en *discordante paren*. Als concordante paren overheersen is er een positieve samenhang en als discordante paren in de meerderheid zijn, is er een negatieve samenhang. Een paar is concordant als de ene onderzoekseenheid op beide variabelen hoger scoort dan de andere onderzoekseenheid. Een paar is discordant als een onderzoekseenheid op de ene variabele hoger en op de andere variabele lager scoort dan de andere onderzoekseenheid.

6.2.1 *Interpretatie*

Je doet onderzoek naar de leesfrequentie van tijdschriften en wilt onder andere weten of er een samenhang bestaat tussen hoe vaak vrouwen de glossy *Linda* lezen hoe vaak zij de *Happinez* lezen. Je houdt een enquête onder alleen vrouwen (of je stelt de vraag ook aan mannen en geeft later met *Select Cases* aan dat je alleen vrouwen als onderzoekseenheden wilt selecteren) en gaat na of zij deze bladen nooit (1), soms (2) of vaak (3) lezen. Omdat beide variabelen ordinaal zijn en er sprake is van een symmetrische relatie (je weet niet welke variabele welke beïnvloedt) is gamma een geschikte maat om vast te stellen of er een verband is. We beginnen met het berekenen van de kolompercentages in een kruistabel om een eerste indruk van een mogelijke samenhang te krijgen.

Tabel 6.4 Kruistabel van frequentie *Linda* en frequentie *Happinez* lezen (SPSS-output)

Happinez * Linda Crosstabulation

			Linda			
			1 nooit	2 soms	3 vaak	Total
Happinez	3 vaak	Count	31	35	24	90
		% within Linda	10,4%	28,7%	38,7%	18,6%
	2 soms	Count	32	47	19	98
		% within Linda	10,7%	38,5%	30,6%	20,3%
	1 nooit	Count	236	40	19	295
		% within Linda	78,9%	32,8%	30,6%	61,1%
Total		Count	299	122	62	483
		% within Linda	100,0%	100,0%	100,0%	100,0%

Aan de percentages is al te zien dat er een positief verband is tussen de frequentie *Linda* lezen en de frequentie *Happinez* lezen. Zo leest 78,9% van de vrouwen in dit onderzoek nooit *Linda* en ook nooit *Happinez* (cel (1,1)). Ook in de volgende kolommen zien we de hoogste percentages in de cellen met dezelfde waarden op de twee variabelen (cellen (2,2) en (3,3)). De waarde van gamma bevestigt dit sterke positieve verband ($\gamma = 0{,}622$, zie tabel 6.5).

Tabel 6.5 Gamma van leeskans *Linda* en *Happinez* (SPSS-output)

Symmetric Measures

		Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
Ordinal by Ordinal	Gamma	,622	,046	10,166	,000
N of Valid Cases		483			

We kunnen dus zeggen dat er een positieve sterke samenhang is tussen de frequentie van het lezen van *Linda* en de frequentie van het lezen van *Happinez*. Dat de samenhang positief is, wil zeggen dat wanneer een vrouw vaker de *Linda* leest, zij ook vaker de *Happinez* zal lezen. Omdat de relatie symmetrisch is (in onze onderzoeksvraag is er immers geen afhankelijke variabele) kunnen we dit ook andersom zeggen: als vrouwen vaak de *Happinez* lezen, lezen ze ook vaak de *Linda*.

De samenhang tussen twee variabelen is positief als er meer concordante dan discordante paren onder de onderzoekseenheden zijn. Wat betreft de frequentie van het lezen van *Linda* en *Happinez* zijn er veel concordante paren te vinden in deze kruistabel. Wat concordante en discordante paren precies zijn en hoe deze zijn te tellen, wordt in de volgende paragraaf uitgelegd.

We bekijken eerst nog een ander voorbeeld. Nu willen we weten of er onder vrouwen een samenhang is tussen de leeskans van het opinieblad *Elsevier* en de glossy *Happinez*.

Tabel 6.6 Kruistabel frequentie *Elsevier* en frequentie *Happinez* lezen (SPSS-output)

Happinez * Elsevier Crosstabulation

			Elsevier			Total
			1 nooit	2 soms	3 vaak	
Happinez	3 vaak	Count	137	3	2	142
		% within Elsevier	55,5%	8,3%	6,5%	45,2%
	2 soms	Count	108	15	9	132
		% within Elsevier	43,7%	41,7%	29,0%	42,0%
	1 nooit	Count	2	18	20	40
		% within Elsevier	0,8%	50,0%	64,5%	12,7%
Total		Count	247	36	31	314
		% within Elsevier	100,0%	100,0%	100,0%	100,0%

Uit tabel 6.6 blijkt een sterke, negatieve samenhang. Wanneer een vrouw nooit *Elsevier* leest, leest zij vaak de *Happinez*, en andersom (en minder vaak de *Happinez* lezen betekent vaker de *Elsevier* lezen). Ook hier is dat te zien aan de percentages. Er is een groot percentage onderzoekseenheden dat nooit *Elsevier* leest en vaak *Happinez* (55,5% in cel (1,3)) en evenzo een groot percentage dat hoog scoort op leeskans *Elsevier* en laag op leeskans *Happinez* (64,5% in cel (3,1)). De richting van de samenhang wordt aangegeven door het minteken. De sterkte van de samenhang komt tot uitdrukking in de grootte van het getal ($\gamma = -0{,}880$: een sterk, negatief verband tussen de twee variabelen).

Ook bij het interpreteren van een ordinale associatiemaat zoals gamma noemen we in de conclusie altijd de waarde van het verband (afgerond op twee decimalen), het aantal onderzoekseenheden, de sterkte van het verband (volgens de richtlijnen zoals beschreven in paragraaf 5.2.1), de richting van het verband (aangegeven door al dan niet een minteken), wat deze richting betekent en de variabelen waar het verband over gaat. Indien bekend worden de onderzoekseenheden genoemd, en ook hier worden minimaal twee percentages (naar eigen inzicht) uit de kruistabel genoemd om het verband toe te lichten.

In een onderzoeksverslag of publicatie zou je op basis van de analyse van tabel 6.6 de volgende tekst kunnen gebruiken:

> *Er blijkt een sterke negatieve samenhang te zijn tussen de frequentie dat vrouwen de Elsevier lezen en de frequentie dat zij de Happinez lezen ($\gamma = -0{,}88$, $n = 314$). Hoe vaker zij de Elsevier lezen, hoe minder vaak zij de Happinez lezen, en andersom. Zo leest 55,5% van de vrouwen die nooit de Happinez lezen vaak de Elsevier, en leest 64,5% van de vrouwen die vaak de Happinez leest nooit de Elsevier.*

6.2.2 Berekening

Voor het berekenen van gamma kijk je naar de *verhouding tussen de waarden van de variabelen*, en niet naar de absolute waarden. In beide eerdere voorbeelden over de leeskans van tijdschriften varieerden de waarden van de variabelen van 1 (laag) tot 3 (hoog). Er is een rangordening, maar de absolute waarden zijn niet relevant. De onderzoeker had ook de waarden 0 (laag), 3 (midden) en 8 (hoog) kunnen kiezen. Gamma was dan op exact hetzelfde getal uitgekomen. Voor de uitleg van de berekening van gamma beginnen we met een eenvoudige 2 × 2-tabel. Hierin kunnen we concordante en discordante paren vinden. Die paren staan centraal in de formule voor gamma:

$$\gamma = \frac{Nc - Nd}{Nc + Nd}$$

Formule voor gamma

Nc staat voor het totale aantal concordante paren, *Nd* voor het totale aantal discordante paren.

Stel, je hebt een 2 × 2 tabel, waarbij je kijkt of het soort baan dat iemand heeft (parttime/fulltime) samenhangt met het inkomen dat die persoon heeft (laag/hoog). Beide variabelen hebben twee waarden waartussen een rangordening bestaat. Voor het gemak hebben we parttime en laag allebei de waarde 0 gegeven, en fulltime en hoog allebei de waarde 1. Er zijn in een 2 × 2-tabel vier cellen, namelijk mensen met een parttimebaan en een laag inkomen (0,0), mensen met een parttimebaan en een hoog inkomen (0,1), mensen met een fulltimebaan en een laag inkomen (1,0) en mensen met een fulltimebaan en een hoog inkomen (1,1).
Om de concordante paren te berekenen, neem je als startpunt een onderzoekseenheid die op beide variabelen (dus zowel op soort baan als op inkomen) de laagste waarde heeft. In dit geval is dat (0,0): iemand met een parttimebaan en een laag inkomen. Hierbij zoeken we iemand die op beide variabelen hoger scoort. In dit geval dus iemand uit cel (1,1), iemand met fulltimebaan en een hoog inkomen. Deze twee onderzoekseenheden vormen een concordant paar. Als van een paar de ene onderzoekseenheid op *beide variabelen hoger scoort* dan de andere onderzoekseenheid, is het een concordant paar. Met andere woorden, onderzoekseenheid B (zie tabel 6.7) scoort op beide variabelen hoger dan onderzoekseenheid A. A en B vormen een concordant paar. Voor personen in de cellen (0,1) en (1,0) kunnen we geen personen vinden die op beide variabelen hoger scoren, omdat ze op een van de twee variabelen al de hoogste score hebben.

Tabel 6.7 Concordante paren

Baan / Inkomen	0 (parttime)	1 (fulltime)
1 (hoog)		B (1,1)
0 (laag)	A (0,0)	

Bij discordante paren moeten de onderzoekseenheden op de *ene variabele hoger, en op de andere variabele lager scoren.* Het startpunt is hier de cel waar onderzoekseenheden op de ene variabele het laagst en op de andere variabele het hoogst scoren, hier bijvoorbeeld (0,1), iemand met een parttimebaan en een hoog inkomen. De onderzoekseenheden die hiermee discordant zijn, moeten aan de voorwaarde voldoen dat de eerste waarde (in de kolom) hoger is dan 0, en de tweede waarde (in de rij) lager is dan 1. Dat is in het voorbeeld van een 2 × 2 tabel alleen in cel (1,0), iemand met een fulltimebaan en een laag inkomen. C en D in tabel 6.8 vormen dus een discordant paar.

Tabel 6.8 Discordante paren

Inkomen \ Baan	0 (parttime)	1 (fulltime)
1 (hoog)	C (1,0)	
0 (laag)		D (0,1)

Nu we weten hoe we de concordante en discordante paren kunnen vinden in een kruistabel, gaan we het toepassen op een kruistabel met meer onderzoekseenheden. Stel, je hebt twaalf mensen ondervraagd, waarvan er zes fulltime werken en zes parttime. Van deze mensen hebben zes mensen een hoog inkomen en zes mensen een laag inkomen (zie tabel 6.9). Voor deze kruistabel willen we gamma uitrekenen. De eerste stap is het berekenen van de concordante paren.

Tabel 6.9 Berekenen van de concordante paren

Inkomen \ Baan	0 (parttime)	1 (fulltime)
1 (hoog)	2	4
0 (laag)	4	2

Als startpunt voor het berekenen van de concordante paren kiezen we de cel met onderzoekseenheden die op beide variabelen (baan en inkomen) de laagste waarden hebben, dus (0,0). In deze cel zitten vier onderzoekseenheden. Er zijn vier mensen die een parttimebaan hebben en een laag inkomen. Deze vier onderzoekseenheden zijn concordant met de vier onderzoekseenheden in cel (1,1), de mensen met een fulltimebaan en een hoog inkomen; deze personen scoren op beide variabelen hoger. Er zijn dus in totaal 4 * 4 = 16 concordante paren. Meer concordante paren kun je met deze twaalf onderzoekseenheden niet maken.
Voor het berekenen van de discordante paren beginnen we met de cel waar op de ene variabele het laagst is gescoord en op de andere variabele het hoogst: (0,1). In deze cel zitten twee onderzoekseenheden, namelijk twee respondenten die een parttimebaan hebben en een hoog inkomen. Deze twee onderzoekseenheden zijn discordant met de twee onderzoekseenheden in cel (1,0). Er zijn dus 2 * 2 = 4 discordante paren. Meer discordante paren zijn er niet.

Tabel 6.10 Berekenen van de discordante paren

Inkomen \ Baan	0 (parttime)	1 (fulltime)
1 (hoog)	2	4
0 (laag)	4	2

Wanneer er meer concordante paren dan discordante paren zijn, is er een positieve samenhang. Indien we alleen concordante paren hadden gehad, was er een perfecte positieve samenhang geweest (+1). Dat is in dit voorbeeld niet het geval. Hoe sterk de samenhang dan wel is, zien we als we de formule voor gamma invullen.

$$\gamma = \frac{Nc - Nd}{Nc + Nd} = \frac{16-4}{16+4} = 0{,}60$$

Gamma geeft de verhouding tussen concordante en discordante paren weer. In dit geval is er een vrij sterke positieve samenhang tussen baan en inkomen. Wanneer iemand een fulltimebaan heeft, heeft diegene vaker een hoger inkomen dan iemand met een parttimebaan; en andersom, iemand met een hoog inkomen heeft vaker een fulltimebaan dan iemand met een laag inkomen.

Bij een grotere tabel, bijvoorbeeld 3 × 3, is het iets ingewikkelder, maar het principe blijft hetzelfde. We gebruiken als voorbeeld het onderzoek naar de frequentie waarmee respondenten *Linda* en *Happinez* lezen. We kijken eerst eens naar de kruistabel zonder de geobserveerde frequenties:

Tabel 6.11 Kruistabel van frequentie Linda en frequentie Happinez

Happinez \ Linda	(1) nooit	(2) soms	(3) vaak
(3) vaak	A	B	C
(2) soms	D	E	F
(1) nooit	G	H	I

In dit geval zijn er geen vier, maar negen cellen waarin je concordante paren en discordante paren kunt vinden. Er zijn namelijk mensen die nooit *Linda* lezen en vaak *Happinez* (onderzoekseenheden A in cel (1,3)); mensen die nooit *Linda* lezen en soms *Happinez* (onderzoekseenheden D in cel (1,2)) enzovoort.
Om de concordante paren uit te rekenen beginnen we weer bij de onderzoekseenheden die op beide variabelen het laagst scoren, in dit geval zijn dat de

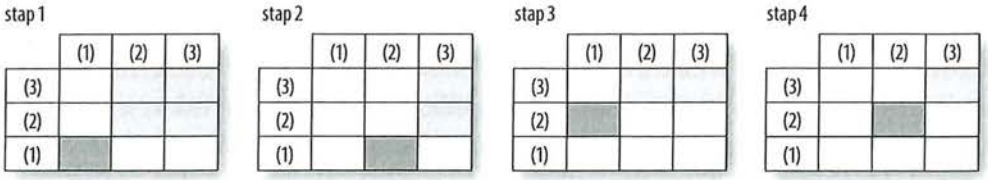
onderzoekseenheden G in (1,1), die nooit de *Linda* lezen en nooit de *Happinez* lezen.

Tabel 6.12 Concordante paren berekenen in 3 × 3 tabel

Linda / Happinez	(1) nooit	(2) soms	(3) vaak
(3) vaak	A	B ♫ ☯	C ♫ ◎ ☯ ♥
(2) soms	D ☯	E ♫ ♥	F ♫ ◎
(1) nooit	G ♫	H ◎	I

De onderzoekseenheden G in cel (1,1) zijn concordant met onderzoekseenheden E, F, B en C (aangegeven met een ♫). De onderzoekseenheden met de waarden (2,2), (3,2), (2,3) en (3,3) scoren namelijk op beide variabelen hoger dan de onderzoekseenheden in het startpunt (1,1).
Maar er zijn meer concordante paren. Onderzoekseenheden H in cel (1,2), zijn concordant met de onderzoekseenheden F in (3,2) en C (3,3) (aangegeven met een ◎). De onderzoekseenheden in F en C scoren namelijk op de ene variabele hoger dan 1, en op de andere variabele hoger dan 2.
Met de onderzoekseenheden I in (3,1) zijn geen concordante paren te maken: er is namelijk geen waarde boven de 3 in deze tabel. (Dat wil zeggen: er kan niet hoger gescoord worden dan 'vaak' een tijdschrift lezen.)
De onderzoekseenheden D in (1,2) kunnen wel concordante paren vormen, namelijk met onderzoekseenheden B en C (aangegeven met een ☯). Tot slot zijn de onderzoekseenheden E concordant met de eenheden C (aangegeven met ♥).

In figuur 6.1 is in vier stappen nogmaals aangegeven hoe de concordante paren te vinden zijn. Let er wel op dat je dit 'patroon' alleen op deze manier kunt gebruiken wanneer de waarden van de variabelen ook op deze manier zijn gerangschikt. In de kolommen lopen de waarden op van links naar rechts en in de rijen lopen de waarden af van boven naar beneden. Als de waarden van de variabelen in een andere volgorde staan, zijn de concordante paren te vinden door na te gaan in welke cellen de waarden van beide variabelen hoger zijn dan de waarden in de cel waarmee wordt gestart. Vanaf de cel waarmee wordt gestart werk je dan systematisch de hele kruistabel door.



Figuur 6.1 Concordante paren in een 3 × 3-tabel

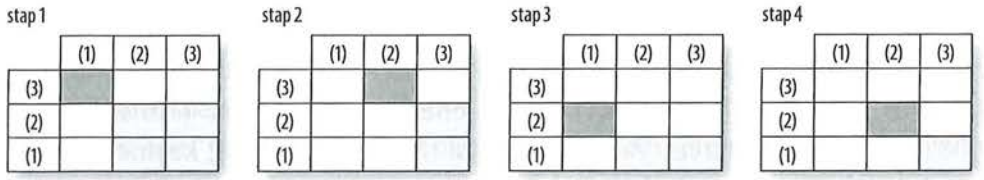
Nu nog de discordante paren. We beginnen nu bij de onderzoekseenheid met de laagste waarde op de ene, en de hoogste waarde op de andere variabele. In dit geval is dat onderzoekseenheid A in cel (1,3), mensen die nooit *Linda* lezen en vaak *Happinez*.

Tabel 6.13 Discordante paren berekenen in 3 x 3 tabel

Happinez \ Linda	(1) nooit	(2) soms	(3) vaak
(3) vaak	A ♫	B ◎	C
(2) soms	D ☯	E ♫ ♥	F ♫ ◎
(1) nooit	G	H ♫ ☯	I ♫ ◎ ☯ ♥

- A is discordant met E, F, H en I (aangegeven met een ♫). Vergeleken met de onderzoekseenheden in het startpunt (1,3) scoren al deze onderzoekseenheden hoger op de 1 (kolomvariabele) en lager op de 3 (rijvariabele), namelijk (2,2), (3,2), (2,1) en (3,1).
- B (2,3) is discordant met F (3,2) en I (3,1) (aangegeven met een ◎).
- C (3,3) is met geen van de onderzoekseenheden discordant, er is namelijk geen waarde hoger dan 3.
- D (1,2) is discordant met H (2,1) en I (3,1) (aangegeven met een ☯).
- E (2,2) discordant met I (3,1) (aangegeven met een ♥).

In figuur 6.2 is in vier stappen aangegeven hoe de discordante paren te vinden zijn (als de waarden van de variabelen in deze volgorde in de kruistabel staan).



Figuur 6.2 Discordante paren in een 3 × 3-tabel

Nu kunnen we gamma gaan berekenen. De gegevens in tabel 6.14 zijn dezelfde als gebruikt in tabel 6.4, met dit verschil dat de percentages en de randtotalen niet zijn overgenomen. Die hebben we voor de berekening van gamma niet nodig.

Tabel 6.14 Berekenen van gamma Linda – Happinez

Linda / Happinez	(1) nooit	(2) soms	(3) vaak
(3) vaak	31	35	24
(2) soms	32	47	19
(1) nooit	236	40	19

We tellen eerst de concordante paren, waarbij we beginnen met de 236 onderzoekseenheden in cel (1,1). Deze zijn concordant met de onderzoekseenheden in (2,2), (2,3), (3,2) en (3,3). Er zijn dus 236 * (47 + 19 + 35 + 24) = 29500 concordante paren te maken met de onderzoekseenheden in cel (1,1).
Daarna gaan we naar de volgende cel in de kruistabel. De 40 onderzoekseenheden in cel (2,1) zijn concordant met de 19 eenheden in (3,2) en 24 in (3,3). Oftewel: 40 * (19 + 24) = 1720. Op dezelfde manier rekenen we ook de andere concordante paren in deze kruistabel uit.

236 * (47 + 19 + 35 + 24)	=	29500
40 * (19 + 24)	=	1720
32 * (35 + 24)	=	1888
47 * (24)	=	1128
		+
	Totaal	34236

Nc (aantal concordante paren) is 34236.

Daarna berekenen we de discordante paren, waarbij we beginnen met de 31 onderzoekseenheden in cel (1,3). Deze zijn discordant met de onderzoekseenheden in (2,2), (3,2), (2,1) en (3,1). Dan hebben we dus 31 * (47 + 19 + 40 + 19) = 3875 discordante paren. Vanuit de cellen (2,3), (1,2) en (2,2) kunnen we ook nog discordante paren vinden. Op deze manier rekenen we ook de rest van de discordante paren uit.

31 * (47 + 19 + 40 + 19)	=	3875
35 * (19 + 19)	=	1330
32 * (40 + 19)	=	1888
47 * (19)	=	893
		+
	Totaal	7986

Nd (aantal discordante paren) is dus 7986.

Aan de hand van de berekeningen is te zien dat er een positieve samenhang zal zijn tussen de twee variabelen; we hebben immers meer concordante paren dan

discordante paren. Om te zien hoe sterk deze positieve samenhang is, vullen we de formule voor gamma in.

$$\gamma = \frac{Nc - Nd}{Nc + Nd} = \frac{34236 - 7986}{34236 + 7986} = \frac{26250}{42222} = 0{,}622$$

De waarde komt exact overeen met de waarde zoals berekend door SPSS (tabel 6.5). We kunnen concluderen dat er onder vrouwen een sterke positieve samenhang is tussen de frequentie van het lezen van de *Linda* en de *Happinez* ($\gamma = 0{,}62$, $n = 483$). Naarmate vrouwen meer de ene glossy lezen, lezen ze ook meer de andere glossy, en andersom.

6.3 Somers' d

De volgende associatiemaat die we bespreken wanneer we een vraagstelling of hypothese met twee ordinale variabelen hebben is Somers' d (*dyx*). Deze lijkt sterk op gamma. Gamma is echter het meest geschikt bij symmetrische verbanden en Somers' d is het meest geschikt bij asymmetrische verbanden. Om Somers' d te berekenen moeten we dus weten wat de afhankelijke en wat de onafhankelijke variabele is.[1] Een verschil in de berekening met gamma is dat Somers' d naast concordante en discordante paren, ook rekening houdt met *geknoopte paren*. Geknoopte paren hebben op één van de twee variabelen dezelfde waarde.

6.3.1 Interpretatie

We nemen als voorbeeld de hypothese 'hoe meer educatieve spelletjes peuters op een tablet spelen, hoe groter hun woordenschat zal zijn'. De onderzoekseenheden zijn hier peuters, en de variabelen zijn 'hoeveelheid educatieve spelletjes spelen op een tablet' en 'grootte van de woordenschat'. Afhankelijk van de manier waarop deze variabelen gemeten zijn, is het meetniveau ordinaal, interval of ratio. In dit hoofdstuk laten we verbanden zien aan de hand van ordinale variabelen, en daarom hebben we de variabele 'hoeveelheid educatieve spelletjes spelen op een tablet' voor het gemak ingedeeld naar 1 = niet/weinig en 2 = veel. De variabele 'grootte woordenschat' hebben we ingedeeld in drie categorieën, namelijk 1 = kleine woordenschat, 2 = medium woordenschat, 3 = grote woordenschat. Deze categorieën zullen in 'werkelijk' onderzoek onderbouwd moeten worden bij de beschrijving van de operationalisatie van je begrippen in variabelen. In dit geval is sprake van een asymmetrisch verband, we verwachten namelijk dat de hoeveelheid educatieve spellen (onafhankelijke variabele, *x*) invloed zal hebben op de grootte van de woordenschat (afhankelijke variabele, *y*). Omdat we twee ordinale variabelen hebben en uitgaan van een asymmetrisch verband, is Somers' d de meest geschikte maat.

Een onderzoeker heeft in een observationele studie bijgehouden hoe vaak per week een peuter op een tablet educatieve spellen heeft gespeeld, en geturfd hoeveel verschillende woorden de peuter in een week uitspreekt. Aan de hand daarvan stelt hij de volgende kruistabel op:

Tabel 6.15 Kruistabel woordenschat naar hoeveelheid educatieve spellen op tablet (SPSS-output)

woordenschat * spel op tablet Crosstabulation

			spel op tablet		
			1 weinig/ niet	2 veel	Total
woordenschat	3 groot	Count	1	11	12
		% within spel op tablet	5,3%	68,8%	34,3%
	2 medium	Count	7	2	9
		% within spel op tablet	36,8%	12,5%	25,7%
	1 klein	Count	11	3	14
		% within spel op tablet	57,9%	18,8%	40,0%
Total		Count	19	16	35
		% within spel op tablet	100,0%	100,0%	100,0%

Uit de kolompercentages in tabel 6.15 is af te lezen dat peuters die veel educatieve spellen op een tablet ook een grotere woordenschat hebben dan peuters die weinig of niet op de tablet spellen spelen. We verwachten op basis van deze percentages een positieve samenhang. De analyse (zie *Directional Measures* in tabel 6.16) bevestigt deze verwachting.

Tabel 6.16 Somers' d van invloed hoeveelheid educatieve spellen op tablet op woordenschat (SPSS-output)

Directional Measures

			Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
Ordinal by Ordinal	Somers'd	Symmetric	,550	,126	4,352	,000
		woordenschat Dependent	,638	,145	4,352	,000
		tablet Dependent	,483	,112	4,352	,000

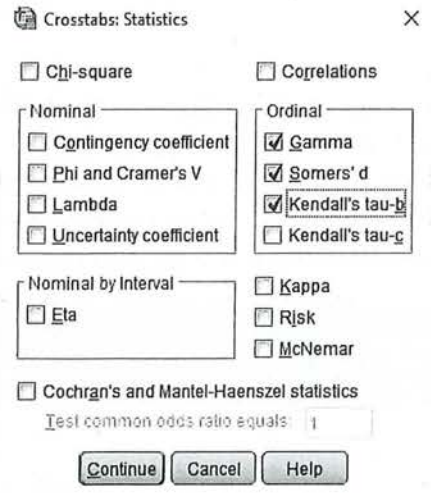
Evenals bij de asymmetrische nominale maten (lambda en Goodman en Kruskals tau) vind je in de SPSS-output meerdere waarden van deze associatiemaat. SPSS weet immers niet welke variabele wij als onafhankelijk en welke wij als afhankelijk hebben benoemd. Aangezien in dit geval de woordenschat de afhankelijke variabele is, gebruiken we de waarde voor Somers' d die achter 'woordenschat Dependent' te vinden is.

We kunnen de volgende conclusie rapporteren:

Er is een sterke positieve samenhang tussen de hoeveelheid educatieve spellen spelen op een tablet en de grootte van de woordenschat bij peuters ($d_{yx} = 0{,}64$, $n = 35$). Hoe vaker zij educatieve spellen spelen op een tablet, hoe groter hun woordenschat is. 68,8% van alle peuters die vaak een spel speelt heeft een grote woordenschat, terwijl maar 5,3% van de peuters die weinig of niet de tablet voor deze doeleinden gebruikt een grote woordenschat heeft.

SPSS Berekenen van gamma, Somers' d en Kendalls tau-b

Om gamma en Somers' d te laten berekenen, volg je dezelfde stappen als bij de associatiematen op nominaal niveau. Eerst maak je een kruistabel. Percenteer op de kolommen via de knop *Cells*. Via de knop *Statistics* kun je aangeven welke associatiemaat je bij deze kruistabel wilt laten uitdraaien, zoals gamma, Somers' d of Kendalls tau-b.



Figuur A Statistics-venster: ordinale associatiematen

Kader 6.1

6.3.2 Berekening

De formule van Somers' d lijkt sterk op die van gamma, met als toevoeging in de noemer de geknoopte paren op de afhankelijke variabele:

$$d_{yx} = \frac{Nc - Nd}{Nc + Nd + Ty}$$

Formule voor Somers' d

We zoeken voor Somers' d de paren die geknoopt zijn op de afhankelijke variabele, op *y*. *Ty* betekent *ties* (knopen) op *y*. In ons voorbeeld in paragraaf 6.2.1 is de afhankelijke variabele de grootte van de woordenschat. We willen immers weten hoe de afhankelijke variabele *y* (woordenschat) varieert met de hoeveelheid educatieve spellen dat op een tablet wordt gespeeld (onafhankelijke variabele, *x*). In tabel 6.17 zie je dezelfde informatie als in kruistabel 6.15, maar weer zonder totalen en percentages.

Tabel 6.17 Woordenschat naar hoeveelheid educatieve spellen op tablet (berekenen van concordante en discordante paren)

Tablet / woordenschat	(1) weinig/niet	(2) vaak
(3) groot	1	11
(2) medium	7	2
(1) klein	11	3

Eerst berekenen we het aantal concordante paren (*Nc*), waarbij we als startpunt de onderzoekseenheden (hier: peuters) nemen die op beide variabelen het laagst scoren, dus die weinig/niet de tablet gebruiken en een kleine woordenschat hebben (1,1). Dit zijn er 11 en deze zijn concordant met de onderzoekseenheden in de cellen die op beide variabelen hoger scoren: de 2 peuters in cel (2,2) en de 11 peuters in cel (2,3). Daarna hebben we nog één concordant paar, namelijk de onderzoekseenheden in cel (1,2) met de onderzoekseenheden in cel (2,3).

```
11 * (2 + 11)  =      143
 7 * (11)      =       77
               ----------- +
               Totaal 220
```

Het aantal concordante paren (*Nc*) is 220.

Voor het aantal discordante paren (*Nd*) beginnen we in cel (1,3), waar 1 onderzoekseenheid op de ene variabele (aantal educatieve spellen op tablet) het laagst scoort en op de andere variabele (woordenschat) het hoogst scoort. Deze is discordant met de cellen waar op de onafhankelijke variabele hoger wordt gescoord en op de afhankelijke variabele lager, dus met (2,2) en (2,1). Tot slot vormen (1,2) en (2,1) nog een discordant paar.

```
1 * (2 + 3)  =       5
7 * (3)      =      21
             ---------- +
             Totaal 26
```

In totaal zijn er 26 discordante paren (*Nd*).

Hadden we voor dit voorbeeld een symmetrisch verband verwacht (dat wil zeggen: geen duidelijke afhankelijke variabele), dan hadden we met deze informatie gamma kunnen uitrekenen. Die was in dit geval 0,789 geweest. We hebben echter wél een duidelijke verwachting, en rekenen daarom Somers' d uit, waarbij we ook rekening houden met de knopen op de afhankelijke variabele.

Een paar is geknoopt op de afhankelijke variabele (y) als de onderzoekseenheden op die afhankelijke variabele dezelfde waarde hebben, maar op de onafhankelijke variabele een andere waarde.
We beginnen te knopen bij cel (1,1) (peuters die weinig of geen tabletspellen spelen en een kleine woordenschat hebben). Hierin bevinden zich 11 onderzoekseenheden die geknoopt zijn met de 3 onderzoekseenheden daarnaast. Deze 3 hebben namelijk dezelfde waarde op de afhankelijke variabele y (namelijk 1, een kleine woordenschat), maar een andere waarde op de onafhankelijke variabele x (namelijk 2, vaak op een tablet spellen spelen). Voor de rij waarbij $y = 1$ zijn derhalve 11 * (3) = 33 paren geknoopt (op y).
Vervolgens doen we hetzelfde voor de rij waarbij $y = 2$. Hiervoor geldt dat (1,2) geknoopt is met (2,2). Wederom blijft de waarde voor y dus hetzelfde (hier: gemiddelde woordenschat), maar verandert de waarde van x. Hetzelfde doen we voor de derde rij (y = 3). NB: Uiteraard maakt het niet uit of je begint te 'knopen' in de bovenste of in de onderste rij!

Tabel 6.18 Knopen op de afhankelijke variabele y

Woordenschat \ Tablet	(1) weinig/niet	(2) vaak
(3) groot	1 →	11
(2) medium	7 →	2
(1) klein	11 →	3

Het berekenen Ty gaat dus als volgt:

$y = 1$; 11 * (3) = 33
$y = 2$; 7 * (2) = 4
$y = 3$; 1 * (11) = 11
——— +
Ty = 58

Er zijn 58 geknoopte paren op y.

Het invullen van de formule levert uiteraard dezelfde uitkomst op als die we eerder vonden in de SPSS-output:

$$d_{yx} = \frac{Nc - Nd}{Nc + Nd + Ty} = \frac{220 - 26}{220 + 26 + 58} = \frac{194}{304} = 0{,}638$$

Er is een sterke positieve samenhang tussen het aantal educatieve spellen dat peuters op een tablet speelt en de grootte van hun woordenschat.

Bij een 3 × 3-tabel (of groter) werkt het principe hetzelfde. Laten we naar een eenvoudige tabel kijken om dit te illustreren. We gaan kijken of onder fantasy-liefhebbers (de onderzoekseenheden) de waardering van het boek *The Lord of the Rings* invloed heeft op de waardering van de film *The Lord of the Rings*. Drie sterren staat voor goed, twee voor matig en één voor slecht.

Tabel 6.19 Waardering film – waardering boek

Waardering film \ Waardering boek	★	★★	★★★
★	2	5	10
★★	2	5	5
★	10	5	2

Eerst bereken je de concordante en discordante paren.

Het berekenen van de concordante paren:

10 * (5 +5 + 5 +10) = 250
5 * (5 + 10) = 75
2 * (5 + 10) = 30
5 * (10) = 50
+
Nc = 405

Het berekenen van de discordante paren:

2 * (5 + 5 + 5 + 2) = 34
5 * (5 + 2) = 35
2 * (5 + 2) = 14
5 * (2) = 10
+
Nd = 93

Er zijn 405 concordante paren en 93 discordante paren.

Nu kijken we naar de geknoopte paren op *y*. We beginnen met de mensen die de film (afhankelijke variabele) slecht vinden en met één ster waarderen. Er zijn tien mensen die zowel de film als het boek slecht vinden. Zij zijn geknoopt

met de mensen die de film ook slecht vinden, maar waarbij de waardering voor het boek beter is, namelijk matig (twee sterren) of goed (drie sterren). Dat zijn 10 * (5 + 2) = 70 geknoopte paren. Maar er is voor dezelfde waarde van y nog een mogelijkheid. De vijf onderzoekseenheden die de film slecht vinden en het boek matig in cel (★★, ★) zijn namelijk geknoopt met de twee onderzoekseenheden die de film slecht vinden en het boek goed in cel (★★★, ★). Dus dat zijn nog eens 5 * (2) = 10 geknoopte paren. Weer is het zo dat de afhankelijke variabele y (waardering voor de film) gelijk blijft, maar de onafhankelijke variabele x (waardering voor het boek) verschilt. In totaal zijn er 70 + 10 = 80 geknoopte paren op y = ★(1 ster). Dit doen we vervolgens voor alle rijen.

Het berekenen van *Ty*:

y = ★;	10 * (5 + 2)	=	70
	5 * (2)	=	10
y = ★★;	2 * (5 + 5)	=	20
	5 * (5)	=	25
y = ★★★;	2 * (5 + 10)	=	30
	5 * (10)	=	50
			+
	Ty	=	205

In totaal zijn er 205 geknoopte paren op y.

Nu we alle gegevens hebben, kunnen we de formule voor Somers' d invullen:

$$d_{yx} = \frac{Nc - Nd}{Nc + Nd + Ty} = \frac{405 - 93}{405 + 93 + 205} = 0{,}444$$

Somers' d is 0,444. De conclusie die we trekken is:

> *We vinden een redelijk sterke positieve samenhang tussen de waardering van het boek en de film The Lord of the Rings (d_{yx} = 0,44, n = 46). Hoe hoger het boek door fantasyliefhebbers wordt gewaardeerd, hoe hoger ze de film zullen waarderen.*

6.4 Kendalls tau-b

Kendalls tau-b (τ_b) is net als gamma een maat voor symmetrische relaties. Evenals Somers' d houdt Kendalls tau-b rekening met geknoopte paren. Het verschil met Somers' d is dat Kendalls tau-b geen onderscheid maakt tussen de afhankelijke en de onafhankelijke variabele (want symmetrisch), en de geknoopte paren op beide variabelen bij de berekening meetellen (bij Somers' d waren het alleen de geknoopte paren op de afhankelijke variabele). Kendalls tau-b komt

het meest tot zijn recht bij vierkante tabellen, omdat tau anders de waarden +1 en –1 niet kan bereiken.

Aan de hand van tabellen 6.20 en 6.21 is te zien dat het knopen van de paren veel verschil kan maken in de uitkomst van de associatiemaat. Er wordt gekeken of er een verband is tussen de mate waarin het televisienieuws wordt gekeken en de mate waarin de krant wordt gelezen (hier beide ingedeeld in de twee categorieën (1) weinig en (2) veel).

Tabel 6.20 Kruistabel krant en nieuws (γ en $\tau_b = 1$)

krant \ nieuws	(1) weinig	(2) veel	Totaal
(2) veel	0	10	10
(1) weinig	10	0	10
Totaal	10	10	20

Tabel 6.21 Kruistabel krant en nieuws ($\gamma = 1$; $\tau_b = 0{,}471$)

krant \ nieuws	(1) weinig	(2) veel	Totaal
(2) veel	7	5	12
(1) weinig	8	0	8
Totaal	15	5	20

In tabel 6.20 zijn zowel gamma als tau-b gelijk aan 1, er zijn alleen concordante paren (namelijk 10 * 10 = 100) en geen discordante paren (0 * 0 = 0) en geen geknoopte paren. In tabel 6.21 zijn er nog steeds geen discordante paren (7 * 0 = 0), waardoor gamma weer de maximale waarde van 1 bereikt. Er zijn echter wel geknoopte paren (namelijk 56 geknoopte paren op de onafhankelijke variabele, en 35 geknoopte paren op de afhankelijke variabele, de berekening van geknoopte paren bij tau zullen we in paragraaf 6.4.2 laten zien), waardoor Kendalls tau-b een veel lagere waarde heeft dan gamma (namelijk 0,471). Bij tabel 6.21 zouden we dus aan de hand van gamma concluderen dat er een perfecte positieve samenhang is, terwijl we aan de hand van tau-b concluderen dat er slechts een redelijk positieve samenhang is.

6.4.1 *Interpretatie*

We bekijken Kendalls tau-b aan de hand van het eerdere voorbeeld van de waardering van het boek en de film van *The Lord of the Rings* (tabel 6.19). We hebben nu echter niet de verwachting dat de waardering van het boek de waardering van de film beïnvloedt. Wellicht is het andersom, en beïnvloedt de waardering van de film juist de waardering van het boek. We weten ook niet of de fantasyliefhebbers (want dat waren de onderzoekseenheden) het boek voor of na het zien van de film hebben gelezen. We onderzoeken een symmetrische relatie tussen twee ordinale variabelen, die evenveel waarden hebben (namelijk allebei drie), we hebben dus een vierkante kruistabel, en daarom is Kendalls tau-b een geschikte maat.

Tabel 6.22 Kruistabel van waardering boek met waardering film en Kendalls tau-b (SPSS-output)

film * boek Crosstabulation

			boek			Total
			1 slecht	2 redelijk	3 goed	
film	3 goed	Count	2	5	10	17
		% within boek	14,3%	33,3%	58,8%	37,0%
	2 redelijk	Count	2	5	5	12
		% within boek	14,3%	33,3%	29,4%	26,1%
	1 slecht	Count	10	5	2	17
		% within boek	71,4%	33,3%	11,8%	37,0%
Total		Count	14	15	17	46
		% within boek	100,0%	100,0%	100,0%	100,0%

Symmetric Measures

		Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
Ordinal by Ordinal	Kendall's tau-b	,446	,112	4,010	,000
N of Valid Cases		46			

De conclusie aan de hand van de kruistabel en de associatiemaat is:

> *Er is een redelijk sterke positieve samenhang tussen de waardering van het boek en de waardering van de film ($\tau_b = 0{,}45$, $n = 46$). Wanneer fantasyliefhebbers het boek meer waarderen, waarderen ze ook de film meer, en omgekeerd. Uit de kruistabel blijkt bijvoorbeeld dat 58,8% van de fantasyliefhebbers die het boek als goed waardeerden, ook de film als goed waardeerden, en 71,4% van de liefhebbers die het boek als slecht waardeerden, dit ook van de film vonden.*

NB: Het maakt bij dit voorbeeld niet uit of we de waardering van het boek in de kolommen zetten of de waardering van de film, het is immers een symmetrische veronderstelling die we analyseren. Omdat we wel altijd op de kolommen percenteren, hebben we deze percentages vermeld in de conclusie. Hadden we de waardering voor de film in de kolommen gezet, dan hadden hier andere percentages gestaan, maar de waarde van de Kendalls tau-b zou hetzelfde zijn geweest.

6.4.2 Berekening

Net als Somers' d maakt Kendalls tau-b gebruik van knopen, maar de associatiemaat knoopt zowel op de rijvariabele (y) als op de kolomvariabele (x). Dat wordt ook duidelijk in de formule:

$$\tau_b = \frac{Nc - Nd}{\sqrt{(Nc + Nd + Tx)(Nc + Nd + Ty)}}$$

Formule voor Kendalls tau-b

We berekenen nu met de hand de waarde van Kendalls tau-b die hoort bij de kruistabel van de waardering van het boek en de waardering van de film *The Lord of the Rings* (tabel 6.19 en 6.22). De concordante en discordante paren hadden we al berekend bij het berekenen van Somers' d, net als het aantal geknoopte paren op de y-variabele. We hadden al gevonden: $Nc = 405$, $Nd = 93$ en $Ty = 205$. We hoeven dus alleen nog de knopen op x te berekenen.
Het idee is hetzelfde als bij de geknoopte paren op y, alleen blijft nu x constant en variëren de waarden van y. We starten bij het paar met de laagste waarden (1,1) (slechte waardering boek, slechte waardering film). Deze is geknoopt op x met (1,2) en (1,3). De waarde van x blijft gelijk, de waarden van y verschillen. Het gaat steeds om de onderzoekseenheden die weinig waardering hebben voor het boek (waarde 1 op x), maar variëren op de waarde van y. Dit geeft 10 * (2 + 2) = 40 geknoopte paren vanuit cel (1,1). Vervolgens kijken we naar cel (1,2), die ook nog geknoopt is met (1,3). Dat zijn 2 * (2) = 4 geknoopte paren.
In totaal heeft de kolom met $x = 1$ dus 40 + 4 = 44 geknoopte paren. Hetzelfde doen we voor de volgende kolommen $x = 2$ en $x = 3$.

Tabel 6.23 Waardering film naar waardering boek

Waardering film \ Waardering boek	★ (1)	★★ (2)	★★★ (3)
★★★ (3)	2	5	10
★★ (2)	2	5	5
★ (1)	10	5	2

Het berekenen van de geknoopte paren op x gaat hetzelfde in zijn werk als knopen op y. Alleen kijken we nu in verticale richting in de tabel. Omdat we als startpunt hebben gekozen voor de onderzoekseenheden met op allebei de variabelen de laagste waarde (1,1) kijken we dus naar alle knopen van onder naar boven in de kolom.

$$
\begin{array}{lrcl}
x = 1; & 10 * (2 + 2) & = & 40 \\
& 2 * (2) & = & 4 \\
x = 2; & 5 * (5 + 5) & = & 50 \\
& 5 * (5) & = & 25 \\
x = 3; & 2 * (5 + 10) & = & 30 \\
& 5 * 10 & = & 50 \\
\hline
& & Tx = & 199
\end{array} +
$$

Er zijn 199 geknoopte paren op x.
Nu we alle gegevens hebben, kunnen we Kendalls tau-b berekenen.

$$\tau_b = \frac{Nc - Nd}{\sqrt{(Nc + Nd + Tx)(Nc + Nd + Ty)}} = \frac{405 - 93}{\sqrt{(405 + 93 + 199)(405 + 93 + 205)}}$$

$$= \frac{312}{699{,}994} = 0{,}446$$

Dit komt overeen met de eerdere gegevens van SPSS (tabel 6.22).

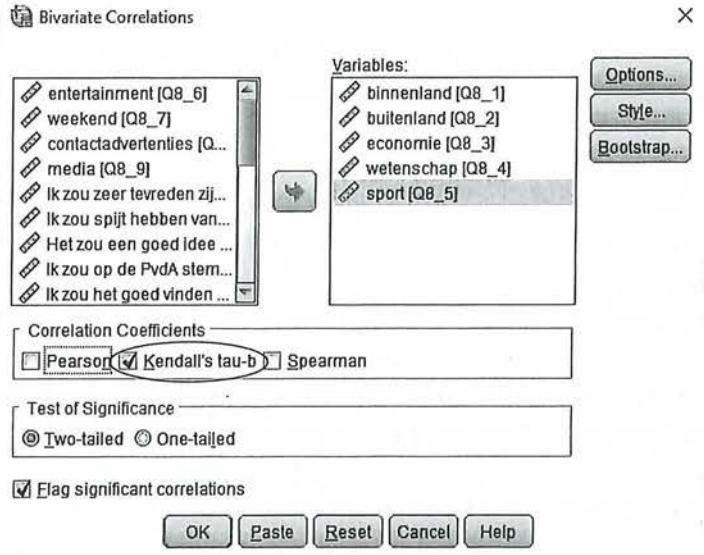
6.4.3 *Kendalls tau-b in een correlatiematrix*

Het uitrekenen van een associatiemaat is een bivariate analyse. In SPSS kun je ook meerdere associatiematen tegelijk uitrekenen. Die worden dan in een correlatiematrix gezet. Deze matrix geeft dan de resultaten van meerdere bivariate analyses. Dit is mogelijk met Kendalls tau-b (kader 6.2 geeft aan hoe je dit in SPSS kunt uitvoeren). Dit is handig wanneer je wilt weten of meerdere variabelen al dan niet sterk met elkaar samenhangen.
Stel dat je in een enquête vijf vragen hebt gesteld die allemaal moeten meten in welke mate krantenlezers de verschillende katernen binnen de krant interessant vinden. De vragen worden ingeleid als stelling ('ik vind het binnenlandkatern interessant', 'ik vind het buitenlandkatern interessant' enzovoort) en de antwoordcategorieën zijn bij alle vijf de variabelen: 1 (niet mee eens), 2 (niet mee oneens of eens) en 3 (mee eens). Je bent benieuwd of er samenhang is tussen de interesse in het ene en interesse in het andere katern. Het gaat hier dus om vijf symmetrische relaties op ordinaal niveau. In de correlatiematrix kun je dan zien of alle onderlinge correlaties inderdaad hoog zijn en welke variabelen het sterkst samenhangen.

SPSS Correlatiematrix met Kendalls tau-b

Het uitrekenen van Kendalls tau-b's in een correlatiematrix wijkt af van de hiervoor getoonde methode om ordinale associatiematen te berekenen. In hoofdstuk 8 laten we zien dat op deze manier ook een correlatiematrix op interval- en rationiveau wordt berekend.

Ga voor het berekenen van Kendalls tau-b over meerdere variabelen via *Analyze* → *Correlate* naar *Bivariate*. Selecteer de variabelen die je nodig hebt en vink 'Kendall's tau-b' aan. Standaard staat de optie 'Pearson' aan. Zet deze voor de overzichtelijkheid uit. De Pearson correlatiecoëfficiënt komt aan bod in hoofdstuk 8.



Figuur A Bivariate Correlations-venster: Kendall's tau-b

Kader 6.2

In tabel 6.24 is te zien dat SPSS alle variabelen tegenover elkaar afzet. De diagonaal in de correlatiematrix is altijd 1, dit is de correlatie van die variabele met zichzelf. Wat ook opvalt, is dat alle correlatiecoëfficiënten twee keer voorkomen. Dit is omdat de correlatie door SPSS voor beide richtingen berekend wordt. Uiteraard is de samenhang tussen binnenland en buitenland, dezelfde als de correlatie tussen buitenland en binnenland. Je hoeft dus maar één kant van de diagonaal te bekijken om de juiste correlaties te vinden.

We zien dat alle correlaties positief zijn: hoe hoger op de ene variabele wordt gescoord, hoe hoger op de andere variabele wordt gescoord. Ook zien we dat alle correlaties redelijk sterk tot sterk zijn: de laagste correlatie wordt gevonden tussen 'binnenland' en 'sport' (tau-b = 0,479). We concluderen in dat geval: *onder krantlezers is er een redelijk sterke positieve samenhang tussen de interesse in het binnenlandkatern en het sportkatern* ($\tau_b = 0{,}48$, $n = 135$).

Tabel 6.24 Correlatiematrix van interesse in verschillende krantenkaternen, Kendalls tau-b (SPSS-output)

Correlations

			Q8_1 binnen-land	Q8_2 buiten-land	Q8_3 econo-mie	Q8_4 weten-schap	Q8_5 sport
Kendall's tau-b	Q8_1 binnenland	Correlation Coefficient	1,000	,761**	,560**	,572**	,479**
		Sig. (2-tailed)	.	,000	,000	,000	,000
		N	135	135	135	135	135
	Q8_2 buitenland	Correlation Coefficient	,761**	1,000	,639**	,589**	,558**
		Sig. (2-tailed)	,000	.	,000	,000	,000
		N	135	135	135	135	135
	Q8_3 economie	Correlation Coefficient	,560**	,639**	1,000	,591**	,652**
		Sig. (2-tailed)	,000	,000	.	,000	,000
		N	135	135	135	135	135
	Q8_4 weten-schap	Correlation Coefficient	,572**	,589**	,591**	1,000	,488**
		Sig. (2-tailed)	,000	,000	,000	.	,000
		N	135	135	135	135	135
	Q8_5 sport	Correlation Coefficient	,479**	,558**	,652**	,488**	1,000
		Sig. (2-tailed)	,000	,000	,000	,000	.
		N	135	135	135	135	135

**. Correlation is significant at the 0.01 level (2-tailed).

Welke variabelen hangen nu het sterkst met elkaar samen? We hebben tot nu toe steeds gekeken naar de samenhang tussen twee variabelen. Maar stel, we willen weten welke drie variabelen het sterkst met elkaar samenhangen. We gaan dan eerst op zoek naar de sterkste samenhang in de correlatiematrix. Dat is de samenhang tussen 'binnenland' en 'buitenland' (tau-b = 0,761). Vervolgens gaan we kijken welke variabele het sterkst samenhangt *met een van deze twee variabelen*. We gaan dus alleen kijken in de kolommen van 'binnenland' en van 'buitenland'. We zien dan dat de variabele 'economie' de sterkste samenhang heeft (namelijk met 'buitenland', tau-b = 0,639). Hoewel de samenhang tussen 'sport' en 'economie' sterker is (tau-b = 0,652), laten we deze variabele toch buiten beschouwing als we echt alleen maar de drie sterkst samenhangende variabelen willen hebben. We kunnen aan de hand van vorenstaande correlatiematrix dus concluderen dat 'binnenland', 'buitenland' en 'economie' het sterkst met elkaar samenhangen.

6.5 Spearmans rho

De *rangcorrelatiecoëfficiënt* van Spearman is een maat voor de correlatie tussen rangnummers. *Spearmans rho* (ρ of r_s) kun je niet alleen gebruiken om de samenhang tussen variabelen op ordinaal niveau te bepalen, maar ook voor variabelen die een interval- of rationiveau hebben. Rho is, net als gamma en Kendalls tau-b, een symmetrische maat. Maar anders dan gamma en Kendalls tau-b is Spearmans rho niet gebaseerd op concordante en discordante paren,

maar op verschillen in de rangorde van de waarden. We zullen eerst een voorbeeld bespreken waarbij Spearmans rho wordt gebruikt bij ordinale variabelen, daarna bij interval- en ratiovariabelen.

6.5.1 *Interpretatie*

Anders dan de voorgaande associatiematen die we besproken hebben, wordt Spearmans rho niet berekend aan de hand van een kruistabel, maar vanuit een datamatrix. De interpretatie van deze maat is niet anders dan van gamma en Kendalls tau-b, maar de maat wordt gebruikt wanneer de variabele relatief veel waarden heeft en de rangordening van de waarden van belang is. Als er veel waarden zijn, is een kruistabel onhandig en onduidelijk. Voor de berekening van Spearmans rho starten we dan ook niet met een kruistabel.

We gaan na of er een verband bestaat tussen het aantal uur dat adolescenten gebruikmaken van 'social network sites' (sns) en hoe gelukkig ze zichzelf vinden. Het aantal uur dat zij doorbrengen op 'social network sites' is een ordinale variabele waarbij klassen zijn aangebracht. Geluk wordt gemeten met een rapportcijfer, een intervalvariabele.

Tabel 6.25 Correlatiematrix Spearmans rho tussen geluk en sns in uren (SPSS-output)

Correlations

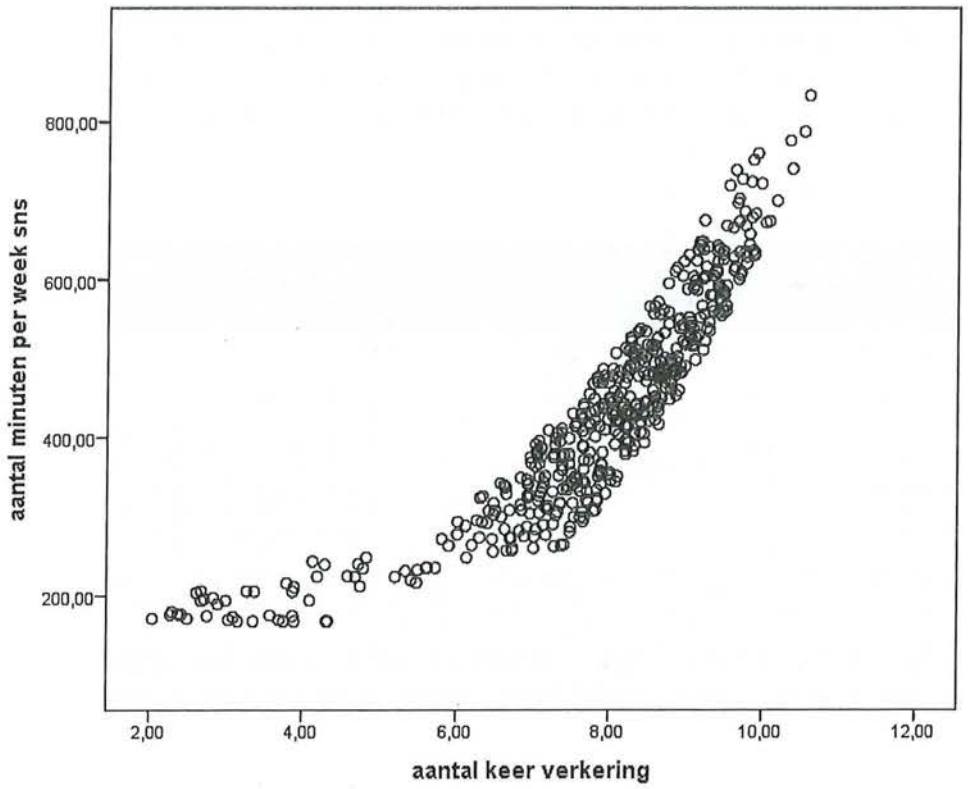
			uur_sns	geluk
Spearman's rho	uur_sns	Correlation Coefficient	1,000	-,829
		Sig. (2-tailed)	.	,042
		N	6	6
	geluk	Correlation Coefficient	-,829	1,000
		Sig. (2-tailed)	,042	.
		N	6	6

We concluderen op basis van tabel 6.25:

> *Uit de correlatie van Spearmans rho blijkt een sterke negatieve samenhang (ρ_s = -0,83, n = 6) tussen het rapportcijfer dat adolescenten zichzelf geven voor de mate van geluk en het aantal uur dat zij per week besteden aan 'social network sites'. Hoe meer tijd zij aan deze sns besteden, hoe minder gelukkig zij zijn, en hoe gelukkiger zij zijn, hoe minder tijd zij aan sns besteden.*

Spearmans rho kan dus bij correlaties worden gebruikt waarbij minimaal één van de variabelen ordinaal is, maar wordt veelal gebruikt wanneer twee interval- of ratiovariabelen met elkaar in verband worden gebracht waarvan tenminste één van de twee of beide niet normaal verdeeld is (zijn) of wanneer een kromlijnig verband tussen de twee bestaat. Zoals we later in hoofdstuk 8 zullen

zien, is rechtlijnigheid (of: het afwezig zijn van een kromlijnig verband) een van de voorwaarden voor het uitvoeren van een correlatie bij twee interval- of ratio-variabelen. Is er wel een kromlijnig verband, dan is het alsnog mogelijk om Spearmans rho te gebruiken, omdat deze niet naar de waarden zelf in de matrix kijkt, maar rangnummers gebruikt voor de berekeningen. We hadden al gezien in paragraaf 3.6.1 dat extreme waarden of outliers ervoor kunnen zorgen dat een verdeling niet meer normaal verdeeld is. Zo kunnen extreme waarden er ook voor zorgen dat correlaties tussen twee variabelen veel hoger of veel lager uitvallen dan wanneer deze extreme waarden er niet in zouden zitten. Bepaalde waarden van onderzoekseenheden trekken dan letterlijk 'het verband uit het lood', zoals te zien is in figuur 6.3.



Figuur 6.3 Voorbeeld van kromlijnig verband tussen aantal keer verkering en aantal minuten per week sns

In bovenstaand *spreidingsdiagram* is gekeken naar het verband tussen hoe vaak tieners verkering hebben gehad (hier op de x-as) en hoeveel minuten zij per week aan sociale netwerken besteden. In het spreidingsdiagram is te zien dat er een aantal tieners is dat op beide variabelen laag scoort, waardoor de kromming ontstaat. Bij Spearmans rho maken deze extreme waarden echter niet uit omdat daar rangordeningen worden gebruikt (die variëren van 1 tot en met n) en niet met de oorspronkelijke data wordt gerekend. In paragraaf 8.1.1 zullen we dieper ingaan op het maken en interpreteren van een spreidingsdiagram.

6.5.2 Berekening

Tot nu toe waren de besproken associatiewaarden te berekenen op basis van een kruistabel. Dat is bij Spearmans rho niet het geval. Voor de berekening van rho gebruiken we de kolommen van de datamatrix.

$$\rho_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Formule voor Spearmans rho

In de formule is d (van het Engelse *difference*) het verschil tussen de rangnummers van waarden van de variabelen x en y. Het berekenen is eenvoudig: de verschillen tussen de rangnummers voor x en y kwadrateer je en tel je daarna bij elkaar op. Om Spearmans rho te berekenen moet je dus eerst die rangnummers bepalen. We laten het zien aan hand van het voorbeeld over de samenhang tussen geluk en aantal uur sociale netwerken per week gebruiken.

Tabel 6.26 Rang van rapportcijfer geluk en aantal uur sns per week

Respondent	Rapportcijfer geluk	Rang geluk	Aantal uur per week sns	Rang uur sns
Sarah	9	1	0 tot 1 uur	6
Madelief	5	5	5 tot 7 uur	4
Lente	7	3	9 tot 11 uur	3
Thomas	8	2	2 tot 3 uur	5
Olivier	4	6	14 tot 15 uur	1
Sam	6	4	12 tot 13 uur	2

In tabel 6.26 is voor zes tieners informatie gegeven over het rapportcijfer dat ze zichzelf geven voor hoe gelukkig ze zijn en de tijd dat zij aan 'social network sites' besteden in uren per week. Sarah scoort van alle respondenten het hoogst op 'geluk' (9), en het laagst op 'uur sns' (0 tot 1 uur). Zij heeft dus 'de eerste plaats' (of: de eerste rang) op de variabele 'geluk', en de laatste plaats (of: de laatste rang) op de variabele 'uur sns'. Olivier brengt de meeste tijd op 'social network sites' door (14 tot 15 uur), dus bezet daarmee de eerste rang, en is het minst gelukkig, en krijgt daarom voor die variabele de zesde rang. Lente krijgt met haar rapportcijfer van een 7 de rang 3 voor geluk, en met haar 9 tot 11 uur sns gebruiken rang 3 voor 'uur sns'. De waarde van de rangorde voor elke persoon loopt voor elk van deze twee variabelen van 1 tot en met n, in dit geval dus van 1 tot en met 6.

Als de rangordeningen zijn vastgesteld, kun je het verschil tussen de rangnummers van de waarden van de variabelen x en y berekenen en kwadrateren (de d^2 in de formule). We gaan dus voor elke respondent zijn of haar rangnummer op y (rangnummer uur per week sns) aftrekken van zijn of haar rangnummer op x (rangnummer rapportcijfer geluk). Vervolgens kwadrateren we per respondent de uitkomst van deze berekening, en tellen we deze uitkomsten bij elkaar op:

Tabel 6.27 Berekenen van het verschil tussen de rangnummmers van rapportcijfer geluk en aantal uur sns per week

Respondent	Rang geluk (*rang x*)	Rang uur sns (*rang y*)	*d* (= *rang x* – *rang y*)	d^2
Sarah	1	6	-5	25
Madelief	5	4	1	1
Lente	3	3	0	0
Thomas	2	5	-3	9
Olivier	6	1	5	25
Sam	4	2	2	4
Σ				64

Nu we dat hebben gedaan kunnen we de formule van Spearmans rho invullen, en zien we dat deze uitkomst dezelfde is als in tabel 6.25:

$$\rho_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6(64)}{6(6^2-1)} = 1 - \frac{384}{210} = 1 - 1{,}829 = -0{,}829$$

Maar wat doen we als mensen een gedeelde plaats hebben? In het vorige voorbeeld verschillen alle respondenten van elkaar met betrekking tot het rapportcijfer voor geluk en de hoeveelheid tijd die zij besteden aan sociale netwerken, zodat je de rangnummers gemakkelijk kunt bepalen. In de praktijk zal het vaak zo zijn dat meerdere respondenten het rapportcijfer '7' geven, of '0 tot 1 uur gebruik van sns maken'. In dat geval hebben ze dus een gedeelde positie. Dit laten we zien aan de hand van een ander voorbeeld.

Tabel 6.28 Gedeelde rang van aantal keer verkering

Respondent	Aantal minuten sns (*x*)	Rang *x*	Aantal keer verkering (*y*)	Rang *y*
A	1200	1	6	2,5
B	850	2	3	6
C	240	3	8	1
D	210	4	3	6
E	180	5	6	2,5
F	160	6	3	6
G	120	7	5	4
H	60	8	2	8
I	30	9	0	10
J	0	10	1	9

In tabel 6.28 is te zien dat voor het aantal minuten sns geen gedeelde rangen zijn. Respondent A en B scoren wel erg hoog in vergelijking met de rest van de respondenten, maar dat maakt bij Spearmans rho niet uit, omdat we niet met de data zelf rekenen, maar met de rangen die we daaraan toewijzen. Hier is dus duidelijk te zien dat ook met extreme waarden gemakkelijk gerekend kan worden. Wat betreft de afhankelijke variabele 'aantal keer verkering' is wel een aantal geknoopte rangen te zien. Zo zien we dat respondenten A en E allebei zes keer verkering hebben gehad, en daarom hebben ze een gedeelde tweede en derde plaats. Ook respondenten B, D en F hebben even vaak verkering gehad (namelijk drie keer). Om het rangnummer te bepalen, neem je de gemiddelde rangpositie. Voor respondenten A en E is dat de som van hun rangposities (2 + 3), gedeeld door het aantal 'knopen' (2).

$$\frac{2+3}{2}=2{,}5$$

Voor respondenten B, D en F geldt dezelfde berekening:

$$\frac{5+6+7}{3}=6$$

Nu de rangordening bepaald is, kun je Spearmans rho uitrekenen.

Tabel 6.29 Berekening van het verschil tussen de rangnummers van minuten sns en verkering

Respondent	Rang *x*	Rang *y*	*d*	d^2
A	1	2,5	-1,5	2,25
B	2	6	-4	16
C	3	1	2	4
D	4	6	-2	4
E	5	2,5	2,5	6,25
F	6	6	0	0
G	7	4	3	9
H	8	8	0	0
I	9	10	-1	1
J	10	9	1	1
Σ				43,5

$$\rho_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6*43,5}{10*(10^2-1)} = 1 - \frac{261}{990} = 1 - 0,264 = 0,736$$

De conclusie die we trekken aan de hand van deze analyse is:

> *Er is sprake van een sterke positieve samenhang tussen de tijd die tieners aan sociale netwerksites besteden en hoe vaak zij verkering hebben gehad ($\rho_s = 0,74$, $n = 10$). Tieners die vaak verkering hebben gehad besteden meer tijd aan sns, en andersom: tieners die meer tijd aan sns besteden hebben vaker verkering.*

Uiteraard bevestigt SPSS dit (zie tabel 6.30).[2]

Tabel 6.30 Spearmans rho correlatie tussen het minuten sns en verkering (SPSS-output)

Correlations

			minuten_sns	verkering
Spearman's rho	minuten_sns	Correlation Coefficient	1,000	,732*
		Sig. (2-tailed)	.	,016
		N	10	10
	verkering	Correlation Coefficient	,732*	1,000
		Sig. (2-tailed)	,016	.
		N	10	10

*. Correlation is significant at the 0.05 level (2-tailed).

6.6 Samenvatting

De associatiematen gamma, Somers' d, Kendalls tau-b en Spearmans rho kunnen waarden aannemen die liggen tussen –1 (perfecte negatieve samenhang) en +1 (perfecte positieve samenhang). Naast de sterkte van een verband tussen twee variabelen geven ordinale associatiematen de richting van de samenhang aan.
Gamma, Kendalls tau-b en Spearmans rho zijn symmetrische maten, waarbij je bij de berekening geen rekening houdt met een eventuele (on)afhankelijke variabele. Somers' d is asymmetrisch en komt, net als tau-b, het beste tot zijn recht bij vierkante tabellen.
Als er sprake is van een symmetrische relatie tussen ordinale variabelen, kan de onderzoeker kiezen uit drie associatiematen: gamma, Kendalls tau-b en Spearmans rho. Afhankelijk van de specifieke kenmerken en berekeningswijze van de associatiemaat heeft in sommige situaties de ene en in andere situaties de andere associatiemaat de voorkeur. Bij de berekening van Kendalls tau-b gebruik je meer informatie dan bij de berekening van gamma. Kendalls tau-b houdt niet alleen rekening met concordante en discordante paren, maar ook met geknoopte paren. Gamma is dan ook een grovere maat dan Kendalls tau-b. Kendalls tau-b is echter weer minder geschikt als een kruistabel niet vierkant is Spearmans rho is vooral geschikt als er relatief veel waarden zijn en de rangorde van belang is. Deze associatiemaat kies je vaak als het gaat om interval- of ratiovariabelen, die erg scheef verdeeld zijn. Spearmans rho is minder geschikt als relatief veel onderzoekseenheden dezelfde waarden hebben en een rangnummer moeten delen.

Tabel 6.31 Samenvatting associatiematen

	Nominaal	Ordinaal
Symmetrisch	Cramers V phi	gamma Kendalls tau-b Spearmans rho
Asymmetrisch	Goodman en Kruskals tau lambda	Somers' d

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

Noten

1 Somers' d komt het best tot zijn recht bij vierkante tabellen (dus 2 × 2, 3 × 3, 4 × 4 enzovoort), omdat Somers' d bij niet-vierkante kruistabellen de waarde +1 en –1 niet kan bereiken. Bij de interpretatie van Somers' d in een niet-vierkante tabel houd je daar rekening mee.
2 Het afrondingsverschil komt doordat wij met drie decimalen achter de komma rekenen en SPSS met meer.