Hoofdstuk 7

Bivariate associatiematen voor nominale en ordinale variabelen

1. Inleiding

In dit hoofdstuk bespreken we de belangrijkste parameters voor de beschrijving van een verband tussen kenmerken van onderzoekseenheden op nominaal en ordinaal niveau. Deze technieken zijn belangrijk, want veel eigenschappen waarin criminologen geïnteresseerd zijn, zijn van het nominale en ordinale niveau.

2. Het percentageverschil als associatiemaat op nominaal niveau

In een kruistabel of contingentietabel staan de variabelen met hun frequenties paarsgewijs horizontaal en verticaal geplaatst, elk met een aantal (nominale) kenmerken, verdeeld over verschillende categorieën. Kruistabellen zijn tweedimensionale tabellen, verdeeld over kolommen en rijen. We maken een onderscheid tussen 2*2 tabellen en r*k tabellen, waarbij r en k staan voor het aantal rijen en kolommen. R*k tabellen zijn een extensie van de eenvoudige 2*2 tabel. Bij kruistabellen is het zo dat in elke cel het geobserveerde aantal staat bij een bepaalde combinatie van kenmerken. Kruistabellen worden gebruikt wanneer we de relatie tussen twee nominale kenmerken willen bestuderen. De verschillen in de afhankelijke variabele worden vergeleken over de verschillende categorieën van de onafhankelijke variabele. We gebruiken hiervoor vaak het *percentageverschil* en de *odds ratio*.

Stel dat we onderzoek doen naar de betrokkenheid bij een problematische jeugdgroep en geslacht. We willen weten of jongens in sterkere mate betrokken zijn bij een problematische jeugdgroep dan meisjes In een kruistabel kunnen we deze informatie eenvoudig weergeven. De samenhang tussen twee nominale variabelen kunnen we omschrijven als de manier waarop een kenmerk verdeeld is binnen de categorieën van een ander kenmerk. In het voorbeeld kijken we dus naar de manier waarop betrokkenheid bij geweld in groepsverband verdeeld is naar geslacht in een eenvoudige 2*2 tabel.

Betrokkenheid bij een problematische jeugdgroep en geslacht

		Geslacht		Totaal
		Meisjes	Jongens	
Betrokkenheid bij problematische jeugdgroep	Niet betrokken	1208	1120	2328
	Betrokken	45	101	146
Totaal		1253	1221	2474

1208, 1120, 45 en 101 noemen we de **celfrequenties**: ze geven aan hoeveel keer een bepaalde combinatie van categorieën voorkomt. (1208+1120) of 2328 en (45+101) of 146 noemen we de **rijtotalen**. (1208+45) of 1253 en (1120+101) of 1221 noemen we de **kolomtotalen**. Rijtotalen en kolomtotalen worden ook wel de **marginalen** genoemd. De marginalen geven de univariate frequentieverdeling van de respectievelijke variabelen weer. De som van alle celfrequenties is het totaal aantal waarnemingen, de *populatieomvang* of de *steekproefomvang*. Aangezien het interpreteren van absolute aantallen in de categorieën vaak moeilijk is (doordat de groepen die we vergelijken niet hetzelfde totaal aantal hebben), moeten de absolute aantallen omgerekend worden naar proporties of percentages (herinner je: percentages zijn proporties vermenigvuldigd met honderd). Zeker indien we de associatie tussen twee variabelen wensen te vergelijken, is het werken met proporties of percentages noodzakelijk.

Een belangrijk probleem hierbij is de *richting waarin we percenteren*. We kunnen binnen de rijen, binnen de kolommen of ten opzichte van het totaal aantal waarnemingen percenteren. De richting van de associatie wordt bepaald door theoretische verwachtingen die men heeft. Statistisch dienen we bij het maken van een bivariate kruistabel het onderscheid te maken tussen een *onafhankelijke* en een *afhankelijke* variabele.

- Een afhankelijke variabele is een kenmerk dat beïnvloed wordt door één of meerdere andere kenmerken.
- Een onafhankelijke variabele is een kenmerk dat invloed uitoefent op een ander kenmerk.

In het voorbeeld beschouwen we geslacht als een onafhankelijke variabele en het betrokken zijn bij een problematische jeugdgroep als een afhankelijke variabele. Het moge nogmaals duidelijk zijn dat geslacht nooit een afhankelijke variabele kan zijn. Of men jongen of meisje is, wordt niet bepaald door het lidmaatschap van een problematische jeugdgroep. Dit is een

inhoudsloze uitspraak. Werken met kruistabellen impliceert, zoals elke bivariate (en multivariate) analyse, een gedegen criminologische argumentatie waarom bepaalde kenmerken als afhankelijk dan wel als onafhankelijk worden beschouwd. In het onderstaande voorbeeld beschouwen we geslacht als onafhankelijke variabele. We hebben de gewoonte om de afhankelijke variabele in de rijen te plaatsen en de onafhankelijke variabele in de kolommen. Dit is echter niet verplicht. Als de afhankelijke variabele in de rijen wordt gepresenteerd, dan dienen kolompercentages gemaakt te worden en dienen deze percentages te worden vergeleken voor de verschillende categorieën van geslacht. Als de afhankelijke variabele in de kolommen wordt geplaatst, dienen de rijpercentages worden berekend. We gaan in dit handboek echter steeds de afhankelijke variabele in de rijen plaatsen.

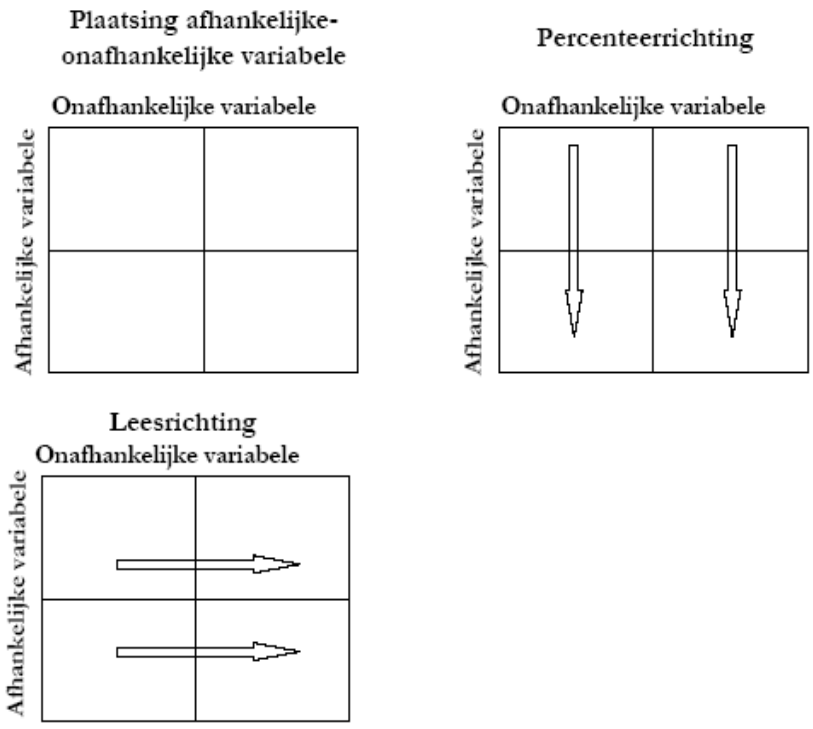
Betrokkenheid bij een problematische jeugdgroep en geslacht

			geslacht		Totaal
			meisje	jongen	
Betrokkenheid bij problematische jeugdgroep	Niet betrokken	Absoluut aantal	1208	1120	2328
		Kolompercentage binnen geslacht	96,4%	91,7%	94,1%
	Betrokken	Absoluut aantal	45	101	146
		Kolompercentage binnen geslacht	**3,6%**	**8,3%**	**5,9%**
Totaal		Absoluut aantal	1253	1221	2474
		Kolompercentage binnen geslacht	100,0%	100,0%	100,0%

In het voorbeeld zien we dat slechts 3.6% van de meisjes lid is van een problematische jeugdgroep en 8.3% van de jongens. Dit geeft een verschil van 4.7 **percentagepunten**. De logica van een analyse van kruistabellen is helder: bekijk de relatieve frequenties van de variabele “betrokkenheid bij een problematische jeugdgroep” nu eens als waren het twee aparte univariate frequentieverdelingen. Dan zie je dat 3.6% van de meisjes betrokken is bij een problematische jeugdgroep en dan zie je dat 8.3% van de jongens uit de steekproef betrokken is bij een problematische jeugdgroep. Stel nu eens dat er helemaal geen associatie was tussen de twee kenmerken. Wat zouden we dan verwachten? Weet je het niet? Volg dan intuïtief je verstand en redeneer mee. Als er geen verband zou bestaan tussen de beide kenmerken, dan zou het toch logisch zijn dat de frequentieverdelingen gelijk lopen? Als er geen verband bestaat, dan verwachten we dat de situatie bij jongens en bij meisjes niet zou afwijken van de totale frequentieverdeling: het percentage van 5.9% zou dan zowel bij meisjes als bij jongens moeten geobserveerd worden. Maar dat doen we niet. We observeren dat de percentages verschillend zijn, en wel in het nadeel van de jongens. Jongens hebben dus

een grotere kans om betrokken te zijn bij een problematische jeugdgroep. Het verband tussen geslacht en betrokkenheid bij problematische jeugdgroepen, gangs, bendes, georganiseerde misdaad, enz. is bijzonder stabiel. De meeste studies tonen dergelijk verband aan. Hoe we het verband dienen te interpreteren, is nog maar de vraag. Geslacht op zich is geen oorzakelijk mechanisme dat een gebeurtenis kan teweegbrengen, maar geslacht kan wel een markeerder zijn voor een ander ongemeten kenmerk, zoals het testosteron-gehalte. Het kan ook zijn dat er geen biologische maar een sociaalpsychologisch mechanisme aan de basis ligt van de geobserveerde samenhang. Mogelijks is het zo dat er minder toezicht is op jongens dan op meisjes en is het zo dat jongens er een meer risicovolle levensstijl op na houden die hen meer met bendes in contact brengt. Hierdoor hebben ze een grotere kans om door bendes gerekruteerd te worden. De cijfers spreken dus niet voor zichzelf. De criminoloog dient de interpretatie te maken en dat is vaak moeilijker dan de eenvoudige berekening van een statistisch verband.

Laten we nu nog eenmaal terugkeren op de richting voor het percenteren omdat studenten daar vaak tegen zondigen. Onderstaande figuur vat de wijze waarop kruistabellen dienen gelezen te worden nog eens samen.



De wijze waarop we percenteren is volledig afhankelijk van de plaats die afhankelijke en onafhankelijke variabelen in een tabel innemen. Het is niet steeds mogelijk om onafhankelijke en afhankelijke variabelen van elkaar te onderscheiden. Dit is het geval waar twee variabelen worden vergeleken en waarbij de ene de verklaring kan vormen voor de andere en omgekeerd. Het valt dus wel eens voor dat er theoretische discussie bestaat over de richting van de relatie tussen twee kenmerken. Nemen we bijvoorbeeld religie en onverdraagzaamheid: het is mogelijk dat het aanhangen van een bepaalde religie leidt tot een hogere mate van onverdraagzaamheid. Anderzijds kan het evengoed zijn dat onverdraagzame mensen een bepaalde soort religie gaan aanhangen. In dat geval dient de criminoloog een keuze te maken en die te verduidelijken. Dit kan door bijvoorbeeld te verwijzen naar wat er in de literatuur gezegd wordt.

Naast de richtlijnen bij het percenteren van de tabellen dienen volgende regels te worden gevolgd:

1. Vermeld altijd het *totaal aantal waarnemingen* indien enkel de percentages of proporties in een tabel werden opgenomen.
2. Bereken *nooit een percentage wanneer het aantal gevallen kleiner is dan 30*. Wanneer we de waarnemingen in een bepaalde categorie met één eenheid verminderen dan betekent dit een verlaging met 3%. Terwijl een verschil van 3% bij een groot aantal waarnemingen significant mag genoemd worden, is dit bij minder dan 30 waarnemingen niet meer het geval. De kans op foute interpretaties wordt dan ook heel groot. In dit geval is het beter enkel de absolute waarden weer te geven.

3. De odds ratio als associatiemaat op nominaal niveau

Een andere manier om de relatie tussen twee nominale kenmerken te bestuderen is door de **odds ratio** te berekenen. De odds is een maat om de *verhouding tussen het voorkomen van een gebeurtenis en het niet voorkomen van een gebeurtenis* uit te drukken. (vb. Het aantal 12 jarigen in de steekproef dat overgaat tot het plegen van een delict gedeeld door het aantal 12 jarigen in die steekproef dat niet overgaat tot het plegen van een delict). De odds ratio is de verhouding tussen twee odds en wordt ook wel de *kruisproduct ratio* genoemd. De odds ratio is een a**symmetrische associatiemaat** die heel gemakkelijk te berekenen is. Verder is het zo dat de odds ratio een aantal heel interessante eigenschappen heeft. Odds ratio's zijn namelijk niet zo gevoelig voor de marginale verdelingen. We merkten hierboven op dat het percenteren

niet zonder gevaren is in kleine steekproeven. De odds ratio kan een alternatief zijn. De odds ratio is tevens de basis voor de logistische regressie-analyse die verder nog aan bod komt. De interpretatie van de odds ratio is relatief eenvoudig. **De odds ratio neemt de waarde aan van 1 bij afwezigheid van een verband en wijkt af van 1 naarmate het verband sterker wordt. De afwijking gebeurt naar 0 of naar + oneindig.**[6]

Hernemen we de hierboven gepresenteerde tabel over de relatie tussen geslacht en betrokkenheid bij een problematische jeugdgroep, dan kunnen we de verhoudingen tussen het betrokken zijn en niet betrokken zijn eerst afzonderlijk berekenen voor jongens en meisjes. Deze afzonderlijke verhoudingen noemen we **odds**.

Uitgewerkt rekenvoorbeeld:

		geslacht		Totaal
		meisje	jongen	
Betrokkenheid bij problematische jeugdgroep	Geen betrokkenheid	1208	1120	2328
	Betrokkenheid	45	101	146
Totaal		1253	1221	2474

De verhouding tussen het niet-betrokken zijn bij een problematische jeugdgroep en het betrokken zijn bij een problematische jeugdgroep (of de odds voor het niet-betrokken zijn bij een problematische jeugdgroep) ziet er voor meisjes als volgt uit: (1208/45) of 26,84. *Dit betekent dat meisjes 26,84 keer meer kans hebben om niet betrokken te zijn bij een problematische jeugdgroep dan wel betrokken te zijn bij een problematische jeugdgroep.* Onder jongens ziet deze verhouding er als volgt uit: (1120/101) of 11,08.

Wanneer we nu de verhouding nemen van deze twee verhoudingen, kennen we de relatieve verhouding naar geslacht. De odds ratio of de verhouding tussen twee odds naar geslacht ziet

[6] De presentatie van odds kan verwarrend zijn omdat de waarde een wordt aangenomen bij geen verband en omdat een odds-ratio niet negatief kan zijn. Dit kan worden opgelost door met je rekenmachine de odds-ratio om te zetten in een logaritmische schaal (natuurlijk logaritme). Het natuurlijk logaritme van 1 is nul, en nul wordt dan het nulpunt en betekent geen verband. Odds-ratio's die lager zijn dan 1, hebben een negatieve log-odds waarde en odds-ratio's die hoger zijn dan 1 hebben een positieve waarde. De multivariate analysetechnieken voor categorische data-analyse, zoals de logistische regressieanalyse, is gebaseerd op de odds-ratio's en log-odds. Dit is niet belangrijk in een verkennend handboek, waar enkel de basis gekend dient te zijn, maar in latere jaren wordt deze materie zeker behandeld.

er als volgt uit: 26,84 / 11,08 of 2,4. *Dit betekent dat meisjes 2,4 keer meer kans hebben om niet betrokken te zijn bij een problematische jeugdgroep dan jongens.* Er is dus een associatie.

Bij wijze van voorbeeld tonen we in onderstaande tabel een reeks van risico- en beschermingsfactoren en hun relatie met regelovertredend gedrag. De resultaten zijn afkomstig uit de zelfrapportagestudie van eerstegraadsleerlingen te Antwerpen, waarbij 2486 jongeren werden bevraagd over hun regelovertredend gedrag. Odds-ratio's worden vaak gepresenteerd in onderzoek naar **risicofactoren** van regelovertredend gedrag. De odds-ratio is dus een belangrijke maat in criminologisch onderzoek.

De tabel is een goede manier om enerzijds verbanden te leren interpreteren en aflezen en anderzijds na te denken over de betekenis van de verbanden. Alle variabelen zijn gebaseerd op meerdere indicatoren (vragen) en de antwoorden werden daarna herleid tot twee categorieën: hoog of laag. Neem het voorbeeld van de werkloosheid van de moeder. Je ziet dat er een positief verband bestaat tussen de werkloosheid van de moeder en ernstige delinquentie (OR van 1.56). Jongeren met een werkloze moeder hebben 1.56 meer kans om ernstig delinquent gedrag te vertonen dan jongeren die geen werkloze moeder hebben. Het verband is er wel, maar is lang niet zo sterk als andere verbanden die in de tabel af te lezen zijn. Neem nu lage zelfcontrole: jongeren met een lage zelfcontrole (dus jongeren die er niet goed in slagen zichzelf te beheersen) hebben 4.15 keer meer kans op het vertonen van ernstige delinquentie dan jongeren met een hoge zelfcontrole. Bekijk de tabel eens en tracht voor jezelf en tracht alle verbanden te lezen. Welke vind je zelf opmerkelijk? Welke verbanden had je niet verwacht?

Odds-Ratio's voor ernstige delinquentie, gewelddadige delinquentie, veelplegers, niet-plegers en betrokkenheid bij een gewelddadige jeugdgroep (VYG)

Risicofactoren of beschermingsfactoren	Ernstige delinquentie	Gewelddadige delinquentie	Veelplegers	Betrokkenheid bij een problematische jeugdgroep
Jongen	**2.84**	**2.78**	**2.83**	**2.12**
Beide ouders Belgisch	**0.48**	**0.48**	**0.54**	**0.44**
Blijven zitten	**1.97**	**1.51**	**2.02**	**2.00**
Eenoudergezin	1.24	1.17	**1.53**	1.29
Werkloosheid vader	**2.84**	**1.38**	1.21	1.11
Werkloosheid moeder	**1.56**	**1.28**	**1.48**	1.19
Armoede	**1.41**	1.18	1.33	**1.65**
Residentiële stabiliteit buurt	**0.61**	**0.71**	**0.65**	**0.56**
Ooit opgepakt door politie	**7.65**	**5.77**	**10.92**	**5.48**
Lid gewelddadige jeugdgroep	**7.51**	**6.43**	**13.40**	**--**
Ouderlijke gehechtheid	**0.49**	**0.48**	**0.37**	**0.34**
Ouderlijk toezicht	**0.33**	**0.39**	**0.18**	**0.15**
Integratie in de klas	0.85	0.95	1.04	0.76
Studiebetrokkenheid	**0.37**	**0.43**	**0.22**	**0.36**
Lage moraliteit	**4.66**	**4.91**	**10.34**	**8.82**
Externe locus of control	**2.46**	**2.39**	**3.10**	**3.88**
Impulsiviteit	**3.43**	**3.70**	**6.98**	**5.37**
Woedebeheersing	**3.27**	**3.81**	**4.25**	**3.37**
Relatieve deprivatie	1.00	**1.21**	0.89	0.95
Criminele geneigdheid	**5.80**	**7.51**	**16.71**	**14.32**
Lage zelfcontrole	**4.15**	**4.73**	**9.13**	**5.58**
Criminele leeftijdsgenoten	**3.98**	**4.85**	**7.89**	**7.88**
Ongestructureerde vrijetijd	**2.65**	**2.85**	**4.40**	**3.65**
Risicovolle leefstijl	**5.51**	**5.46**	**12.14**	**11.21**

Bron: Pauwels (2007) de vetgedrukte resultaten zijn statistisch significant ($p < 0.05$)

4. Chi-kwadraat (X^2) als associatiemaat op nominaal niveau

Een **chi-kwadraattoets** wordt in de statistiek gebruikt om te zien of waargenomen celfrequenties systematisch afwijken van verwachte celfrequenties indien geen associatie zou bestaan tussen twee kenmerken. Een chi-kwadraattoets wordt veel gebruikt om kruistabellen te analyseren. De chi-kwadraat kan beschouwd worden als een maat voor de sterkte van een

relatie tussen twee variabelen gemeten op **nominaal** meetniveau, of tussen een nominale en een ordinale variabele. De waarde van de chi-kwadraat neemt toe naarmate de associatie tussen de variabelen sterker is. Het is echter geen eenduidig te interpreteren maat. Het is m.a.w. moeilijk deze associatiemaat te interpreteren. *De chi-kwadraat varieert van 0 bij afwezigheid van een verband tot zeer hoge waarden.* Er is **geen absolute begrenzing** aan de waarden die chi-kwadraat kan aannemen, vandaar dat het moeilijk is de sterkte van het verband te interpreteren. Daarom zijn er door statistici op chi-kwadraat gebaseerde associatiematen bedacht, zoals ***Phi,*** de ***contingentiecoëfficient C*** **en** ***Cramer's V,*** die eenvoudiger te interpreteren zijn.

Chi-kwadraat is gebaseerd op een vergelijking van twee kruistabellen. De eerste kruistabel bestaat uit de ***geobserveerde frequenties (de werkelijke waarden)***, deze wordt vergeleken met een hypothetische tabel (of het theoretische model) waarin de ***verwachte frequenties bij statistische onafhankelijkheid*** worden berekend. Dit vereist een beetje meer uitleg. Wat bedoelen we met deze beide begrippen? Het is cruciaal dat je ze kent en goed begrijpt, anders slaag je er niet in de chi-kwadraat waarde uit te rekenen. We gaan hier uit van een redenering die een beetje te vergelijken valt met de situatie van het percentageverschil: we hebben daar vastgesteld dat er een verband is tussen twee variabelen als de frequentieverdeling voor de beide categorieën van de onafhankelijke variabelen niet dezelfde was: er was een verschil tussen jongens en meisjes in termen van hun betrokkenheid bij een problematische jeugdgroep. Als er geen verband zou zijn tussen de beide kenmerken, dan verwachten we dat de totale frequenties niet verschillend zijn bij jongens en meisjes. Maar dat waren ze in het voorbeeld dus duidelijk wel. Welnu, bij de berekening van chi-kwadraat passen we opnieuw een analoge redenering toe. We vertrekken vanuit de geobserveerde celfrequenties en gaan deze situatie vergelijken met de verwachte celfrequenties die we zouden vinden in de hypothetische situatie dat er geen verband bestaat tussen de beide kenmerken. Het komt er dus op aan die "verwachte celfrequenties in de situatie dat er geen verband bestaat tussen beide kenmerken" zelf te gaan berekenen.

Als je deze waarde hebt gevonden, kan je die waarde in rekening brengen en het verschil berekenen tussen de werkelijke waarde of de geobserveerde waarde in een cel en de verwachte waarde in de situatie dat er geen verband is. De wetenschappelijke benaming hiervoor is de verwachte situatie bij statistische onafhankelijkheid. De verwachte frequenties bij statistische onafhankelijkheid worden berekend op basis van de marginalen van de

geobserveerde waarden en de frequenties in de tabel worden zodanig verspreid over de verschillende waarden dat beide variabelen geen enkele associatie hebben. De formule voor de berekening van chi-kwadraat ziet er als volgt uit:

$$\chi^2 = \sum \frac{(f(o)_{ij} - f(e)_{ij})^2}{f(e)_{ij}} \qquad \chi^2 = \sum \frac{(geobserveerd - verwacht)^2}{verwacht}$$

Chi-kwadraat is evenwel erg gevoelig voor het aantal meeteenheden in onze tabel. Als het aantal respondenten in een steekproef verdubbelt, dan verdubbelt ook de chi-kwadraat waarde in vergelijking met de oorspronkelijke chi-kwadraat. Daarom wordt aangeraden een sterktemaat te berekenen die de chi-kwadraat ongevoelig maakt voor de grootte van de steekproef.

Opgelet! Om **chi-kwadraat** te kunnen gebruiken, dienen een aantal **voorwaarden** vervuld te zijn:

- **Ten eerste**, de data in de contingentietabel dienen ruwe frequenties te zijn, geen scores noch percentages.
- **Ten tweede**, de onderzochte variabelen dienen categorisch te zijn en de meetwaarden dienen elkaar uit te sluiten. Elke observatie (meeteenheid) mag slechts in één cel thuishoren.
- **Ten derde**, de X^2 waarde mag maar geïnterpreteerd worden indien aan een aantal voorwaarden is voldaan. Deze houden in dat maximaal 20% van de cellen een verwachte frequentie bevatten van < 5, en geen enkele een verwachte frequentie van 0. Indien dit wel het geval is, dienen cellen samengetrokken te worden.[7]

[7] Statistische verwerkingspakketten zoals SPSS geven ons altijd een waarschuwing of aan deze voorwaarde voldaan is: 0 cells (0%) have expected count less than 5. The minimum expected count is 8,50.

Betrokkenheid bij problematische jeugdgroep: geobserveerde versus verwachte celfrequenties

			Geslacht		Totaal
			Meisjes	Jongens	
Betrokkenheid bij problematische jeugdgroep	**Niet betrokken**	**geobserveerd**	**1185**	**1090**	**2275**
		verwacht	*1151,5*	*1123,0*	*2275,0*
		% binnen geslacht	**94,6%**	**89,2%**	**91,9%**
	Betrokken	**geobserveerd**	**68**	**132**	**200**
		verwacht	*101,5*	*99,0*	*200,0*
		% binnen geslacht	**5,4%**	**10,8%**	**8,1%**
Totaal		**geobserveerd**	**1253**	**1222**	**2475**
		verwacht	*1253,0*	*1222,0*	*2475,0*
		% binnen geslacht	100,0%	100,0%	100,0%

Een uitgewerkt rekenvoorbeeld

Indien er geen associatie zou zijn tussen het al dan niet betrokken zijn bij een problematische jeugdgroep en geslacht, zouden een even grote proportie jongens als meisjes niet betrokken zijn bij een problematische jeugdgroep en een even grote proportie jongens als meisjes wel betrokken zijn bij een problematische jeugdgroep. Hiervoor kijken we naar de laatste kolom met de totalen.

- We zien dat er 91,9% van de respondenten (ongeacht het geslacht) niet betrokken is bij een problematische jeugdgroep.
- Bij afwezigheid van associatie zouden we dus kunnen verwachten dat zowel 91,9% van de meisjes als 91,9% van de jongens niet betrokken is bij een problematische jeugdgroep.
- De verwachte waarde (bij statistische onafhankelijkheid) voor het **niet betrokken** zijn bij een problematische jeugdgroep **bij meisjes** berekenen we als volgt:

  91.9%*1253= 1151.5
- De verwachte waarde (bij statistische onafhankelijkheid) voor het **niet betrokken** zijn bij een problematische jeugdgroep **bij jongens** berekenen we als volgt:

  91,9*%*1222= 1123.0*
- De verwachte waarde (bij statistische onafhankelijkheid) voor het **betrokken** zijn bij een problematische jeugdgroep **bij meisjes** berekenen we als volgt:

  8.1%*1253= 101.5

- De verwachte waarde (bij statistische onafhankelijkheid) voor het **betrokken** zijn bij een problematische jeugdgroep **bij jongens** berekenen we als volgt:

  8.1%*1222= 99.0

Chi kwadraat wordt dan zo berekend, we nemen de formule er nog eens bij:

$$\chi^2 = \sum \frac{(geobserveerd - verwacht)^2}{verwacht}$$

Chi-kwadraat = ((1185-1151.5)² / 1151.5) + ((1090-1123.0)² / 1123.0) + ((68-101.5)² / 101.5) + ((132-99)² / 99) = 0.97 + 0.97 + 11.06 + 11= 24,00

De chi-kwadraat waarde bedraagt voor deze kruistabel 24. **Dit moet je als volgt interpreteren:** er is wel degelijk een verschil tussen de geobserveerde celfrequenties en de verwachte celfrequenties indien er geen associatie zou bestaan tussen beide kenmerken**. Je voelt intuïtief aan dat je hier niet echt wijzer van wordt. Het is immers moeilijk een verband te interpreteren zonder richtlijnen. Richtlijnen helpen omdat ze bijvoorbeeld de bovengrens en ondergrens aangeven van een verband. Chi-kwadraat heeft geen vaste bovengrens.** Indien onze steekproef dubbel zo groot zou zijn, zou de waarde van chi-kwadraat bij eenzelfde associatie verdubbelen. Hoe moeten we nu precies vaststellen hoe sterk de associatie is? Als er wel degelijk een verband bestaat, dan moet dat verband onafhankelijk van de steekproefgrootte kunnen worden bepaald. Het mag er niet toe doen of de steekproef nu uit vijfhonderd respondenten bestaat of uit duizend respondenten. Statistici kenden dit probleem en hebben daar een handige oplossing voor bedacht. Ze vonden een manier om de associatiemaat chi-kwadraat te gaan normeren. De mogelijkheden om te normeren zijn niet oneindig, maar een handige manier om daar toch in te slagen is om de chi-kwadraat waarde te gaan relateren aan de grootte van de steekproef. Op die manier is het alvast niet meer zo dat de steekproefgrootte het verband gaat beïnvloeden. Er zijn twee varianten bedacht van associatiematen die allebei op chi-kwadraat gebaseerd zijn: de ene associatiemaat is Phi en de andere associatiemaat is Cramer's V. Deze maten corrigeren voor de problemen die aanwezig zijn bij het gebruik van chi-kwadraat als associatiemaat.

5. Phi

Phi is een associatiemaat die gebaseerd is op chi-kwadraat en neemt de *waarde nul aan bij geen associatie en de waarde van één bij een perfecte statistische associatie*. Phi wordt gebruikt bij de berekening van de associatie tussen kenmerken in een 2*2 tabel.

De formule ziet er als volgt uit:

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

Toegepast op ons voorbeeld:

$$\sqrt{\frac{24}{2475}} = 0.0985$$

= 0.1

In het voorbeeld hierboven bedraagt phi 0.1. Het gaat hier dus om een zwakke associatie.

6. Cramer's V

Cramer's V is o.i. de meest aangewezen coëfficiënt van die associatiematen die steunen op chi-kwadraat. Cramer's V is belangrijk want deze associatiemaat corrigeert voor de gevoeligheid van chi-kwadraat voor de steekproefgrootte en het aantal rijen en kolommen van variabelen. Cramer's V varieert eveneens van nul tot één en wordt gebruikt bij de berekening van associaties tussen kenmerken in een r*k tabel. In het voorbeeld hierboven bedraagt Cramers's V 0.1. Het gaat hier dus om een zwakke associatie.

Volledigheidshalve geven we ook de formule weer:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

$$(k = \min(r, k))$$

k = 2 (want 2*2 tabel)

$$v = \sqrt{\frac{24}{2475(2-1)}} = 0.0985$$

V= 0.1

7. Gamma als associatiemaat op ordinaal niveau

Gamma wordt nog regelmatig gebruikt om de samenhang tussen variabelen van het ordinale niveau te berekenen. De waarde van gamma verloopt van *min één tot plus één*, waarbij min één wijst op een perfect negatief verband, plus één op een perfect positief verband en nul op de afwezigheid van een verband. Hoe verder van nul, hoe sterker de associatie!
Heeft de associatiemaat een waarde kleiner dan nul, dan betekent het dat hoge waarden op de ene variabele samengaan met lage waarden op de andere variabele. We spreken van een **negatieve relatie.** Heeft de associatiemaat een waarde groter dan nul, dan betekent het dat hoge waarden op de ene variabele samengaan met hoge waarden op de andere variabele. Er is sprake van een **positieve relatie**.

Ordinale variabelen, zoals vragen uit een vragenlijst waarbij de antwoordcategorieën variëren van "helemaal niet akkoord" tot "helemaal akkoord" of waarbij de antwoordcategorieën variëren van "nooit" over "soms", "vaak" tot "zeer vaak", zijn eigenlijk strikt genomen van het ordinale niveau: de afstand tussen helemaal niet akkoord en helemaal akkoord is niet dezelfde als de afstand tussen akkoord en helemaal akkoord.[8] Beschouw onderstaande 3*3 kruistabel. **Gamma** is gebaseerd op de verhouding tussen *consistente paren* en *inconsistente paren*. **Een consistent paar** is ieder paar van elementen waarbij de waarde van het ene element op beide variabelen hoger is dan van het andere element. **Een inconsistent** paar is ieder paar van elementen waarbij de waarde van het ene element op de ene variabele hoger is en op de andere variabele lager is dan van het andere element.

[8] Hoewel numerieke waarden gebruikt worden bij de invoer van de gegevens in de gegevensmatrix, vormen deze numerieke waarden geen representatie van een echte metrische waarde en daarom dient eigenlijk een ordinale associatiemaat berekend te worden om het verband te kennen tussen twee ordinale kenmerken.

Tabel: Consistente en inconsistente paren bij de berekening van gamma

Delinquent gedrag	**Sociaal economische status**		
	Laag	**Midden**	**Hoog**
Laag	A	B	C
Midden	D	E	F
Hoog	G	H	I

Wanneer we het aantal consistente paren en inconsistente paren in bovenstaande 3*3 tabel identificeren, bekomen we het volgende:

CP (Consistente paren) zijn : (A*E) + (A*F) + (A*H) + (A*I) + (B*F) + (B*I) + (D*H) + (D*I) + (E*I)

IP (Inconsistente paren) zijn : (C*E) + (C*D) + (C*H) + (C*G) + (B*D) + (B*G) + (F*H) + (F*G) + (E*G)

Gamma wordt als volgt berekend: **(CP-IP) / (CP + IP).** Dit getal situeert zich tussen min één en plus één. Dit is de waarde van gamma. Laten we een voorbeeld geven uit criminologisch onderzoek. De waarde van Gamma moet je niet met de rekenmachine zelf kunnen uitrekenen.

Kruistabel: regels zijn gemaakt om te breken en spijbelen van vrienden

		Regels zijn gemaakt om te breken					**Totaal**
		helemaal oneens	**beetje oneens**	**noch eens, noch oneens**	**beetje eens**	**helemaal eens**	
Spijbelen van vrienden	**Geen vriend**	811	261	265	185	94	1616
	Eén vriend	221	154	159	130	83	747
	Twee vrienden	11	10	15	16	15	67
	Drie of meer vrienden	9	5	9	4	11	38
Totaal		1052	430	448	335	203	2468

Deze kruistabel bevat twee variabelen uit een zelfrapportagestudie onder jongeren. We vroegen aan de jongeren in welke mate ze het eens waren met de uitspraak *"regels zijn gemaakt om te breken"*. Deze uitspraak is een indicator voor lage moraliteit. We vroegen ook aan diezelfde jongeren hoeveel van hun beste vrienden spijbelen. Wat verwachtten de onderzoekers nu? Wel, er wordt in de criminologische literatuur nogal eens gewezen op het

verband tussen leeftijdsgenoten en eigen waarden en normen. Je kan dat verband op twee manieren interpreteren: leeftijdsgenoten beïnvloeden de eigen waarden en normen, maar je zou ook het omgekeerde kunnen verwachten: de eigen waarden en normen spelen mee bij de keuze voor de eigen vriendenkring. In beide gevallen verwachten we echter dat de bivariate samenhang positief is: dat wil zeggen: hoge waarden op de ene variabele moeten samengaan met hoge waarden op de andere variabele. Anders gezegd: hoe positiever men staat tegenover regelovertreding, hoe meer spijbelende vrienden men heeft. Het probleem van de kip of het ei lossen we met deze statistische analyse niet op. Wel kunnen we vaststellen of de idee van samenhang klopt. Deze blijkt wel degelijk te kloppen. Gamma bedraagt hier 0.328. Er is dus een matig verband tussen de ordinale variabelen "regels zijn gemaakt om te breken" en het aantal spijbelende vrienden. Gamma kan ook berekend worden voor metrische variabelen die men ordinaal maakt door deze variabelen in categorieën te brengen. Echter, het is mogelijk dat de samenhang gebaseerd op de maat gamma ietwat verschilt van de samenhang die gebaseerd is op de metrische correlatiecoëfficiënt. Dat is mogelijk omdat we informatie verliezen: als we variabelen hercoderen en een zeer gedetailleerde metrische variabele herleiden tot een ordinale variabele met vijf categorieën, gooien we toch wat informatie weg. Als we metrische variabelen in categorieën onderverdelen, moeten we dat op een verstandige manier doen. We kunnen ons bijvoorbeeld baseren op de kwartielen om dat te doen. Als we nu een groot verschil vaststellen tussen het verband gemeten op basis van de ordinale maat gamma en de originele correlatiecoëfficiënt, dan is dat een teken dat er iets aan de hand is wat nadere inspectie vereist. Het kan betekenen dat de associatie niet rechtlijnig is. De beste manier om dat na te gaan is om een puntenwolk op te vragen.

8. De rangcorrelatiecoëfficiënt van Spearman en Kendall's Tau-*b*

Wanneer onze data gemeten zijn op ordinaal niveau en de waarnemingen van onze studie kunnen gerangschikt worden in twee onderscheiden reeksen, kunnen we gebruik maken van **rangcorrelatiecoëfficiënten**. In dit geval toetsen we de nulhypothese dat beide rangordeningen niet van elkaar verschillen. Wat wordt hiermee bedoeld? We kunnen de respondenten rangordenen van laag naar hoog op basis van de antwoorden op de variabele "regels zijn gemaakt om te breken". We kunnen hetzelfde doen voor de variabele "spijbelen van vrienden". Vervolgens kunnen we ons de vraag stellen of deze beide ordeningen samenhangen: iemand die hoog gerangschikt is op variabele X, is die ook hoger gerangschikt op variabele Y? De meest bekende rangcorrelatiecoëfficiënten zijn deze van Spearman (*Rho* genoemd) en deze van Kendall (*Tau-b*). We nemen opnieuw dezelfde associatie tussen twee

enquêtevragen over spijbelen van vrienden en de attitude tegenover regelovertreding als voorbeeld.

Kruistabel: regels zijn gemaakt om te breken en spijbelen van vrienden

		Regels zijn gemaakt om te breken					Totaal
		helemaal oneens	beetje oneens	noch eens, noch oneens	beetje eens	helemaal eens	
Spijbelen van vrienden	Geen vriend	811	261	265	185	94	1616
	Eén vriend	221	154	159	130	83	747
	Twee vrienden	11	10	15	16	15	67
	Drie of meer vrienden	9	5	9	4	11	38
Totaal		1052	430	448	335	203	2468

De rangcorrelatiecoëfficiënt van Spearman is afgeleid van de **product-moment correlatiecoëfficiënt van Pearson**. Deze coëfficiënt varieert van min één tot plus één. Nul wijst op de afwezigheid van een verband. Min één wijst op een perfect negatieve associatie en plus één wijst op een perfect positieve associatie. Om de rangcorrelatiecoëfficiënt te berekenen dienen we de respondenten die we bevraagd hebben te gaan rangordenen op deze beide kenmerken. Om ***Rho*** te berekenen bepalen we per eenheid het verschil tussen de beide rangordeningen. De achterliggende idee is dat samenhang tussen de beide reeksen perfect zal zijn indien de rangordeningen niet van elkaar verschillen, en dat wanneer dit laatste wel het geval is, de samenhang minder sterk zal zijn. ***Kendall's Tau-b*** is eveneens een symmetrische associatiemaat die varieert tussen min één en plus één. Bij perfecte statistische associatie worden de waarden plus één of min één slechts bereikt onder de conditie dat het aantal rijen even groot is als het aantal kolommen. Een analyse aan de hand van SPSS toont ons volgende resultaten.

Tabel: Rangcorrelaties tussen twee variabelen

		Regels zijn gemaakt om te breken
Kendall'sTau-b	spijbelen vrienden	0.203
Spearman's Rho	spijbelen vrienden	0.226

Er is dus een matige positieve associatie tussen beide ordinale kenmerken.

9. Leerdoelen

In dit hoofdstuk werd de opmerking gemaakt dat er een zeer belangrijk verband bestaat tussen het kiezen van een associatiemaat en het meetniveau van variabelen. Daar waar men in de univariate beschrijvende statistiek enkel en slechts enkel met het meetniveau van één kenmerk dient rekening te houden, is dit iets complexer in de bivariate statistiek. Hier dienen we rekening te houden met het meetniveau van de twee kenmerken. **Als regel geldt dat wanneer twee meetniveaus verschillen, we het laagste meetniveau kiezen.** We hebben in dit hoofdstuk een reeks van associatiematen gezien die van toepassing zijn op de studie van relaties tussen kernmerken gemeten op lagere meetniveaus. We kunnen daarbij het onderscheid tussen het ordinale en nominale meetniveau maken. We hebben gezien dat nominale associatiematen vooral gebaseerd zijn op het percentageverschil en ook op chi-kwadraat. Je kan echter twee percentageverschillen berekenen. Je moet dus goed redeneren: welke variabele beschouw je als onafhankelijke variabele en welke variabele beschouw je als afhankelijke variabele? Inhoudelijk moet je heel goed weten wat een chi-kwadraat waarde betekent. Ook de formule en de berekening van een chi-kwadraat worden verwacht gekend te zijn. We leggen ook de klemtoon op het correct interpreteren van de associatie en het trekken van een besluit, i.e. het beantwoorden van een onderzoeksvraag waar twee variabelen bij betrokken zijn. Je dient ook de mogelijkheden en beperkingen van associatiematen voor categorische variabelen te kennen. Eveneens moet je goed het onderscheid tussen zulke associatiematen begrijpen. Je dient te weten wanneer je Phi dan wel Cramer's V gebruikt, of wanneer je een contingentietabel analyseert aan de hand van een percentageverschil. Ook dien je de odds en odds-ratio te kennen. Deze dien je zelf met behulp van een rekenmachine te kunnen uitrekenen. Tot slot dien je te weten wanneer je kiest voor gamma of Kendall's tau-b. Gamma, Rho en Kendall's tau-b moeten niet zelf kunnen berekend worden. De theoretische begrippen die werden behandeld moeten echter wel gekend zijn. Tot slot: denk steeds na wat de resultaten betekenen in functie van de onderzoeksvraag die altijd aan de basis ligt van een statistische bivariate analyse.