Beschrijvende statistiek

Het berekenen en interpreteren van tabellen en statistieken

Bregje van Groningen
Connie de Boer

Vierde druk

Boom

inclusief website!

Met behulp van onderstaande unieke activeringscode kunt u toegang krijgen tot www.beschrijvendestatistiek.nl voor extra materiaal. Deze code is persoonsgebonden en gekoppeld aan de 4e druk. Na activering van de code is de website 2 jaar toegankelijk. De code kan tot zes maanden na het verschijnen van een volgende druk geactiveerd worden.

9823-XL-53-LG

Omslagontwerp: Cunera Joosten, Amsterdam
Opmaak binnenwerk: Nu-nique grafische vormgeving, Goor

ISBN 9789089539823
ISBN 9789461278432 (e-book)
NUR 916

www.boomuitgeversamsterdam.nl
www.beschrijvendestatistiek.nl

Woord vooraf

Voor veel studenten is statistiek een moeilijk vak. Daarom vinden we het belangrijk dat de uitleg zo duidelijk mogelijk is. Dat stelt hoge eisen aan de leesbaarheid. Ter wille van de leesbaarheid hebben we nauwelijks verwijzingen naar literatuur in de tekst opgenomen. In de literatuurlijst achter in het boek staan de publicaties waarop wij onze kennis hebben gebaseerd. Deze boeken gaan op sommige statistieken veel uitgebreider in. De literatuurlijst kun je daardoor ook beschouwen als aanbevolen literatuur voor een verdere verdieping van de kennis.

Wij denken dat een goed begrip van statistiek pas ontstaat als je weet hoe je statistische gegevens berekent. Daarom is de uitleg mede gebaseerd op de formules die je gebruikt voor een kengetal of statistische analyse. In de praktijk zul je statistieken zelden met de hand, op basis van formules, uitrekenen. Daarvoor bestaan heel geschikte rekenprogramma's, zoals SPSS. Voor het interpreteren van de resultaten is het echter van groot belang dat je weet wat de betekenis is van de statistische gegevens. Dan is het nuttig deze een paar keer zelf te hebben uitgerekend. Daarna kun je ook oefenen met behulp van een computerprogramma. In aparte kaders is uitgelegd hoe je de besproken analyses door SPSS kunt laten uitvoeren. Deze combinatie van uitleg van berekeningen met de hand en berekeningen met SPSS geeft je de mogelijkheid zelf sommen en oefeningen te maken waarbij je de berekeningen met de hand controleert aan de hand van de resultaten van SPSS.

Eerdere versies van dit boek zijn gebruikt in het onderwijs over onderzoeksmethoden en statistiek aan propedeusestudenten Communicatiewetenschap aan de Universiteit van Amsterdam. De reacties en het commentaar van de studenten en de docenten van deze cursussen hebben geleid tot vele verbeteringen van de tekst.

De volgende personen willen wij expliciet noemen als dank voor hun inhoudelijke en tekstuele bijdragen: Tiede Bijlsma, Sanneke Schouwstra, Rob de Lange, Wouter de Nooy, Mieke Sillekens, Reza Kartosen, Floris Müller, Johannes von Engelhardt, Rob Erven, Carel van Wijk, Nadine Bol, Jeroen Jonkman, Marianne Ouwehand, Gert van Driel en Rhianne Hoek. Verder hebben verschillende docenten en studenten van de cursus 'Methoden van Communicatie Onderzoek en Statistiek' commentaar geleverd op eerdere versies van deze tekst.

Nieuw in de vierde druk

In de vierde druk gaan we nog meer in op het belang van de operationalisatie, het kiezen van de juiste meetniveaus en de mogelijke consequenties van keuzes die gemaakt (moeten) worden in het analyseproces. Voorbeelden zijn aangepast aan (op het moment van schrijven) meer actuele zaken.

In hoofdstuk 3 is een aantal paragrafen toegevoegd waarin we onder andere ingaan op de normale verdeling als kansverdeling en *z*-scores bij het berekenen van kansen. Omdat het een boek over *beschrijvende* statistiek is, zullen we dat niet koppelen aan het toetsen van hypothesen en significantie.

Daarnaast is een nieuw hoofdstuk toegevoegd (hoofdstuk 4) waarin het ver- en bewerken van data in SPSS centraal staat, en de consequenties daarvan voor het meetniveau van de variabelen en de analyses die mogelijk zijn. In hoofdstuk 8 wordt uitgebreider dan in de derde druk ingegaan op het berekenen en interpreteren van een variantieanalyse. Hoofdstuk 9 en 10 zijn samengevoegd en uitgebreid. Het heet nu *Schaalconstructie*, en we bespreken daarin zowel de theoretische als statistische overwegingen bij het maken van een valide en betrouwbare schaal. In bijna alle hoofdstukken wordt meer stilgestaan bij de voorwaarden van bepaalde analyses en de consequenties wanneer niet aan de voorwaarden wordt voldaan.

Tot slot is de vierde druk aangepast aan de ontwikkelingen in de SPSS-software. Alle voorbeelden waarin SPSS wordt gebruikt, zijn nu uitgevoerd met SPSS 23.

Bregje van Groningen
Connie de Boer

Inhoud

Kaders

Inleiding

In dit boek staat de beschrijvende statistiek centraal. Maar wat is beschrijvende statistiek? Sociale wetenschappers hebben als doel kennis te genereren over de sociale werkelijkheid. Dat kan nieuwe kennis zijn, maar ook kennis waaruit blijkt dat wat we 'wisten' niet of niet helemaal klopt. Kennis kun je verkrijgen door het uitvoeren van een onderzoek. Wanneer je een onderzoek hebt uitgevoerd, wil je de resultaten van het onderzoek zo duidelijk mogelijk weergeven. Dat kan op een korte en overzichtelijke manier gebeuren door middel van kengetallen, tabellen of grafieken. Dit is precies waar het in de beschrijvende statistiek om gaat: het samenvattend beschrijven van de kenmerken van een groep onderzoekseenheden.
Een voorbeeld van een kengetal is het rekenkundig gemiddelde. Door de gemiddelde leeftijd van een groep uit te rekenen geef je een samenvattende beschrijving van een specifiek kenmerk van die groep, namelijk hun leeftijd. Je kunt dit kenmerk ook beschrijven door een tabel of een grafiek te gebruiken, zoals in tabel 1 is gedaan.

Tabel 1 Beschrijving van leeftijd in tabel

Leeftijd	Absolute frequentie	Percentage
18	5	25
19	6	30
20	1	5
21	2	10
22	2	10
24	4	20
Totaal	20	100

Om kengetallen, tabellen en grafieken te produceren moet een onderzoeker een aantal handelingen verrichten. Ten eerste moet hij de gegevens verzamelen en die verzamelde data onderbrengen in een datamatrix. Daarna moet hij de data verwerken (kengetallen bepalen en tabellen en grafieken maken). Deze gegevens vermeldt hij vervolgens in een onderzoeksverslag, waarbij het belangrijk is dat hij de statistieken op de juiste manier weergeeft en interpreteert. Deze fasen, de datamatrix, de data-analyse en de interpretatie van statistieken, worden in dit boek besproken.

In de komende hoofdstukken zal steeds over het algemeen eerst de theoretische achtergrond worden besproken. Het gaat daarbij om de vraag waarom je de analyses nodig hebt. Dit wordt geïllustreerd aan de hand van voorbeelden en waar mogelijk met een grafische weergave. De analyses zijn uitgevoerd met het computerprogramma SPSS, een programma dat in de sociale wetenschap veel wordt gebruikt. In deze vierde druk is gebruikgemaakt van SPSS 23 (Windows-versie) en Windows 10. Het is mogelijk dat de kaders en tabellen er wat betreft lay-out iets anders uitzien wanneer een andere versie van SPSS of Windows-versie, of een Mac is gebruikt. De inhoud is uiteraard steeds hetzelfde. We zullen in aparte kaders uitleggen hoe je met behulp van SPSS de besproken analyses zelf kunt uitvoeren.
In een wetenschappelijke tekst mag je nooit SPSS-tabellen opnemen, ook niet in de bijlagen. Je maakt als onderzoeker zelf tabellen waarin alleen de voor het onderzoek relevante gegevens staan. In dit boek wordt van deze regel afgeweken, omdat we willen laten zien hoe je een SPSS-output kunt lezen en interpreteren. In dit boek zijn de data, wanneer niet zelf verzonnen, afkomstig van databestanden die worden gebruikt door collega's van de afdeling Communicatiewetenschap, scripties van Masterstudenten, en enquêtes die zijn afgenomen door studenten van de module 'Methoden en Technieken voor Communicatieonderzoek'. Veel van deze data zijn bewerkt door de auteurs om de voorbeelden in het boek zo duidelijk mogelijk te maken.
Naast de interpretatie vanuit SPSS zal de handmatige berekening van verschillende analyses worden uitgelegd. Het zelf uitrekenen van statistieken op basis van de formules maakt duidelijk welke betekenis je eraan kunt toekennen. Dit zal het interpreteren van de gegenereerde kengetallen, tabellen en grafieken vergemakkelijken. Deze aspecten – nut en toepassing, voorbeelden, interpretatie en handmatige berekening – zullen in de komende hoofdstukken aan de orde komen bij de behandeling van de verschillende statistieken. Om zo dicht mogelijk in de buurt te komen van de waarden zoals ze in SPSS gepresenteerd zijn, rekenen wij altijd met drie decimalen achter de komma. In teksten over het interpreteren van de maten zullen we, zoals dat ook in de meeste wetenschappelijke artikelen wordt gedaan, ons beperken tot twee decimalen achter de komma.

In de beschrijvende statistiek gebruik je de gegevens van een groep onderzoekseenheden. Die groep kan een *steekproef* of een *populatie* zijn. De gebruikte statistieken zeggen dan iets over de samenstelling van die steekproef of over die populatie, afhankelijk van welke gegevens je gebruikt. Als het om een steekproef gaat, zegt de beschrijvende statistiek dus iets over de steekproef en niet over de populatie waaruit die steekproef afkomstig is. *Beschrijvende* statistiek onderscheidt zich van de *inferentiële* statistiek doordat je je bij de beschrijvende statistiek beperkt tot uitspraken over een groep onderzoekseenheden waarvan je de gegevens hebt. Wanneer je niet de hele populatie hebt onderzocht, kun je over de populatie niets met zekerheid zeggen. Op basis van de steekproefgegevens kun je echter wel met een bepaalde mate van waarschijnlijkheid uitspraken

doen over de populatie. Dit gebeurt in de *inferentiële statistiek*. Op basis van steekproefgegevens maak je dan een schatting van een populatiewaarde.

De statistieken of kengetallen uit de beschrijvende en de inferentiële statistiek geef je aan met symbolen, met letters. Zo duid je het gemiddelde in de populatie aan met een μ (mu), en een gemiddelde in een steekproef met een $\bar{x}$ (*x* streep) of in een wetenschappelijk artikel met *M* (mean). De standaarddeviatie in een populatie geef je aan met een σ (sigma). Als je de standaarddeviatie van de steekproef (*s*) gebruikt als schatter van de standaarddeviatie in de populatie, ben je met inferentiële statistiek bezig. De formules van σ (standaarddeviatie van de populatie) en *s* (standaarddeviatie van de steekproef) vertonen een klein verschil. Bij σ wordt door *n* (aantal onderzoekseenheden) gedeeld, en bij *s* door $n - 1$ (aantal vrijheidsgraden[1]). Als je door *n* zou delen, dan schat je de standaarddeviatie in de populatie iets te laag. Ditzelfde geldt niet alleen voor de standaarddeviatie, maar ook voor andere kengetallen die in dit boek worden behandeld.
Bij de behandeling van de beschrijvende statistiek zou het logisch zijn om te kiezen voor de kengetallen waarbij door *n* wordt gedeeld. De statistieken worden immers niet gebruikt als schatter voor de populatieparameters. Hier hebben we niet voor gekozen, omdat we naast het met de hand uitrekenen van statistische gegevens aandacht besteden aan SPSS. Je kunt de met de hand uitgerekende statistieken controleren door de gegevens in een SPSS-bestand in te voeren. Het is ook mogelijk om op basis van een SPSS-bestand zelf oefeningen te maken voor het met de hand uitrekenen van statistieken. Aangezien je SPSS ook gebruikt voor de inferentiële statistiek, zou er – zeker als het gaat om een klein aantal onderzoekseenheden – een verschil kunnen bestaan tussen de output van SPSS en de zelf uitgerekende gegevens als je de formules uit de beschrijvende statistiek gebruikt. Daarom hebben we ervoor gekozen de formules te gebruiken die horen bij de inferentiële statistiek.

In hoofdstuk 1 staan de basiselementen van de beschrijvende statistiek centraal. De centrummaten (modus, mediaan en gemiddelde) en spreidingsmaten worden respectievelijk in hoofdstuk 2 en 3 beschreven. In hoofdstuk 3 wordt ook de normale verdeling als kansverdeling besproken. Hoofdstuk 4 geeft informatie over hoe je in SPSS je data kunt bewerken. De hoofdstukken 5, 6, 8 en 9 gaan over de associatiematen op nominaal, ordinaal en interval- en rationiveau. In hoofdstuk 7 wordt aandacht besteed aan tabelsplitsing. Hoofdstuk 10 gaat over schaalconstructie en behandelt factoranalyse en betrouwbaarheidsanalyses.

Website

Bij dit boek hoort een website met aanvullend materiaal: www.beschrijvendestatistiek.nl. Op deze website staan per hoofdstuk opdrachten. Studenten kunnen zo oefenen met de lesstof van het hoofdstuk. Op pagina 4 van dit boek staat een persoonlijke inlogcode voor toegang tot de website.

Noot

1 Dit is het aantal keren dat een waarde 'vrijelijk kan variëren'. Als je van een groep mensen de gemiddelde leeftijd weet, en je de leeftijden weet van alle individuen uit die groep op één na ($n-1$), staat de leeftijd van die laatste persoon vast. Die kan niet meer variëren. De leeftijd van deze laatste persoon kun je precies uitrekenen op basis van de leeftijden van de anderen uit de groep en de gemiddelde leeftijd.

Basiselementen 1

Je doet onderzoek om iets over de werkelijkheid te weten te komen. Op basis van dat onderzoek kun je dan uitspraken doen over de werkelijkheid. Daarbij moet duidelijk worden over wie of wat je op basis van het onderzoek een uitspraak doet. Dat zijn de objecten of *onderzoekseenheden*, de personen of zaken over wie je iets zegt.
Als je op basis van een onderzoek bijvoorbeeld de conclusie trekt dat de gemiddelde leeftijd van de eerstejaarsstudenten van een universiteit 19,7 jaar is, zeg je op basis van dit onderzoek iets over eerstejaarsstudenten. Dat zijn dan de onderzoekseenheden. Van deze studenten beschrijf je een kenmerk, namelijk de leeftijd. In het onderzoek kun je meer kenmerken van de studenten hebben verzameld, zoals geslacht, vooropleiding en studiekeuze. Deze kenmerken (leeftijd, geslacht, vooropleiding en studiekeuze) zijn de *variabelen* in het genoemde onderzoek.
Onderzoekseenheden hoeven niet altijd personen te zijn. Je kunt op basis van een onderzoek ook uitspraken doen over de lengte van voorpagina-artikelen in dagbladen. In dat geval zijn de voorpagina-artikelen de onderzoekseenheden, de objecten waarover je een uitspraak doet. De lengte is hier een kenmerk van de onderzochte artikelen en is daarom een variabele in het onderzoek. Met behulp van statistiek kun je precieze uitspraken doen over de kenmerken van onderzoekseenheden, zoals 'de gemiddelde leeftijd van eerstejaarsstudenten is 19,7 jaar' en 'de meeste voorpagina-artikelen zijn korter dan twee kolommen'.
Een ander voorbeeld: 'televisiekijkers met een hoge opleiding kijken vaker naar het nieuws dan televisiekijkers met een lage opleiding'. In deze uitspraak wordt iets gezegd over televisiekijkers. In het onderzoek zijn televisiekijkers de onderzoekseenheden. In het voorbeeld zijn de kenmerken van die televisiekijkers: opleiding en de frequentie waarmee naar het nieuws wordt gekeken. 'Opleiding' en 'frequentie nieuws kijken' zijn de variabelen in het onderzoek. Meer voorbeelden van onderzoekseenheden en variabelen staan in paragraaf 1.4.

1.1 Datamatrix

Variabelen, de kenmerken van onderzoekseenheden, kunnen verschillende waarden hebben. Bij sommige kenmerken zijn de waarden al een getal, bij andere kenmerken zou je voor de voorkomende categorieën een getal kunnen verzinnen. De waarden van bijvoorbeeld het kenmerk leeftijd zijn getallen die direct gerelateerd zijn aan de werkelijkheid. Als een persoon 21 jaar oud is, is

het logisch dat deze persoon de waarde 21 krijgt voor de variabele 'leeftijd in jaren'. Maar bijvoorbeeld de variabele geslacht heeft geen vaststaande numerieke waarde. Om in de statistiek toch op een geordende wijze iets te kunnen zeggen over de onderzoekseenheden, krijgen de categorieën 'man' en 'vrouw' waarin de variabele 'geslacht' kan worden onderverdeeld, wel een numerieke waarde om de dataverwerking te vergemakkelijken. Je zou kunnen besluiten vrouwen de waarde 1 te geven en mannen de waarde 2. Op die manier kun je alle onderzoekseenheden voorzien van een numerieke waarde voor het kenmerk 'geslacht'. Al deze kenmerken van onderzoekseenheden kun je onderbrengen in een *datamatrix*. Een datamatrix is een spreadsheet waarin per onderzoekseenheid alle kenmerken als afzonderlijke variabelen worden beschreven. De onderzoekseenheden staan in de rijen van de datamatrix en de variabelen in de kolommen (tabel 1.1).

Stel dat je lekker op een terrasje zit met je vrienden. Je vraagt je af of je vrienden naar dezelfde televisiezenders kijken als jij. Om dat uit te vinden, maak je een lijstje met een aantal televisiezenders waarnaar je zelf regelmatig kijkt en gaat iedereen vragen hoeveel uur zij ongeveer per week naar die zender kijken. Daarbij schrijf je ook op welke leeftijd die persoon heeft en of het een man of een vrouw is. Omdat je alles in getallen wilt uitdrukken, stel je dat als iemand vrouw is zij de waarde 1 krijgt, en als iemand man is de waarde 2 (we hadden ook vrouw de waarde 2 kunnen geven en man de waarde 1, of een symbool kunnen toevoegen, waarbij vrouw = ♀ en man = ♂). De datamatrix ziet er dan als volgt uit.

Tabel 1.1 Voorbeeld van een datamatrix

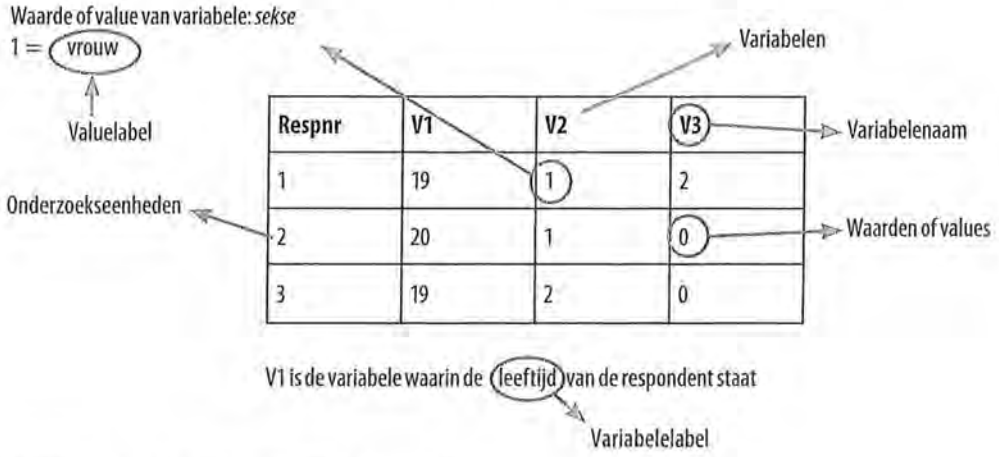
Persoon	Leeftijd	Sekse	NPO1	NPO3	RTL4	RTL5	Net5	BBC
1	19	1	2	0	2	2	4	0
2	20	1	0	0	1	2	2	1
3	19	2	0	1	1	1	0	3
4	22	1	3	2	2	3	3	2
5	24	2	1	0	3	1	1	1
6	21	2	0	3	2	0	1	1
7	20	1	2	2	1	2	3	2

In de eerste rij staan per kolom de namen van de variabelen die je in je onderzoek hebt gemeten. In de cellen daaronder staan de waarden die de respectievelijke onderzoekseenheden hebben op die variabelen. Persoon 1 is dus een 19-jarige vrouw, die twee uur naar NPO1, RTL4 en RTL5 kijkt, vier uur naar Net5, en niet naar NPO3 en de BBC.

Voor elke afzonderlijke variabele kun je een *frequentieverdeling* maken om uitspraken te doen over de percentuele verdeling van de onderzoekseenheden over de waarden van die variabele. In dit voorbeeld, waarbij we ons hebben beperkt tot zeven onderzoekseenheden, is 42,9% man (drie van de zeven personen) en

57,1% vrouw (vier van de zeven personen). Voor elke variabele (elke kolom in de datamatrix) kun je een dergelijke frequentieverdeling maken.

Om met de terminologie wat vertrouwd te raken laten we nogmaals een klein gedeelte van de datamatrix zien, en zullen we stilstaan bij de verschillende begrippen die hierbij horen. We hadden al gezien dat in de rijen de onderzoekseenheden staan, en dat deze variëren op een aantal kenmerken (vandaar ook de naam: variabelen). In de datamatrix behorende bij tabel 1.1 is duidelijk te zien wat er met de variabelen bedoeld wordt. Met 'leeftijd' wordt de leeftijd van de respondent bedoeld, en met 'sekse' het geslacht van de respondent. Met 'NPO1' kunnen echter verschillende dingen worden bedoeld. Is de onderzoekseenheden gevraagd hoeveel uur ze per week kijken? Of per dag? Of misschien is hun wel een schaal voorgelegd waarop ze konden antwoorden van 0 = nooit tot 5 = heel vaak.
De namen die boven de kolommen van een datamatrix staan, zijn de *variabelennamen*. We zouden ook voor heel andere variabelennamen kunnen kiezen, zoals voor V1, V2 en V3, omdat deze informatie bijvoorbeeld correspondeert met respectievelijk vraag 1, vraag 2 en vraag 3 in een vragenlijst (zie figuur 1.1).



Figuur 1.1 Terminologie bij datamatrix

V3 is hier dus een variabelennaam. Wat je dan feitelijk bedoelt met die variabelennaam, maak je kenbaar in het *variabelenlabel*. Het label van V1 is in voorstaand voorbeeld dus leeftijd in jaren. Het label van V2 is sekse, en van V3 is het label aantal uur dat naar NPO1 wordt gekeken. De getallen die in de matrix staan, noemen we *waarden* of, in het Engels, *values*. De waarden die bij de variabele V1 horen (leeftijd), zijn gemakkelijk te interpreteren: 19 betekent dat deze respondent 19 jaar oud is. De waarde 1 van de variabele V2 is al moeilijker te interpreteren. Daarom worden ook de waarden voorzien van een label: het *valuelabel*. Daarin wordt aangegeven wat bedoeld wordt met de waarden. In ons voorbeeld is het valuelabel van de waarde 1 van de variabele V2: vrouw. Het valuelabel van de waarde 2 van de variabele V2 is hier man. Om te begrijpen wat er met de verschillende variabelennamen, variabelenlabels en valuelabels

wordt bedoeld, is het nodig om deze informatie ergens te vermelden. Dit kan bijvoorbeeld in een codeboek. In SPSS kun je deze informatie zelf toevoegen in het tabblad 'Variable View' (zie kader 1.1).

1.2 Frequentietabellen

Een variabele (kenmerk) met de daarbij behorende waarden kun je op een overzichtelijke manier presenteren in een *frequentietabel.* Stel dat je in de zomer weer met wat vrienden op een terrasje zit en jij bent aangewezen om de drankjes te gaan halen. Je kunt proberen alles te onthouden, maar op een bierviltje de drankjes turven is gemakkelijker. Door te turven maak je een overzicht van het aantal keer dat een waarde voorkomt. Het tellen van de streepjes brengt je op de *absolute frequentie.*

Tabel 1.2 Turven en tellen van drankjes

Drankje	Aantal (geturfd)	Absolute frequentie	Percentage
Bier	𝍸 III	8	47,1
Rosé	𝍸	5	29,4
Cola Light	I	1	5,9
Cappuccino	III	3	17,6
Totaal		17	100

Dit is de basis voor het opstellen van een frequentietabel. Uit tabel 1.2 blijkt dat van de zeventien mensen acht een biertje willen, vijf een rosé enzovoort.

Een absolute frequentie kan moeilijk te interpreteren zijn, zeker wanneer je meerdere frequentieverdelingen met elkaar wilt vergelijken. Daarom is het handig om naast de absolute frequenties percentages te geven. De percentages bereken je door de absolute frequentie waarmee een specifieke waarde voorkomt te delen door het totaal aantal eenheden. In vorenstaand voorbeeld heeft $\frac{8}{17} = 0{,}471 = 47{,}1\%$ van je vrienden een biertje besteld.

De week daarop zit je weer op een terras, maar nu met 22 vrienden. Het aantal bier, rosé en cola light is hetzelfde, maar in plaats van drie, worden nu acht cappuccino's besteld.

Absoluut gezien worden dezelfde hoeveelheden bier, rosé en cola light besteld, maar relatief gezien (kijkend naar de percentages, rekening houdend met het totale aantal drankjes) wordt er minder bier, rosé en cola light besteld.

Uit tabel 1.3 blijkt dat nu 36,4% een biertje bestelt: $\frac{8}{22} = 0{,}364$.

Tabel 1.3 Aantal drankjes (absoluut en in percentages)

Drankje	Aantal (geturfd)	Absolute frequentie	Percentage
Bier	𝍸 III	8	36,4
Rosé	𝍸	5	22,7
Cola Light	I	1	4,5
Cappuccino	𝍸 III	8	36,4
Totaal		22	100

1.2.1 Hoe ziet dit eruit in SPSS?

Stel, je wilt van de zeventien vrienden tijdens het eerste terrasbezoek weten hoe oud ze zijn. Je pakt weer je bierviltje, vraagt ieders leeftijd en gaat turven. Later die dag voer je je gegevens in SPSS in, en je laat SPSS een frequentietabel maken (zie kader 1.1). Zoals is te zien in tabel 1.4, geeft SPSS naast de absolute frequentie (*Frequency*) en het percentage (*Percent*) ook het geldige percentage (*Valid Percent*) en het cumulatieve percentage (*Cumulative Percent*).

Tabel 1.4 Frequentieverdeling van de variabele leeftijd (SPSS-output)

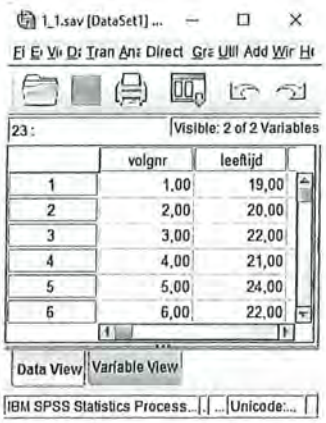
Leeftijd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	19	4	23,5	23,5	23,5
	20	5	29,4	29,4	52,9
	21	3	17,6	17,6	70,6
	22	4	23,5	23,5	94,1
	24	1	5,9	5,9	100,0
	Total	17	100,0	100,0	

In de eerste kolom (*Valid*) zie je de waarden die de variabele leeftijd in ons voorbeeld heeft. Je vrienden zijn 19, 20, 21, 22 of 24 jaar oud. In de tweede kolom *(Frequency)* staan de absolute frequenties, het aantal keer dat een bepaalde waarde voorkomt. In dit geval zijn de percentages in de vierde kolom *(Valid Percent)* identiek aan de percentages in de derde kolom *(Percent).* In paragraaf 1.2.2 zullen we bespreken in welke situaties dit niet het geval is. In deze kolommen kunnen we aflezen dat 23,5% van de onderzoekseenheden 22 jaar is. In de kolom *Cumulative Percent* worden de percentages van elke volgende waarde bij de voorgaande opgeteld (23,5 + 29,4 = 52,9 enzovoort). Je zou aan de hand van deze kolom kunnen concluderen dat 52,9% van de onderzoekseenheden (in dit geval je vrienden op het terrasje) 20 jaar of jonger is.

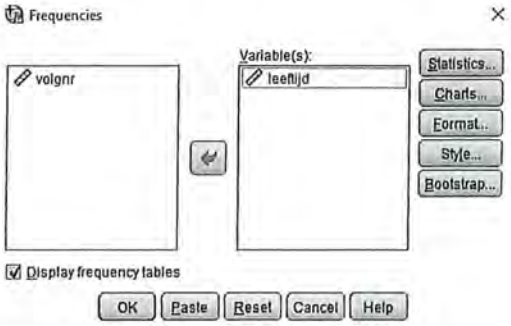
SPSS Invoeren van gegevens en het maken van een frequentietabel

Het invoeren van gegevens gebeurt in de Data View van SPSS. Per onderzoekseenheid kun je bijvoorbeeld na een uniek volgnummer voor elke onderzoekseenheid de waarde voor de variabele leeftijd intikken (zie figuur A). De namen van de variabelen (hier: 'volgnr' en 'leeftijd') kun je invoeren op de pagina die achter deze datamatrix ligt (klik linksonder op Variable View).



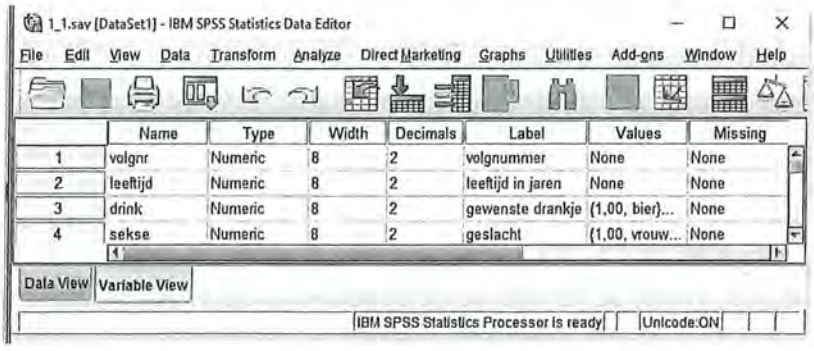
Figuur A Datamatrix

Als de data in SPSS zijn ingevoerd, kun je om een frequentietabel vragen. Ga via Analyze → Descriptive Statistics → Frequencies om vervolgens in het Frequencies-venster (zie figuur B) de variabelen te selecteren waar je een frequentietabel van wilt. SPSS maakt zelf het outputbestand waarin je deze tabel kunt vinden.



Figuur B Frequencies-venster

Door in SPSS links onderin op 'Variable View' te klikken, krijg je een overzicht van jouw variabelen te zien.



Figuur C Variable View

1.2.2 Missing values

Je vrienden willen je best vertellen hoe oud ze zijn. Wanneer je echter willekeurig mensen op een terras gaat vragen hoe oud ze zijn, kan het gebeuren dat ze je dat niet willen vertellen. Dan heb je voor die personen (onderzoekseenheden) dus geen informatie over het kenmerk leeftijd. Deze ontbrekende waarden noem je *missing values*. Je neemt deze mensen wel mee in je onderzoek naar de kenmerken van de personen op een terras. Maar als je wilt weten met welk percentage elke leeftijd vertegenwoordigd is, wil je soms niet dat deze onderzoekseenheden met onbekende leeftijden meetellen bij de berekening van de percentages. In tabel 1.5 is te zien hoe dit er in SPSS uitziet.

Tabel 1.5 Frequentieverdeling naar leeftijd met missing values (SPSS-output)

Leeftijd

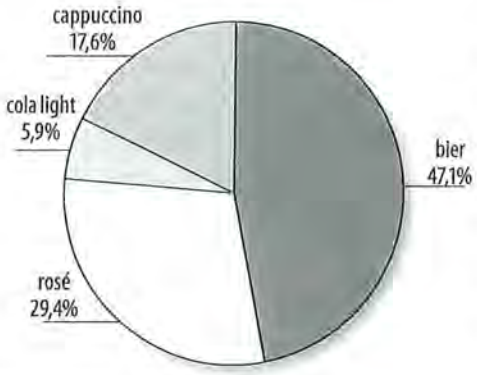
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	19	4	20,0	23,5	23,5
	20	5	25,0	29,4	52,9
	21	3	15,0	17,6	70,6
	22	4	20,0	23,5	94,1
	24	1	5,0	5,9	100,0
	Total	17	85,0	100,0	
Missing	geen antwoord	3	15,0		
Total		20	100,0		

In totaal zitten twintig personen op het terras. Daarvan hebben drie mensen geen antwoord willen geven op de vraag naar hun leeftijd. Er is nu een verschil tussen Percent en Valid Percent. Bij Percent worden deze mensen namelijk wel meegerekend (15% van de ondervraagden heeft geen antwoord gegeven). Het percentage 19-jarigen van alle mensen op het terras is 20%. Of dat een zinnig percentage is, is nog maar de vraag, want de drie die geen antwoord hebben gegeven zouden ook 19 jaar kunnen zijn, maar informatie daarover ontbreekt. Daarom kun je in dit geval beter kijken in de kolom met Valid Percent, het geldige percentage. Hierin worden de mensen die geen antwoord hebben gegeven niet meegerekend. Dan kun je constateren dat 23,5% van de mensen op het terras die deze vraag hebben beantwoord 19 jaar is. In paragraaf 4.2 wordt verder ingegaan op hoe je in SPSS waarden missing kunt maken en wat daar de consequenties van kunnen zijn in je onderzoek.

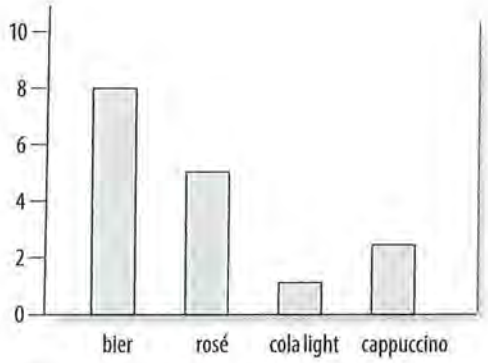
1.2.3 Grafieken

Een frequentietabel kun je ook grafisch weergeven. Een grafiek geeft geen extra informatie, maar kan visueel snel duidelijk maken wat de verdeling is van de waarden van een variabele. Bij een frequentieverdeling is het mogelijk om een

taart-, cirkel- of staafdiagram te gebruiken. In figuren 1.2 en 1.3 wordt visueel duidelijk gemaakt welke dranken door veel en welke dranken door weinig van je vrienden zijn besteld. De diagrammen geven de verdeling van de percentages of de frequenties weer.



Figuur 1.2 Taartdiagram drankjes



Figuur 1.3 Staafdiagram drankjes

In het taartdiagram kun je zien dat de meeste mensen een biertje hebben besteld (47,1%) en dat cola light door het kleinste aantal mensen is besteld (5,9%). Ditzelfde is af te lezen in het staafdiagram: acht mensen bestelden een biertje, en één persoon een cola. Een taart- of cirkeldiagram wordt meestal gebruikt om aan te geven wat relatief vaak voorkomt (percentages), terwijl staafdiagrammen meestal worden gebruikt om absolute aantallen weer te geven. Beide worden alleen gebruikt wanneer een variabele niet zo heel veel waarden heeft. Wanneer je in een enquête aan 200 onderzoekseenheden naar hun leeftijd vraagt en daar vijftig verschillende leeftijden uitkomen, is het niet erg overzichtelijk om daar een cirkeldiagram van te maken.

SPSS Het maken van grafieken

SPSS kan op verschillende manieren een grafiek voor je maken. De eenvoudigste manier is om dit te doen bij het maken van een frequentietabel. Via *Analyze → Descriptive Statistics → Frequencies* geef je eerst aan dat je een frequentieverdeling wilt maken. In het betreffende scherm (zie kader 1.1 figuur B) kun je nu Charts aanklikken. Vervolgens kun je in het Charts-venster (zie figuur A) aangeven of je een staafdiagram (Bar charts), een taartdiagram (Pie charts) of een histogram (Histograms) wilt. Histogrammen zullen besproken worden in hoofdstuk 3. Onder Chart Values geef je aan of dit in absolute frequenties of in percentages moet worden weergegeven.
Door in de output van SPSS dubbel te klikken op het figuur is het nog mogelijk om de tekst te bewerken of de arceringen te veranderen.

Figuur A Charts-venster

Een andere manier om grafieken te maken is via Graphs. Deze optie geeft veel verschillende soorten grafieken, die hier niet nader worden besproken.

Kader 1.2

1.3 Kruistabellen

Tot nu toe hebben we bij het maken van frequentieverdelingen steeds naar één enkele variabele afzonderlijk gekeken. Je kunt ook twee variabelen tegelijk gebruiken in je analyses. Wanneer je dat doet, maak je bijvoorbeeld een *kruistabel*. In feite ga je weer turven, maar nu gebruik je de waarden van twee variabelen tegelijkertijd. Hoeveel vrouwen zijn er 19 jaar? En hoeveel mannen? In tabel 1.6 is aan de hand van de datamatrix (tabel 1.5) een kruistabel gemaakt van de variabelen leeftijd en sekse. Van de twee 19-jarigen is er één man en één vrouw. De twee 20-jarigen zijn vrouwen. Zo ga je het hele rijtje af.

Tabel 1.6 Verdeling van leeftijd naar sekse (absolute frequenties)

Sekse / Leeftijd	Vrouw	Man	Totaal
19	1	1	2
20	2	0	2
21	0	1	1
22	1	0	1
24	0	1	1
Totaal	4	3	7

Ook bij kruistabellen is het overzichtelijk om de percentages erbij te geven. Dat kan in een kruistabel op drie manieren.

Totaalpercentages
Je stelt het totaal aantal onderzoekseenheden op 100% en berekent dan de celpercentages.

Tabel 1.7 Verdeling van leeftijd en sekse, percentages (gepercenteerd op totaal)

Sekse / Leeftijd	Vrouw	Man	Totaal
19	14,3	14,3	28,6
20	28,6	0	28,6
21	0	14,3	14,3
22	14,3	0	14,3
24	0	14,3	14,3
Totaal	57,1	42,9	100

Je kunt uit deze tabel aflezen dat 14,3% van alle onderzochte personen 19 jaar en vrouw is en 14,3% 19 jaar en man is. Op deze manier interpreteer je alle percentages in deze tabel.

Kolompercentages
De tweede manier om percentages in een kruistabel te berekenen is door de onderzoekseenheden in de kolommen op 100% te stellen en dan de percentages te berekenen.

Tabel 1.8 Verdeling van leeftijd naar sekse, percentages (gepercenteerd op sekse)

Sekse / Leeftijd	Vrouw	Man	Totaal
19	25,0	33,3	28,6
20	50,0	0	28,6
21	0	33,3	14,3
22	25,0	0	14,3
24	0	33,3	14,3
Totaal	100	100	100

In tabel 1.8 is sekse, de kolomvariabele, op 100% gesteld. De percentages in de kolommen tellen op tot 100%. Je ziet dat deze kolompercentages in de cellen anders zijn dan de percentages in tabel 1.7, de interpretatie is ook anders. Je redeneert vanuit de onderzoekseenheden die op 100% zijn gesteld: 25% van alle vrouwen is 19 jaar en 33,3% van alle mannen is 19 jaar. In dit boek gebruiken we in de regel kolompercentages.

Rijpercentages
De laatste manier om percentages in een kruistabel te berekenen is door de onderzoekseenheden in de rijen op 100% te stellen, en dan de percentages te berekenen. Wanneer je leeftijd, de rijvariabele, tot 100% laat optellen, krijg je weer andere percentages in de kruistabel, met weer een andere interpretatie.

Tabel 1.9 Verdeling van sekse naar leeftijd, percentages (gepercenteerd op leeftijd)

Sekse / Leeftijd	Vrouw	Man	Totaal
19	50	50	100
20	100	0	100
21	0	100	100
22	100	0	100
24	0	100	100
Totaal	57,1	42,9	100

Uit tabel 1.9 blijkt dat 50% van alle 19-jarigen vrouw is. Van alle 21-jarigen is 100% man.

Welke manier van percenteren je kiest, is afhankelijk van welke uitspraken je wilt doen op basis van je onderzoek. Als je iets wilt zeggen over de man-vrouw-verdeling naar leeftijd, dan moet je percenteren over de variabele leeftijd. Als je iets wilt zeggen over de leeftijden van mannen in vergelijking met vrouwen, percenteer je over de variabele sekse. Als er sprake is van een afhankelijke en een onafhankelijke variabele, percenteer je over de onafhankelijke variabele[1] (zie paragraaf 1.4).

Hoe ziet dit eruit in SPSS?
Je kunt SPSS kruistabellen laten maken met absolute frequenties en alle drie de genoemde percentuele berekeningen. Dan krijg je in elke cel erg veel getallen. Daarom is het beter om een keuze te maken en aan te geven op welke variabele je wilt percenteren (welke variabele je op 100% zet). De output in tabel 1.10 komt overeen met tabel 1.8 (gepercenteerd op sekse, kolompercentages).

In een SPSS-kruistabel worden achter *Count* de absolute frequenties gegeven. Er is één vrouw en één man van 19 jaar. In totaal zijn er twee 19-jarigen. In de onderste rij kun je lezen dat er in totaal vier vrouwen en drie mannen zijn. Onder *Count* staat '% within sekse'. Dit houdt in dat je gepercenteerd hebt naar sekse (ook te zien aan de 100% in de onderste rij van de kolommen). Van de vrouwen is 25% 19 jaar en van de mannen is 33,3% 19 jaar.

Als je in SPSS een kruistabel maakt, zie je de basiselementen terug zoals die besproken zijn in figuur 1.1. Je ziet bijvoorbeeld zowel de variabelenaam als het variabelelabel boven de kruistabel staan, en ook in de kolommen en de rijen komt deze informatie terug. In tabel 1.11 zie je bijvoorbeeld dat gekeken is of mannen en vrouwen van elkaar verschillen in de mate van interesse in kunst.

Tabel 1.10 Verdeling van leeftijd naar sekse (SPSS-output)

leeftijd * sekse Crosstabulation

			sekse		Total
			vrouw	man	
leeftijd	19	Count	1	1	2
		% within sekse	25,0%	33,3%	28,6%
	20	Count	2	0	2
		% within sekse	50,0%	,0%	28,6%
	21	Count	0	1	1
		% within sekse	,0%	33,3%	14,3%
	22	Count	1	0	1
		% within sekse	25,0%	,0%	14,3%
	24	Count	0	1	1
		% within sekse	,0%	33,3%	14,3%
Total		Count	4	3	7
		% within sekse	100,0%	100,0%	100,0%

Boven de kruistabel staat 'V57 interesse in kunst * V49 sekse Crosstabulation'. V57 is de variabelenaam van de variabele die als label 'interesse in kunst' heeft, geslacht heeft als variabelenaam 'V49' en als variabelelabel 'sekse'. De variabele die interesse in kunst meet heeft drie waarden, die waarden hebben als labels 'sterk in geïnteresseerd', 'tamelijk sterk in geïnteresseerd' en 'niet zo in geïnteresseerd'. De variabele die de sekse van de respondent meet, heeft twee waarden met als labels 'man' en 'vrouw'.

Tabel 1.11 Kruistabel van interesse in kunst naar sekse (SPSS-output)

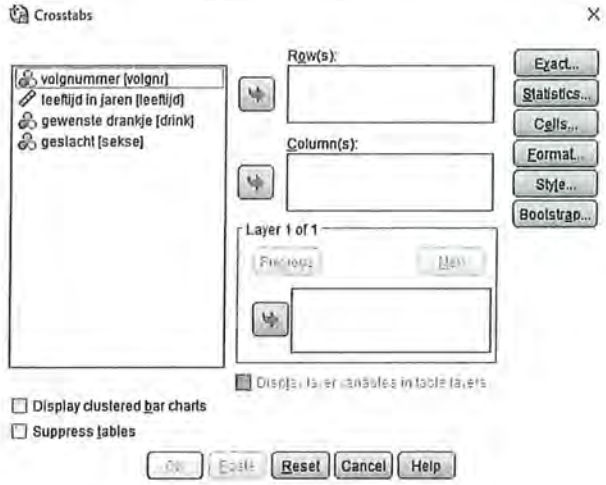
V57 Interesse in kunst *V49 sekse Crosstabulation

			V49 sekse		Total
			1 man	2 vrouw	
V57 Interesse in kunst	1 sterk in geïnteresseerd	Count	62	97	159
		& within V49 sekse	6,9%	10,3%	8,6%
	2 tamelijk sterk in geïnteresseerd	Count	190	267	457
		& within V49 sekse	21,2%	28,3%	24,9%
	3 niet zo in geïnteresseerd	Count	644	579	1223
		& within V49 sekse	71,9%	61,4%	66,5%
Total		Count	896	943	1839
		& within V49 sekse	100,0%	100,0%	100,0%

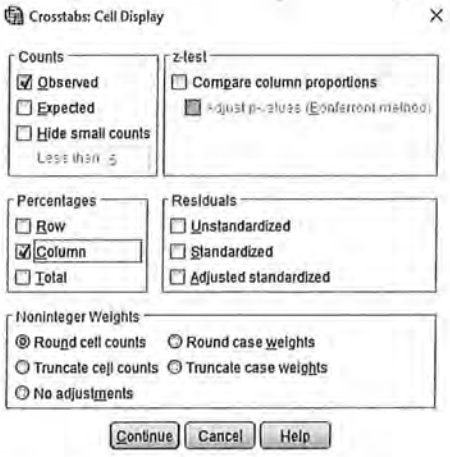
Bovenstaande kruistabel, die gepercenteerd is op de kolommen, laat bijvoorbeeld zien dat van de 896 mannen die aan deze enquête hebben deelgenomen 6,9% sterk in kunst is geïnteresseerd, tegenover 10,3% van de 943 vrouwen.

SPSS Kruistabellen maken

Om een kruistabel te maken in SPSS ga je via Analyze → Descriptive Statistics naar Crosstabs. In het Crosstabs-venster (zie figuur A) kun je aangeven welke variabele in de rijen (Row(s)) en welke variabele in de kolommen (Column(s)) moet komen. Doorgaans kiezen wij ervoor om de onafhankelijke variabele in de kolommen te zetten en de afhankelijke variabele in de rijen.
Om aan te geven of er op de kolommen of de rijen gepercenteerd moet worden, klik je op Cells, om vervolgens in de Cell Display (zie figuur B) aan te geven of je op de rijen of de kolommen wilt percenteren. Via Format kun je eventueel aangeven dat je de waarden op- of aflopend wilt presenteren.



Figuur A Crosstabs-venster



Figuur B Cell Display-venster

1.4 Variabelen

We hebben gezien dat variabelen de kenmerken van onderzoekseenheden zijn. Laten we om dit nog verder te verduidelijken eens kijken naar het artikel in kader 1.4.

Voorbeeld Gedrag van mensen is bepalend voor koffiekeuze

Cappuccinodrinkers vertonen eerder obsessief gedrag dan mensen die café latte drinken. Lattedrinkers zijn eerder geneigd anderen tevreden te stellen, terwijl mensen die hun koffie zwart drinken sneller humeurig, direct en in zichzelf gekeerd zijn.
Dit blijkt uit onderzoek waarbij duizend koffiedrinkers werden geobserveerd. Tijdens de observaties en de enquête werd gekeken naar zowel persoonlijke als psychologische eigenschappen waaronder introversie/extraversie, geduld, perfectionisme en gevoeligheid.
Tijdens de enquête moesten de ondervraagden zich inleven in tal van verschillende situaties. Achteraf werd gevraagd wat voor soort koffie zij het liefst dronken. Op basis hiervan kon Dokter Ramani Durvasula een link leggen tussen het gedrag van mensen en de koffie die zij drinken.
Bron: nu.nl, 20 september 2013, lifestylepagina

Kader 1.4

Zoals gesteld hebben onderzoekseenheden verschillende kenmerken die we variabelen noemen. In het nieuwsartikel 'gedrag van mensen is bepalend voor koffiekeuze' zijn de onderzoekseenheden koffiedrinkers. Mensen die geen koffie drinken worden in dit onderzoek namelijk niet meegenomen in de resultaten. Er zijn duizend mensen (we zeggen ook wel: n = 1000) geobserveerd (een van de methodes van onderzoek) en er is bij hen een enquête afgenomen (een andere methode van onderzoek). De eerste variabele waarop de koffiedrinkers van elkaar verschillen, is 'het soort koffie dat zij het liefst drinken'. Deze variabele heeft in dit geval drie categorieën, drie waarden, zij zijn namelijk cappuccinodrinkers, lattedrinkers, of zij drinken hun koffie zwart. Deze informatie kan zowel door de observatie zijn verkregen, als door de enquête die zij hebben ingevuld.
In het artikel worden meerdere variabelen genoemd, namelijk de mate van obsessief gedrag (sommige koffiedrinkers zijn meer obsessief dan anderen), de mate van anderen tevreden willen stellen (bijvoorbeeld gemeten op een schaal van in zeer kleine mate tot en met in zeer grote mate), de mate van humeurigheid, de mate van directheid en de mate van in zichzelf gekeerd zijn.

De manier waarop je de variabelen meet, is bepalend voor de rest van je onderzoek. Meestal baseer je dit op voorgaand (wetenschappelijk onderzoek), en/of op theorieën die al gebruikt zijn over het onderwerp. Het meetbaar maken van je variabelen in één of meerdere vragen wordt *operationalisatie* genoemd. In

het methodendeel van je onderzoek neem je een gedeelte op waarbij je verantwoordt welke vragen je stelt om welke variabelen te meten en welke antwoorden (waarden) je opneemt. Sommige variabelen (zoals demografische kenmerken) hoeven niet altijd helemaal onderbouwd te worden. Een variabele als 'geslacht' kan immers alleen de waarden 'man' en 'vrouw' aannemen, dat hoeft niet per se verder verantwoord te worden. Je hoeft niet uit te leggen wat je dan precies met 'man' en 'vrouw' bedoelt. Maar bij sommige begrippen is dat moeilijker. Wanneer je een inhoudsanalyse doet naar de mate van seksisme in tijdschriften, moet je eerst duidelijk maken wat je onder seksisme verstaat, en vervolgens moet je daar een of meer items bij maken waarmee je de mate van seksisme wilt vaststellen. Voor de operationalisatie van seksisme in tijdschriften gebruik je bijvoorbeeld twee items (variabelen) waarmee je aangeeft of er in het tijdschrift:
- vrouwen in een 'typisch vrouwelijk' beroep zijn afgebeeld (0 = nee, 1 = ja);
- mannen in een 'typisch mannelijk' beroep zijn afgebeeld (0 = nee, 1 = ja).

NB: Voor de codeur die dit moet gaan verwerken is het dan ook handig een lijst met 'typisch vrouwelijke' en 'typisch mannelijke' beroepen te hebben.

In wetenschappelijk onderzoek waarbij kwantitatieve methoden worden gebruikt, wordt veelal nagegaan wat de verbanden zijn tussen kenmerken van onderzoekseenheden. Soms is er niet alleen sprake van een verband of samenhang, maar oefenen de kenmerken invloed uit op elkaar. Als je twee variabelen met elkaar in verband brengt, kan de één afhankelijk en de ander onafhankelijk zijn.
Je zou je kunnen voorstellen dat welke tijdschriften je leest, afhankelijk is van je leeftijd, of dat het geloof dat je hebt invloed uitoefent op de krant die je leest, of de partij waarop je stemt. De variabele die invloed uitoefent, is de *onafhankelijke variabele* en wordt in de beschrijvende statistiek meestal aangegeven door een x. Als de waarde van deze x verandert, heeft dat gevolgen voor de andere variabele. De variabele die wordt beïnvloed, is de *afhankelijke variabele* (meestal aangegeven door een y). In het voorbeeld van leeftijd en welk tijdschrift je leest, is leeftijd dus de onafhankelijke variabele en het type tijdschrift de afhankelijke variabele. De keuze voor een tijdschrift wordt (mede) beïnvloed door de leeftijd. In dit geval is het moeilijk voor te stellen dat het andersom zou kunnen zijn. De keuze voor een tijdschrift kan immers nooit je leeftijd beïnvloeden. Maar je kunt wel beredeneren dat kinderen niet *Elsevier* en wel *Donald Duck* lezen, terwijl dat bij ouderen eerder andersom is.
Als we nog eens kijken naar het voorbeeld van de koffiedrinkers, zien we dat in dit onderzoek 'soort koffiedrinker' de onafhankelijke variabele is, en dat de andere variabelen (obsessief gedrag, humeurigheid enzovoort), de afhankelijke variabelen zijn. Er wordt immers gesteld dat als iemand een cappuccinodrinker is, iemand eerder obsessief gedrag vertoont dan wanneer iemand een lattedrinker is. De variabele 'mate van obsessief gedrag' wordt hier dus beïnvloed door het soort koffiedrinker dat een persoon is.

Er is niet altijd een onderscheid in onafhankelijk en afhankelijk te maken bij de analyse van het verband tussen variabelen. Als je kijkt of er een verband is tussen het aantal uren dat iemand achter de computer zit en het aantal uren dat iemand televisiekijkt, is niet duidelijk wat nu wat beïnvloedt. Wanneer je veel achter de computer zit, heb je minder tijd om televisie te kijken. Maar andersom is het ook waar: als je veel televisie kijkt, heb je minder tijd om achter de computer te zitten.
Sommige variabelen zijn bijna altijd onafhankelijk, zoals leeftijd en sekse. Er zijn namelijk maar weinig factoren die je leeftijd kunnen beïnvloeden, of je geslacht. Hoe vaak je ook achter de computer zit, je leeftijd of geslacht zal er immers nooit door veranderen.
Wat de afhankelijke en wat de onafhankelijke variabele is, zal vaak blijken uit wat de onderzoeker wil weten. De volgende voorbeelden van onderzoeksvragen maken dat duidelijk:

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

Bij de eerste vraag ga je ervan uit dat woonplaats invloed heeft op het inkomen dat iemand verdient. Het inkomen is hoger of lager voor mensen met een verschillende woonplaats. Je gaat daarbij impliciet uit van een theorie die de hoogte van het inkomen verklaart door de woonplaats. Woonplaats is hier dan ook de onafhankelijke variabele (x), en inkomen de afhankelijke variabele (y). In de tweede vraag is het inkomen juist de onafhankelijke variabele (x). De vraag is of de krant die mensen lezen anders is voor de verschillende inkomensgroepen. In dit geval heeft de onderzoeker kennelijk overwegingen die de invloed van het inkomen op de keuze voor een krant aannemelijk maken. In de laatste vraag is er geen (on)afhankelijke variabele. Er is alleen sprake van een verband tussen favoriete televisieserie en boekgenre. De onderzoeker zag geen aanleiding in de vraagstelling een richting aan te geven in het verband tussen televisieserie en boekgenre. De favoriete serie van iemand zou het favoriete boekgenre kunnen beïnvloeden, maar het omgekeerde kan ook het geval zijn. Mensen die een voorkeur hebben voor een bepaald boekgenre zullen televisieseries kijken die daarbij aansluiten.

1.5 Meetniveaus

De manier waarop je een kenmerk meet, bepaalt ook het meetniveau van de variabele. Het meetniveau van de variabele bepaalt onder andere welke analyses wel en welke analyses niet mogelijk zijn. Hoe hoger een meetniveau, hoe meer mogelijkheden er zijn. Er zijn vier meetniveaus: nominaal, ordinaal, interval en ratio.

1.5.1 Nominaal meetniveau

Het meest elementaire meetniveau kenmerkt zich doordat je niet kunt rekenen met de waarden die je aan de variabelen hebt gegeven. De numerieke waarde is slechts een naamgeving en heeft als getal geen betekenis. De drankjes die geturfd zijn (tabel 1.2) zijn een voorbeeld van een nominaal meetniveau. Je kunt voor de verschillende drankjes een waarde kiezen (1 = bier, 2 = rosé, 3 = cola light, 4 = cappuccino), maar je had net zo goed andere getallen, letters of een symbool kunnen gebruiken  = bier,  = rosé enzovoort). De gekozen waardeaanduidingen onderscheiden de verschillende soorten drankjes. Je kunt niet spreken van een rangordening in die drankjes. Bier (1) is niet meer of minder dan rosé (2), cola light (3) of cappuccino (4). De volgorde in deze getallen kwam toevallig zo uit, omdat bier het eerste drankje was waarvoor je een waarde moest kiezen.
Andere voorbeelden van een nominaal meetniveau zijn geslacht, politieke partij, beroep, religie, woonplaats, favoriete televisiezender, type koffiedrinker en de krant die je leest.

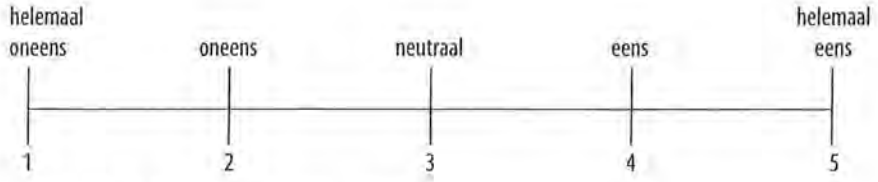
1.5.2 Ordinaal meetniveau

Bij een ordinaal meetniveau is wel sprake van een rangordening. De intervallen tussen de waarden hebben bij ordinale variabelen echter geen betekenis. Een voorbeeld van een ordinale variabele is opleiding, waarbij de respondenten de hoogste opleiding opgeven die ze hebben afgerond.

Je hebt ervoor gekozen om de variabele opleiding de volgende waarden toe te kennen:
1 vmbo;
2 havo;
3 vwo;
4 hbo;
5 wo.

Iemand met een vmbo-opleiding heeft een lagere opleiding genoten dan iemand met een havo-opleiding, die weer een lagere opleiding heeft dan iemand die vwo heeft gedaan. De volgorde is bij een ordinaal meetniveau dus wel van belang. De intervallen zijn echter niet gelijk. De afstand tussen vmbo (1) en havo (2) is niet net zo groot als de afstand tussen havo (2) en vwo (3). Het is ook niet zo dat wanneer je een wo-diploma hebt, je vijf keer hoger opgeleid bent dan iemand met een vmbo-diploma, omdat wo hier toevallig de waarde 5 heeft en vmbo de waarde 1. In plaats van de waarden 1, 2, 3, 4 en 5 had je ook de waarden 1, 4, 5, 8 en 9 kunnen kiezen voor de respectievelijke opleidingsniveaus. Wel is het

belangrijk dat de waarden op blijven lopen: een hogere opleiding moet ook een hogere waarde hebben.
Wanneer je in een vragenlijst voor de beantwoording van een vraag een schaal gebruikt (zie figuur 1.4), heeft ook deze variabele een ordinaal meetniveau. Ook hier hebben de afstanden tussen de waarden geen betekenis. In dit geval geldt: hoe hoger je scoort, hoe meer je het met de vraag eens bent. Maar het verschil tussen 'oneens' en 'neutraal' is niet net zo groot als tussen 'eens' en 'helemaal eens', bijvoorbeeld. Je kunt ook niet zeggen dat als je op deze schaal '4' scoort, je het dan twee keer zoveel eens bent met de stelling in vergelijking met iemand die op deze schaal '2' scoort. Dat komt omdat de afstanden tussen de waarden betekenisloos zijn.



Figuur 1.4 Ordinale variabele: schaal

Andere voorbeelden van variabelen die op ordinaal niveau gemeten zijn, zijn: inkomensklassen (minder dan 1000, 1000 tot en met 2000, en meer dan 2000 euro per maand), leeftijdsgroepen (jonger dan 25, 25 tot 44, 45 tot 64 en 65 jaar en ouder), frequentie bioscoopbezoek (als dit niet in absolute aantallen maar als volgt is gemeten: eenmaal per week, twee- à driemaal per maand, eenmaal per maand, minder vaak dan eenmaal per maand).

Variabelen met een nominaal of ordinaal meetniveau noemen we ook wel 'categorisch'. Het gaat hierbij namelijk voornamelijk om de verschillende categorieën die je kunt onderscheiden door middel van verschillende cijfers, maar je kunt met deze cijfers niet rekenen.

1.5.3 *Interval meetniveau*

Als variabelen op intervalniveau gemeten zijn, is er niet alleen sprake van rangordening, maar hebben de intervallen tussen de verschillende waarden die een variabele aan kan nemen ook een betekenis. Een veelgebruikt voorbeeld is temperatuur. Het verschil tussen 5 en 10 °C is even groot als het verschil tussen 10 en 15 °C (namelijk 5 °C). Er is sprake van een vaste meeteenheid waarbij de waarden voor de graden betekenis toekennen aan de afstanden tussen de graden.
Wat je echter niet kunt zeggen is dat 20 °C twee keer zo warm is als 10 °C. Dit komt door het ontbreken van een natuurlijk (of absoluut) nulpunt. Het nulpunt bij graden Celsius is namelijk arbitrair. Er zijn meer manieren om temperatuur te meten, zoals door middel van graden Fahrenheit. Bij meting in graden

Fahrenheit is er een ander nulpunt en zijn de intervallen tussen de graden anders dan bij graden Celsius. Wanneer het in Amerika tien graden warmer wordt, is die temperatuur in de regel gemeten in Fahrenheit. Wanneer in Nederland de temperatuur met tien graden stijgt, is dit niet dezelfde warmtestijging, omdat wij hier in graden Celsius rekenen.

Variabelen als inkomensklassen en leeftijdsgroepen kun je ook op intervalniveau meten als je ervoor zorgt dat de afstanden tussen de waarden altijd even groot zijn. Stel, je kiest de waarden voor de variabele leeftijdsgroepen als volgt:
1 21 – 25 jaar;
2 26 – 30 jaar;
3 31 – 35 jaar;
4 36 – 40 jaar;
enzovoort.

De afstanden tussen de waarden zijn in dit voorbeeld steeds even groot. De intervallen hebben daarmee een betekenis. Je kunt zeggen dat het verschil tussen klasse 3 en klasse 4 net zo groot is als het verschil tussen klasse 2 en 3. Er is geen absoluut of natuurlijk nulpunt. Je kunt niet zeggen dat iemand uit groep 4 vier keer zo oud is als iemand uit groep 1. Met deze indeling van de leeftijdsgroepen hebben we dus een variabele die op intervalniveau gemeten is.
Een ander voorbeeld van een intervalschaal is 'geboortejaar'. Geboortejaar is arbitrair, wij rekenen met een andere jaartelling dan bijvoorbeeld de Chinezen of Boeddhisten. Er is geen absoluut nulpunt. We hebben een nulpunt 'afgesproken'.

1.5.4 Ratio meetniveau

Het niveau waarbij sprake is van rangordening en waarbij de intervallen betekenis hebben én er een natuurlijk nulpunt aanwezig is, is het rationiveau. Op dit niveau is nul ook werkelijk een absoluut, niet-arbitrair nulpunt. Te denken valt aan lengte, gemeten in aantal centimeters. Er is dan een absoluut nulpunt: iets met een lengte van nul heeft geen lengte. Nu hebben niet alleen de verschillen tussen de afzonderlijke waarden betekenis, maar ook het quotiënt (het resultaat van een deling). Een krantenartikel met een kolomlengte van 15 centimeter is drie keer langer dan een artikel met een kolomlengte van 5 centimeter. Het verschil in kolomlengte tussen deze twee krantenartikelen is 10 centimeter.
Wanneer je zou willen weten hoeveel Facebook-vrienden iemand heeft, en je naar het aantal vrienden vraagt, is dat ook een voorbeeld van een ratio meetniveau. Iemand die negentig vrienden heeft, heeft er drie keer meer dan iemand die dertig vrienden heeft. Ook is er een absoluut nulpunt: minder dan nul Facebook-vrienden kun je niet hebben, het is niet mogelijk om min tien vrienden te hebben. Andere voorbeelden van variabelen op rationiveau zijn leeftijd, aantal uren televisiekijken, gewicht, hoeveelheid studenten in een collegezaal.

Variabelen op interval- en rationiveau noemen we ook wel 'numeriek', omdat bij dit meetniveau de variabelen een numerieke eigenschap hebben waar je mee kunt rekenen. In SPSS worden interval en ratio variabelen onder de noemer *Scale* geschaard.

1.5.5 Criteria

Om te bepalen wat het meetniveau is van een variabele zijn er vier criteria waarop je moet letten: de classificatie, de rangordening, de betekenis van het interval en het natuurlijke nulpunt (zie tabel 1.12).

Tabel 1.12 Criteria meetniveaus

			Ratio
		Interval	absoluut nulpunt
	Ordinaal	'vaste' meeteenheid	'vaste' meeteenheid
Nominaal	rangorde	rangorde	rangorde
classificatie	classificatie	classificatie	classificatie

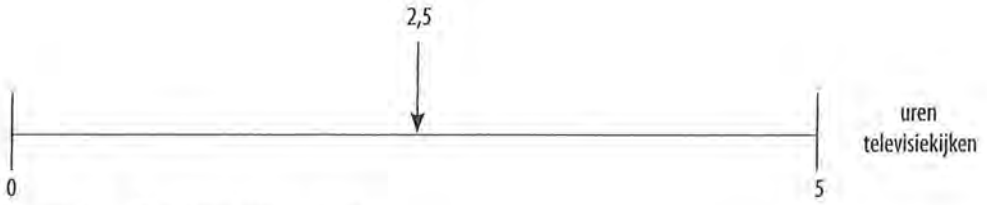
1.6 Waarden van variabelen

De wijze waarop een kenmerk is gemeten, bepaalt dus het meetniveau. Daar heb je als onderzoeker veelal zelf de hand in. Je kunt er zelf voor kiezen om naar iemands leeftijd te vragen (ratio meetniveau), of te vragen in welke leeftijdsklasse ze vallen (ordinaal). Meestal kies je als onderzoeker ervoor om een zo hoog mogelijk meetniveau te nemen, zodat je daar later meer analyses mee kunt uitvoeren.
Sommige variabelen hebben een vaststaand meetniveau. De variabele sekse bijvoorbeeld is altijd nominaal. Het is belangrijk om, voordat je een onderzoek gaat uitvoeren en daarbij de variabelen kiest (om op te nemen in je enquête of codeboek), na te denken over de manier waarop je iets gaat meten en wat daarmee het meetniveau van je variabele is.

1.6.1 Continue en discrete meetschalen

Naast het meetniveau zijn variabelen ook te onderscheiden in continue en discrete variabelen. Bij een *continue meetschaal* kunnen alle mogelijke waarden de uitkomst zijn van de meetprocedure: niet alleen bijvoorbeeld de waarden 1 en 2, maar ook 1,2 en 1,75. Dit in tegenstelling tot een *discrete meetschaal,* die beperkt is tot een telbaar aantal waarden, waarvoor je in de regel gehele getallen gebruikt. De tussenliggende waarden hebben bij discrete meetschalen geen

betekenis (mensen hebben 1, 2 of 3 televisies in huis, maar geen 1,5 televisie). Een continu verschijnsel wordt vaak toch met een discrete meetschaal gemeten, maar dan hebben tussenliggende waarden wel een betekenis. De onderzoeker heeft dan de keuze gemaakt om een kenmerk met een beperkte precisie te meten, bijvoorbeeld om leeftijd in jaren te meten en niet in jaren plus aantal maanden, of nog preciezer in aantal dagen.
Voorbeelden van continue verschijnselen zijn (leef)tijd en afstanden. Bijvoorbeeld, bij tijd kun je naast een onderscheid in uren een onderscheid maken in minuten, seconden, of zelfs nanoseconden. Ook de tussenliggende waarden hebben dan een betekenis (zie figuur 1.5).



Figuur 1.5 Continue meetschaal

Voorbeelden van discrete verschijnselen zijn geslacht, partij waarop iemand stemt, aantal kinderen in een gezin, aantal televisietoestellen in een huis.

Figuur 1.6 Discrete meetschaal

Je hebt niet 1,5 televisietoestel, of een half kind.

1.7 Univariate, bivariate en multivariate analyses

Wanneer je in de statistiek uitspraken wilt doen over de onderzoekseenheden met betrekking tot een aantal kenmerken/variabelen, hangt de formulering van die uitspraak onder andere af van hoe je de variabelen gemeten hebt. Heb je op een vijfpuntsschaal gevraagd hoe vaak iemand sociale media gebruikt? Of: heb je gevraagd hoeveel uur iemand sociale media gebruikt? In het eerste geval zul je alleen gebruik kunnen maken van analyses die geschikt zijn voor een ordinaal meetniveau, in het tweede geval ook van analyses die geschikt zijn voor een rationiveau.
Naast het meetniveau is het van belang *hoeveel* variabelen je in je onderzoek betrekt. We onderscheiden daarbij drie analyseniveaus, namelijk univariaat (één variabele), bivariaat (twee variabelen) en multivariaat (meer dan twee variabelen).

1.7.1 *Univariate analyses*

Univariate analyses kenmerken zich doordat er een uitspraak over één variabele wordt gedaan. Voorbeelden van vragen of hypothesen waarvoor je univariate analyses gebruikt, zijn:

1. Hoeveel uur per dag maken ouderen gebruik van sociale media?
2. Welke winkelketen wordt het meest bezocht in Nederland?
3. In De Telegraaf zijn de meeste artikelen sensatiegericht.

In het eerste voorbeeld zijn 'ouderen' de onderzoekseenheden, en de variabele die wordt gemeten is 'aantal uur per dag gebruikmaken van sociale media'. We kunnen hier ook al zien dat het voor de hand ligt dat deze variabele op rationiveau wordt gemeten, want er wordt een uitspraak verwacht over het aantal uren dat ouderen gebruikmaken van sociale media. Het exacte meetniveau moet meestal blijken uit de bijgeleverde enquête of het codeboek.
In het tweede voorbeeld zijn de onderzoekseenheden minder duidelijk: het zou kunnen gaan om Nederlanders, maar het zou ook kunnen gaan om toeristen. De precieze onderzoekseenheden moeten daarom beschreven worden in het onderzoeksverslag zelf. Het zal in ieder geval om 'mensen' gaan. De variabele die hier wordt gemeten is 'favoriete winkelketen' (of: 'soort winkelketen dat het meest bezocht wordt'). Het meetniveau van deze variabele is nominaal: er moet worden gevraagd aan de respondenten welke winkel zij het meest bezoeken. Je krijgt immers een lijstje met categorieën van verschillende winkels: 1 = Hema, 2 = Bijenkorf, 3 = Zara, etc.).
In het laatste voorbeeld zijn de onderzoekseenheden 'artikelen in De Telegraaf' en wordt gekeken naar de mate van sensatiegerichtheid. Dit zou onderzocht kunnen worden in een inhoudsanalyse, waarbij je een schaal maakt van sensatiegerichtheid. In paragraaf 1.7.2 komt deze terug, als vergelijking.
In paragraaf 1.1 zagen we al dat je bij een univariate analyse een frequentietabel kunt maken, en daarbij kun je vervolgens de centrum- en spreidingsmaten berekenen. Deze staan centraal in de hoofdstukken 2 en 3.

1.7.2 *Bivariate analyses*

Wanneer je twee variabelen met elkaar vergelijkt, zoals we deden bij de kruistabellen (paragraaf 1.3), spreek je van een bivariate analyse. Voorbeelden van uitspraken waarvoor een bivariate analyse nodig is, zijn:

1. Vrouwen kijken vaker naar het journaal dan mannen.
2. Hoogopgeleide vrouwen kijken vaker naar het journaal dan laagopgeleide vrouwen.
3. Wat is het verband tussen het soort koffiedrinkers en de mate van humeurigheid?
4. De artikelen in populaire kranten zijn meer sensatiegericht dan in kwaliteitskranten.

In het eerste voorbeeld zijn er twee variabelen, namelijk geslacht (iemand is of man, of vrouw) en de frequentie waarmee het journaal wordt gekeken. Geslacht is een nominale variabele en bovendien de onafhankelijke variabele, de frequentie journaalkijken is de afhankelijke variabele. Het meetniveau van deze variabele hangt af van de manier waarop je dat gemeten hebt in de vragenlijst. Heb je een schaal gebruikt variërend van '1 = nooit tot en met 5 = elke dag' dan is deze ordinaal, heb je gevraagd: 'hoeveel dagen per week kijkt u naar het journaal?' dan is deze gemeten op rationiveau.
Het tweede voorbeeld bevat ook twee variabelen, namelijk opleidingsniveau en de frequentie waarmee het journaal wordt gekeken. Geslacht is hier nu geen variabele, omdat deze hypothese alleen over vrouwen gaat, en er geen vergelijking wordt gemaakt tussen mannen en vrouwen. Vrouwen zijn hier dus de onderzoekseenheden. Het opleidingsniveau is in deze hypothese de onafhankelijke variabele, want volgens de hypothese heeft opleidingsniveau invloed op de frequentie waarmee naar het journaal wordt gekeken, de afhankelijke variabele.
Bij het voorbeeld over koffiedrinkers is er geen duidelijke onafhankelijke variabele, maar zijn er twee variabelen die elkaar kunnen beïnvloeden. Er is de variabele 'soort koffiedrinker' (nominaal) en de variabele 'mate van humeurigheid' (afhankelijk van hoe deze gemeten is, is deze ordinaal, interval of ratio).
In het laatste voorbeeld zijn de twee variabelen 'soort krant', waarbij een krant ofwel in de categorie populaire krant, ofwel in de categorie kwaliteitskrant valt. Het is daarmee een nominale variabele. De tweede variabele, de afhankelijke variabele, is 'mate van sensatiegerichtheid', en ook hier wordt het meetniveau bepaald door de manier waarop je dat als onderzoeker gemeten hebt.
Bij bivariate analyses kun je kijken naar verschillen, naar samenhang of naar verbanden. In hoofdstukken 5, 6, 8 en 9 worden bivariate analyses behandeld.

1.7.3 Multivariate analyses

Wanneer je meer dan twee variabelen gebruikt, voer je een multivariate analyse uit. In dit boek zullen we bij analyses met meerdere variabelen altijd kiezen voor één afhankelijke variabele, en meerdere onafhankelijke variabelen. Analyses waarin meerdere afhankelijke variabelen worden gebruikt zijn wel mogelijk, maar zullen hier niet behandeld worden. Voorbeelden van uitspraken waarvoor multivariate analyses nodig zijn, zijn:

1. Hoogopgeleide vrouwen kijken vaker naar het journaal dan laagopgeleide mannen.
2. Het effect van soort koffiedrinker op de mate van humeurigheid is voor mannen anders dan voor vrouwen.
3. De mate van tevredenheid met het eigen leven hangt af van de leeftijd, het aantal dagen in de week en het aantal uren per dag dat een tiener achter de pc mag zitten, en de mate waarin de tiener gameverslaafd is.

Anders dan in voorbeeld 2 bij de bivariate analyses, worden hier bij de eerste uitspraak wel drie variabelen gebruikt. De afhankelijke variabele is weer 'frequentie journaalkijken', maar er zijn nu twee onafhankelijke variabelen die daar invloed op kunnen uitoefenen, namelijk geslacht en opleidingsniveau. Ook in het tweede voorbeeld zijn er drie variabelen, namelijk geslacht (onafhankelijk, nominaal), soort koffiedrinker (onafhankelijk, nominaal) en de mate van humeurigheid (de afhankelijke variabele; het meetniveau hangt af van de manier waarop je deze gemeten hebt).
In het derde voorbeeld zijn er vijf variabelen. De onderzoekseenheden zijn hier tieners, de afhankelijke variabele is hier de mate van tevredenheid met het eigen leven. Bij deze hypothese vermoedt de onderzoeker dat de mate van tevredenheid met het eigen leven wordt beïnvloed door zowel de leeftijd van de tiener, het aantal dagen in de week dat achter de pc wordt gezeten, het aantal uur dat per dag achter de pc wordt gezeten, en de mate waarin de tiener gameverslaafd is.
Multivariate analyses komen aan bod in hoofdstuk 7, en in de hoofstukken 8 en 9 waar zowel bi- als multivariate analyses worden besproken.

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

Noot

1 Wij gebruiken kolompercentages en we zetten de onafhankelijke variabele daarom in de kolommen.

Centrummaten 2

In hoofdstuk 1 is het meetniveau van variabelen besproken (nominaal, ordinaal, interval of ratio). Verder zijn de begrippen continu en discreet toegelicht. Zoals gezegd is dit van belang voor de verschillende analyses die mogelijk zijn met die variabelen. In dit hoofdstuk staan de meest elementaire analyses centraal: centrummaten.
Een centrummaat is een getal dat aangeeft rond welke (centrale) waarde de uitkomsten van een serie waarnemingen liggen. Je onderscheidt drie centrummaten: modus, mediaan en gemiddelde.

2.1 Modus

De modus is de waarde die het meest voorkomt, de waarde met de hoogste frequentie. De modus kun je bij alle meetniveaus gebruiken, maar geeft niet altijd zinnige informatie. Op nominaal niveau is de modus de meest geschikte en de enig mogelijke centrummaat.
Laten we nog eens kijken naar het aantal drankjes dat geturfd is op het terrasje.

Tabel 2.1 Centrummaat op nominaal niveau: de modus

Drankje	Aantal (geturfd)	Absolute frequentie
1: bier	𝍸 III	⑧
2: rosé	𝍸	5
3: cola light	I	1
4: cappuccino	III	3
Totaal		17

Uit tabel 2.1 blijkt dat de meeste mensen bier willen (acht mensen), de modus is dus 1, omdat dat de waarde is die we aan bier hebben toegekend. Nogmaals, de waarden hebben op nominaal niveau geen getalsmatige betekenis.
Je kunt de modus bepalen aan de hand van een frequentieverdeling. We hebben bijvoorbeeld in een databestand een variabele die aangeeft wat de favoriete televisieserie is van de respondenten. Favoriete televisieserie is een nominale variabele omdat er slechts sprake is van classificatie (naamgeving), maar niet van een rangordening. In de output van SPSS ziet de frequentieverdeling eruit als in tabel 2.2.

Tabel 2.2 Frequentieverdeling van de variabele televisieserie (SPSS-output)

Serie favoriete tvserie

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 True Detective	22	33,8	33,8	33,8
	2 Game of Thrones	20	30,8	30,8	64,6
	3 Dr. Who	23	35,4	35,4	100,0
	Total	65	100,0	100,0	

Zowel aan de absolute frequentie als aan het percentage is te zien dat de meeste mensen in dit onderzoek *Dr. Who* als favoriete serie hebben. *Dr. Who* heeft in dit onderzoek de waarde 3. De modus is dus 3.

Een modus mag je ook uitrekenen voor variabelen die op ordinaal, interval- of rationiveau zijn gemeten. Zo zou je kunnen kijken welke opleiding (lager, middelbaar of hoger onderwijs – ordinaal) het meest voorkomt of hoe oud (leeftijd in jaren – ratio) de meeste van je respondenten zijn.
Een nadeel van de modus is dat dit kengetal geen informatie geeft over de overige waarden van een variabele. Daardoor geeft de modus soms geen informatie waar je wat aan hebt. Als we bijvoorbeeld kijken naar de leeftijdsverdeling van vijftig mensen die variëren van 18 tot 81 jaar, dan zou de modus 18 kunnen zijn. Het is mogelijk dat maar vijf personen die leeftijd hebben, en dat van alle andere leeftijden er steeds vier of minder zijn. De modus is ook 18 als veertig van de vijftig personen 18 jaar zijn, maar dan is er sprake van een geheel andere leeftijdsverdeling binnen die groep. Een modus van 18 geeft in dit geval beperkte en niet erg nuttige informatie.

2.2 Mediaan

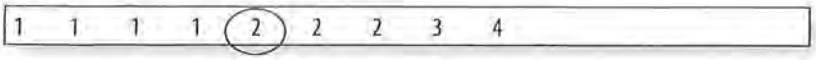
De mediaan is de middelste waarneming na rangordening van de data van laag naar hoog. Het is de waarneming waar 50% van de onderzoekseenheden onder ligt en 50% boven. Uit deze definitie blijkt al dat je de mediaan niet op nominaal niveau kunt gebruiken, want er moet een rangorde in de waarden zitten. Je gebruikt de mediaan op ordinaal of hoger niveau. De mediaan is in de regel de meest geschikte centrummaat voor ordinale variabelen.
Stel, negen personen hebben aangegeven hoeveel televisie ze per dag kijken. Televisiekijken is in dit geval op een ordinale schaal gemeten, namelijk:
1: één uur of minder;
2: meer dan één, maar minder dan anderhalf uur;
3: anderhalf tot twee uur;
4: meer dan twee, maar minder dan tweeënhalf uur;
5: tweeënhalf tot drie uur;
6: meer dan drie uur.

NB: De categorieën moeten elkaar uitsluiten!

Dit leverde de volgende gegevens op voor de meting van de televisiekijktijd:

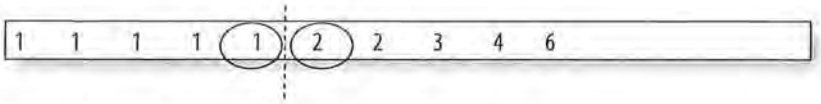
1 2 1 4 2 1 2 1 3

De eerste stap is het rangschikken van de gegevens. De middelste waarneming is de mediaan (zie figuur 2.1).



Figuur 2.1 Mediaan bij oneven aantal waarnemingen

De mediaan is hier 2. Dat wil zeggen dat 50% van de personen een score had van 2 of lager en dat 50% van de personen een score had van 2 of hoger. Met andere woorden, 50% van de personen kijkt minder dan anderhalf uur per dag televisie, en 50% van de personen kijkt meer dan één uur per dag televisie. De waarde 2 is immers 'meer dan één, maar minder dan anderhalf uur'.
Wanneer er een even aantal waarnemingen is, is er geen middelste waarneming. Toch kun je wel een mediaan berekenen. Eerst sorteer je de waarnemingen weer op grootte. Vervolgens tel je de twee middelste waarden bij elkaar op en deel je dat getal door twee (zie figuur 2.2).



Figuur 2.2 Mediaan bij even aantal waarnemingen

De mediaan is hier (1 + 2) / 2 = 1,5. Dat wil zeggen dat de mediaan tussen de 1 en de 2 ligt. Dit betekent dat 50% van de onderzoekseenheden voor die variabele een waarde heeft van 2 of meer en 50% een waarde van 1 of minder. Als het weer gaat om de eerder gebruikte ordinale schaal voor kijktijd, betekent dit dat 50% een uur of minder en 50% langer dan een uur per dag televisiekijkt. Het gebruik van de mediaan bij interval- en ratiovariabelen kan zinnig zijn omdat deze centrummaat ongevoelig is voor uitschieters, terwijl het rekenkundig gemiddelde daar wel gevoelig voor is (zie paragraaf 2.3).
Ook de mediaan is op basis van een frequentietabel in de SPSS-output eenvoudig te bepalen. SPSS geeft bijvoorbeeld de in tabel 2.3 weergegeven frequentietabel van de variabele opleiding.

Aan het cumulatieve percentage is af te lezen dat bij de waarde 5 (vwo) de 50%-grens wordt gepasseerd. De mediaan is derhalve 5. Dat wil zeggen dat 50% van de respondenten een opleidingsniveau heeft van vwo of lager en dat 50% een opleidingsniveau heeft van vwo of hoger.
De mediaan kan dus nooit berekend worden bij variabelen op nominaal niveau, je kunt immers niet zeggen: '50% van de respondenten leest het liefst detectives of minder, en 50% van de respondenten leest het liefst detectives of meer'. In

de variabele 'favoriete genre boek' zit immers geen rangordening en heeft de mediaan geen betekenis. De mediaan kan wel op een hoger niveau dan ordinaal worden berekend, dus bij numerieke variabelen (interval en ratio). Wanneer je 'aantal uur per week een boek lezen' hebt gemeten (een ratiovariabele), is het geoorloofd om te zeggen: '50% van de jongeren leest 5 uur of minder per week een boek en 50% van de jongeren leest 5 uur of meer per week een boek'.

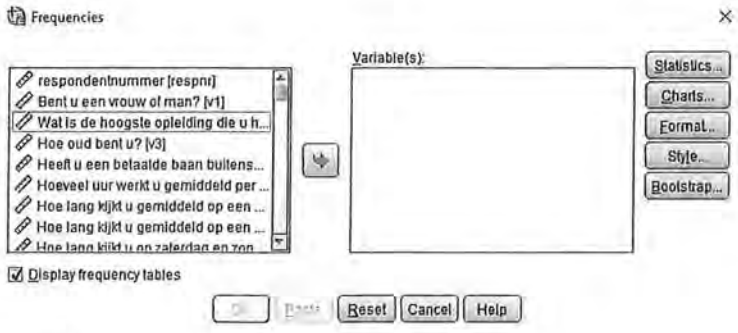
Tabel 2.3 Frequentieverdeling van de variabele opleiding (SPSS-output)

v2 opleiding

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 lager onderwijs	26	3,6	3,6	3,6
	2 mavo	45	6,2	6,2	9,8
	3 mbo	83	11,4	11,5	21,3
	4 havo	100	13,8	13,8	35,1
	5 vwo	209	28,8	28,9	64,0
	6 hbo	158	21,8	21,8	85,8
	7 universiteit	103	14,2	14,2	100,0
	Total	724	99,9	100,0	
Missing	System	1	,1		
Total		725	100,0		

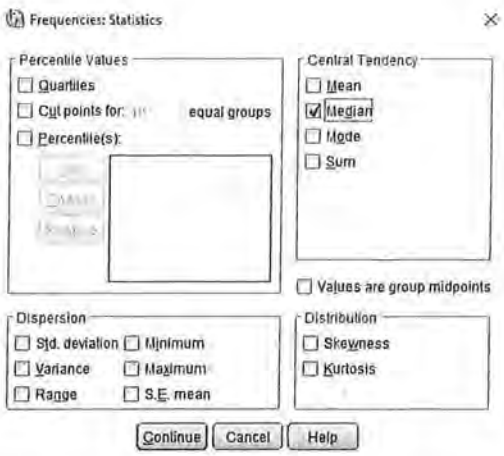
SPSS Centrummaten

Voor het berekenen van centrummaten in SPSS ga je eerst weer naar *Frequencies (Analyze → Descriptive Statistics → Frequencies)*. Hier selecteer je de variabele waarvan je een centrummaat wilt laten berekenen, bijvoorbeeld opleiding (zie figuur A).



Figuur A Frequencies-venster

Via *Statistics* kun je door SPSS een centrummaat laten berekenen (zie figuur B).



Figuur B Statistics-venster

In het *Statistics*-venster kun je aanvinken welke centrummaat (onder *Central Tendency*) je wilt laten berekenen. Afhankelijk van het meetniveau en je eigen wensen kun je kiezen voor het gemiddelde (*Mean*), de mediaan (*Median*) of de modus (*Mode*). In dit voorbeeld gaat het om de variabele opleiding. Opleiding is een ordinale variabele, de meest geschikte centrummaat is dan de mediaan. Overigens raden wij aan om het gemiddelde op een andere manier door SPSS te laten berekenen, zie daarvoor kader 3.1 in hoofdstuk 3.

Om de analyse uit te voeren kun je klikken op OK of op PASTE. In hoofdstuk 4 (Bewerken van je data) leggen we uit waarom je beter op PASTE kunt klikken.

Kader 2.1

2.3 (Rekenkundig) gemiddelde

De laatste centrummaat is het gemiddelde. Het gemiddelde gebruik je alleen op interval- en rationiveau. Bij nominale en ordinale variabelen is het uitrekenen van het gemiddelde niet geoorloofd. Je kunt niet zeggen dat het gemiddelde opleidingsniveau 3,4 is, aangezien de voor de afzonderlijke opleidingen gekozen waarden en de intervallen tussen die waarden geen betekenis hebben. Bij het interval- en rationiveau hebben de getallen van de waarden wel een betekenis. Met die waarden mogen we dan ook rekenen.
Het rekenkundig gemiddelde bereken je door alle waarnemingen bij elkaar op te tellen en te delen door het totaal aantal waarnemingen (n). Het symbool voor het gemiddelde is $\bar{x}$ (x streep). In wetenschappelijke artikelen wordt ook vaak de letter M, van het Engelse woord *Mean*, gebruikt om het gemiddelde aan te geven.
In formulevorm ziet de berekening er als volgt uit:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Formule voor gemiddelde

De Griekse hoofdletter sigma (Σ) wordt gebruikt als sommatieteken en betekent: neem de som van. Boven de sigma staat een n, en onder de sigma i = 1. Dit betekent: neem van elke i-de onderzoekseenheid, vanaf de eerste (i =1) tot en met de n-de (i = n), de waarde van x. De sigma betekent dat je al die waarden bij elkaar moet optellen.[1] De n staat voor het totaal aantal waarnemingen. De gehele formule zegt dus: om $\bar{x}$ te berekenen neem je de som van alle waarden van x (van waarneming 1 tot en met n) en deel je deze door n (het totaal aantal waarnemingen).

Laten we eens kijken naar tien personen die een statistiektest hebben afgelegd met daarin twintig vragen. De reeks hierna toont het aantal fouten dat gemaakt is.

0 0 1 1 1 4 4 4 6 6

Voor de berekening van het gemiddelde tellen we alle *x*'en (alle individuele scores) bij elkaar op en delen deze door 10.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0+0+1+1+1+4+4+4+6+6}{10} = \frac{27}{10} = 2{,}7$$

Het gemiddeld aantal fouten dat de personen hebben gemaakt in de statistiektest is dus 2,7.
Ook vanuit een frequentietabel is het gemiddelde te berekenen. De formule is dan iets anders, maar het principe werkt hetzelfde.

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j f_j}{n}$$

Formule gemiddelde berekend op basis van groepsfrequenties

In plaats van een *i* staat er nu een *j* onder de sigma en een *k* erboven. Achter de *x* is een *f* toegevoegd. De *j* betekent 'de groep' en er zijn in totaal *k* groepen. De f_j is de frequentie waarmee een waarde van *x* in groep *j* voorkomt. Om het gemiddelde te berekenen vermenigvuldig je voor elke groep (*j* = 1 tot en met *k*) de waarde *x* met de frequentie (f_j = het aantal onderzoekseenheden in die groep); de producten van alle groepen tel je vervolgens bij elkaar op en deel je door het totale aantal onderzoekseenheden (*n*).
Een voorbeeld zal dit duidelijker maken. De frequenties van het 'aantal fouten in de test' staan in tabel 2.4.

Tabel 2.4 Frequentietabel 'fouten in test' in absolute aantallen (n = 10)

Fouten (x)	Frequentie (f)
0	2
1	3
4	3
6	2
	10

Tabel 2.4 laat zien dat er twee keer nul fouten werden gemaakt, drie keer één fout enzovoorts. Om het gemiddelde te berekenen zou je volgens de eerste formule (waarbij achter de x een i stond) alle x'en (de fouten van alle individuen) uit moeten schrijven, en bij elkaar optellen. Je krijgt dan de eerder gegeven rij cijfers: 0 0 1 1 1 4 4 4 6 6. Het is eenvoudiger om de x te vermenigvuldigen met de frequentie f en daarna de uitkomsten daarvan bij elkaar op te tellen, zoals in tabel 2.5 is gedaan.

Tabel 2.5 Berekenen van gemiddeld 'aantal fouten' (n=10)

Fouten (x)	Frequentie (f)	$x_j * f_j$
0	2	0 * 2 = 0
1	3	1 * 3 = 3
4	3	4 * 3 = 12
6	2	6 * 2 = 12
		Σ 27

De formule is nu eenvoudig in te vullen:

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j f_j}{n} = \frac{27}{10} = 2{,}7$$

Bij tien onderzoekseenheden is het nog mogelijk om de waarden van alle eenheden uit te schrijven en bij elkaar op te tellen. In het volgende voorbeeld gaat het om een groter aantal onderzoekseenheden. We willen nu de gemiddelde leeftijd weten van 73 personen. Gegeven is de frequentietabel (tabel 2.6).

Tabel 2.6 Frequentietabel van leeftijd

Leeftijd (*x*)	Frequentie (*f*)
21	1
22	6
23	22
24	19
25	20
26	3
28	1
31	1
	73

Het is te veel werk om alle scores apart uit te gaan schrijven. Je zou dan een lange rij met getallen krijgen: 21 22 22 22 22 22 22 23 23 23 23 23 enzovoort. We kiezen er daarom voor om het gemiddelde te berekenen op basis van de groepen (tabel 2.7).

Tabel 2.7 Berekenen van gemiddelde leeftijd ($n = 73$)

Leeftijd (*x*)	Frequentie (*f*)	$x_j * f_j$		
21	1	21 * 1	=	21
22	6	22 * 6	=	132
23	22	23 * 22	=	506
24	19	24 * 19	=	456
25	20	25 * 20	=	500
26	3	26 * 3	=	78
28	1	28 * 1	=	28
31	1	31 * 1	=	31
	73	Σ	=	1752

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j f_j}{n} = \frac{1752}{73} = 24$$

De gemiddelde leeftijd van deze 73 respondenten is 24 jaar.

2.4 Keuze tussen centrummaten

Als een variabele op nominaal niveau is gemeten, is de enige centrummaat die je kunt gebruiken de modus. Een andere keuze is er niet. Bij variabelen op ordinaal niveau geven modus en mediaan informatie die voor een onderzoeker nuttig kan zijn. Bij ratio- en intervalvariabelen is het gebruik van alle drie de centrummaten mogelijk, maar die informatie is niet altijd zinnig.
Stel, je wilt door middel van een kengetal iets zeggen over de leeftijden van de leden van een huishouden. Het huishouden bestaat uit een man, vrouw, vier kinderen (waarvan een tweeling) en een inwonende grootmoeder. Hun leeftijden zijn:

15 15 16 17 45 50 80

In dit geval is de modus 15, de mediaan 17 en de gemiddelde leeftijd 34. Welke centrummaat geeft in dit geval de meest zinnige informatie?
Als niet de jongste maar de oudste toevallig een tweeling was geweest, was de modus 17 geweest. De modus als kengetal zegt hier niet zoveel over de leeftijdsverdeling. Aan de mediaan hebben we meer. Dat 50% 17 jaar of jonger en 50% 17 jaar of ouder is, geeft een aardige indruk van de leeftijdsverdeling. De gemiddelde leeftijd van 34 zegt weer minder, want dit gemiddelde wordt sterk beïnvloed door de leeftijd van de grootmoeder. Zonder deze uitschieter zou de gemiddelde leeftijd 26,3 jaar zijn.

15 15 16 17 45 50

Zonder de leeftijd van de grootmoeder verandert de modus niet en ligt de mediaan tussen de 16 en 17 (16,5). De keuze voor een centrummaat is dus afhankelijk van de informatie die je nodig hebt, en de uitschieters die eventueel in een verdeling voorkomen.

2.5 Samenvatting

De eerste stap voor het kiezen van een centrummaat is het vaststellen van het meetniveau van de variabelen. Het meetniveau beperkt de keuzemogelijkheden. Als we de kenmerken van het meetniveau maximaal willen benutten, is de meest geschikte centrummaat voor interval- en ratiovariabelen het rekenkundig gemiddelde, voor ordinale variabelen is het de mediaan en voor nominale variabelen de modus.

Tabel 2.8 Meetniveaus en centrummaten

Nominaal	Ordinaal	Interval	Ratio
modus	*modus*	*modus*	*modus*
	mediaan	*mediaan*	*mediaan*
		gemiddelde	***gemiddelde***

In tabel 2.8 staat cursief voor 'geoorloofd'. Cursief en vet staat voor 'meest geschikt', dat wil zeggen dat er maximaal gebruik wordt gemaakt van de kenmerken van het meetniveau.

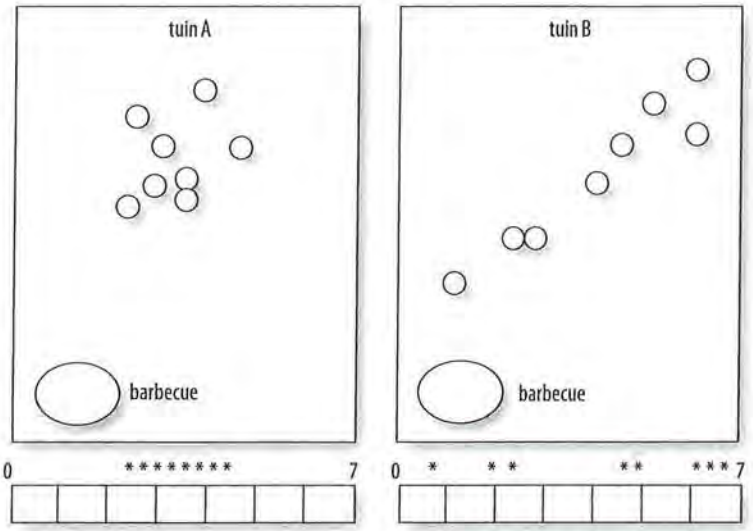
Ga naar de website om de opdrachten bij dit hoofdstuk te maken.

Noot

1 We zullen bij het uitwerken van de formules niet altijd deze informatie rondom de sigma vermelden; over het algemeen wordt ook uit de formule zelf al duidelijk wat er gedaan moet worden.

Spreiding 3

In dit hoofdstuk staan de spreidingsmaten centraal. In het vorige hoofdstuk zagen we dat centrummaten aangeven rond welke waarde op de meetschaal de waarnemingen zich centreren, oftewel rond welke centrale waarde op de meetschaal de waarnemingen verspreid zijn. Spreidingsmaten geven aan hoe sterk de waarden zich concentreren: liggen deze dicht bij elkaar of zijn ze juist erg verdeeld?
Wat is nu eigenlijk spreiding? Spreiding is in feite niets anders dan de afstand tussen de verschillende waarnemingen. Stel, je bent op een tuinfeestje met in een hoek van de tuin een barbecue (zie figuur 3.1).



Figuur 3.1 Voorbeeld van tuin A, waarin er weinig spreiding is van de personen over de tuin, en tuin B, waarin de personen veel meer verspreid zijn, terwijl de gemiddelde afstand tot de barbecue ongeveer gelijk is

Wanneer iedereen dicht bij elkaar staat, al of niet bij de barbecue, is er weinig spreiding (tuin A). Wanneer iedereen is verdeeld over de hele tuin, is er sprake van veel spreiding (tuin B). Omdat de afstand tussen de waarden centraal staat in veel spreidingsmaten, zijn deze maten vooral van belang bij metingen op interval- of rationiveau. Op nominaal niveau kun je niet spreken over afstanden tussen de waarden. Bij nominale variabelen is het aantal mogelijke waarden een manier om een indicatie van spreiding te geven. Er werden bijvoorbeeld vier verschillende soorten drankjes besteld. Als er een volgende keer meer verschillende soorten drankjes besteld worden, is de spreiding gemeten in het *aantal categorieën* (het

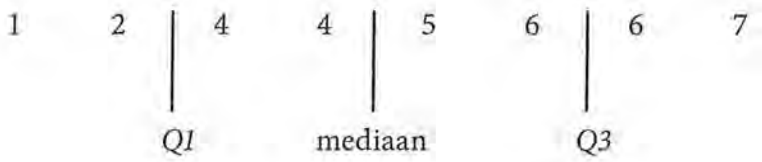
aantal verschillende drankjes) groter. Een andere manier om bij nominale variabelen een indicatie van de spreiding te geven is de *variatieratio*, waarmee je het aandeel van de onderzoekseenheden aangeeft dat niet in de modale categorie valt. De meest simpele indicatie van spreiding bij variabelen vanaf ordinaal niveau is de *range*. Dit is het verschil tussen de hoogste en laagste waarde van de variabele.

3.1 Kwartielen

Bij variabelen van minimaal ordinaal niveau zijn er meer mogelijkheden om iets over spreiding te zeggen, omdat de waarden die je gebruikt een bepaalde rangordening hebben. Plaats je de uitkomsten van een serie waarnemingen in oplopende volgorde, dan ontstaat een geordende getallenreeks. In het vorige hoofdstuk hebben we al gezien dat de 50%-grens van deze getallenreeks de mediaan is. We kunnen de getallenreeks ook in vier stukken verdelen, die elk 25% van de waarnemingen bevatten. Deze stukken noem je de *kwartielen*. Informatie over deze kwartielen geeft een indruk van de spreiding bij ordinale variabelen.
Het eerste kwartiel (aangeduid als *Q1*) is de waarde waarbij 25% van de onderzoekseenheden een kleinere of gelijke waarde heeft, en 75% een gelijke of grotere waarde. Het tweede kwartiel is de mediaan: 50% is gelijk of kleiner dan die waarde en 50% is gelijk of groter dan die waarde. Het derde kwartiel (*Q3*) is de waarde waarbij 25% van de onderzoekseenheden grotere of gelijke waarden heeft (en daarmee heeft 75% een gelijke of kleinere waarde).
Het verschil tussen het eerste en derde kwartiel noem je de *interkwartielafstand* (*Q3 – Q1*). Er moet sprake zijn van waarden die groter of kleiner zijn, en omdat voor het berekenen van de interkwartielafstand het verschil tussen twee waarden (*Q1* en *Q3*) wordt bepaald, moet dat verschil ook betekenis hebben. Deze spreidingsmaat is dus alleen geschikt als het meetniveau minimaal interval is.
Een voorbeeld: op het tuinfeestje zijn de mensen verspreid over de hele tuin (die 7 meter lang is) en is de afstand tussen acht mensen en de barbecue berekend in meters. Deze gegevens zijn vervolgens gerangschikt van laag naar hoog.

1 2 4 4 5 6 6 7 (afstand in meters ten opzichte van de barbecue)

Er is één persoon die één meter van de barbecue staat, er is ook één persoon op een afstand van twee meter, er zijn twee mensen die vier meter van de barbecue staan enzovoort. De mediaan is hier 4,5 (= (4 + 5) ÷ 2). Om de mediaan te vinden heb je de rij in twee gelijke stukken opgedeeld. Het eerste kwartiel ligt op de helft van het eerste stuk, het derde kwartiel ligt op de helft van het tweede stuk.



Het eerste kwartiel (*Q1*) is hier 3 (= (2 + 4) ÷ 2) en het derde kwartiel 6 (= (6 + 6) ÷ 2).
De interkwartielafstand is dan: $Q3 - Q1 = 6 - 3 = 3$.

De waarde van de interkwartielafstand is beter te interpreteren wanneer je deze vergelijkt met een andere interkwartielafstand. Stel dat je na een paar uur nog een meting doet, en dan de volgende data krijgt:

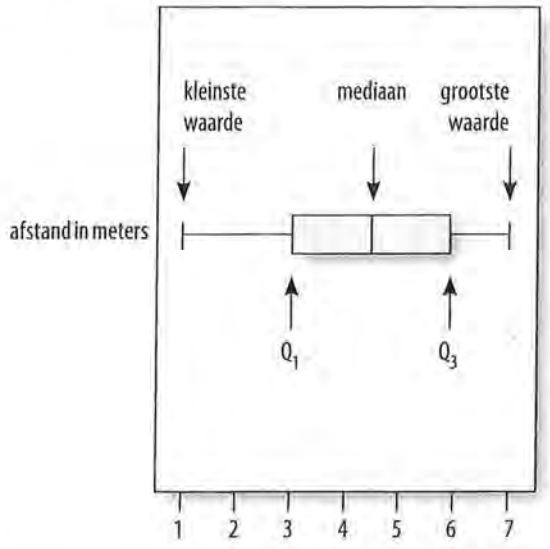
1 4 | 4 4 | 5 5 | 6 7

Q1 mediaan *Q3*

Het eerste kwartiel (*Q1*) is nu 4 (= (4 + 4) ÷ 2) en het derde kwartiel 5,5 (= (5 + 6) ÷ 2).
De interkwartielafstand is dan: $Q3 - Q1 = 5{,}5 - 4 = 1{,}5$ meter.
De interkwartielafstand is kleiner geworden (gedaald van 3 naar 1,5). Je kunt nu een vergelijking maken tussen de interkwartielafstanden. De spreiding is dus eerder op de avond groter dan later op de avond.

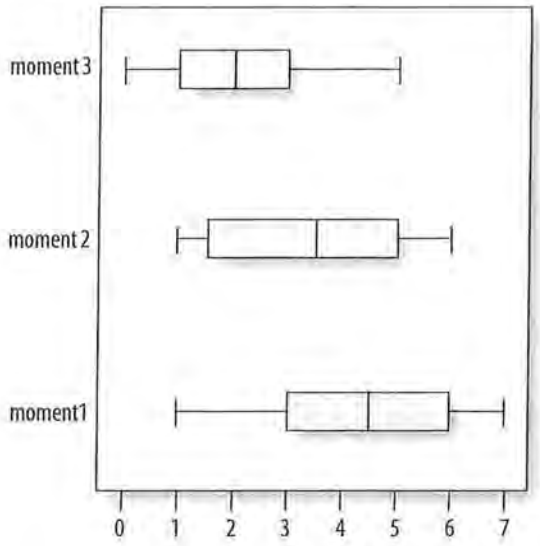
3.1.1 Boxplot

Een *boxplot* is een grafische weergave van de kwartielen. Het is een handig hulpmiddel om zowel de centrale tendentie (centrummaat) als de spreiding in één oogopslag te zien. In een boxplot worden de laagste waarde, het eerste kwartiel, de mediaan, het derde kwartiel, en de hoogste waarde weergegeven. Figuur 3.2 geeft een boxplot van de positie van de gasten op de barbecue 'eerder op de avond'.



Figuur 3.2 Boxplot van de afstand tot de barbecue 'eerder op de avond'

Gedurende de avond lopen deze acht mensen door de tuin heen, waardoor de spreiding kan variëren. Wanneer je op drie momenten de posities van deze mensen bijhoudt, kun je door middel van een boxplot snel een overzicht krijgen van het verloop van de avond.



Figuur 3.3 Boxplot van drie momenten

Naarmate de avond vordert, wordt de spreiding kleiner (op moment 3) en neemt langzamerhand de afstand tot de barbecue af.

3.2 Variantie

Door een centrummaat te berekenen geef je een beschrijving van een groep onderzoekseenheden. Een spreidingsmaat voegt hier belangrijke informatie aan toe.
We zagen al eerder aan het voorbeeld van de barbecue op het tuinfeest dat spreiding iets zegt over de afstand van de onderzoekseenheden ten opzichte van een bepaald centrum. De meest gebruikte manier om iets over spreiding te zeggen is de standaarddeviatie. Deze wordt berekend aan de hand van de variantie en de variatie. We zullen daarom eerst in deze paragraaf de *variatie* en *variantie* bespreken.

We kijken naar een voorbeeld waarin aan acht jongeren is gevraagd hoeveel uur zij per week online het nieuws lezen. Dit zijn de gemeten waarden:

0 1 2 2 4 6 6 8

Op interval- en rationiveau is de meest geschikte centrummaat het rekenkundig gemiddelde. Het gemiddeld aantal uur dat jongeren online het nieuws lezen is 3,625.[1]

$$\bar{x} = \frac{\sum x_i}{n} = \frac{0+1+2+2+4+6+6+8}{8} = 3{,}625$$

We kunnen nu per persoon (per onderzoekseenheid) kijken in hoeverre hij of zij in afstand verschilt (of afwijkt) van het gemiddelde. Persoon 1 leest geen nieuws online, hij of zij heeft de waarde 0. Deze persoon leest dus 3,625 minder uur online het nieuws dan het gemiddelde van deze acht personen. Oftewel: 0 – 3,625 = –3,625.
Persoon 2 leest 1 uur per week online het nieuws. Het verschil met het gemiddelde is 1 – 3,625 = –2,625. Dit verschil kan voor elke onderzoekseenheid berekend worden. De notatie hiervoor is $(x_i - \bar{x})$. Letterlijk staat hier: van elke individuele x (en x is aantal uur online nieuws kijken voor alle afzonderlijke personen) wordt het gemiddelde van x afgetrokken.
Wanneer we dat voor elke onderzoekseenheid doen, hebben we acht verschillende afstanden ten opzichte van het gemiddelde, de 'individuele verschillen met de gemiddelde afstand' (tabel 3.1). Dit zegt nog steeds niet zoveel. Over die verschillen met de gemiddelde afstand kun je ook een gemiddelde berekenen: het gemiddelde verschil met de gemiddelde afstand. Om dit te berekenen, zou je alle individuele verschillen ten opzichte van het gemiddelde bij elkaar op willen tellen $\sum(x_i - \bar{x})$ en delen door n (het totaal). Het probleem daarbij is dat deze som (Σ) *altijd* uitkomt op nul.[2]

Tabel 3.1 Individuele verschillen met de gemiddelde afstand

x_i	$(x_i - \bar{x})$
0	(0 – 3,625) = –3,625
1	(1 – 3,625) = –2,625
2	(2 – 3,625) = –1,625
2	(2 – 3,625) = –1,625
4	(4 – 3,625) = 0,375
6	(6 – 3,625) = 2,375
6	(6 – 3,625) = 2,375
8	(8 – 3,625) = 4,375
Σ	0

Om van de nul af te komen kunnen we elk verschil kwadrateren. Door te kwadrateren raken we het minteken kwijt, zodat we de negatieve waarden bij de eerste vier onderzoekseenheden kwijt zijn. Daarna kunnen we alle kwadraten bij elkaar optellen, sommeren. In formulevorm ziet dat er als volgt uit:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Formule voor variatie

Deze kwadratensom noem je de *variatie*. Maar dit getal is als spreidingsmaat moeilijk te interpreteren, omdat het getal sterk afhankelijk is van het aantal onderzoekseenheden. Hoe meer onderzoekseenheden er zijn, hoe hoger de waarde van de variatie wordt. Dat maakt interpretatie van het getal moeilijk. Je lost dit probleem op door de *variantie* te berekenen. De variantie is een soort gemiddelde kwadratische afwijking ten opzichte van het gemiddelde. Het symbool voor variantie is s^2. De formule is:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Formule voor variantie

Letterlijk staat er: elke individuele x min het gemiddelde van x wordt gekwadrateerd, de kwadraten van alle onderzoekseenheden worden bij elkaar opgeteld, en de som wordt gedeeld door het totaal aantal waarnemingen (n) min 1.[3]
Om de variantie uit te rekenen berekenen we dus eerst de variatie (de teller in de formule voor variantie).

Tabel 3.2 Berekenen van de variatie ($n = 8$)

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
0	(0 – 3,625) = –3,625	13,141
1	(1 – 3,625) = –2,625	6,891
2	(2 – 3,625) = –1,625	2,641
2	(2 – 3,625) = –1,625	2,641
4	(4 – 3,625) = 0,375	0,141
6	(6 – 3,625) = 2,375	5,641
6	(6 – 3,625) = 2,375	5,641
8	(8 – 3,625) = 4,375	19,141
Σ	0	55,878

De variatie is hier 55,878.
Vervolgens vullen we de formule in om de variantie te berekenen:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} = \frac{55{,}878}{8-1} = 7{,}983$$

De variantie is hier 7,983. Maar ook dit getal is moeilijk te interpreteren. We hebben immers een kwadratensom gebruikt. Daardoor verduidelijkt de uitkomst van de variantie niet direct hoe het aantal uur dat online nieuws wordt gelezen is verspreid over de acht jongeren.

Wat we wel kunnen doen, is de variantie van deze groep jongeren vergelijken met de variantie van een groep ouderen. We vragen aan acht ouderen hoe vaak zij per week online het nieuws lezen, en berekenen eerst weer het gemiddelde.

$$\bar{x} = \frac{0+0+2+2+2+6+6+8}{8} = 3{,}25$$

Vervolgens berekenen we de variatie. Deze is 63,504 (zie tabel 3.3).

Tabel 3.3 Berekenen van de variatie (voor ouderen)

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
0	(0 – 3,25) = –3,25	10,563
0	(0 – 3,25) = –3,25	10,563
2	(2 – 3,25) = –1,25	1,563
2	(2 – 3,25) = –1,25	1,563
2	(2 – 3,25) = –1,25	1,563
6	(6 – 3,25) = 2,75	7,563
6	(6 – 3,25) = 2,75	7,563
8	(8 – 3,25) = 4,75	22,563
Σ	0	63,504

De variantie is hier:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} = \frac{63{,}504}{8-1} = 9{,}072$$

De variantie is voor de groep ouderen dus groter dan voor de groep jongeren. Bij ouderen is er dus meer spreiding in het aantal uur dat online nieuws wordt gelezen dan bij jongeren. Het gemiddeld aantal uur dat online nieuws gelezen wordt, verschilt echter niet veel van elkaar (3,63 voor de jongeren en 3,25 voor de ouderen).

Laten we nog een voorbeeld bekijken. Je zoekt een kamer in Amsterdam en bekijkt de prijzen van zes kamers.
kamer 1: € 175
kamer 2: € 180
kamer 3: € 190
kamer 4: € 240
kamer 5: € 350
kamer 6: € 550

Wat is nu de variantie van deze kamerprijzen? De eerste stap is het berekenen van het gemiddelde.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{175+180+190+240+350+550}{6} = 280{,}833$$

De gemiddelde kamerprijs van deze zes kamers in Amsterdam is dus € 280,833. Vervolgens neem je de kwadratensom van de verschillen van elke individuele kamerprijs ten opzichte van het gemiddelde (de variatie).

Tabel 3.4 Berekenen van de variatie (kamerprijzen Amsterdam)

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
175	(175 – 280,833) = –105,833	11200,624
180	(180 – 280,833) = –100,833	10167,294
190	(190 – 280,833) = –90,833	8250,634
240	(240 – 280,833) = –40,833	1667,334
350	(350 – 280,833) = 69,167	4784,074
550	(550 – 280,833) = 269,167	72450,874
Σ	0	108520,834

Nu kun je de formule voor variantie invullen.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} = \frac{108520{,}834}{6-1} = 21704{,}167$$

De variantie in kamerprijzen is in dit geval € 21.704,167. Dit is natuurlijk een raar getal: hoe moet je dit interpreteren in combinatie met de gemiddelde prijs van € 280,833?
De standaarddeviatie is een kengetal dat je wel gemakkelijk kunt interpreteren. Deze wordt uitgelegd in de volgende paragraaf.

3.3 Standaarddeviatie

Door de afstand tussen de individuele score en het gemiddelde te kwadrateren, krijgen we getallen die moeilijk te interpreteren zijn. Het is daarom nuttig om het kwadraat in de variantie op te heffen. Dit doe je door de wortel te trekken uit de waarde die we voor de variantie hebben berekend. Op die manier berekenen we de *standaarddeviatie*, ook wel standaardafwijking genoemd.
De formule voor de standaarddeviatie (aangeduid met de letter *s*) is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

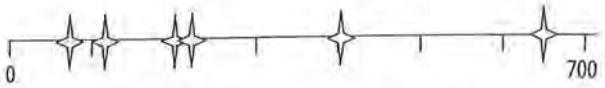
Formule voor de standaarddeviatie

In het eerste voorbeeld bij variantie (het aantal uren dat door jongeren online nieuws wordt gelezen) was de variantie 7,983. Om daar de standaarddeviatie van te berekenen, nemen we de wortel van dat getal: $\sqrt{7,983} = 2,825$. De standaardafwijking is dus 2,83. Dit getal is beter te interpreteren in combinatie met het gemiddelde dan de variantie. Jongeren lezen gemiddeld 3,63 uur per week online het nieuws, en wijken daar gemiddeld 2,83 uur van af.
Dit wordt helemaal duidelijk in het voorbeeld van de kamerprijzen. De variantie van de kamerprijzen was € 21704,167, een getal dat in geen enkele verhouding staat tot het gemiddelde van € 280,833.
De standaarddeviatie is veel beter te interpreteren: $\sqrt{21704,167} = 147,323$. De standaarddeviatie is de gemiddelde afwijking ten opzichte van het gemiddelde en die is hier € 147,32. De kamerprijzen in Amsterdam zijn gemiddeld € 280,833 en daar wordt gemiddeld € 147,32 van afgeweken.

Hoe lager de standaarddeviatie, hoe dichter de individuele scores zich rondom het gemiddelde concentreren. En andersom, hoe hoger de standaarddeviatie, hoe verder (hoe meer verspreid) de individuele scores van het gemiddelde af liggen.
Laten we eens twee steden met elkaar vergelijken qua kamerprijzen. Behalve naar kamers in Amsterdam kijk je ook naar kamers in Utrecht. In Utrecht treffen we de volgende kamerprijzen aan:
kamer 1: € 75
kamer 2: € 130
kamer 3: € 200
kamer 4: € 230
kamer 5: € 400
kamer 6: € 650

Het gemiddelde van deze kamerprijzen is hetzelfde als in Amsterdam, namelijk € 280,83. Enkel op basis van het gemiddelde zou je denken dat er geen verschil is tussen de prijzen van kamers in Amsterdam en Utrecht. Maar de standaarddeviatie van de kamerprijzen in Utrecht verschilt wel van de standaarddeviatie van de kamerprijzen in Amsterdam, deze is in Utrecht namelijk € 212,00. Hoewel het gemiddelde in beide steden dus gelijk is, is de spreiding in Utrecht groter dan de spreiding in Amsterdam. Dit is ook goed te zien in de figuren 3.4 en 3.5.

Figuur 3.4 Spreiding kamerprijzen Amsterdam



Figuur 3.5 Spreiding kamerprijzen Utrecht

De statistische resultaten van een univariate analyse vermeld je doorgaans in de tekst. Je rapporteert altijd het totale aantal onderzoekseenheden, de (meest geschikte) centrummaat en, zo mogelijk, de spreidingsmaat. Bijvoorbeeld: 'Aan het onderzoek deden meer meisjes (56%) dan jongens mee (44%). De 180 jongeren keken gemiddeld ongeveer 2,5 keer per week naar soaps (M = 2,58, SD = 2,26).'

Wil je meerdere centrummaten en spreidingsmaten tegelijk rapporteren, dan is het handig om een overzichtstabel te presenteren. Onderstaande tabel (tabel 3.5) is een voorbeeld van hoe zo'n overzichtstabel er in een wetenschappelijk artikel uit kan zien. In dit artikel is een experiment uitgevoerd waarin werd gekeken naar de voorkeur voor omslagen van kinderboeken, waar steeds een realistisch omslag werd vergeleken met een niet realistisch omslag en een niet complex omslag met een complex omslag. De waarden die in de tabel staan, geven de gemiddelde voorkeur weer voor het boekomslag. Uit de tabel is onder andere af te lezen dat boekomslagen met een realistische foto hoger worden gewaardeerd dan boekomslagen met een onrealistische foto.[4]

Tabel 3.5 Gemiddelde scores voor realisme en complexiteit per set omslagen, in artikel van Hartman et al. (2014)[5]

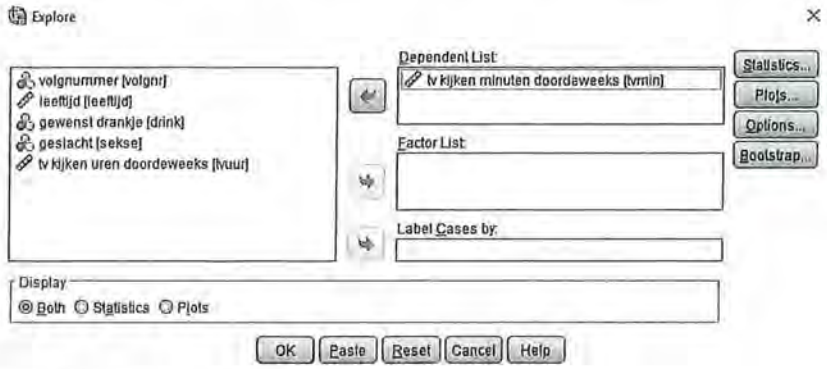
Set	Realisme				Complexiteit			
	Onrealisme		Realistisch		Niet complex		Complex	
	M	SD	M	SD	M	SD	M	SD
Met voeten	1.87*	1.05	4.23*	.95	2.45*	1.19	3.56*	1.12
Met detectives	2.32*	1.13	3.65*	1.10	2.25*	1.00	3.42*	1.10
Met springende persoon	2.14*	1.12	3.92	1.21	2.04*	.99	3.61	1.05

3.4 Centrum- en spreidingsmaten in SPSS

Net als de modus en de mediaan kan het gemiddelde in SPSS berekend worden door een frequentietabel uit te laten draaien. Hierbij kan er voor gekozen worden om de juiste spreidingsmaten aan te vinken. Omdat interval- en ratiovariabelen over het algemeen erg veel waarden hebben, is het echter niet altijd overzichtelijk om een frequentieverdeling te maken. Het is ook mogelijk om via het commando *Descriptives* de informatie op te vragen, of via het commando *Explore*. Wij raden deze laatste manier aan bij het beschrijven van numerieke variabelen (zie ook kader 3.1).

SPSS Centrum- en spreidingsmaten

Bij het beschrijven van numerieke variabelen (interval- of ratio meetniveau) wordt gebruikgemaakt van *Analyze → Descriptive Statistics → Explore*. In het venster dat verschijnt kun je de variabele(n) die je wilt beschrijven invoeren in de *Dependent List*. Eventueel kan onder *Plots* gekozen worden om een histogram te laten maken.



Figuur A: Explore-venster

Kader 3.1

Je hebt in een enquête gevraagd hoe vaak respondenten per week naar de televisie kijken, en je hebt dat gemeten in het aantal minuten dat ze dat doen (zie ook paragraaf 4.3 voor het samenstellen van variabelen). Wanneer je door middel van *Explore* deze variabele gaat beschrijven, krijg je eerst de volgende twee tabellen:

Tabel 3.6 Beschrijving van de variabele minuten tv-kijken via Explore (SPSS-output)

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
minutentv	1468	100,0%	0	0,0%	1468	100,0%

Descriptives

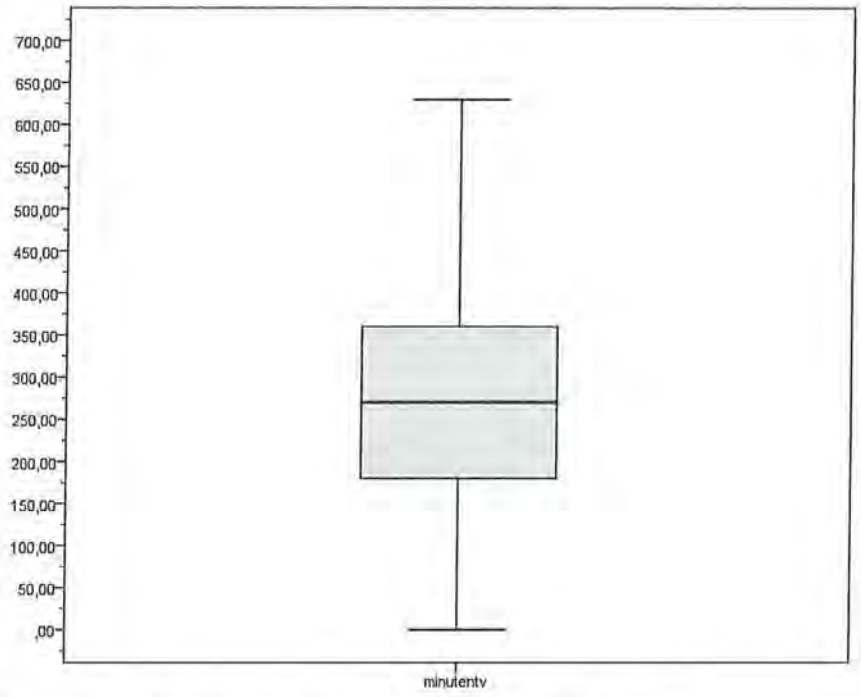
			Statistic	Std. Error
minutentv	Mean		284,1335	3,62596
	95% Confidence Interval for Mean	Lower Bound	277,0209	
		Upper Bound	291,2461	
	5% Trimmed Mean		279,6783	
	Median		270,0000	
	Variance		19300,652	
	Std. Deviation		138,92678	
	Minimum		,00	
	Maximum		630,00	
	Range		630,00	
	Interquartile Range		180,00	
	Skewness		,464	,064
	Kurtosis		-,434	,128

Je ziet hierin bijna alle informatie die je nodig hebt. In de tabel *Case Processing Summary* zien we dat 1468 respondenten deze vraag hebben beantwoord en dat er geen *missing values* zijn. In de bovenste rij van de tabel *Descriptives* staat het gemiddelde: gemiddeld kijken de onderzoekseenheden 284,13 minuten per week naar de televisie. Iets verder in de tabel zien we de standaarddeviatie: de gemiddelde afstand ten opzichte van het gemiddelde is 138,93. Dat wil dus zeggen dat er een grote mate van spreiding is, er zijn mensen die veel lager en mensen die veel hoger dan het gemiddelde scoren. De standaarddeviatie werd berekend door de wortel te trekken uit de variantie (die staat daarboven vermeld: 19300,65). De minimumwaarde die op deze variabele is gescoord is nul (er zijn mensen die niet televisiekijken) en de maximumwaarde is 630 (de maximale tijd dat per week televisie wordt gekeken is 10,5 uur). De mediaan is 270: 50% van de respondenten kijkt 270 minuten of minder per week naar de televisie, 50% van de respondenten kijkt 270 minuten of meer per week naar de televisie. Tot slot kun je de range uit de tabel aflezen (het verschil tussen de hoogste en laagste waarde), en de interkwartielafstand, die is hier 180.

Van de gepresenteerde spreidingsmaten geeft de standaarddeviatie in dit geval de meeste informatie, vooral als deze gekoppeld wordt aan het gemiddelde. De interkwartielafstand geeft pas nuttige informatie wanneer je bijvoorbeeld de

interkwartielafstand van vrouwen zou vergelijken met de interkwartielafstand van mannen, of een andere vergelijking zou maken.

Behalve een overzichtelijke beschrijvende tabel, krijg je bij het uitvoeren van deze analyse een boxplot in je output (zie figuur 3.6).



Figuur 3.6 Boxplot van aantal minuten tv-kijken (via Explore)

3.5 Standaardiseren (z-scores)

Z-scores zijn gestandaardiseerde scores van een variabele die we voor elke onderzoekseenheid apart kunnen uitrekenen. Door standaardisatie zijn de waarden van variabelen die een verschillende meeteenheid hebben met elkaar te vergelijken. De z-scores zijn gebaseerd op de standaarddeviatie en het gemiddelde van een variabele. Aangezien de z-scores gebaseerd zijn op het rekenkundig gemiddelde, kunnen ze alleen maar uitgerekend worden voor variabelen die op interval- of rationiveau zijn gemeten. De z-score geeft aan hoeveel maal de standaarddeviatie de waarde van de betreffende onderzoekseenheid afwijkt van het gemiddelde van een variabele. Een negatieve z-score betekent dat de waarde van de onderzoekseenheid voor die variabele kleiner is dan het gemiddelde van de groep, een positieve z-score betekent dat deze waarde groter is dan het gemiddelde van de groep. Zo betekent $z = 1$ dat de waarde van de onderzoekseenheid op de variabele één standaarddeviatie groter is dan het gemiddelde, en $z = -2$ betekent dat de waarde twee standaarddeviaties kleiner is dan het gemiddelde.

De formule voor z is:

$$z = \frac{x - \bar{x}}{s}$$

Formule voor de z-score

Om z uit te rekenen moet dus eerst het gemiddelde ($\bar{x}$) worden berekend, en de standaarddeviatie (s). Stel, je wilt twee gegevens van een aantal personen met elkaar vergelijken: hun intelligentie (IQ) en hun inkomen. Deze variabelen hebben een verschillende meeteenheid. Het inkomen is gemeten in euro's per week, en de intelligentie met de scores van een IQ-test. De waarden van de variabelen zijn in eerste instantie moeilijk met elkaar te vergelijken. Dit wordt eenvoudiger als je per onderzoekseenheid z-scores berekent.

Tabel 3.7 Datamatrix inkomen en IQ

	IQ	Inkomen
A	115	460
B	85	340
C	100	400
$\bar{x}$	100	400
s	15	60

De twee variabelen hebben een interval/ratio meetniveau, we kunnen dus het gemiddelde en de standaarddeviatie berekenen. Voor het IQ is het gemiddelde 100 en de standaarddeviatie 15, voor inkomen is het gemiddelde 400 en de standaarddeviatie 60 (zie tabel 3.8). Met deze informatie kun je per onderzoekseenheid de z-score uitrekenen. Voor persoon A geldt bijvoorbeeld voor IQ een z-score van

$$z = \frac{x - \bar{x}}{s} = \frac{115 - 100}{15} = 1$$

en voor inkomen een z-score van

$$z = \frac{x - \bar{x}}{s} = \frac{460 - 400}{60} = 1$$

Op deze manier kun je voor de drie onderzoekseenheden per variabele een z-score berekenen zoals weergegeven is in tabel 3.8.

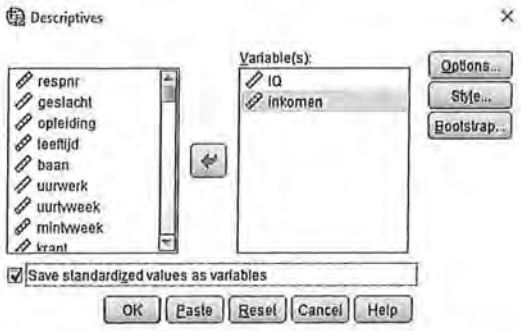
Tabel 3.8 Z-scores voor IQ en inkomen

	IQ	Inkomen	Z-IQ	Z-Inkomen
A	115	460	1	1
B	85	340	–1	–1
C	100	400	0	0
$\bar{x}$	100	400		
s	15	60		

Uit tabel 3.8 blijkt dat onderzoekseenheid A één standaarddeviatie boven het gemiddelde scoort voor zowel IQ als inkomen, en dat voor onderzoekseenheid C de waarden voor beide variabelen gelijk zijn aan de gemiddeldes. Door naar de z-scores te kijken wordt meteen duidelijk dat er een sterke samenhang is tussen IQ en inkomen. Dit is bij de werkelijke waarden van de variabelen minder direct te zien.

SPSS Berekenen van z-scores

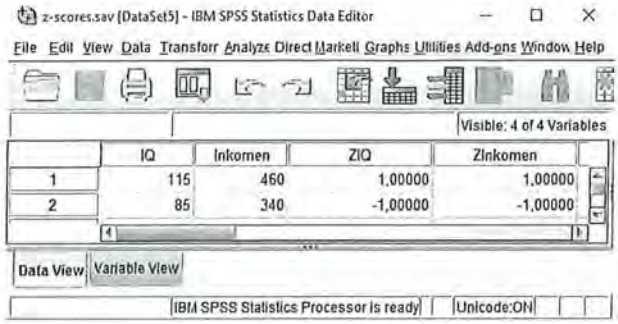
Wanneer je in SPSS z-scores berekent, worden deze aan de datamatrix toegevoegd. Om de z-scores te laten berekenen volg je de volgende procedure: *Analyze → Descriptive Statistics → Descriptives*. Hier voer je de variabelen in waar je z-scores van wilt hebben en vink je vervolgens het vakje *Save standardized values as variables* aan.



Figuur A Z-score via Descriptives

Wanneer je nu op PASTE klikt en de syntax runt (zie paragraaf 4.1), zijn de z-scores aan de datamatrix toegevoegd (zie figuur B).

SPSS geeft zelf een naam aan de nieuwe variabelen, in dit geval ZIQ voor de z-scores van IQ en ZInkomen voor de z-scores van inkomen.

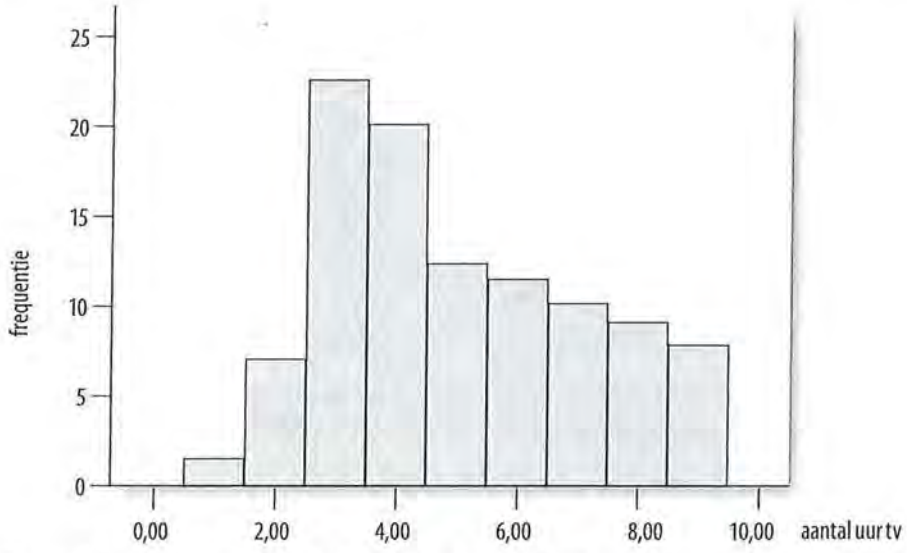


Figuur B Data View met z-scores

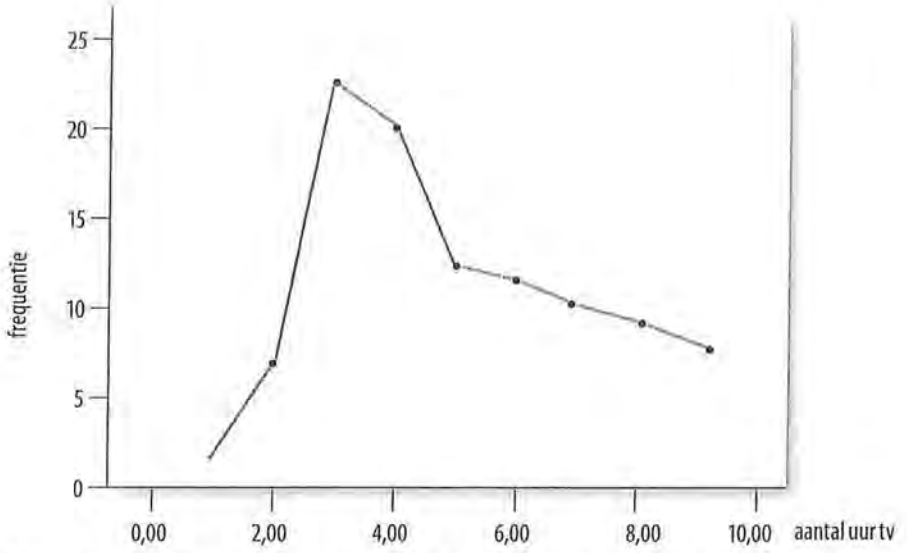
Kader 3.2

3.6 Normale en scheve verdelingen

In hoofdstuk 1 hebben we gezien dat je een frequentieverdeling grafisch kunt weergeven in een taartdiagram of in een staafdiagram (paragraaf 1.2.3). Deze figuren gebruik je beide voor nominale variabelen en/of wanneer het aantal waarden van een variabele beperkt is. Wanneer het meetniveau minimaal ordinaal is, is een histogram of frequentiepolygoon mogelijk. Een histogram is een grafische weergave van de frequentieverdeling van (in klassen) geordende data. Hieruit blijkt al dat de meetschaal dan minimaal op ordinaal niveau moet zijn. Een histogram toont in kolommen hoe vaak een waarde voorkomt (zie figuur 3.7).



Figuur 3.7 Histogram van aantal uur televisiekijken

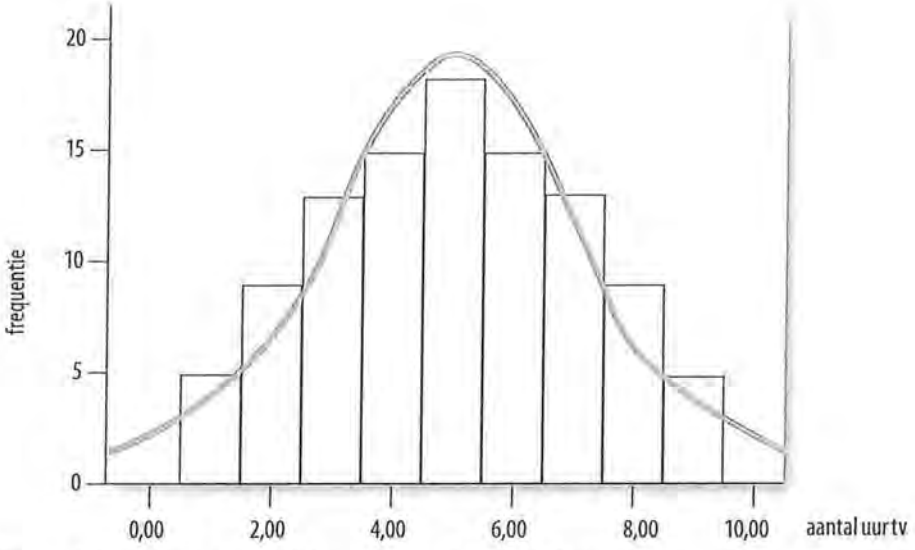


Figuur 3.8 Frequentiepolygoon van het aantal uur televisiekijken (behorende bij histogram figuur 3.7)

Zoals te zien is in figuur 3.7, wordt in een histogram de waarde van de variabele op de *x*-as in het midden van de kolombreedte (klassenmidden) aangegeven. De middens van de klassen boven in de kolommen zou je door middel van een lijn met elkaar kunnen verbinden. Op die manier ontstaat een frequentiepolygoon (zie figuur 3.8). In een histogram en een frequentiepolygoon is te zien hoe de frequentieverdeling eruitziet.

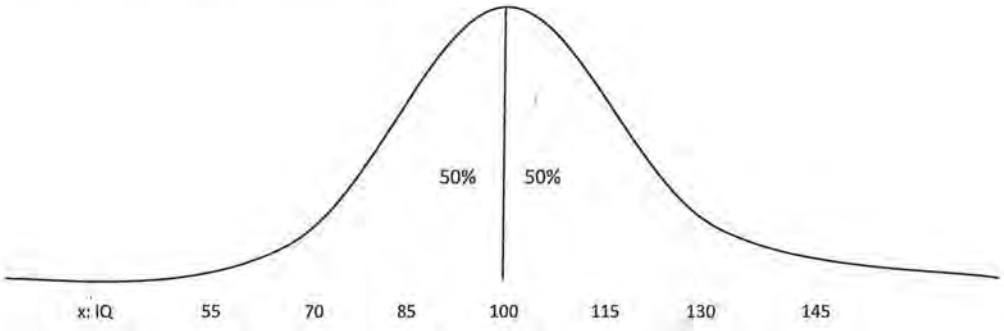
3.6.1 *Normale verdeling*

Behalve de centrummaten en de spreiding van een frequentieverdeling is ook de symmetrie dan wel scheefheid een kenmerk van verdelingen. Wanneer de verdeling symmetrisch is, spreken we van een *normale verdeling*. Wanneer we een vloeiende lijn tekenen, krijgen we een klokvormige figuur. In een normale verdeling is er één top waar het gemiddelde, de mediaan en de modus samenvallen. In figuur 3.9, waar een verdeling wordt gegeven van aantal uur televisiekijken, is dat goed te zien. Het gemiddelde, de mediaan en de modus zijn hier 5 (uur). De mensen die 'extreem' scoren, zitten in de staartjes van de verdeling.



Figuur 3.9 Normale verdeling van uur televisiekijken

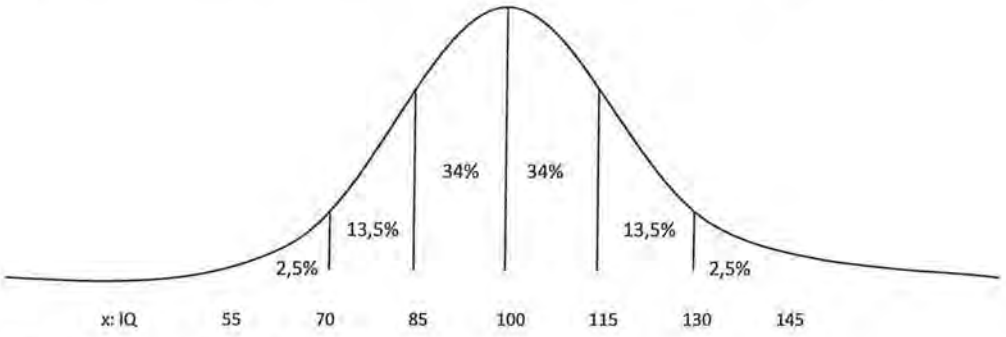
Stel dat je de frequentieverdeling van de IQ-scores van alle metingen onder Nederlandse volwassenen in een grafiek zou weergeven, dan zou deze de normale verdeling benaderen (zie figuur 3.10). Dat wil dus zeggen dat er weinig mensen zijn met een erg lage IQ-score, er weinig mensen zijn met een erg hoge IQ-score, en dat de meeste mensen in de verdeling niet erg ver boven of onder het gemiddelde van 100 zouden zitten. De meeste mensen (modus) hebben een IQ van 100, en omdat de verdeling symmetrisch is, is ook de mediaan 100 en kunnen we zeggen dat 50% van de volwassenen een IQ heeft van 100 of lager en 50% een IQ van 100 of hoger.



Figuur 3.10 Normale verdeling van IQ-scores

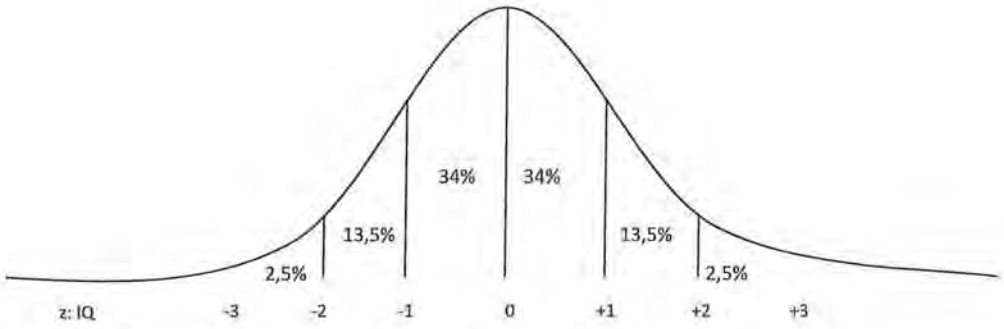
We kunnen aan de hand van een normale verdeling iets zeggen over de kans dat een bepaalde waarde voorkomt. Wanneer een variabele normaal verdeeld is, kun je de kans berekenen dat bepaalde waarden voorkomen.

Bij een normale verdeling kunnen we stellen dat 50% hoger en 50% lager scoort dan het gemiddelde. De kans dat een willekeurige volwassene een IQ heeft van 100 of hoger (dit is zowel de mediaan als het gemiddelde) is 50% (zie figuur 3.10). Het is ook mogelijk om de afzonderlijke helften weer verder op te delen in kansen (figuur 3.11). Bij een normale verdeling ligt ongeveer 34% van de scores altijd tussen het gemiddelde en één standaarddeviatie lager of hoger verwijderd van het gemiddelde. Metingen die een tot twee standaarddeviaties verwijderd zijn van het gemiddelde komen in ongeveer 13,5% van de gevallen voor. Ongeveer 2,5% van de scores is minimaal drie standaarddeviaties hoger of lager dan het gemiddelde. Dit wordt de *empirische regel* genoemd (figuur 3.11).



Figuur 3.11 De empirische regel

In figuur 3.10 en 3.11 is te zien dat IQ verdeeld is in stapjes van 15. Dat is omdat in dit voorbeeld het gemiddelde IQ 100 was, met een standaarddeviatie van 15. Op basis van de empirische regel (figuur 3.11) zouden we kunnen stellen dat ongeveer 34% van de volwassen mensen in deze steekproef een IQ heeft tussen de 100 en 115, en ongeveer 34% een IQ heeft tussen de 85 en 100. We kunnen ook zeggen dat 2,5% een IQ heeft van 130 of hoger. Oftewel: de kans dat iemand een IQ heeft van 130 of hoger is 2,5%. Dit is de *overschrijdingskans*.
Voor elke waarde van IQ kunnen we de overschrijdingskans bepalen. Hiervoor maken we gebruik van de z-scores. We standaardiseren de waarden van het IQ in z-scores die aangeven hoeveel maal de standaarddeviatie van die waarde afwijkt van het gemiddelde. We zetten de normale verdeling van de waarden van het IQ om in een verdeling van z-scores. Dit is de *standaardnormale verdeling*.
In een standaardnormale verdeling (figuur 3.12) zijn de waarden van de oorspronkelijke variabele gestandaardiseerd door middel van z-scores. Het gemiddelde van deze z-scores is altijd nul, en de standaarddeviatie is altijd 1. Dit geldt voor elke variabele die we standaardiseren. In het voorbeeld van IQ was de standaarddeviatie 15. Zoals we konden zeggen dat ongeveer 68% van de volwassenen een IQ heeft tussen de 85 en 115, kunnen we ook zeggen: 68% van de volwassenen wijkt maximaal één standaarddeviatie, één z-score af van het gemiddelde. Voor elke waarde van het IQ kunnen we nu bepalen hoe groot de kans is dat een willekeurige volwassene een IQ heeft dat hoger is dan die waarde. Daarvoor kun je een tabel gebruiken met de overschrijdingskansen voor mogelijke z-scores (zie tabel 3.9).



Figuur 3.12 Standaardnormale verdeling

Hieronder vind je een gedeelte van deze tabel. De gehele tabel is te vinden na het formuleblad in de bijlage.

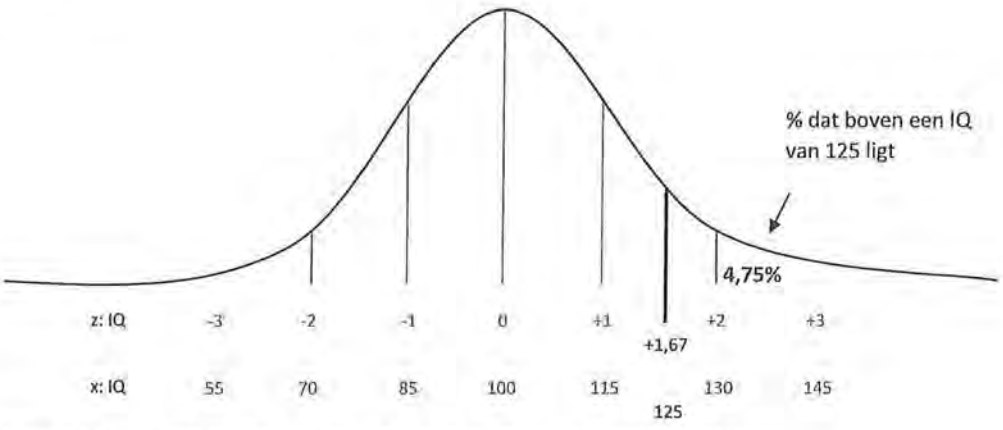
Tabel 3.9 Tabel met rechter overschrijdingskansen in de standaardnormale verdeling

z	$P_R(z)$	z	$P_R(z)$	z	$P_R(z)$	z	$P_R(z)$
0,00	0,5000	0,38	0,3520	0,76	0,2236	1,14	0,1271
0,01	0,4960	0,39	0,3483	0,77	0,2206	1,15	0,1251
0,02	0,4920	0,40	0,3446	0,78	0,2177	1,16	0,1230
0,03	0,4880	0,41	0,3409	0,79	0,2148	1,17	0,1210
0,04	0,4840	0,42	0,3372	0,80	0,2119	1,18	0,1190
0,05	0,4801	0,43	0,3336	0,81	0,2090	1,19	0,1170
0,06	0,4761	0,44	0,3300	0,82	0,2061	1,20	0,1151
0,07	0,4721	0,45	0,3264	0,83	0,2033	1,21	0,1131
0,08	0,4681	0,46	0,3228	0,84	0,2005	1,22	0,1112
0,09	0,4641	0,47	0,3192	0,85	0,1977	1,23	0,1093
0,10	0,4602	0,48	0,3156	0,86	0,1949	1,24	0,1075
0,11	0,4562	0,49	0,3121	0,87	0,1922	1,25	0,1056
0,12	0,4522	0,50	0,3085	0,88	0,1894	1,26	0,1038
0,13	0,4483	0,51	0,3050	0,89	0,1867	1,27	0,1020
0,14	0,4443	0,52	0,3015	0,90	0,1841	1,28	0,1003
0,15	0,4404	0,53	0,2981	0,91	0,1814	1,29	0,0985
0,16	0,4364	0,54	0,2946	0,92	0,1788	1,30	0,0968
0,17	0,4325	0,55	0,2912	0,93	0,1762	1,31	0,0951
0,18	0,4286	0,56	0,2877	0,94	0,1736	1,32	0,0934
0,19	0,4247	0,57	0,2843	0,95	0,1711	1,33	0,0918
0,20	0,4207	0,58	0,2810	0,96	0,1685	1,34	0,0901
0,21	0,4168	0,59	0,2776	0,97	0,1660	1,35	0,0885
0,22	0,4129	0,60	0,2743	0,98	0,1635	1,36	0,0869
0,23	0,4090	0,61	0,2709	0,99	0,1611	1,37	0,0853
0,24	0,4052	0,62	0,2676	1,00	0,1587	1,38	0,0838
0,25	0,4013	0,63	0,2643	1,01	0,1562	1,39	0,0823
0,26	0,3974	0,64	0,2611	1,02	0,1539	1,40	0,0808
0,27	0,3936	0,65	0,2578	1,03	0,1515	1,41	0,0793
0,28	0,3897	0,66	0,2546	1,04	0,1492	1,42	0,0778
0,29	0,3859	0,67	0,2514	1,05	0,1469	1,43	0,0764
0,30	0,3821	0,68	0,2483	1,06	0,1446	1,44	0,0749
0,31	0,3783	0,69	0,2451	1,07	0,1423	1,45	0,0735
0,32	0,3745	0,70	0,2420	1,08	0,1401	1,46	0,0721
0,33	0,3707	0,71	0,2389	1,09	0,1379	1,47	0,0708
0,34	0,3669	0,72	0,2358	1,10	0,1357	1,48	0,0694
0,35	0,3632	0,73	0,2327	1,11	0,1335	1,49	0,0681
0,36	0,3594	0,74	0,2296	1,12	0,1314	1,50	0,0668
0,37	0,3557	0,75	0,2266	1,13	0,1292	1,51	0,0655

Boven de tabel zie je staan: 'Rechter overschrijdingskansen in de standaardnormale verdeling'. Dat betekent dat je uit deze tabel alleen maar de rechterkant van de tabel kunt berekenen. Omdat de standaardnormale verdeling geheel symmetrisch is, geldt de rechterkant van de verdeling echter ook voor de linkerkant van de verdeling, maar dan zijn de z-scores negatief. Kijk je bijvoorbeeld bij een z-score van 0,00, dan is de kans dat iemand hoger of lager dan die waarde scoort, 50%. Kijk je bij een z-score van 1, dan zie je dat de rechter overschrijdingskans 0,1587 is. Met andere woorden: de kans dat iemand minimaal één standaarddeviatie hoger of lager scoort dan het gemiddelde, is 15,87%. De kans dat iemand minimaal twee standaarddeviaties hoger of lager scoort dan het gemiddelde, is maar 2,28%.
Stel, je hebt je eigen IQ laten berekenen, dat is 125, en je wilt weten hoeveel procent van de volwassenen een hoger IQ heeft dan jij. Je moet dan eerst je IQ-score omzetten in een z-score

$$z = \frac{(x - \bar{x})}{s} = \frac{125 - 100}{15} = \frac{25}{15} = 1{,}667$$

Jij scoort dus 1,67 standaarddeviaties boven het gemiddelde. Deze waarde opzoeken in de tabel leert ons dat 4,75% van de volwassenen deze waarde of hoger scoort. Dat betekent dus ook dat 100 – 4,75 = 95,25% een lager IQ dan jij heeft.



Figuur 3.13 Berekenen van kansen via de standaardnormale verdeling

Hoe meer standaarddeviaties een onderzoekseenheid afwijkt van het gemiddelde, hoe minder groot de kans is dat die waarde vaak voorkomt. Het is waarschijnlijker dat je iemand treft met een IQ van 115 of hoger (namelijk 15,87%) dan dat je iemand treft met een IQ van 130 of hoger (namelijk 2,28%). We spreken van een *extreme waarde* wanneer een onderzoekseenheid vijf standaarddeviaties ($z < -5$ of $z > 5$) onder of boven het gemiddelde scoort, en van een *uitbijter* (*outlier*) wanneer een onderzoekseenheid drie standaarddeviaties onder of boven het gemiddelde scoort ($z < -3$ of $z > 3$). Een extreme waarde of uitbijter kan ervoor zorgen dat de verdeling scheef wordt.

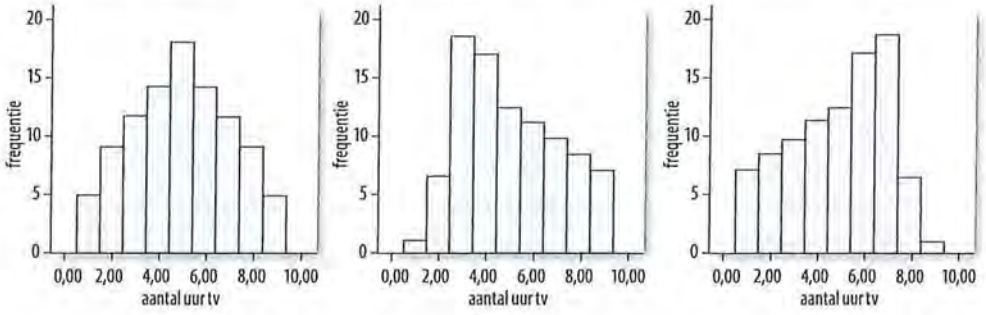
3.6.2 *Scheve verdelingen*

Een *scheve verdeling* is een verdeling die niet symmetrisch is. Scheve verdelingen kunnen ofwel scheef naar rechts, ofwel scheef naar links zijn. Scheefheid (in het Engels *skewness*) ontstaat wanneer ten opzichte van de modus aan één kant van de verdeling meer afwijkende waarden voorkomen dan aan de andere kant. Er zijn dan aan één kant extreme waarden, ofwel waarden die ver afliggen van de modus. Liggen de extreme waarden aan de linkerkant, dan is de verdeling scheef naar links, en liggen de extreme waarden rechts, dan is de verdeling scheef naar rechts (zie figuur 3.14).

Uur tv	Frequentie
1	5
2	9
3	12
4	15
5	18
6	15
7	15
8	9
9	5
N	100
Modus	5
mediaan	5
gemiddelde	5
skewness	0

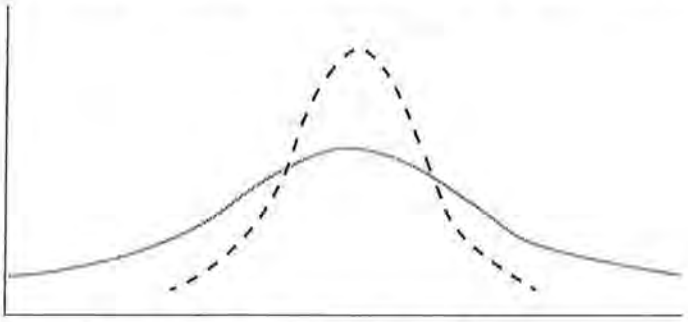
Uur tv	Frequentie
1	1
2	7
3	22
4	20
5	12
6	11
7	10
8	9
9	8
N	100
Modus	3
mediaan	4,5
gemiddelde	5
skewness	0,412

Uur tv	Frequentie
1	8
2	9
3	10
4	11
5	12
6	20
7	22
8	7
9	1
N	100
Modus	7
mediaan	5,5
gemiddelde	5
skewness	−0,412



Figuur 3.14 Histogrammen met bijbehorende frequentieverdeling van een normaal verdeelde variabele, een verdeling die scheef is naar rechts en een verdeling die scheef is naar links

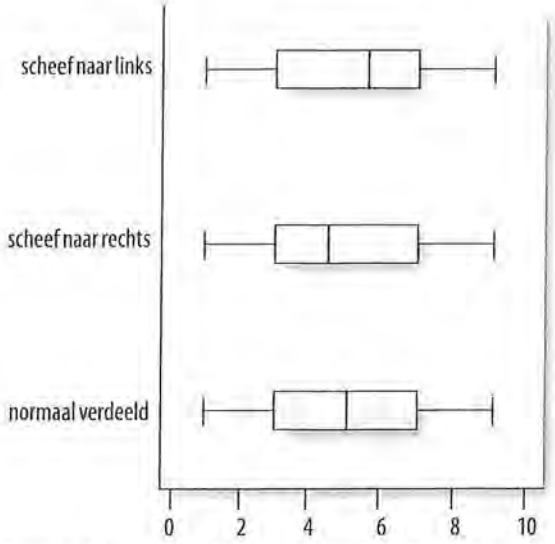
Behalve over de scheefheid van de verdeling kun je ook iets zeggen over de gewelfdheid van de verdeling, dat wil zeggen hoe plat of spits de verdeling is. Deze gewelfdheid noem je *kurtosis*. Een hoge kurtosis wijst op een verdeling met een sterke piek, een lage kurtosis wijst op een platte verdeling (zie figuur 3.15).



Figuur 3.15 Een platte en een spitse verdeling

Hoe spitser de verdeling (in figuur 3.16 de gestippelde lijn), hoe minder spreiding er is, en hoe minder extreme waarden zich in de verdeling bevinden.
De mate van scheefheid en kurtosis kunnen we door SPSS laten berekenen. In ons voorbeeld van aantal uur televisiekijken is sprake van een normale verdeling. Modus, mediaan en gemiddelde zijn 5. De standaarddeviatie is 2,17. De *skewness* is nul, dat wil zeggen dat de verdeling geheel symmetrisch is. Als de *skewness* een positieve waarde heeft, is de verdeling scheef naar rechts en als de *skewness* een negatieve waarde heeft, is de verdeling scheef naar links (figuur 3.14[6]). Over het algemeen hanteren we een marge van 1 (zowel naar links als naar rechts) om te bepalen of een verdeling te scheef is of niet. Wanneer een variabele te scheef is verdeeld, gaat de empirische regel niet meer op en is het niet meer zomaar mogelijk om kansen te berekenen aan de hand van een standaardnormaalverdeling.

In een histogram en een boxplot is het verschil tussen normale en scheve verdelingen snel te herkennen. In de onderste boxplot in figuur 3.16 zien we een normaal verdeelde variabele voor aantal uur televisiekijken, in de twee boxplots daarboven zien we scheve verdelingen voor de variabele aantal uur televisiekijken. De boxplot bovenaan is iets scheef naar links, de boxplot daaronder is iets scheef naar rechts. In alle drie de gevallen zijn honderd respondenten ondervraagd over hun kijkgedrag (uren televisiekijken). De laagste en hoogste waarden zijn in alle drie de boxplots gelijk. In de scheve verdeling naar links zitten meer extreme waarden in de linkerkant van de verdeling, en zit de mediaan meer rechts van het midden. Andersom zitten in de scheve verdeling naar rechts meer extreme waarden in de rechterkant van de verdeling, en zit de mediaan meer links van het midden. Bij de normale verdeling is er een symmetrische verdeling van waarden rond de mediaan.



Figuur 3.16 Boxplots van normale en scheve verdelingen

3.7 Samenvatting

Spreiding is de mate waarin de waarden van een variabele variëren. Bij de variatie, de variantie en de standaarddeviatie bereken je dit op basis van de afstand ten opzichte van het gemiddelde. Daarom gebruik je deze spreidingsmaten enkel op interval- en rationiveau. Ook voor de berekening van de interkwartielafstand is minimaal een meting op intervalniveau nodig.
De variatie is de kwadratensom van de afstanden van alle onderzoekseenheden tot het gemiddelde. Bij de variantie deel je deze kwadratensom door het aantal onderzoekseenheden minus 1. Omdat een kwadratensom moeilijk te interpreteren is (de waarden die verkregen worden, staan niet in verhouding tot de oorspronkelijke waarden), wordt voor de berekening van de standaarddeviatie de wortel getrokken uit de berekende variantie: $s = \sqrt{s^2}$

Om variabelen met verschillende meeteenheden met elkaar te kunnen vergelijken, worden de waarnemingen bij de onderzoekseenheden op deze variabelen gestandaardiseerd. Dit gebeurt door middel van een z-score. Deze score geeft aan hoeveel standaarddeviaties de waarneming van het gemiddelde afligt. Met een z-score kan in de tabel voor standaardnormaalverdelingen een overschrijdingskans worden opgezocht die aangeeft hoeveel kans er is dat een waarde boven (of onder) die z-score wordt gevonden. Een voorwaarde voor deze manier van kansberekening is dat de variabele normaal verdeeld is, en niet te scheef. Wanneer er te veel extreme waarden zijn, gaat de empirische regel niet meer op.

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

Noten

1 Ter herinnering: in dit boek zullen wij altijd rekenen met drie decimalen achter de komma. Bij het interpreteren van de waarden (meestal in de conclusie) ronden we (pas) af naar twee decimalen.
2 Door afrondingsverschillen is dit soms iets meer of iets minder dan exact nul.
3 Er wordt door $n - 1$ gedeeld en niet door n omdat je meestal wilt dat de variantie in een steekproef een schatting geeft van de variantie in de populatie. De $n - 1$ geeft een correctie die maakt dat de variantie als schatter kan dienen voor de variantie in de populatie. Als je de variantie niet als schatter voor een populatiewaarde gebruikt, dan kun je door n delen (σ^2). Omdat wij onze berekeningen met SPSS willen controleren en in SPSS bij de berekeningen door $n - 1$ wordt gedeeld, gebruiken we s^2.
4 Hartman, L., Okken, V. & Rompay, T. van (2014). Evaluating books by their covers; de invloed van realisme en complexiteit in fotografiegebruik op de waardering van tweens. *Tijdschrift voor Communicatiewetenschap, 42*(2), pp. 221-243.
5 In dit boek zullen we niet stilstaan bij de betekenis van de asterisken bij de cijfers in de tabel. Voor de beschrijvende statistiek volstaat het overzichtelijk weergeven van de gemiddelden en standaarddeviaties.
6 Overigens valt het met de scheefheid van de verdelingen in de voorbeelden in figuur 3.15 en figuur 3.17 nog wel mee. Frequentieverdelingen zijn soms veel schever verdeeld.

Bewerken van je data

4

Als je je data hebt ingevoerd in SPSS, is het belangrijk dat je de datamatrix controleert op fouten. Dat kun je bijvoorbeeld doen door van de variabelen die zich daarvoor lenen (beperkt aantal waarden) frequentietabellen te draaien. Als er in die frequentietabellen onmogelijke waarden voorkomen, ga je terug naar je datamatrix. Probeer erachter te komen hoe de fout is ontstaan, want een fout bij de ene variabele kan een teken zijn dat er bij een van de cases (onderzoekseenheden) ook bij andere variabelen iets mis is gegaan.
Nadat je je data hebt gecontroleerd, kun je niet altijd meteen aan de slag met je analyses. Sommige data zullen nog bewerkt moeten worden voordat ze geschikt zijn om analyses mee uit te voeren. In dit hoofdstuk staan we stil bij drie belangrijke manieren om je data te bewerken, namelijk het aangeven van missende waarden (*missing values*), het maken van een nieuwe variabele op basis van bestaande variabelen (*Compute*), en het aanpassen van de waarden binnen een bestaande variabele (*Recode*). In paragraaf 4.5 staan we stil bij de mogelijkheid om bepaalde subgroepen in je analyses te selecteren en/of juist uit te sluiten door middel van *Select Cases*. Voordat we dat gaan doen, kijken we naar het maken van een syntax in SPSS.

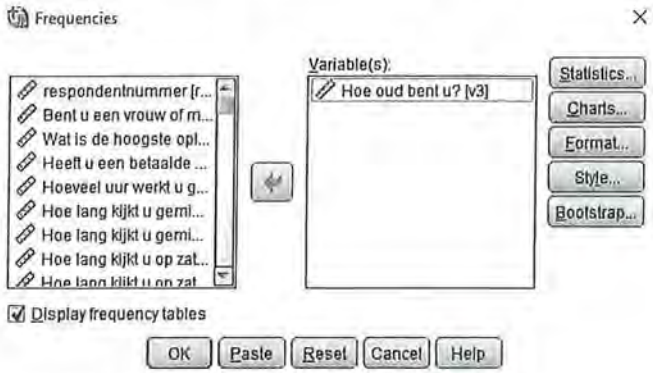
4.1 Syntax

In kader 2.1 (Centrummaten in SPSS) schreven we al dat bij het uitvoeren van commando's in SPSS beter op de PASTE-knop dan op de OK-knop geklikt kan worden. Dat is omdat er via de PASTE-knop een *syntax* in SPSS gemaakt kan worden. In deze paragraaf leggen we uit wat een syntax is en hoe je deze kunt gebruiken in SPSS. Omdat dit commando centraal staat in deze paragraaf, wordt het niet in een apart kader behandeld.
Een syntax is de besturingstaal van SPSS. We zouden ervoor kunnen kiezen om in plaats van op de knopjes te drukken (zoals *Analyze → Frequencies*), deze 'opdracht' in te typen in de syntax.

Onderstaande syntax laat bijvoorbeeld zien dat we een frequentietabel (*frequencies*) willen uitdraaien van variabele v3, en dat we daar de standaarddeviatie (*stddev*) en het gemiddelde (*mean*) van willen laten berekenen.

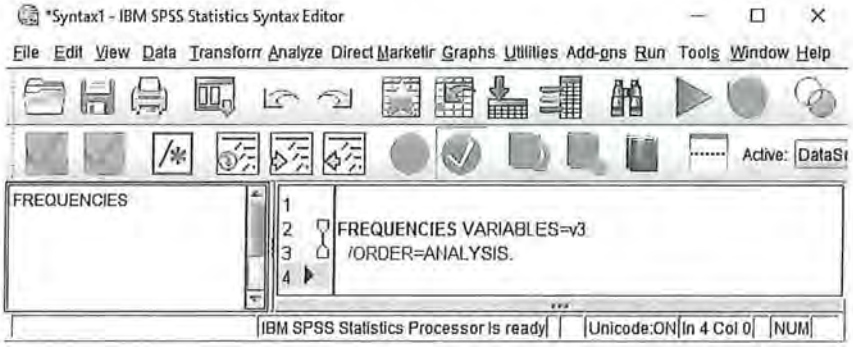
```
FREQUENCIES
VARIABLES=v3
/STATISTICS=STDDEV MEAN
/ORDER= ANALYSIS.
```

Het is een handige manier om overzicht te houden van welke bewerkingen en analyses je allemaal hebt uitgevoerd. Syntaxen kun je apart opslaan en elke keer over je databestand draaien (we noemen dat dan *runnen*) om de eerdere bewerkingen opnieuw te laten uitvoeren.
Een syntax maak je door elke analyse of bewerking in SPSS af te sluiten door op PASTE te drukken (in plaats van op OK).



Figuur 4.1 Het maken van een syntax in SPSS

Wanneer je dat doet, wordt de syntax van die bewerking in een nieuw venster gezet:

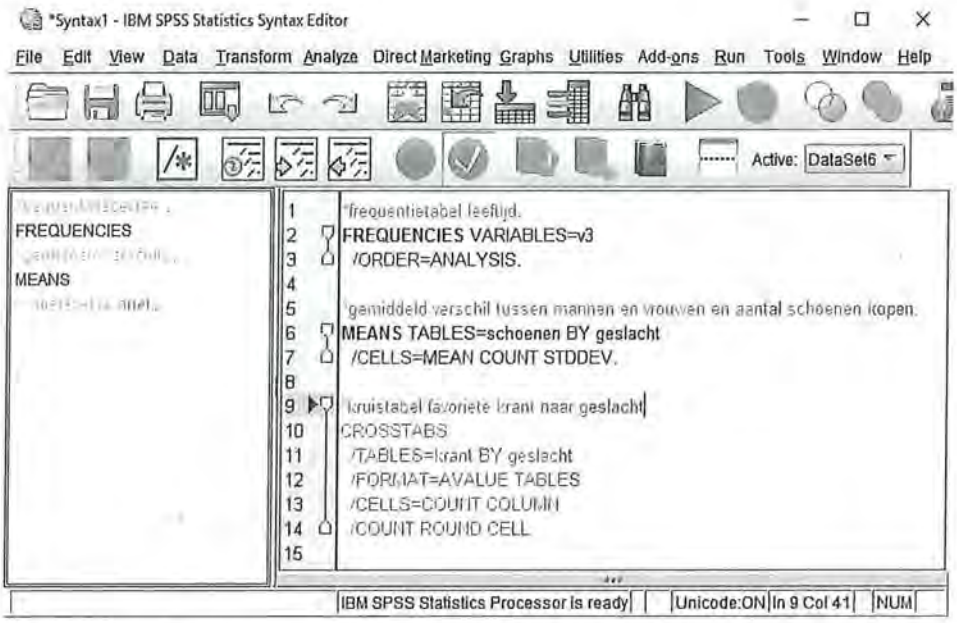


Figuur 4.2 Syntax in syntaxscherm

Het is belangrijk om te weten dat op dat moment nog niet aan SPSS het commando is gegeven om deze opdracht daadwerkelijk uit te voeren! Om dat te laten gebeuren moet je eerst op *Run* klikken, of selecteer je het stukje syntax en klik je op de groene pijl. Alleen dan gaat SPSS ook daadwerkelijk de analyse (in dit geval een frequentieverdeling) uitvoeren.
Je kunt ook tekst toevoegen in je syntax, zodat je later precies kunt zien waar je op dat moment mee bezig was en waarvoor de syntax diende. Het is daarbij belangrijk dat je aangeeft dat het om een tekst gaat die je zelf hebt toegevoegd, en dat die tekst niet onderdeel is van de analyse. Aangezien SPSS zal proberen om alles wat er getypt is te vertalen naar commando's die het kent, moet je je commentaar op een bijzondere manier toevoegen. Dit kan op drie manieren:

1. Begin de regel met het teken *. Alles wat er tussen * en de eerstvolgende punt (.) staat, vat SPSS op als commentaar en niet als een commando. Ook hier geldt dat je geen punten binnen het commentaar moet zetten, tenzij je de volgende regel weer begint met een *, en de regel weer afsluit met een punt.
2. Zet je commentaar tussen /* en */. Alles wat tussen deze combinatie van symbolen staat, wordt door SPSS opgevat als commentaar.
3. Begin de regel met het woord COMMENT. Alles wat tussen COMMENT en de eerstvolgende punt (.) staat, vat SPSS op als commentaar in plaats van een commando. NB: Binnen het commentaar mogen dus geen punten voorkomen, anders denkt SPSS dat daar het commentaar al stopt en probeert SPSS de rest van je commentaar als een commando uit te voeren.

Wanneer je op de juiste manier een tekstregel hebt ingevoerd, zal deze zin lichtgrijs worden, en zullen de commando's van de syntax 'helder' en in kleur blijven. Als dat niet het geval is, is er ergens iets misgegaan met het toevoegen van je commentaar. In figuur 4.3 zie je dat de eerste twee toevoegingen van tekst wel goed zijn, maar de onderste niet:



Figuur 4.3 Geschreven tekst in een syntax

Je ziet dat het werken met een afsluitende punt erg belangrijk is. Wanneer je je commentaar niet met een punt afsluit, zal SPSS alles wat erop volgt tot de eerstvolgende punt opvatten als commentaar dat overgeslagen kan worden.

In SPSS kun je op verschillende manieren je data bewerken. Je kunt een waarde bij een variabele uitsluiten bij je analyses. Dat kan bijvoorbeeld het geval zijn als respondenten geen antwoord op de vraag hebben gegeven, of als je een nieuwe variabele wilt maken waarin je een aantal variabelen bij elkaar optelt. Dat doe je bijvoorbeeld als je meer variabelen hebt die (verschillende) aspecten

van eenzelfde verschijnsel meten. De volgende paragrafen gaan over de verschillende manieren van data bewerken en welke consequenties dat kan hebben voor het meetniveau en de analyses die je uitvoert.

4.2 Missing values

Het kan zijn dat een respondent (of een onderzoeker) bij het invullen van een vragenlijst of codeboek een typefout maakt of een antwoord geeft dat je als onderzoeker niet mee wilt nemen in je onderzoek, zoals bijvoorbeeld de optie 'wil niet zeggen' of 'niet van toepassing'. Soms is het niet zo erg als een aantal mensen een typefout maakt, omdat het voor je analyse weinig verschil maakt. Het kan je analyses echter ook behoorlijk verstoren. Bedenk maar eens wat er gebeurt met het berekenen van de gemiddelde leeftijd wanneer een respondent per ongeluk de waarde '404' heeft ingevuld in plaats van '40' of '44'. In dit voorbeeld weet je zeker dat iemand een foutje heeft gemaakt, want een leeftijd van 404 is natuurlijk niet mogelijk. Ook cijfercodes voor de opties 'weet ik niet', 'geen opgave' of 'niet van toepassing' kunnen invloed hebben op de resultaten van je analyses als je deze waarden betrekt bij je berekeningen.
We hebben gezien dat het nominale meetniveau zich kenmerkt door enkel een classificatie van waarden en het ordinale meetniveau daarnaast ook een rangorde heeft. Stel je voor dat je in je onderzoek een variabele hebt opgenomen waarin je vraagt hoeveel interesse iemand heeft in de politiek. Je hebt daarvoor vier antwoordcategorieën onderscheiden, namelijk

1. Geen interesse
2. Matige interesse
3. Veel interesse
9. Weet ik niet

Het meetniveau van deze variabele is nu nominaal. Hoewel in de eerste drie antwoordcategorieën een rangorde zit, maakt de antwoordcategorie '9' dat het geen ordinale variabele is. Je kunt nu immers niet meer zeggen: hoe hoger iemand op deze variabele scoort, hoe meer interesse diegene in de politiek heeft. Voor analyses met deze variabele is de antwoordcategorie '9' niet van belang voor het onderzoek. Als je iets wilt zeggen over de variabele 'interesse in politiek', is het aan te raden om respondenten die de categorie '9' hebben aangekruist niet op te nemen in je onderzoek. Je kunt deze waarde 'missend' maken. Deze wordt dan niet meegenomen in je berekeningen voor deze variabele. Alleen die respondenten worden meegenomen die 1, 2, of 3 scoren, waarbij ze meer interesse in politiek hebben naarmate ze hoger scoren. Wanneer je dat doet, is het meetniveau niet langer nominaal maar ordinaal.
Hieronder zie je twee frequentieverdelingen van de variabele 'interesse in politiek'. In tabel 4.1 zijn er geen waarden *missing* gemaakt; dit is ook te zien in het bovenste tabelletje *Statistics*. Zoals al besproken in paragraaf 1.2.2, is er daardoor geen verschil tussen *Percent* en *Valid Percent*: voor alle onderzoekseenheden

wordt hier een frequentieverdeling gemaakt. Het meetniveau is in dit geval nominaal, er is geen sprake van rangordening. De modus is hier 1: de meeste mensen hebben geen interesse in politiek.

Tabel 4.1 Frequentietabel van variabele zonder missing values (SPSS-output)

Statistics

int_politiek interesse in politiek

N	Valid	2202
	Missing	0

int_politiek interesse in politiek

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 geen interesse	920	41,8	41,8	41,8
	2 matige interesse	685	31,1	31,1	72,9
	3 veel interesse	236	10,7	10,7	83,6
	9 weet ik niet	361	16,4	16,4	100,0
	Total	2202	100,0	100,0	

Tabel 4.2 Frequentietabel van variabele met missing values (SPSS-output)

Statistics

int_politiek interesse in politiek

N	Valid	1841
	Missing	361

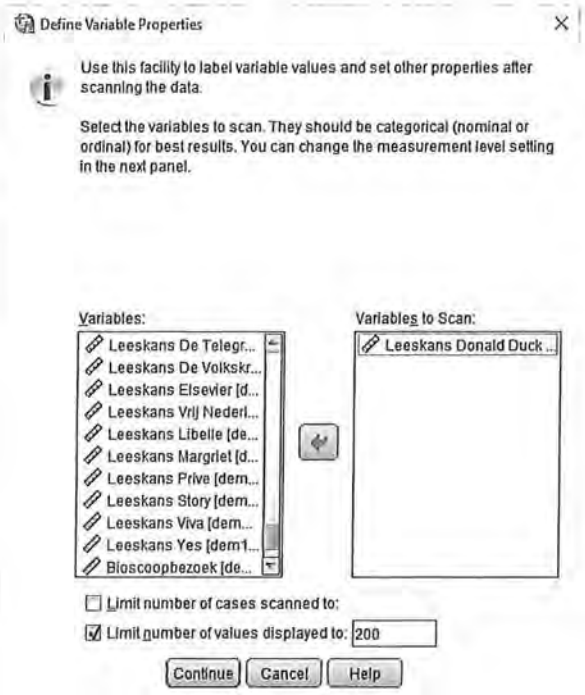
int_politiek interesse in politiek

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 geen interesse	920	41,8	50,0	50,0
	2 matige interesse	685	31,1	37,2	87,2
	3 veel interesse	236	10,7	12,8	100,0
	Total	1841	83,6	100,0	
Missing	9 weet ik niet	361	16,4		
Total		2202	100,0		

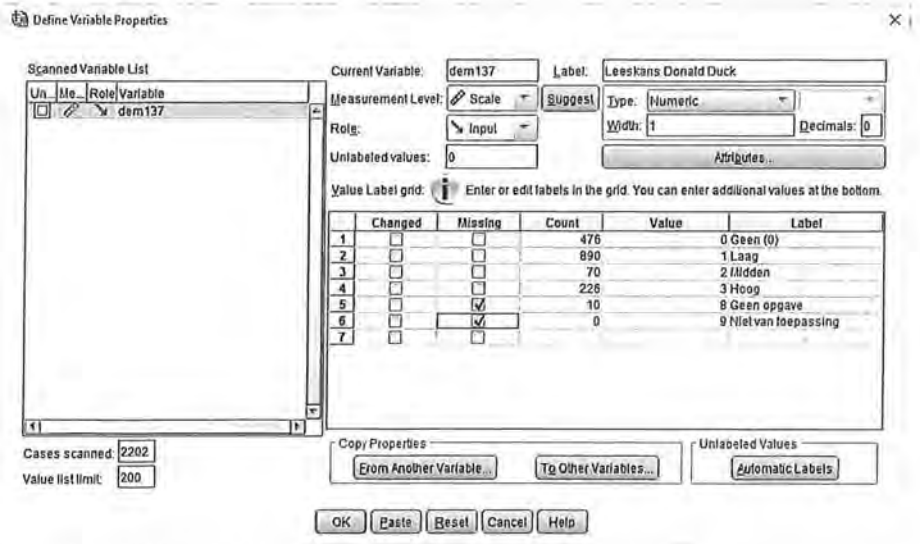
In tabel 4.2 hebben we de waarde 9 wél missing gemaakt (Zie kader 4.1 voor de handeling in SPSS). In deze tabel is te zien dat 361 respondenten de vraag hebben beantwoord met 'weet ik niet', en dat deze mensen niet in de analyse van deze variabele zijn meegenomen. Er is nu ook een verschil tussen de kolom met *Percent* en de kolom met *Valid Percent*. Van de 1841 respondenten die hebben aangegeven hoeveel interesse ze in politiek hebben, heeft 50,0% geen interesse. Het meetniveau van deze variabele is nu ordinaal, wat betekent dat niet alleen de modus, maar ook de mediaan berekend mag worden. Ook de mediaan is 1: meer dan de helft van de onderzoekseenheden scoort hoger dan 'geen interesse'.

SPSS Missing maken van waarden

Je kunt op twee manieren in SPSS aangeven welke waarden je missing wilt maken. De eerste manier is via het tabblad *Variable View* in je databestand, en vervolgens in de kolom *missing* aangeven welke waarden je voor deze variabele missend wilt maken. Het nadeel van deze manier is dat je er geen syntax van kunt maken. Daarom raden wij aan om in SPSS via *Data* naar *Define Variable Properties* te gaan (figuur A). Met deze functie kun je niet alleen waarden missing maken, maar ook valuelabels aanbrengen en variabelenamen aanbrengen (figuur B), en daar vervolgens een syntax van maken.



Figuur A Define Variable Properties-venster



Figuur B Missing maken van waarden in Define Variable Properties-venster

Je ziet in het scherm van SPSS dat je onder Label zowel het label van de variabele hier kunt veranderen, als de valuelabels. Om een waarde missing te maken klik je op het vakje zoals in bovenstaand voorbeeld. Door vervolgens op Paste te drukken, wordt een syntax gemaakt die gerund kan worden (zie paragraaf 4.1).

Kader 4.1

4.3 Compute

Soms wil je in SPSS een nieuwe variabele laten uitrekenen aan de hand van een of meerdere bestaande variabelen. Je hebt bijvoorbeeld naar iemands geboortejaar gevraagd maar wilt werken met de leeftijd van de respondent. Of je hebt gevraagd hoeveel uur iemand naar de publieke omroep kijkt maar wilt dat omzetten in het aantal minuten. Of je wilt een nieuwe schaalvariabele maken waarin verschillende variabelen bij elkaar worden opgeteld. Al deze bewerkingen worden in SPSS met *Compute* (letterlijk: berekenen) uitgevoerd. Met het commando *Compute* kun je variabelen bij elkaar optellen, van elkaar aftrekken, vermenigvuldigen, middelen enzovoort. Een voorwaarde is dan wel dat het meetniveau van de variabele minimaal interval moet zijn om deze berekeningen te kunnen uitvoeren. Een uitzondering hierop vormen ordinale variabelen waarvan de voor de antwoorden gebruikte schaal op interval lijkt.

Wanneer je van een respondent wilt weten hoeveel tijd diegene besteedt aan televisiekijken, zou je ervoor kunnen kiezen om dat te vragen in het aantal uren dat iemand tv kijkt, en het aantal minuten dat iemand tv kijkt. Je krijgt dan een nauwkeuriger antwoord dan wanneer je alleen naar het aantal uren zou vragen of iemand op een schaal van 1) weinig tot 5) veel laat antwoorden. Je zou dan in een enquête de vragen kunnen stellen:
Hoe lang kijkt u op een doordeweekse dag televisie? ..... uur en .... minuten
Een respondent die 3,5 uur televisiekijkt op een doordeweekse dag zou dan dus invullen: 3 uur en 30 minuten. Dit worden in je datamatrix twee variabelen: het aantal uur dat op een doordeweekse dag televisie wordt gekeken (we geven deze variabele even voor het gemak de naam TVUUR) en het aantal minuten dat op een doordeweekse dag televisie wordt gekeken (we noemen deze variabele hier even TVMIN). We willen echter bij het uitvoeren van een analyse dat deze twee variabelen worden samengevoegd, namelijk in ofwel het aantal uur dat iemand televisiekijkt, ofwel het aantal minuten dat iemand televisiekijkt. Met de functie *Compute* kun je deze variabelen dan bij elkaar optellen. Je zou dan de som krijgen: TVUUR + TVMIN = totale tijd televisiekijken. Dat gaat in dit geval niet zomaar: als je 3 uur bij 30 minuten laat optellen, dan zou je bij de bovenstaande respondent de formule krijgen: 3 + 30 = 33, en dit is niet de totale televisiekijktijd. Je kunt niet zomaar uren en minuten bij elkaar optellen. We zullen dus eerst van minuten uren moeten maken, of van uren minuten, voordat we deze twee variabelen bij elkaar op kunnen tellen.

Naast het optellen (of aftrekken, of vermenigvuldigen) van meerdere variabelen, kunnen we bij *Compute* ook een variabele zelf 'veranderen' door ermee te rekenen. In dit geval besluiten we om van uren minuten te maken. We moeten hier dan het aantal uren vermenigvuldigen met 60. De berekening die je dan krijgt is: (TVUUR * 60) + TVMIN.

Compute wordt ook vaak gebruikt om indexscores of gemiddelde schalen te maken. Bij een *indexscore* worden ordinale, interval- of ratiovariabelen bij elkaar opgeteld, bijvoorbeeld het aantal uur dat per week naar NPO1 wordt gekeken + het aantal uur dat per week naar NPO2 wordt gekeken + het aantal uur dat per week naar NPO3 wordt gekeken, om zo de nieuwe variabele te maken: aantal uur dat per week naar de publieke omroep wordt gekeken. Stel dat een persoon 2 uur naar NPO1 kijkt, 3 uur naar NPO2 en 1 uur naar NPO3, dan heeft deze persoon een indexscore van 2 + 3 + 1 = 6 voor het kijken naar de publieke omroep.

Je kunt er ook voor kiezen om gemiddelde schalen te maken. Stel dat je wilt weten hoe het NOS-journaal wordt gewaardeerd, en je vraagt de respondent een aantal rapportcijfers te geven voor de verschillende onderdelen. Je vraagt bijvoorbeeld hoe iemand de hoeveelheid nieuwsitems waardeert, de afwisseling van de items, de kwaliteit van de nieuwslezer, en het decor. In dat geval heb je vier rapportcijfers. Je kunt deze bij elkaar optellen, maar dan krijg je een vreemde waarde. Als iemand respectievelijk de rapportcijfers 6, 7, 8 en 6 zou geven, is die indexscore 27. Het is in dat geval beter om een gemiddelde score te berekenen. Dat zou je kunnen doen door de variabelen bij elkaar op te tellen en te delen door het aantal variabelen: (cijferNOS1 + cijferNOS2 + cijferNOS3 + cijferNOS4) / 4. De gemiddelde waardering voor het NOS-journaal is dan een 6,8. Die waarde is beter te interpreteren dan de waarde 27 die de sommering van de vier variabelen oplevert.

Een nadeel van deze methode is dat wanneer een respondent op één van die variabelen geen antwoord heeft gegeven (*missing value*), er geen uiteindelijke score wordt berekend. Dat komt doordat de opdracht niet wordt uitgevoerd als een van de betrokken variabelen een *missing value* heeft. Dan is het beter om bij het maken van een gemiddelde schaal het commando *MEAN* te gebruiken (zie kader 4.2). Op deze manier wordt een gemiddelde berekend, waarbij rekening wordt gehouden met het aantal variabelen waar de respondent ook daadwerkelijk op heeft geantwoord. Als dat maar bij drie van de vier vragen het geval is, wordt voor die respondent gedeeld door 3 en niet door 4.

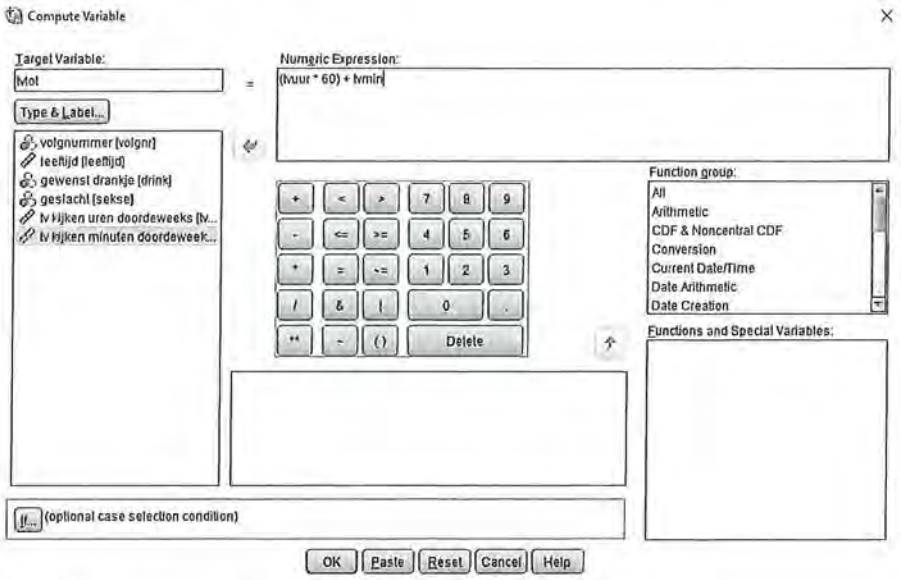
SPSS Nieuwe variabele maken door middel van Compute

Het berekenen van een nieuwe variabele op basis van bestaande variabelen in SPSS gaat via *Transform → Compute Variable...*

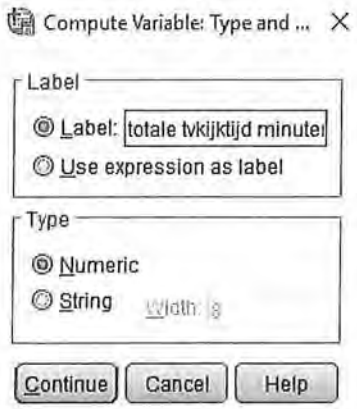
In het *Compute Variable*-venster (*figuur A*) dat dan verschijnt, kies je een nieuwe naam voor je variabele (onder *Target Variable:*). Let er daarbij op dat je geen spaties of leestekens gebruikt (op de *underscore* na herkent SPSS deze namelijk niet). Vervolgens kun je SPSS vertellen hoe die nieuwe variabele berekend moet worden (onder *Numeric Expression*).

De IF en FUNCTION-functies in het Compute-venster zullen wij in dit boek niet gebruiken. Verwar de IF-functie niet met die bij Select Cases (zie paragraaf 4.5)!

Eventueel kan onder Type & Label (*figuur B*) een langere beschrijving van de variabele gegeven worden waarin wel gebruikgemaakt kan worden van spaties en/of leestekens.

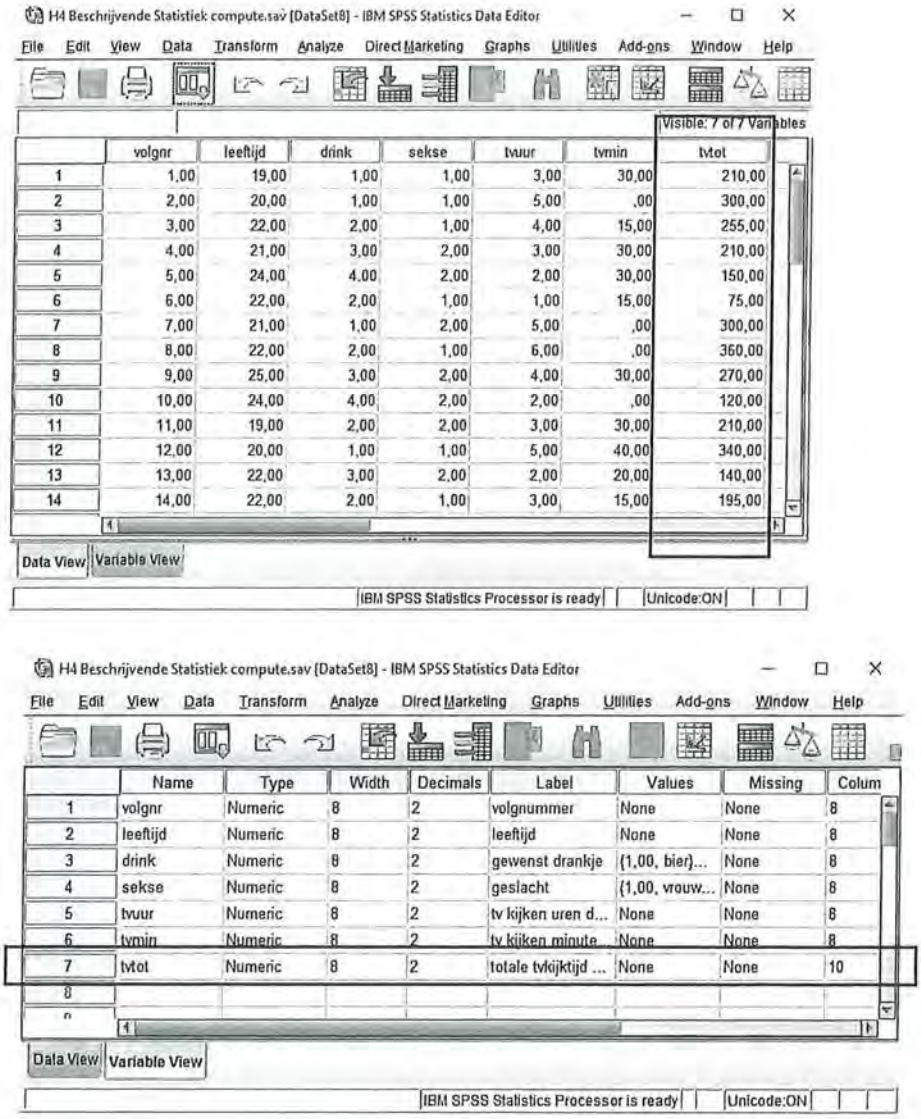


Figuur A Compute Variable-venster



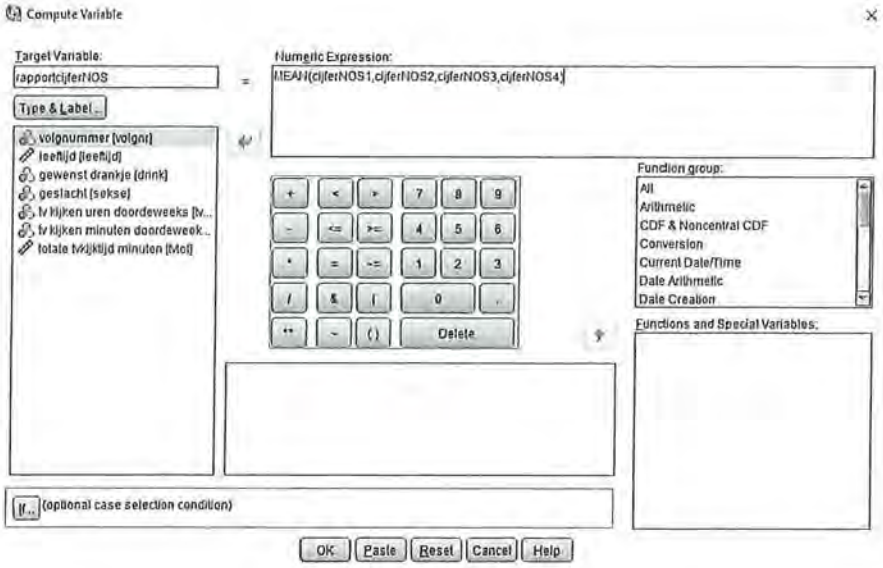
Figuur B Type & Label-venster

De nieuwe variabele verschijnt, na het runnen van de syntax, achteraan in je datamatrix in de *Data View*, en onderaan in de *Variable View* (zie Figuur C).



Figuur C Nieuwe variabele na Compute in Data View en Variable View

Wanneer je een gemiddelde schaal gaat samenstellen aan de hand van het commando *MEAN*, typ je in het *Numeric Expression*-venster zelf het woord *MEAN*, en zet je tussen haakjes de variabelen die je scheidt met komma's (*figuur D*). In dit voorbeeld is het gemiddelde rapportcijfer voor het NOS-journaal gemeten door het commando **MEAN(cijferNOS1, cijferNOS2, cijferNOS3, cijferNOS4)** in te voeren.



Figuur D Gemiddelde schaal maken door middel van Compute

Kader 4.2

4.4 Hercoderen

De opdracht *Recode*, in het Nederlands hercoderen, wordt gebruikt om binnen een bestaande variabele de waarden te herverdelen in verschillende klassen. Je hebt bijvoorbeeld de variabele 'kijktijd publieke omroep' gemaakt, waarin je door middel van *Compute* het aantal minuten dat iemand naar NPO1, NPO2 en NPO3 kijkt bij elkaar hebt opgeteld tot een indexscore. Je zou daar een frequentieverdeling van willen maken. In dit geval wordt dat echter een totaal onoverzichtelijke tabel, want de variabele 'kijktijd publieke omroep' is gemeten op rationiveau, en kan wel eens variëren tussen de 0 en de 1200 (of meer). In het ergste geval krijg je dan een tabel met 1200 of meer rijen. Om toch inzicht te krijgen in de vraag of de respondenten weinig, matig of erg vaak naar de publieke omroep kijken kun je ervoor kiezen om deze variabele te herverdelen in een aantal categorieën, zodat het overzichtelijker wordt om de variabele in een frequentietabel op te nemen. De grootte van de categorieën bepaal je als onderzoeker veelal zelf. Je kunt die keuze baseren op praktische argumenten (bijvoorbeeld om enkele ongeveer gelijke groepen te krijgen) of toont op theoretische gronden aan dat de keuze voor de verdeling voor jouw onderzoek de beste is. Laten we er in dit voorbeeld van uitgaan dat de variabele inderdaad als minimum nul minuten scoort (iemand kijkt niet naar de publieke omroep) en als maximum 1200 minuten. We willen een overzichtelijke kruistabel maken en herverdelen de variabele daarom in drie nieuwe klassen:

0 – 400 = 1
401 – 800 = 2
801 – 1200 = 3

Een respondent die bij de oorspronkelijke variabele 700 scoorde (de persoon keek 700 minuten in de week naar de publieke omroep), valt in deze nieuwe variabele in klasse 2. Een respondent die 200 minuten keek valt nu in klasse 1. Je kunt nu een meer overzichtelijke tabel maken van de kijktijd naar de publieke omroep in drie categorieën.

Tabel 4.3 Frequentietabel van kijktijd publieke omroep in klassen (SPSS-output)

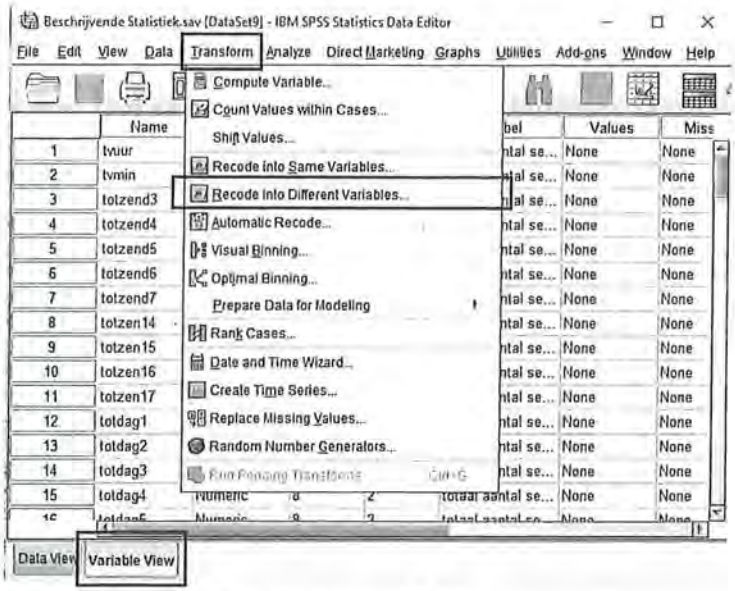
minpoHER minuten pub omr klassen

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 0 - 400 minuten	377	27,4	27,4	27,4
	2 401 - 800 minuten	531	38,6	38,6	65,9
	3 801 - 1200 minuten	469	34,1	34,1	100,0
	Total	1377	100,0	100,0	

Een belangrijke consequentie van het hercoderen van je variabele is dat het meetniveau van je variabele kan veranderen. De oorspronkelijke variabele 'kijktijd publieke omroep' was gemeten op rationiveau, de nieuwe variabele 'kijktijd publieke omroep in klassen' is ordinaal. Dat betekent dus ook dat je nu minder analyses met de variabele kunt uitvoeren. Je kunt nu bijvoorbeeld geen gemiddelde meer uitrekenen, alleen nog een mediaan en een modus.
Anders dan bij *Compute*, waar het meetniveau van de variabelen minimaal interval moet zijn, kun je ook nominale variabelen herverdelen in groepen. Voor alle bewerkingen geldt dat je in je onderzoek moet kunnen verantwoorden waarom je nu juist deze klassen maakt, op basis waarvan je je afweging maakt. Je zou de nominale variabele 'woonplaats' bijvoorbeeld kunnen herverdelen naar verschillende provincies, of de nominale variabele 'partijkeuze' kunnen herverdelen naar een schaal links – midden – rechts. De variabele 'partijkeuze' (nominaal) naar de nieuwe variabele 'politieke oriëntering' zou je nu als ordinaal kunnen beschouwen: hoe hoger iemand op de schaal scoort, hoe meer rechts georiënteerd deze is. Nogmaals, dit soort beweringen moeten altijd onderbouwd worden!

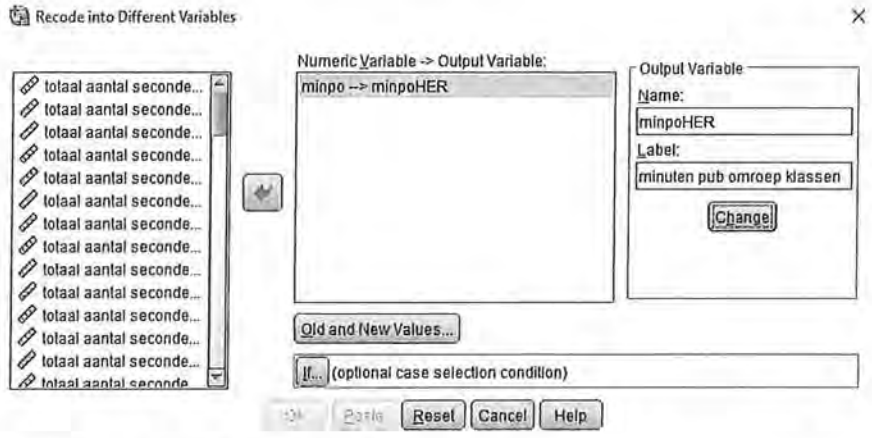
SPSS Hercoderen van variabelen

Ook het hercoderen van variabelen in SPSS gaat via Transform (figuur A). Daarbij is het belangrijk dat je altijd *Recode into Different Variables* kiest, en niet *Recode into Same Variables*. Het verschil is dat wanneer je hercodeert in een nieuwe variabele (*into different*), je je oorspronkelijke variabele blijft behouden. Bij de optie *into Same* vervang je de oorspronkelijke variabele door de nieuwe. Wanneer je nog wilt werken met de oude variabele (bijvoorbeeld omdat je in een andere analyse wel die variabele nodig hebt op rationiveau), kan dat dus alleen maar wanneer je *into different* hebt gekozen.



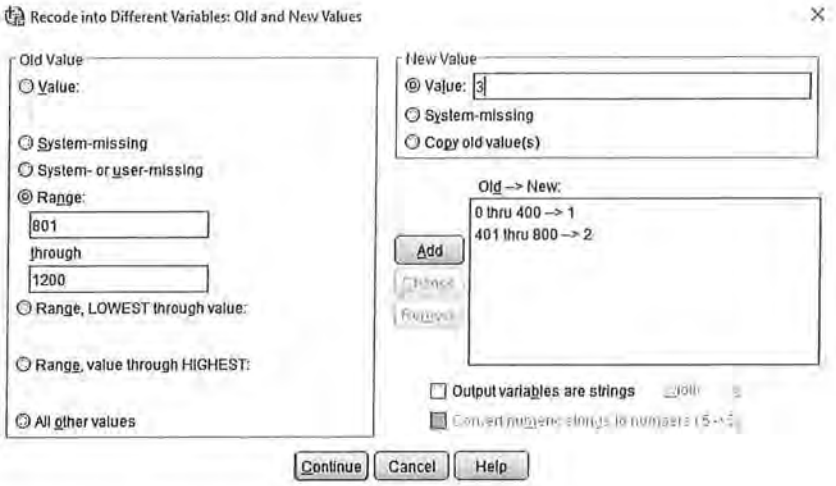
Figuur A Hercoderen van een variabele

In het Recode into Different Variables Window (*figuur B*) selecteer je de variabele die je wilt hercoderen in het venster *Numeric Variable → Output Variable*: Er verschijnt vervolgens een vraagteken achter de variabele. In de vensters onder *Output Variable* kun je bij *Name* de nieuwe naam van de variabelen invoeren (ook hier weer: geen spaties en/of leestekens), bij *Label* kun je een (langere) beschrijving geven van de nieuwe variabele. In eerste instantie is de optie *Paste* (die nodig is om de syntax te maken) uitgeschakeld; deze wordt pas actief wanneer je op *Change* hebt geklikt. Dan verdwijnt ook het vraagteken achter je oorspronkelijke variabele en komt daar de naam van je nieuwe gehercodeerde variabele te staan.



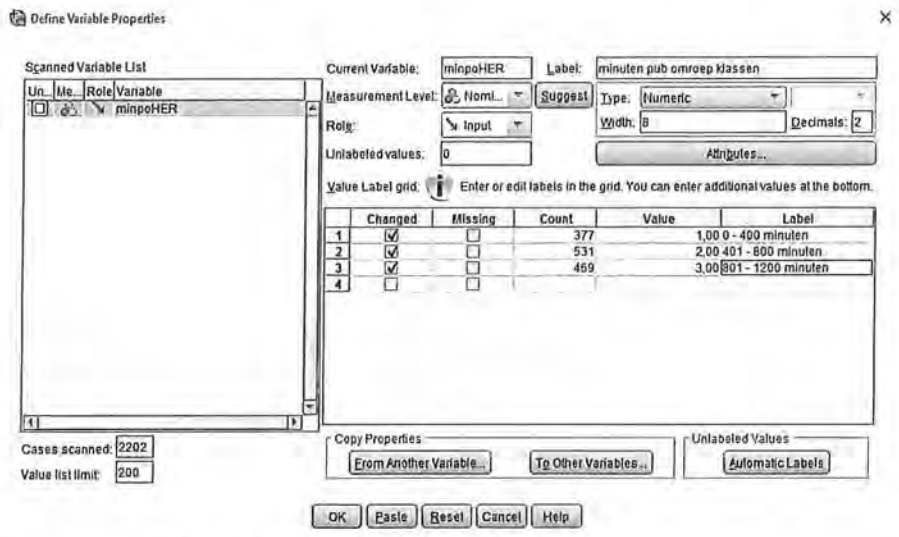
Figuur B Recode into Different Variables-venster

Vervolgens klik je op de knop *Old and New Values ...* In het venster dat verschijnt kun je aangeven wat de oorspronkelijke waarden zijn (Old Values) en wat de nieuwe waarden moeten worden (New Value). Bij ordinale, interval, en ratiovariabele kun je vaak de functie 'Range' gebruiken. Je geeft aan wat de minimumwaarde van een klasse is en de maximumwaarde van die klasse, geeft deze klasse een nieuwe waarde, en klikt vervolens op 'Add'.



Figuur C Old and New Values-venster

Het is ook mogelijk om met *LOWEST through value* te werken (vanaf de laagste waarde *tot en met* een bepaalde waarde, in bovenstaand voorbeeld *Range, Lowest through value: 500*) en *value through HIGHEST* (vanaf een bepaalde waarde *tot en met* de hoogste mogelijke waarde die deze variabele aanneemt, hier zou dat zijn *Range, value through Highest: 1001*). Bij het hercoderen van een nominale variabele kan de optie *Range* uiteraard niet gebruikt worden. Dan is er immers geen sprake van oplopende waarden.
Wanneer je de variabele hebt gehercodeerd en de syntax hebt laten runnen, zal de nieuwe variabele, net als bij *Compute* in de Data View als laatste, en in de Variable View als onderste variabele verschijnen. SPSS heeft de values echter nog niet gelijk ook een label gegeven. Daarvoor dien je eerst bij *Data → Define Variable Properties* zelf de valuelabels in te typen.



Figuur D Define Variable Properties om valuelabels te maken.

4.5 Select Cases

Wanneer je analyses wilt uitvoeren op een bepaalde subgroep in je steekproef, selecteer je die op basis van een bepaalde waarde op een variabele. Je wilt bijvoorbeeld alleen een uitspraak doen over de mannen in je steekproef, of alleen over hoger opgeleiden, of alleen over de respondenten in een bepaalde leeftijdsgroep. In dat geval kun je gebruikmaken van *Select Cases*. Een van de verschillen met *missing values* is dat wanneer je cases selecteert, je de rest van de waarden automatisch uitsluit voor *alle* verdere analyses die je doet.
Ook in deze paragraaf zijn de bewerkingen in SPSS niet in een apart kader gezet maar in de tekst opgenomen.
We hebben een datamatrix met daarin de informatie van twaalf respondenten. Er is informatie over hun opleidingsniveau, hun geslacht en hun leeftijd. Opleidingsniveau is gemeten met de waarden
1 = laag opgeleid
2 = midden opgeleid
3 = hoog opgeleid
Sekse is gemeten met voor vrouw de waarde 1, en voor man de waarde 2. Leeftijd is gemeten door de respondenten te vragen hoe oud ze zijn.

	opleiding	sekse	leeftijd
1	1,00	1,00	19,00
2	1,00	2,00	20,00
3	1,00	1,00	40,00
4	1,00	2,00	39,00
5	2,00	1,00	29,00
6	2,00	2,00	25,00
7	2,00	1,00	20,00
8	2,00	2,00	22,00
9	3,00	1,00	26,00
10	3,00	2,00	26,00
11	3,00	1,00	38,00
12	3,00	2,00	27,00

Figuur 4.4 Datamatrix van opleiding, sekse en leeftijd (N = 12)

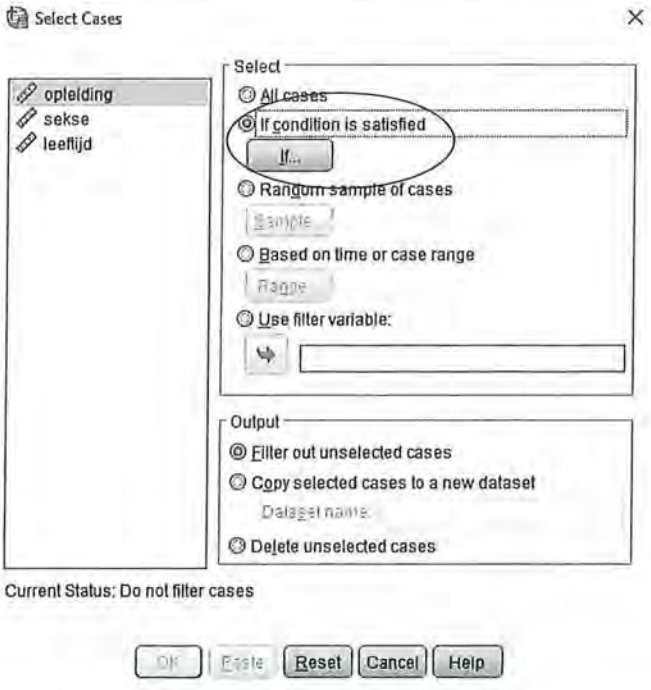
De mediaan van opleidingsniveau is 2 (midden opgeleid), de modus van sekse is 2 (er zijn meer mannen dan vrouwen) en de gemiddelde leeftijd is 27,58 (SD = 7,55).
Wanneer we nu een frequentieverdeling van bijvoorbeeld opleidingsniveau uit zouden draaien, krijgen we daar de informatie te zien van alle twaalf onderzoekseenheden (tabel 4.4).

Tabel 4.4 Frequentieverdeling van opleidingsniveau (N = 12) (SPSS-output)

opleiding

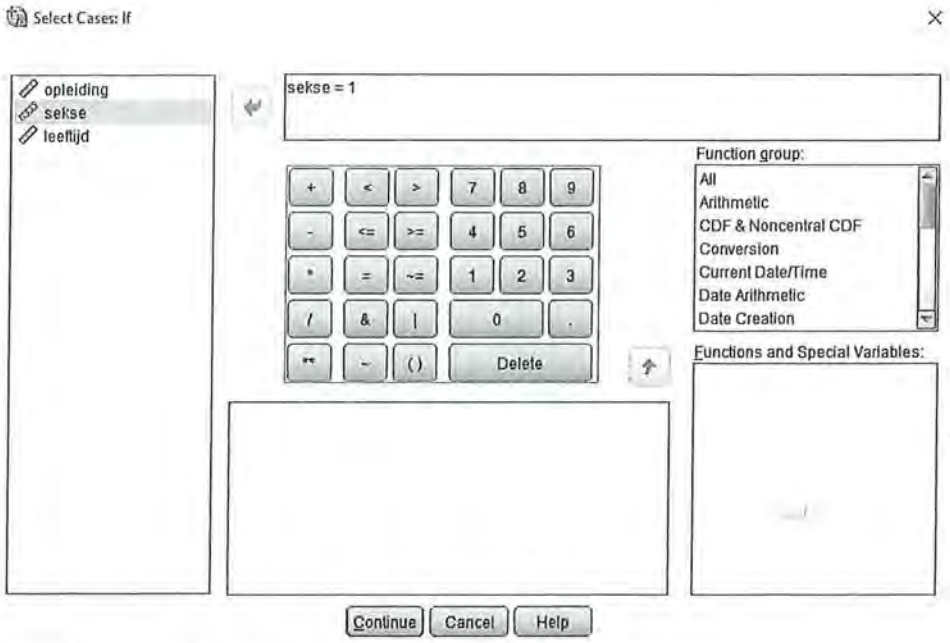
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,00 laag	4	33,3	33,3	33,3
	2,00 midden	4	33,3	33,3	66,7
	3,00 hoog	4	33,3	33,3	100,0
	Total	12	100,0	100,0	

Nu wil je nogmaals naar de verdelingen kijken, maar alleen voor je vrouwelijke respondenten. Je geeft nu in SPSS het commando dat je alleen vrouwen wilt selecteren. Je zegt dus eigenlijk: ik wil alleen die respondenten selecteren wanneer aan de voorwaarde wordt voldaan, dat op deze variabele 1 wordt gescoord (want dat was in ons onderzoek de waarde voor vrouw).
In SPSS kun je dit aangeven via *Data → Select Cases*. Je krijgt dan een venster (zie figuur 4.5) waarin je bepaalde subgroepen kunt selecteren via *if condition is satisfied*.
NB: Je kunt ditzelfde venster ook gebruiken als je weer alle respondenten wilt selecteren (via de bovenste optie: *All cases*).



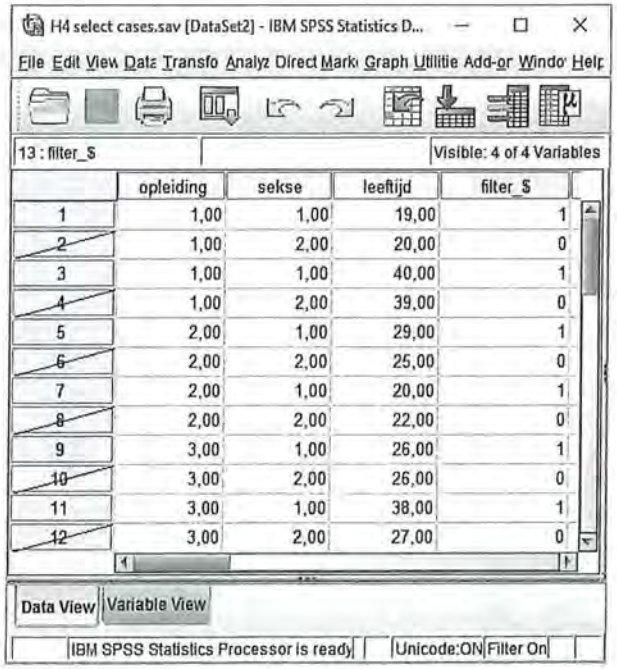
Figuur 4.5 Select Cases-venster

Wanneer je nu op *If* klikt, krijg je een nieuw venster (figuur 4.6) waarin je kunt aangeven van welke variabele je een selectie van waarden wilt maken. In ons voorbeeld willen we alleen de vrouwen selecteren, dus gebruiken we het commando: sekse = 1, zie figuur 4.6.



Figuur 4.6 Selecteren van alleen vrouwen via IF-commando.

Wanneer je dit commando laat uitvoeren in SPSS, zul je zien dat in de datamatrix in de *Data View* alle respondenten die niet aan die voorwaarden voldoen, oftewel de respondenten die man zijn, door SPSS worden weggestreept, zie figuur 4.7.



Figuur 4.7 Datamatrix na select cases van sekse = 1

In de laatste kolom is een nieuwe 'variabele' aangemaakt met de naam *filter_$*. Deze variabele zullen we nooit in de analyses zelf gebruiken! Het is slechts een

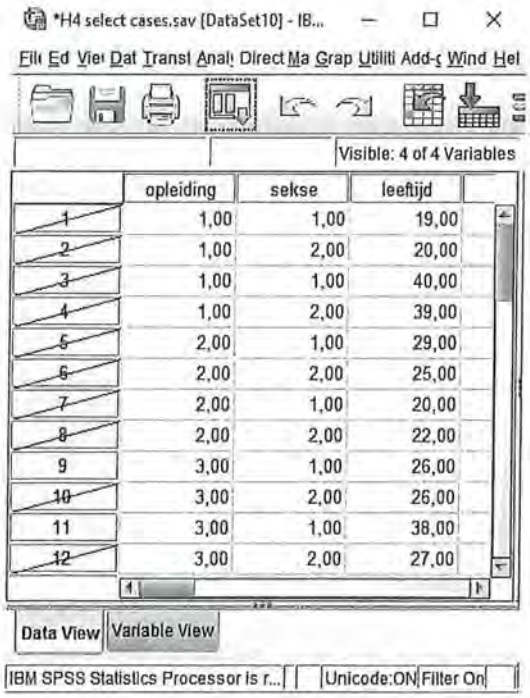
variabele die aangeeft of de respondent wel (waarde 1) of niet (waarde 0) geselecteerd is na het selecteren van onze cases.
Alle analyses die we vanaf dit moment met de variabelen uitvoeren, gaan alleen nog maar over vrouwen. Wanneer je nu een frequentieverdeling maakt van opleidingsniveau, is de mediaan weliswaar nog steeds 2, maar die gaat nu alleen over de respondenten die vrouw zijn:

Tabel 4.5 Frequentieverdeling van opleiding voor vrouwen (n = 5) (SPSS-output)

opleiding

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,00 laag	1	20,0	20,0	20,0
	2,00 midden	2	40,0	40,0	60,0
	3,00 hoog	2	40,0	40,0	100,0
	Total	5	100,0	100,0	

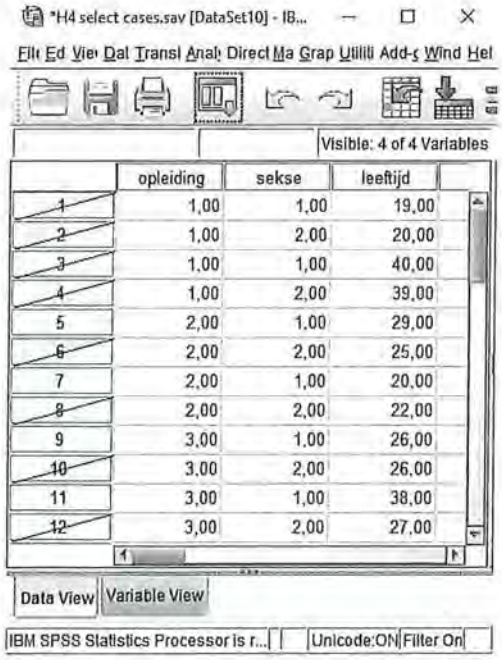
Wanneer je ook het gemiddelde zou berekenen, zou je zien dat waar eerst de gemiddelde leeftijd van alle respondenten 27,58 was, de gemiddelde leeftijd van de vrouwen 30,6 (*SD* = 8,35) is.
In een volgende stap willen we niet alleen vrouwen selecteren, maar willen we alleen die vrouwen selecteren die een hoog opleidingsniveau hebben. We selecteren op dezelfde manier als eerder de vrouwen, maar voegen nu nog een variabele toe, namelijk opleidingsniveau. We typen nu als het *IF*-commando:
sekse = 1 & opleiding = 3



Figuur 4.8 Datamatrix na select cases van sekse = 1 & opleiding = 3

In plaats van het &-teken is het ook mogelijk om het woord AND te typen.
Weer worden alle respondenten die niet aan die voorwaarde voldoen, weggestreept (zie figuur 4.8). In dit geval blijven er nog maar twee vrouwen over, want die hebben allebei een hoge opleiding genoten. Wanneer we van deze twee nu de gemiddelde leeftijd zouden uitrekenen, zal die weer anders zijn dan bij alle vrouwen; vrouwen met een hoog opleidingsniveau zijn gemiddeld 32 jaar oud (*SD* = 8,49).

Op dezelfde manier kunnen we SPSS vertellen dat we alleen vrouwen willen selecteren met een gemiddeld of hoog opleidingsniveau. We selecteren de vrouwen weer op de inmiddels bekende manier, en voegen daaraan toe dat ze óf de waarde 2, óf de waarde 3 moeten scoren. 'Of' kun je in SPSS aangeven met het teken | (dat ook op het numerieke toetsenbord staat in het *Select Cases*-venster, zie figuur 4.6), of door het woord OR te typen. Het commando ziet er dan als volgt uit:
sekse = 1 & (opleiding = 2 | opleiding = 3)
Belangrijk hierbij is dat je de variabelenaam opleiding tussen haakjes zet, én dat je deze voor beide variabelenwaarden herhaalt! Het commando: (opleiding = 2 | 3) wordt niet door SPSS herkend.
Weer worden alle mannen weggestreept, maar ook de vrouwen die op opleiding de score '1' (laag opgeleid) hadden:



Figuur 4.9 Datamatrix na select cases van sekse = 1 & (opleiding = 2 | opleiding = 3)

Wanneer we nu de centrummaat voor leeftijd zouden berekenen, kunnen we zeggen dat vrouwen met een gemiddeld of hoog opleidingsniveau gemiddeld 28,25 jaar oud zijn (*SD* = 7,50).

Het is ook mogelijk om aan te geven dat je alleen maar respondenten in een bepaalde leeftijdscategorie wilt selecteren, bijvoorbeeld alleen de respondenten die 25 jaar of ouder zijn. Het commando dat je intypt zou dan zijn:
Leeftijd > 24[1]
Maar je kunt ook respondenten selecteren met een leeftijd tussen de 19 en 22 en tussen de 38 en 40 jaar. Het commando is dan:
(leeftijd > 18 & leeftijd < 23) | (leeftijd > 37 & leeftijd <41)
Om te controleren of dat goed is gegaan kun je een frequentietabel uitdraaien van de variabele leeftijd. In tabel 4.6 is te zien dat inderdaad alleen de respondenten tussen de 19 en 22 en tussen de 38 en 40 in de analyse worden opgenomen.

Tabel 4.6 Frequentietabel van leeftijd na Select Cases (SPSS-output)

leeftijd

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	19,00	1	14,3	14,3	14,3
	20,00	2	28,6	28,6	42,9
	22,00	1	14,3	14,3	57,1
	38,00	1	14,3	14,3	71,4
	39,00	1	14,3	14,3	85,7
	40,00	1	14,3	14,3	100,0
	Total	7	100,0	100,0	

Uiteraard kun je de commando's zo ingewikkeld maken als je zelf wilt. Je kunt vrouwen selecteren die een gemiddeld of hoog opleidingsniveau hebben en in de leeftijdscategorie 19-22 of in de leeftijdscategorie 38-40 vallen:
sekse = 1 & (opleiding = 2 | opleiding = 3) & ((leeftijd > 18 & leeftijd < 23) | (leeftijd > 37 & leeftijd <41))
Let er in ieder geval goed op dat je de verschillende variabelen die je gebruikt tussen haakjes zet, dat je de juiste manier van AND en OR gebruikt, en dat je steeds de variabelenaam in je commando blijft herhalen.
Vergeet overigens niet om via *Select Cases* bij een volgende analyse weer al je onderzoekseenheden te selecteren!

4.6 Samenvatting

Als je data zijn verzameld en gecontroleerd, zul je voor sommige analyses je data nog moeten bewerken. Dit heeft in bijna alle gevallen consequenties voor het meetniveau van je oorspronkelijke variabele. Bij een ordinaal bedoelde variabele die de optie 'niet van toepassing' heeft, zal deze waarde eerst missing gemaakt moeten worden om het meetniveau ook daadwerkelijk ordinaal te laten zijn. Door middel van *missing values* is het mogelijk om bij een variabele bepaalde waarden niet mee te laten tellen.

Ook bij de functie *Select Cases* laat je bepaalde waarden niet meetellen, maar hierbij gaat het erom dat je subgroepen selecteert, en daarmee ook subgroepen uitsluit. Als je een analyse hebt waarin je alleen van hoogopgeleide vrouwen de gemiddelde leeftijd wilt weten, kun je door middel van *Select Cases* alle onderzoekseenheden die niet aan dat criterium voldoen, uitsluiten. Na het uitvoeren van *Select Cases* gaan alle analyses alleen nog maar over deze selectie van onderzoekseenheden.
Bij *Compute* en *Recode* maak je een nieuwe variabele op basis van (een) bestaande variabele(n). Door middel van *Compute* kun je verschillende variabelen bij elkaar optellen, een gemiddelde schaal maken of een berekening uitvoeren waardoor je bijvoorbeeld van uren minuten maakt (of andersom). *Compute* kan dan ook alleen gebruikt worden bij een variabele met minimaal interval of op interval gelijkend meetniveau. Bij *Recode* maak je een herverdeling van de waarden binnen een bestaande variabele. Ook daarmee kan het meetniveau van je oorspronkelijke variabele veranderen: de ratiovariabele leeftijd (waarbij je hebt gevraagd hoe oud iemand is), wordt ordinaal wanneer je daar leeftijdsgroepen van maakt. *Recode* kan echter ook gebruikt worden om nominale variabelen te herschikken.

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

Noot

1 Je kunt hier ook het groter of gelijk aan teken gebruiken. Het commando zou dan zijn: leeftijd >= 25.

Associatiematen op nominaal niveau 5

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

5.1 Wat zijn associatiematen?

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

5.1.1 Meetniveau van de variabelen

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

5.1.2 Symmetrische en asymmetrische relaties

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

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

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

5.1.3 Samenhang in kruistabellen

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

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

Boekgenre * Serie Crosstabulation

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

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

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

Boekgenre * Serie Crosstabulation

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

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

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

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

5.2 Cramers V

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

5.2.1 Interpretatie

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

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

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

Boekgenre * Serie Crosstabulation

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

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

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

Symmetric Measures

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

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

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

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

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

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

Onze conclusie zou bij bovenstaand voorbeeld dan zijn:

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

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

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

5.2.2 Berekening

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

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

Formule voor Cramers V

En de formule voor χ^2 is:

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

Formule voor Chi-kwadraat

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

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

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

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

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

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

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

Tabel 5.5 Geobserveerde frequenties

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

Tabel 5.6 Verwachte frequenties[3]

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

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

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

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

Tabel 5.7 Het berekenen van χ^2

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

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

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

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

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

5.3 Phi

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

5.3.1 Interpretatie

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

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

Boekgenre * Serie Crosstabulation

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

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

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

Symmetric Measures

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

5.3.2 *Berekening*

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

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

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

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

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

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

Tabel 5.11 Het berekenen van χ^2

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

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

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

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

5.4 Goodman en Kruskals tau

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

5.4.1 Berekening

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

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

Formule voor Goodman en Kruskals tau

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

De formule voor E_1 is

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

Formule voor E_1 bij Goodman en Kruskals tau

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

De formule voor E_2 is

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

Formule voor E_2 bij Goodman en Kruskals tau

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

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

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

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

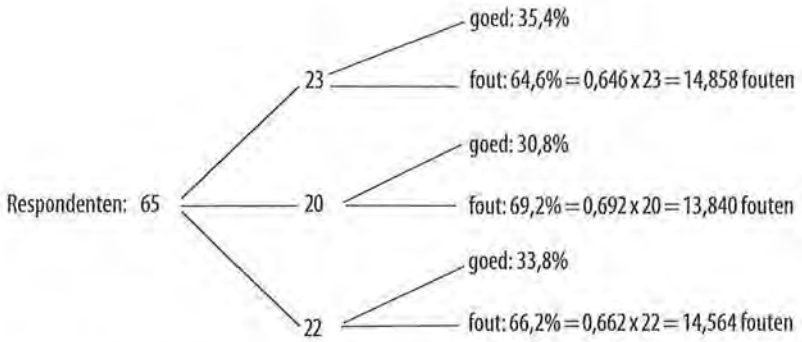
Tabel 5.12 Kruistabel van favoriete boekgenre naar sekse (SPSS-output)

Boekgenre * Sekse Crosstabulation

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

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

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



Figuur 5.1 Berekening E_1 (afhankelijke variabele = boekgenre)

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

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

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

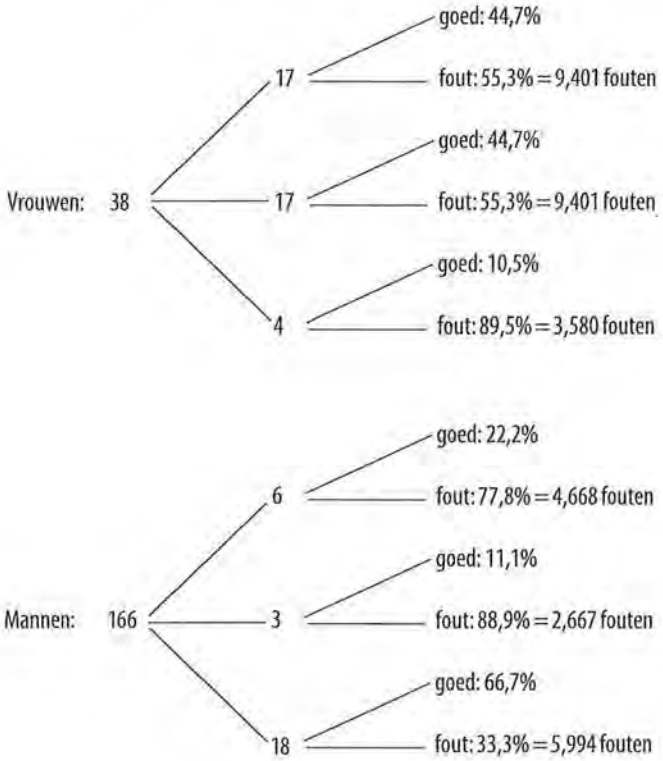
Formule voor E_1 bij tau.

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

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

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

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



Figuur 5.2 Berekening E_2 (onafhankelijke variabele = sekse)

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

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

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

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

Formule voor E_2 bij tau.

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

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

Vrouwen

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

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

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

Mannen

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

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

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

Nu hebben we alle benodigdheden om de E_2 te berekenen:

$E_2 = \sum_j E_{2j}$

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

Nu kunnen we de formule voor tau invullen:

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

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

5.4.2 *Interpretatie*

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

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

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

Directional Measures

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

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

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

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

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

zendervoorkeur * opleiding Crosstabulation

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

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

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

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

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

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

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

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

Laag opgeleiden

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

Midden opgeleiden

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

Hoog opgeleiden

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

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

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

Tot slot vullen we de formule voor tau in:

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

SPSS bevestigt deze uitkomst:

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

Directional Measures

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

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

Onze conclusie zou zijn:

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

5.5 Lambda

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

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

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

Formule voor lambda

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

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

Formule voor E_1 bij lambda

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

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

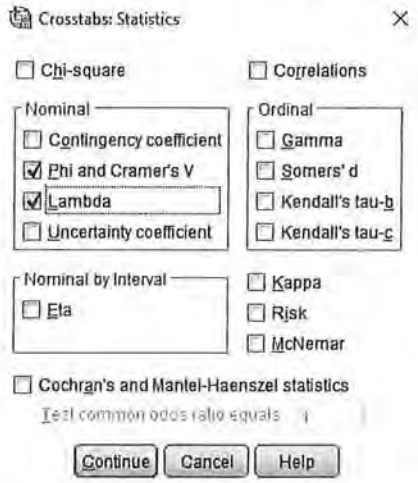
Formule voor E_2 bij lambda

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

SPSS Berekenen van nominale associatiematen

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

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



Figuur A Statistics-venster: nominale associatiematen.

Kader 5.1

5.5.1 Berekening

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

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

praten * Serie Crosstabulation

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

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

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

$E_1 = 65 - 33 = 32.$

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

$E_2 = 65 - 36 = 29.$

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

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

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

5.5.2 *Interpretatie*

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

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

Directional Measures

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

Onze conclusie aan de hand van bovenstaande output is:

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

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

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

5.6 Voorwaarden bij het maken van een kruistabel

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

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

Boekgenre * tvprog Crosstabulation

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

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

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

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

boekgenre * tvprogHER Crosstabulation

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

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

5.7 Samenvatting

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

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

Tabel 5.20 Nominale associatiematen naar relatie

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

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

Noten

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

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

Inkomen \ Baan	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

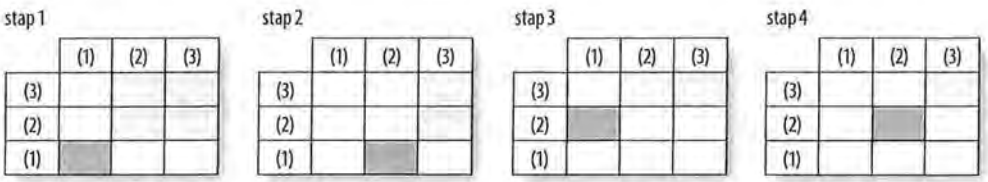
onderzoekseenheden G in (1,1), die nooit de *Linda* lezen en nooit de *Happinez* lezen.

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

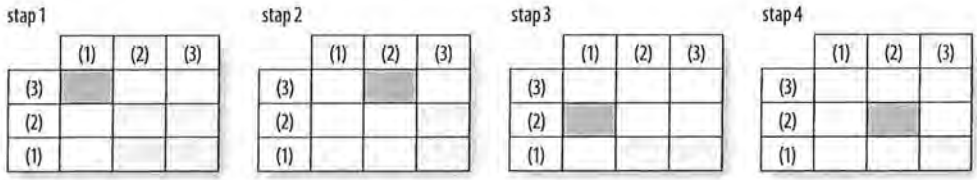
Nu nog de discordante paren. We beginnen nu bij de onderzoekseenheid met de laagste waarde op de ene, en de hoogste waarde op de andere variabele. In dit geval is dat onderzoekseenheid A in cel (1,3), mensen die nooit *Linda* lezen en vaak *Happinez*.

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 x 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

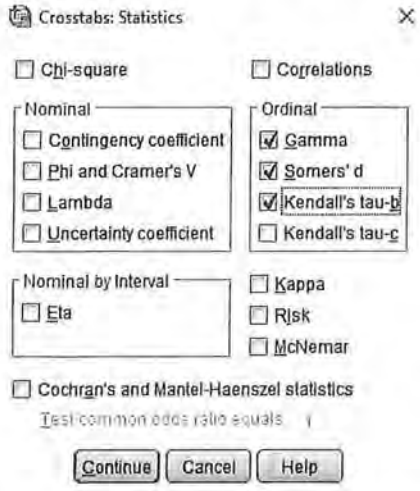
Evenals bij de asymmetrische nominale maten (lambda en Goodman en Kruskals tau) vind je in de SPSS-output meerdere waarden van deze associatiemaat. SPSS weet immers niet welke variabele wij als onafhankelijk en welke wij als afhankelijk hebben benoemd. Aangezien in dit geval de woordenschat de afhankelijke 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 (τ_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).

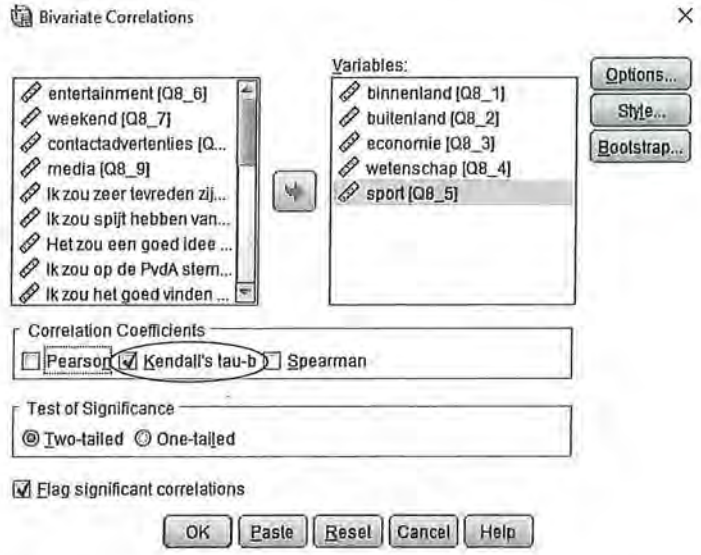
6.4.3 *Kendalls tau-b in een correlatiematrix*

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

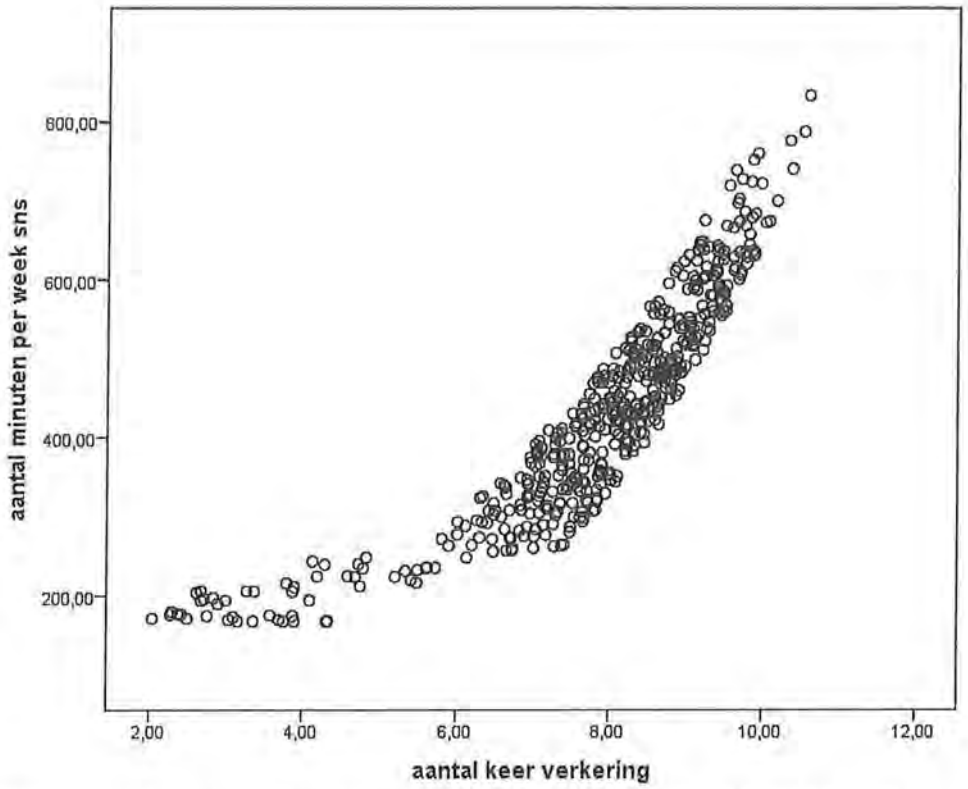
			uur_sns	geluk
Spearman's rho	uur_sns	Correlation Coefficient	1,000	-,829
		Sig. (2-tailed)	.	,042
		N	6	6
	geluk	Correlation Coefficient	-,829	1,000
		Sig. (2-tailed)	,042	.
		N	6	6

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.

Tabelsplitsing 7

In de voorgaande twee hoofdstukken hebben we gekeken naar het verband tussen twee nominale of ordinale variabelen, oftewel naar bivariate analyses. In dit hoofdstuk staan multivariate analyses met nominale en ordinale variabelen centraal, dat wil zeggen dat we meer dan twee variabelen in de analyse zullen betrekken.

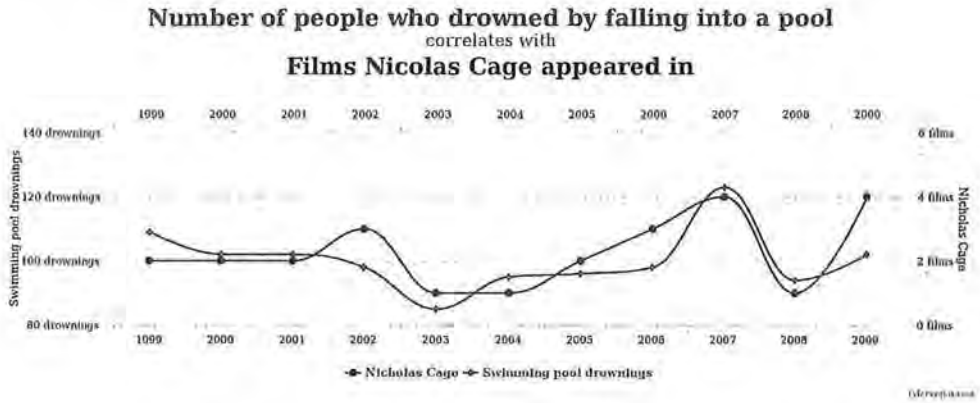
Waarom zou je meer dan twee variabelen in de analyse betrekken? Dat doe je omdat als je een samenhang tussen twee variabelen hebt gevonden, je het inzicht in die samenhang kunt vergroten door naar de rol van een derde variabele te kijken. Het is belangrijk om je te realiseren dat de samenhang tussen twee variabelen niets zegt over *causaliteit*, over oorzaak en gevolg. Om er zeker van te zijn dat veranderingen in de ene variabele (x) de oorzaak zijn van veranderingen in de andere variabele (y) zou je moeten uitsluiten dat een derde variabele verantwoordelijk is voor de veranderingen in zowel x als y. Dat is meestal niet zo eenvoudig. Als er sprake is van een causaal verband, moeten veranderingen in x ook altijd voorafgaan aan veranderingen in y. Zelfs als we dat hebben kunnen vaststellen, is het echter nog altijd mogelijk dat een andere variabele eerst de veranderingen in x heeft veroorzaakt en pas daarna de veranderingen in y. Zelfs als we op theoretische gronden een invloed van x op y verwachten en we daarom kiezen voor een asymmetrische associatiemaat, weten we nog niet zeker dat er een *causaal* verband is. Dat laatste kunnen we alleen aannemelijk maken (dus niet: vaststellen!) door allerlei andere verklaringen uit te sluiten. Als we aan de hand van een Somers' d concluderen dat als peuters meer educatieve spellen op een tablet spelen, zij een hogere woordenschat hebben, dan kunnen we niet zeggen dat deze woordenschat door het spelen van educatieve spellen wordt *veroorzaakt*. We moeten voorzichtig zijn met deze conclusie vanwege twee redenen. Ten eerste kan het zo zijn dat er een derde variabele in het spel is (al dan niet gemeten in het onderzoek) die de resultaten beïnvloedt. Ten tweede geven associatiematen geen uitsluiting over de causaliteit van een verband. We hebben dan als onderzoeker wel bedacht dat in bovenstaand voorbeeld het spelen op de tablet de onafhankelijke variabele was en de woordenschat de afhankelijke variabele, en het is ook intuïtief aannemelijk dat het spelen invloed heeft op de woordenschat, maar statistisch gezien is er geen bewijs voor een causale relatie.

Bij enquêtes en inhoudsanalyses kun je variabelen opnemen waarvoor je kunt controleren of ze de relatie tussen x en y beïnvloeden, maar het is niet mogelijk om alle denkbare van invloed zijnde factoren te meten. Bij een experiment

wordt geprobeerd om de situatie dusdanig te controleren dat een poging tot een uitspraak over causaliteit gedaan kan worden. Dit doen we door bijvoorbeeld een *experiment* in een laboratorium af te nemen waardoor je zeker weet dat x voorafgaat aan y en storende invloeden beperkt en/of onder controle worden gehouden. Bovendien kunnen we door een controlegroep toe te voegen de uitkomsten van die groep vergelijken met de experimentele groep.

7.1 Interpretatie

Wanneer we een samenhang tussen twee variabelen vinden, betekent dat dus niet dat er een oorzakelijk verband tussen de twee variabelen bestaat. Op internet vinden we verschillende (grappige) voorbeelden van zeer sterke verbanden van variabelen die in werkelijkheid niets met elkaar te maken hebben.[1] Zo zien we dat er een zeer sterk positief verband bestaat tussen het aantal films waar Nicholas Cage in speelt en het aantal verdrinkingen in een zwembad.



Figuur 7.1 Correlatie tussen aantal films met Nicholas Cage en aantal verdrinkingen in een zwembad

In werkelijkheid is het vrij onwaarschijnlijk dat het aantal films met Nicholas Cage de oorzaak is van het aantal verdrinkingen in een zwembad (of andersom: dat het aantal verdrinkingen in een zwembad van invloed is op het aantal films waarin Nicholas Cage speelt). Er is dan sprake van een *schijnsamenhang*. Zo'n schijnsamenhang, die je ook wel een *spurieuze samenhang* noemt, wordt meestal veroorzaakt door andere variabele(n) of is soms gewoon toeval. Het is waarschijnlijker dat er een andere oorzaak is voor het aantal verdrinkingen in een zwembad dan het aantal films met Nicholas Cage, bijvoorbeeld de temperatuur in dat jaar. Omdat het in 2007 warmer was dan in 2008, waren er dat jaar meer zwempartijen en daarom meer verdrinkingen. Dat er in 2007 ook meer films met de acteur waren, is dan toeval maar geen oorzaak. Bij controle door de variabele temperatuur blijkt de eerder gevonden samenhang tussen verdrinkingen in een zwembad en het aantal films met Cage spurieus te zijn: een schijnverband.

Ook is het mogelijk dat je bij controle voor een derde variabele constateert dat het eerder gevonden verband voor sommige groepen sterker is dan voor andere groepen. In dat geval is er geen sprake van een schijnsamenhang. De samenhang is er echt, maar je kunt de eerder gevonden samenhang tussen twee variabelen wel verder *specificeren* door gebruik te maken van een derde variabele. Als je in een onderzoek onder werknemers in een bedrijf een samenhang vindt tussen de hoogte van het inkomen en het aantal jaar dat iemand in dienst is, zou je kunnen concluderen dat het inkomen toeneemt naarmate iemand langer in dienst is. Als je die samenhang controleert voor de variabele geslacht, is het mogelijk dat de samenhang tussen inkomen en dienstjaren bij mannen veel sterker is dan bij vrouwen. Je hebt dan het eerder gevonden verband gespecificeerd en een *interactie* met de variabele geslacht aangetoond. Je conclusie is dan dat langer in dienst zijn bij mannen tot een grotere inkomenstoename leidt dan bij vrouwen.
Er is ook nog de mogelijkheid dat je een verband tussen twee variabelen had verwacht dat niet of nauwelijks aanwezig blijkt te zijn. Na controle voor een derde variabele kan blijken dat er voor afzonderlijke groepen wel een verband is. Als voor die groepen de verbanden tussen de twee variabelen tegengesteld zijn, wordt het verband voor alle groepen tezamen *versluierd*.

7.1.1 Spurieuze samenhang

Als de samenhang tussen twee variabelen hoog is maar het onwaarschijnlijk is dat de twee variabelen elkaar beïnvloeden en je ook geen goede verklaring hebt voor de samenhang, is het verstandig te zoeken naar een derde variabele die de samenhang wel kan verklaren. Laten we een fictief onderzoek naar ijsconsumptie en het aantal verdrinkingen als voorbeeld nemen. Er blijkt een sterke samenhang te bestaan tussen de hoeveelheid ijs die per dag wordt gegeten en het aantal verdrinkingen per dag. De onderzoekseenheden zijn hier 'dagen' (zie tabel 7.1).

Tabel 7.1 Kruistabel van ijsconsumptie en verdrinkingen per dag (n = 200)

IJs / Verdrinkingen	0 weinig	1 veel	
1 veel	18	82	100
0 weinig	82	18	100
	100	100	200 (dagen)

We hebben geen idee of ijs eten het aantal verdrinkingen zou beïnvloeden of andersom, en er is sprake van ordinale variabelen, want 'veel' is meer dan 'weinig'. Wanneer we hier gamma zouden berekenen, zou deze uitkomen op 0,91. Onze conclusie zou zijn: er is een zeer sterke, positieve samenhang tussen het aantal ijsjes dat op een dag wordt gegeten en het aantal verdrinkingen dat op een dag plaatsvindt. Op dagen dat er veel ijs wordt gegeten zijn er ook veel verdrinkingen, en als er weinig ijs wordt gegeten zijn er weinig verdrinkingen.

Dit is natuurlijk onzinnige informatie. Waarom zou er zo'n sterk verband zijn tussen deze twee variabelen? Hoe is dat verband te verklaren? IJsconsumptie kan toch niet leiden tot een toename in het aantal verdrinkingen en verdrinkingen hebben ook geen hogere ijsconsumptie tot gevolg. Het is waarschijnlijk dat er een derde variabele in het spel is die invloed heeft op beide variabelen. We voegen daarom een derde variabele toe: de gemiddelde temperatuur per dag. Voor het gemak maken we onderscheid tussen koude dagen en warme dagen. We zouden nu naar de ijsconsumptie en het aantal verdrinkingen kunnen kijken op warme dagen en op de koude dagen afzonderlijk. We krijgen dan twee verschillende tabellen (tabel 7.2), die samen de (oorspronkelijke) tabel vormen. We noemen deze twee tabellen van zo'n tabelsplitsing *deeltabellen* of *partiële tabellen*. Binnen de deeltabellen wordt de derde variabele (temperatuur) constant gehouden. We hebben dus een tabel voor warme dagen en een tabel voor koude dagen, die samen de eerdere tabel (tabel 7.1) vormen.

Tabel 7.2 Kruistabellen van ijsconsumptie en verdrinkingen op warme en koude dagen

IJs / Verdr.	0	1	
1	9	81	90
0	1	9	10
	10	90	100

warme dagen

IJs / Verdr.	0	1	
1	9	1	10
0	81	9	90
	90	10	100

koude dagen

IJs / Verdr.	0	1	
1	18	82	100
0	82	18	100
	100	100	200

Wanneer we voor de partiële tabellen gamma's uitrekenen, komen deze beide uit op 0. Er is dus op warme dagen geen verband tussen de ijsconsumptie en het aantal verdrinkingen, en op koude dagen evenmin. Dit betekent dat de eerder gevonden samenhang een schijnsamenhang is, oftewel: de eerder gevonden samenhang is spurieus. De derde variabele (temperatuur) heeft zowel invloed op de ijsconsumpties als op het aantal verdrinkingen.
Dit kunnen we controleren door te kijken naar de kruistabellen van ijsconsumptie naar temperatuur en verdrinkingen naar temperatuur (tabel 7.3). Voor beide kruistabellen geldt dat er een sterk verband is (gamma = 0,99). Hieruit blijkt dat de temperatuur op een dag zowel samenhangt met de ijsconsumptie als met het aantal verdrinkingen.

Als blijkt dat een derde variabele een eerder gevonden verband verklaart, zijn er twee mogelijkheden. Er is sprake van een *antecedente variabele* of van een *interveniërende variabele* (ook wel *mediërende variabele*).

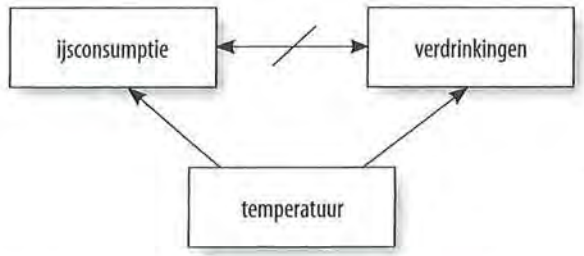
Je spreekt van een antecedente variabele als de derde variabele in de tijd voorafgaat aan de andere twee variabelen. In het voorbeeld ijsconsumptie – verdrinkingen is temperatuur een antecedente variabele, een variabele die voorafgaat aan de ijsconsumptie en aan het aantal verdrinkingen. Wanneer het warmer is,

zwemmen er meer mensen, en wanneer er meer mensen zwemmen, is er meer kans op een verdrinking. Hetzelfde geldt voor de ijsconsumptie: wanneer het warmer is, zullen mensen eerder een ijsje eten dan wanneer het koud is. Deze verbanden zijn verklaarbaar. Dat was bij het eerder gevonden verband tussen ijs eten en verdrinkingen niet het geval (zie figuur 7.2).

Tabel 7.3 Kruistabellen van verdrinkingen en ijsconsumptie naar temperatuur (n = 200)

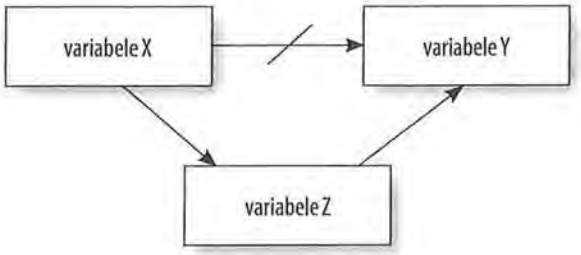
Temp. / Verdrinkingen	0 koud	1 warm	
1 veel	10	90	100
0 weinig	90	10	100
	100	100	200 (dagen)

Temp. / IJs	0 koud	1 warm	
1 veel	10	90	100
0 weinig	90	10	100
	100	100	200 (dagen)



Figuur 7.2 Spurieus verband door antecedente variabele

Het kan ook voorkomen dat de derde variabele beïnvloed is door een van de twee oorspronkelijke variabelen en vervolgens de andere variabele beïnvloedt. Het is dan geen antecedente, maar een interveniërende variabele. De derde variabele Z (die we hier een *mediator* noemen) gaat nu niet vooraf aan X en Y, maar Z wordt beïnvloed door X en Z beïnvloedt vervolgens Y, waardoor er een schijnverband tussen X en Y ontstaat (zie figuur 7.3). Dit wordt ook *mediatie* genoemd.



Figuur 7.3 Spurieus verband met interveniërende variabele

Je vindt bijvoorbeeld een sterke samenhang tussen sekse en het maandelijks inkomen. Mannen hebben een hoger inkomen dan vrouwen. Het is mogelijk dat dit verband verdwijnt als je controleert voor parttime of fulltime werken. Vrouwen werken vaker parttime en hebben daardoor een lager inkomen dan mannen. Parttime of fulltime werken is in dit voorbeeld de interveniërende variabele: sekse (X) → parttime of fulltime werken (Z) → inkomen (Y).

Of een variabele antecedent of interveniërend oftewel mediërend is, moet je meestal op basis van gezond verstand bepalen. Een mediator (zoals in figuur 7.3) moet bijvoorbeeld altijd zowel onafhankelijk als afhankelijk kunnen zijn, omdat deze zowel beïnvloed wordt (door X) als invloed uitoefent (op Y).

Hoe ziet tabelsplitsing er nu uit in SPSS? Eerst bekijken we de kruistabel en de samenhang (gamma) tussen ijsconsumptie en het aantal verdrinkingen zonder de variabele temperatuur (tabel 7.4). Deze tabel is gelijk aan tabel 7.1. Gamma is inderdaad 0,91. Er is een zeer sterke, positieve samenhang tussen ijsconsumptie en het aantal verdrinkingen.

Tabel 7.4 Kruistabel van ijsconsumptie en verdrinkingen (SPSS-output)

Verdrinkingen * ijs Crosstabulation

			ijs		
			0 weing	1 veel	Total
verdrinkingen	1 veel	Count	18	82	100
		% within ijs	18,0%	82,0%	50,0%
	0 weing	Count	82	18	100
		% within ijs	82,0%	18,0%	50,0%
Total		Count	100	100	200
		% within ijs	100,0%	100,0%	100,0%

Symmetric Measures

		Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
Ordinal by Ordinal	Gamma	,908	,032	11,779	,000
N of Valid Cases		200			

Vervolgens splitsen we deze tabellen naar deeltabellen. In deze tabellen is de temperatuur constant gehouden. Er is apart naar de warme en koude dagen gekeken. In de onderste tabel is ook de oorspronkelijke, bivariate tabel te zien. In het onderste deel van tabel 7.5 (*Symmetric Measures*) is te zien dat gamma zowel op warme dagen als op koude dagen 0 is. Er is geen enkel verband tussen ijsconsumptie en het aantal verdrinkingen als de temperatuur constant wordt gehouden. De eerder gevonden samenhang is spurieus, en wordt veroorzaakt door de temperatuur (antecedent).

Tabel 7.5 Tabelsplitsing ijs – verdrinkingen naar warme en koude dagen (SPSS-output)

Verdrinkingen * ijs * temperatuur Crosstabulation

temperatuur				ijs		Total
				0 weing	1 veel	
1 warme dag	verdrinkingen	1 veel	Count	9	81	90
			% within ijs	90,0%	90,0%	90,0%
		0 weing	Count	1	9	10
			% within ijs	10,0%	10,0%	10,0%
	Total		Count	10	90	100
			% within ijs	100,0%	100,0%	100,0%
0 koude dag	verdrinkingen	1 veel	Count	9	1	10
			% within ijs	10,0%	10,0%	10,0%
		0 weing	Count	81	9	90
			% within ijs	90,0%	90,0%	90,0%
	Total		Count	90	10	100
			% within ijs	100,0%	100,0%	100,0%
Total	verdrinkingen	1 veel	Count	18	82	100
			% within ijs	18,0%	82,0%	50,0%
		0 weing	Count	82	18	100
			% within ijs	82,0%	18,0%	50,0%
	Total		Count	100	100	200
			% within ijs	100,0%	100,0%	100,0%

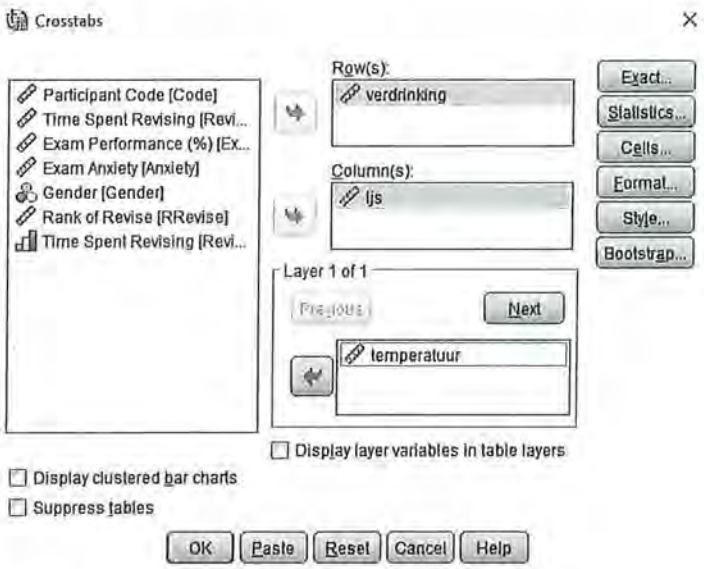
Symmetric Measures

temperatuur				Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
0 koude dag	Ordinal by Ordinal	Gamma	Zero-Order	,000	,556	,000	1,000
	N of Valid Cases			100			
1 warme dag	Ordinal by Ordinal	Gamma	Zero-Order	,000	,556	,000	1,000
	N of Valid Cases			100			
Total	Ordinal by Ordinal	Gamma	Zero-Order	,908	,032	11,779	,000
			First Order Partial	,000			
	N of Valid Cases			200			

SPSS Tabelsplitsing

Tabelsplitsing gebruik je bij kruistabellen. De eerste stappen van tabelsplitsing zijn dan ook dezelfde als bij het maken van een kruistabel. Via *Analyze* → *Descriptive Statistics* → *Crosstabs* kies je eerst de variabelen voor de kolommen en de rijen. Vervolgens plaats je in de *Layer* de derde variabele.

Via *Statistics* kun je nu kiezen voor een bijpassende associatiemaat.



Figuur A Crosstabs-venster: toevoegen van Layer

Kader 7.1

Uiteraard kun je in principe alle associatiematen bij tabelsplitsing gebruiken. Welke associatiemaat je kiest, is afhankelijk van het meetniveau van de variabelen waartussen het bivariate verband wordt berekend. De variabele waar je op 'splitst' kan een nominaal meetniveau hebben, maar dat heeft geen consequenties voor de berekening van de partiële verbanden.

7.1.2 *Specificatie*

Bij tabelsplitsing is het ook mogelijk dat je na controle voor een derde variabele niet constateert dat het oorspronkelijke verband spurieus was, maar dat er wel iets anders aan de hand is.

Stel dat je een onderzoek hebt uitgevoerd onder tieners. Daaruit blijkt dat het bezoeken van online sociale media slecht is voor het zelfbeeld van tieners. Je bent daarom benieuwd of het zien van bepaalde foto's op deze sociale media (voornamelijk zien van selfies, vakantiefoto's of foto's van sportieve prestaties) verband houdt met het zelfbeeld. De variabele 'soort foto' is nominaal en onafhankelijk, en daarom is lambda of Goodman en Kruskalls tau de meest geschikte associatiemaat.

Tabel 7.6 Kruistabel van zelfbeeld naar soort foto (SPSS-output)

Zelfbeeld * Soort_foto Crosstabulation

			Soort_foto			
			1 selfie	2 vakantie-foto	3 sportieve prestatie	Total
Zelfbeeld	1 laag	Count	5	4	1	10
		% within Soort_foto	50,0%	33,3%	12,5%	33,3%
	2 gemiddeld	Count	2	6	3	11
		% within Soort_foto	20,0%	50,0%	37,5%	36,7%
	3 hoog	Count	3	2	4	9
		% within Soort_foto	30,0%	16,7%	50,0%	30,0%
Total		Count	10	12	8	30
		% within Soort_foto	100,0%	100,0%	100,0%	100,0%

Directional Measures

			Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
Nominal by Nominal	Lambda	Symmetric	,189	,170	1,039	,299
		Zelfbeeld Dependent	,211	,175	1,090	,276
		SNS Dependent	,167	,196	,782	,434
	Goodman and Kruskal tau	Zelfbeeld Dependent	,083	,067		,308
		SNS Dependent	,082	,067		,316

Er is een zwak verband tussen het soort foto en het zelfbeeld bij tieners ($\lambda = 0{,}21$, $n = 30$).[2] De helft van de tieners die voornamelijk selfies zien heeft een laag zelfbeeld, en de helft van de tieners die voornamelijk sportieve prestaties op foto's zien, heeft een hoog zelfbeeld.

Maar zou het verband ook aan andere factoren kunnen liggen? We vermoeden dat dit verband voor meisjes anders kan zijn dan voor jongens. We splitsen daarom de tabel op naar de variabele sekse:

Uit tabel 7.7 blijkt dat dit verband alleen bestaat onder meisjes ($\lambda = 0{,}43$), maar niet onder jongens ($\lambda = 0{,}00$). We zien aan de tabel dat er geen meisjes zijn met een 'hoog zelfbeeld' en dat er geen jongens zijn met een 'gemiddeld zelfbeeld'. Waar 71,4% van de meisjes die voornamelijk selfies zien een laag zelfbeeld hebben, heeft 0% van de jongens dat. De meeste jongens hebben een hoog zelfbeeld, maar daarbij maakt het niet zoveel uit welk soort foto's zij voornamelijk op sociale media zien. Het oorspronkelijke verband hebben we nu *gespecificeerd*. We zeggen ook wel dat we een interactie-effect hebben gevonden tussen soort foto en geslacht op het zelfbeeld: het effect van soort foto op zelfbeeld, is voor jongens anders dan voor meisjes.

Tabel 7.7 Tabelsplitsing soort foto – zelfbeeld naar geslacht (SPSS-output)

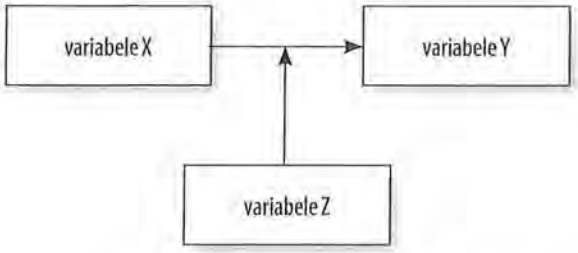
Zelfbeeld * Soort_foto * Sekse Crosstabulation

Sekse				Soort_foto			Total
				1 selfie	2 vakantie-foto	3 sportieve prestatie	
1 meisje	Zelfbeeld	1 laag	Count	5	2	0	7
			% within Soort_foto	71,4%	25,0%	0,0%	38,9%
		2 gemiddeld	Count	2	6	3	11
			% within Soort_foto	28,6%	75,0%	100,0%	61,1%
		Total	Count	7	8	3	18
			% within Soort_foto	100,0%	100,0%	100,0%	100,0%
2 jongen	Zelfbeeld	1 laag	Count	0	2	1	3
			% within Soort_foto	0,0%	50,0%	20,0%	25,0%
		3 hoog	Count	3	2	4	9
			% within Soort_foto	100,0%	50,0%	80,0%	75,0%
		Total	Count	3	4	5	12
			% within Soort_foto	100,0%	100,0%	100,0%	100,0%
Total	Zelfbeeld	1 laag	Count	5	4	1	10
			% within Soort_foto	50,0%	33,3%	12,5%	33,3%
		2 gemiddeld	Count	2	6	3	11
			% within Soort_foto	20,0%	50,0%	37,5%	36,7%
		3 hoog	Count	3	2	4	9
			% within Soort_foto	30,0%	16,7%	50,0%	30,0%
		Total	Count	10	12	8	30
			% within Soort_foto	100,0%	100,0%	100,0%	100,0%

Directional Measures

Sekse				Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
1 meisje	Nominal by Nominal	Lamba	Symmetric	,353	,236	1,279	,201
			Zelfbeeld Dependent	,429	,286	1,177	,239
			Soort_foto Dependent	,300	,221	1,177	,239
		Goodman and Kruskal tau	Zelfbeeld Dependent	,315	,190		,069
			Soort_foto Dependent	,162	,123		,064
2 jongen	Nominal by Nominal	Lamba	Symmetric	,100	,318	,303	,762
			Zelfbeeld Dependent	,000	,667	,000	1,000
			Soort_foto Dependent	,143	,229	,586	,558
		Goodman and Kruskal tau	Zelfbeeld Dependent	,200	,195		,333
			Soort_foto Dependent	,092	,101		,363
Total	Nominal by Nominal	Lamba	Symmetric	,189	,170	1,039	,299
			Zelfbeeld Dependent	,211	,175	1,090	,276
			Soort_foto Dependent	,167	,196	,782	,434
		Goodman and Kruskal tau	Zelfbeeld Dependent	,083	,067		,308
			Soort_foto Dependent	,082	,067		,316

Bij specificatie (wat we ook wel *moderatie* of *interactie* noemen) hanteren we de grove richtlijn van een verschil van minimaal 0,1 tussen de associaties van de verschillende groepen, en ten opzichte van de bivariate associatie (de oorspronkelijk sterkte van de associatiemaat voor de tabelsplitsing).
Wanneer we interactie grafisch zouden weergeven in een conceptueel model, zou dat eruit zien als in figuur 7.4:



Figuur 7.4 Conceptueel model met moderatie / specificatie / interactie

Variabele Z noemen we hier de *moderator*, die het effect van X op Y beïnvloedt.

7.1.3 Versluiering

Tot slot kan tabelsplitsing ons ook helpen als we tegen onze verwachting in *geen* verband tussen twee variabelen vinden. Laten we eens kijken naar tabelsplitsing bij een vierkante tabel (2 × 2) op ordinaal niveau. We kijken naar het verband tussen het opleidingsniveau en algemene kennis (voor dit voorbeeld beide in twee klassen opgedeeld), en hebben de hypothese dat mensen met een hoger opleidingsniveau meer algemene kennis hebben. Een geschikte associatiemaat is dan Somers' d, want de relatie is asymmetrisch. Opleiding is hier de onafhankelijke variabele en kennis de afhankelijke variabele. Wanneer we door SPSS deze maat laten berekenen, levert dit een Somers' d op van –0,02. Er is dus geen verband tussen opleiding en kennis. We voegen nu een derde variabele toe, namelijk televisiekijken (weinig/veel), en krijgen de resultaten die in tabel 7.8 staan.

We zien in tabel 7.8 dat het verband voor mensen die weinig televisiekijken anders is dan voor mensen die veel televisiekijken. Televisiekijken is hier dus een moderator. Waar we in eerste instantie geen verband vonden (d_{yx} = -0,02), vinden we nu bij beide groepen (mensen die weinig televisiekijken en mensen die veel televisiekijken) wél een verband. Bij weinig televisiekijken is het zo dat er een redelijke negatieve samenhang is tussen opleiding en kennis (d_{yx} = -0,39): hoe hoger de opleiding hoe minder de kennis is bij mensen die weinig televisiekijken. Bij mensen die veel televisiekijken is er juist een redelijke positieve samenhang (d_{yx} = 0,34): een hogere opleiding leidt bij mensen die veel televisiekijken tot meer kennis. Hoogopgeleiden steken meer op van veel televisiekijken dan laagopgeleiden.

Tabel 7.8 Tabelsplitsing: kruistabellen van kennis naar opleiding, onder constanthouding van tv-kijken (SPSS-output)

kennis * opleiding * tvkijken Crosstabulation

tvkijken				opleiding		Total
				1 laag	2 hoog	
2 veel	kennis	2 veel	Count	36	35	71
			% within opleiding	53,7%	87,5%	66,4%
		1 weing	Count	31	5	36
			% within opleiding	46,3%	12,5%	33,6%
	Total		Count	67	40	107
			% within opleiding	100,0%	100,0%	100,0%
1 weinig	kennis	2 veel	Count	33	10	43
			% within opleiding	63,5%	25,0%	46,7%
		1 weing	Count	19	30	49
			% within opleiding	36,5%	75,0%	53,3%
	Total		Count	52	40	92
			% within opleiding	100,0%	100,0%	100,0%
Total	kennis	2 veel	Count	69	45	114
			% within opleiding	58,0%	56,3%	57,3%
		1 weing	Count	50	35	85
			% within opleiding	42,0%	43,8%	42,7%
	Total		Count	119	80	199
			% within opleiding	100,0%	100,0%	100,0%

Directional Measures

tvkijken				Value	Asymptotic Standardized Error	Approximate T	Approximate Significance
2 veel	Ordinal by Ordinal	Somers' d	Symmetric	,346	,080	4,115	,000
			kennis Dependent	,338	,080	4,115	,000
			opleiding Dependent	,354	,083	4,115	,000
1 weinig	Ordinal by Ordinal	Somers' d	Symmetric	-,382	,095	-,3,997	,000
			kennis Dependent	-,385	096	-,3,997	,000
			opleiding Dependent	-,380	,095	-,3,997	,000
Total	Ordinal by Ordinal	Somers' d	Symmetric	-,017	,071	-,242	,809
			kennis Dependent	-,017	,072	-,242	,809
			opleiding Dependent	-,017	,070	-,242	,809

We spreken bij deze vorm van moderatie van een *onderdrukt of versluierd* verband.

7.2 Samenvatting

Wanneer we een verband vinden bij een bivariate analyse, kunnen we niets over de causaliteit zeggen. Het kan namelijk zijn dat er een andere factor in het spel is (die je al dan niet gemeten hebt). Wanneer je deze derde variabele constant houdt, kun je in ieder geval meer genuanceerde informatie geven. Wanneer een derde variabele wordt toegevoegd, noemen we dat een multivariate analyse. In dit hoofdstuk zijn de volgende situaties besproken die zich dan kunnen voordoen:

- Bij controle voor een derde variabele verandert er niets. Dan geldt de geconstateerde samenhang dus ook voor elke waarde van de derde variabele.
- Het verband tussen twee variabelen is een schijnsamenhang die wordt verklaard door een derde variabele. Wanneer een verband spurieus is, is er een derde variabele die de samenhang verklaart. Bij controle voor die derde variabele verdwijnt het eerder gevonden verband. Er is dan sprake van een antecedente variabele of mediatie door een interveniërende variabele. Om te bepalen of het een antecedente of een interveniërende variabele is, moet je zoeken naar de logica in de verbanden tussen de variabelen.
- Een derde variabele specificeert een eerder gevonden verband. Als de samenhang voor een bepaalde groep sterker is dan voor een andere groep, spreken we van specificatie, interactie of moderatie. Door middel van tabelsplitsing, waarbij je de derde variabele constant houdt, kun je bepalen of en hoe een derde variabele de samenhang beïnvloedt. Er moet dan een minimaal verschil van 0,1 zijn tussen de partiële associaties, en met de bivariate associatie.
- Een derde variabele kan het verband tussen twee variabelen onderdrukken/versluieren, waardoor een verwacht verband tussen twee variabelen niet wordt gevonden. Na tabelsplitsing blijkt dan dat voor de afzonderlijke waarden van de derde variabele de verbanden tegengesteld zijn.

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

Noten

1 Zie onder andere de website www.tylervigen.com/spurious-correlations.
2 We hadden hier ook voor tau kunnen kiezen.

Associatiematen op interval- en rationiveau

8

In dit hoofdstuk staan de associatiematen op interval- en rationiveau centraal. We behandelen zowel bivariate analyses (correlatie en enkelvoudige regressie) als een multivariate analyse (meervoudige regressie).

In hoofdstuk 6 zagen we dat associatiematen voor variabelen die op minimaal ordinaal niveau zijn gemeten een waarde kunnen aannemen die varieert tussen –1 (perfecte negatieve samenhang) en +1 (perfecte positieve samenhang). Dit is ook het geval bij associatiematen voor variabelen die op interval- of rationiveau zijn gemeten. Er is ook op dit niveau een ordening in de waarden, waardoor je zowel een stijgende als een dalende lijn in de samenhang tussen twee variabelen kunt onderscheiden (zie figuur 8.1).

Figuur 8.1 Grafische weergave van samenhang

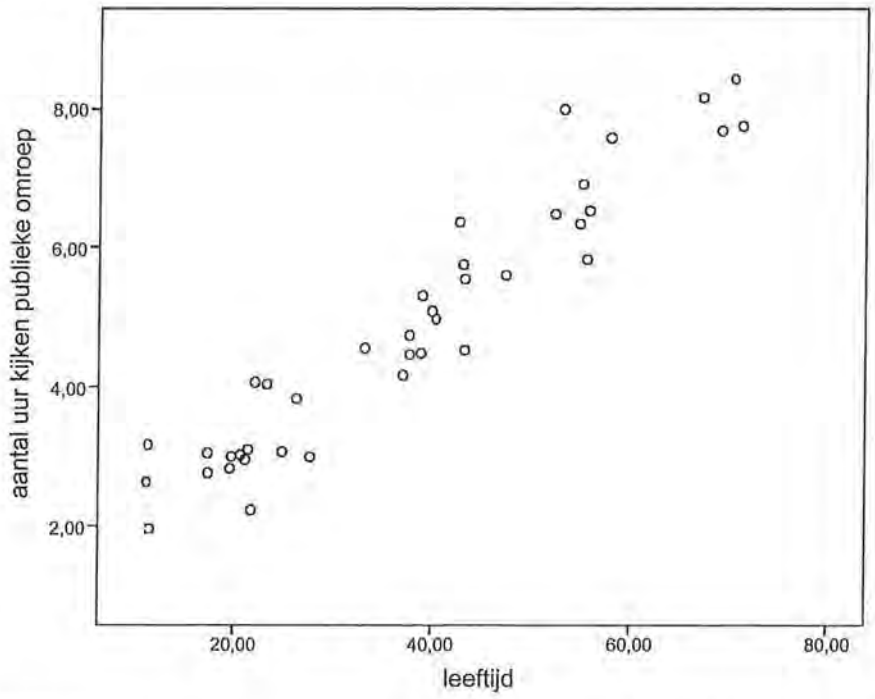
8.1 Pearsons correlatiecoëfficiënt

In hoofdstuk 6 zagen we dat Kendalls tau-b en gamma geschikte associatiematen zijn voor een symmetrische samenhang op ordinaal niveau, en dat we Kendalls tau-b kunnen weergeven in een correlatiematrix. Als maat voor de samenhang tussen twee variabelen die op interval- of rationiveau zijn gemeten, gebruiken we vaak de correlatiecoëfficiënt *r*. Deze correlatiecoëfficiënt is een symmetrische associatiemaat. Bij de berekening is de (on)afhankelijkheid van de variabelen niet van belang. De correlatiecoëfficiënt gebruik je alleen bij variabelen die beide op interval- of rationiveau zijn gemeten. Voluit heet de associatiemaat *Pearson productmoment correlatiecoëfficiënt.* Deze correlatiecoefficiënt duid je aan met de letter *r*.

8.1.1 Grafische weergave

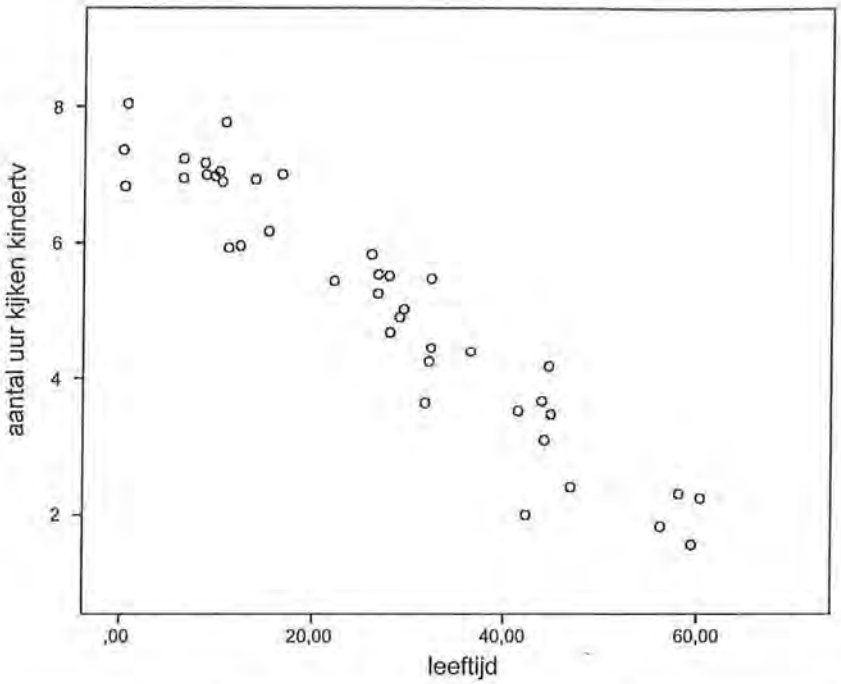
Een samenhang of correlatie is vaak zichtbaar in een *spreidingsdiagram* (*scatterplot*). We zagen deze al kort in de bespreking van Spearmans rho in paragraaf 6.5.1. Een spreidingsdiagram is een puntenwolk die is gebaseerd op de waarnemingen van twee variabelen. Als er sprake is van een onafhankelijke variabele, kiezen we voor deze onafhankelijke variabele de horizontale as, de x-as, en voor de afhankelijke variabele de verticale as, de y-as. Wanneer er geen sprake is van een onafhankelijke en afhankelijke variabele, maakt het niet uit welke variabele je op de x-as en welke je op de y-as plaatst. De waarden die een onderzoekseenheid op de twee variabelen heeft, bepaalt de positie van die onderzoekseenheid binnen dit assenstelsel. Alle onderzoekseenheden samen vormen een puntenwolk.

Wanneer je wilt weten wat de doelgroep van een omroep zou kunnen zijn, zou je bijvoorbeeld kunnen kijken naar de samenhang tussen leeftijd en het aantal uur dat (per week) naar die omroep wordt gekeken.[1] In figuur 8.2 is een spreidingsdiagram gegeven waarin de samenhang is te zien tussen leeftijd en het aantal uur dat iemand per week naar de publieke omroep kijkt. Uit dit spreidingsdiagram blijkt dat er sprake is van een positieve samenhang (positieve correlatie); bij toename van de leeftijd neemt ook de kijktijd naar de publieke omroep toe.
Op basis van deze grafiek zouden we al een voorzichtige conclusie kunnen trekken: naarmate mensen ouder zijn, wordt vaker naar de publieke omroep gekeken. Dit is wellicht voor de publieke omroep aanleiding om voornamelijk programma's voor ouderen uit te zenden.



Figuur 8.2 Spreidingsdiagram van leeftijd en aantal uur naar publieke omroep kijken (SPSS-output)

Op dezelfde manier zouden we kunnen kijken naar de samenhang tussen leeftijd en het kijken naar kindertelevisie. Daar verwachten we eerder een negatieve samenhang; dit is immers gericht op jongeren.

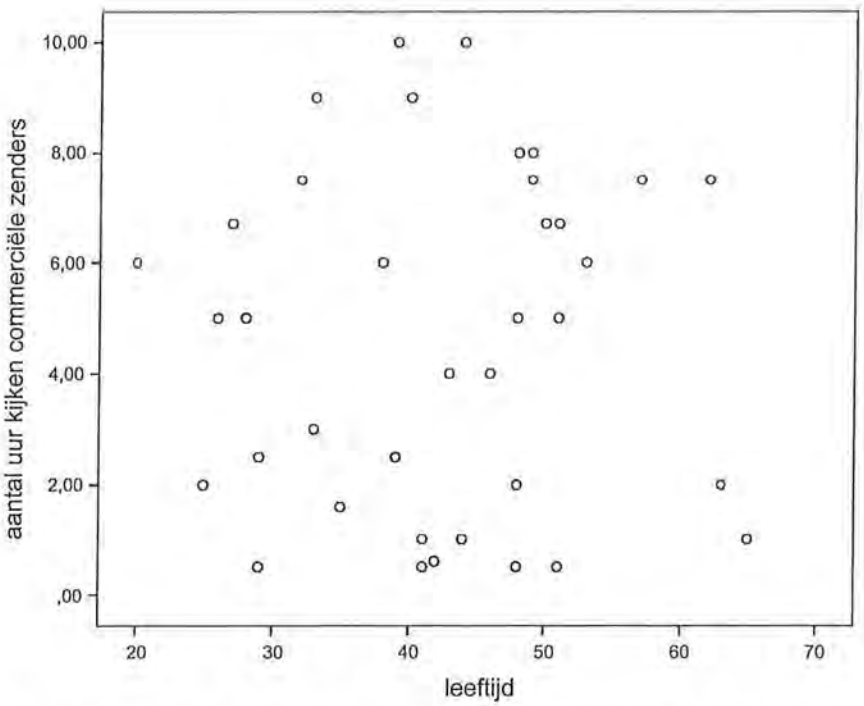


Figuur 8.3 Spreidingsdiagram tussen leeftijd en uren kijken naar kindertelevisie (SPSS-output)

Er blijkt (figuur 8.3) inderdaad een negatieve samenhang uit het spreidingsdiagram. Ouderen kijken minder naar kindertelevisie dan jongeren.

Tot slot een voorbeeld van een spreidingsdiagram waarin de samenhang minder duidelijk is (figuur 8.4). We kijken naar de samenhang tussen leeftijd en het aantal uur dat mensen naar commerciële zenders kijken. We zien dat de puntenwolk meer verdeeld is over alle leeftijden. In dit spreidingsdiagram zien we dus geen samenhang tussen leeftijd en het kijken naar commerciële zenders.

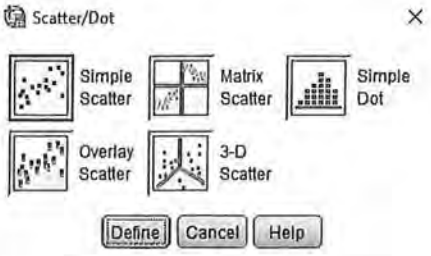
Sommige ouderen kijken veel, anderen weinig, en hetzelfde geldt voor jongeren. We verwachten op basis van dit diagram dus dat de waarde van de maat voor samenhang laag (dat wil zeggen: dicht bij het nulpunt) zal zijn.



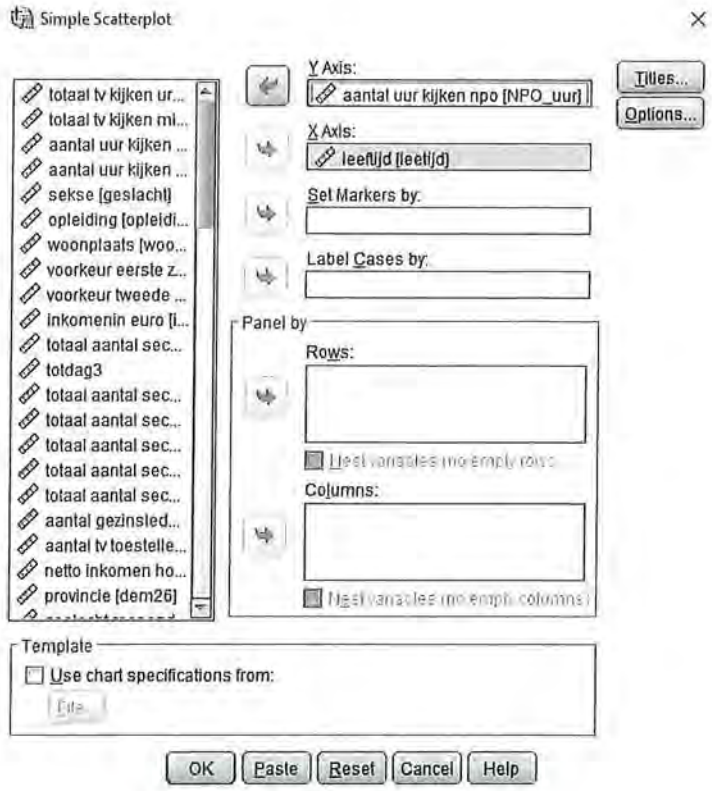
Figuur 8.4 Spreidingsdiagram tussen leeftijd en uren kijken naar commerciële zenders (SPSS-output)

SPSS Het maken van een spreidingsdiagram

Om een spreidingsdiagram te maken zoek je in de menubalk naar *Graphs* en ga je naar *Legacy Dialogs*. Vervolgens kies je in het *Scatter/Dot*-venster (figuur A) voor het eenvoudige spreidingsdiagram (*Simple Scatter*).
In het *Simple Scatterplot* (figuur B) kunnen we dan de variabelen kiezen die we op de x-as en de y-as van het spreidingsdiagram willen hebben. Zet de onafhankelijke variabele op de x-as en de afhankelijke variabele op de y-as.



Figuur A Scatter/Dot-venster



Figuur B Simple Scatterplot-venster

Kader 8.1

8.1.2 *Interpretatie*

Een spreidingsdiagram geeft een eerste indruk van de sterkte en de richting van de samenhang, de waarde van een correlatiecoëfficiënt geeft meer precieze informatie. We bekijken de correlatiecoëfficiënt van het eerste voorbeeld: de samenhang tussen leeftijd en het aantal uur per week dat mensen naar de publieke omroep kijken.

De correlatiecoëfficiënt tussen leeftijd en uren kijken naar de publieke omroep is 0,955 ($r = 0{,}96$). Tabel 8.1 geeft deze waarde tweemaal; één keer voor de samenhang tussen leeftijd en uren kijken naar de publieke omroep en één keer voor de samenhang tussen uren kijken naar de publieke omroep en leeftijd. Uiteraard is deze samenhang hetzelfde, want de berekening van de correlatiecoëfficiënt gaat uit van een symmetrische relatie tussen de twee variabelen.

Tabel 8.1 Correlatie tussen leeftijd en aantal uur kijken naar publieke omroep (SPSS-output)

Correlations

		leeftijd	NPO_uur aantal uur kijken publieke omroep
leeftijd	Pearson Correlation	1	,955**
	Sig. (2-tailed)		,000
	N	40	40
NPO_uur aantal uur kijken publieke omroep	Pearson Correlation	,955**	1
	Sig. (2-tailed)	,000	
	N	40	40

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

De interpretatie is gelijk aan de interpretatie van de overige associatiematen. We zien hier dus een zeer sterke, positieve samenhang tussen leeftijd en uren kijken naar de publieke omroep. Met de leeftijd neemt het kijken naar de publieke omroep toe. De interpretatie van de correlatiecoëfficiënt r is hetzelfde als de interpretatie van de associatiematen op ordinaal niveau.
De correlatiecoëfficiënt kan een waarde aannemen die ligt tussen –1 en +1:

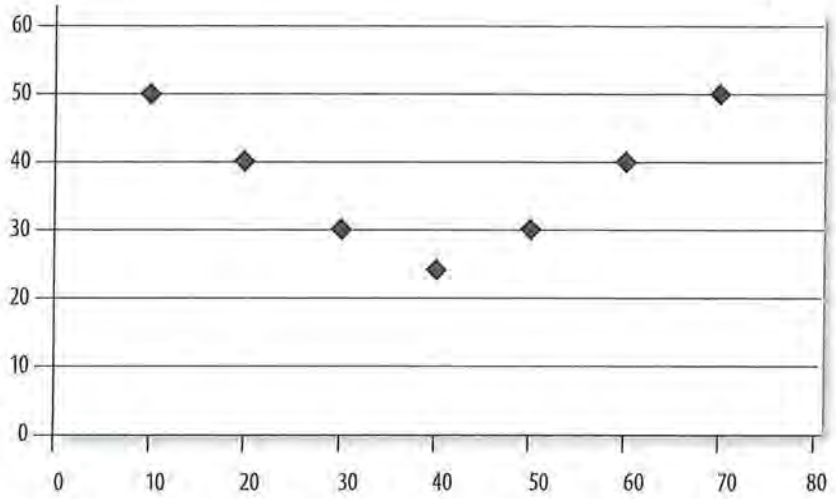
- Correlatiecoëfficiënt = 1: de twee variabelen hangen perfect samen, er is een positief verband, een stijgende lijn (als x stijgt, stijgt y ook, en andersom).
- Correlatiecoëfficiënt = 0: er is geen lineaire (= rechtlijnige) samenhang tussen de twee variabelen.
- Correlatiecoëfficiënt = –1: er is een perfect negatieve samenhang tussen de twee variabelen, er is een dalende lijn (als x stijgt dan daalt y, en andersom).

In dit geval concluderen we dus:

> *Er is een zeer sterke positieve samenhang ($r = 0{,}96$, $n = 40$) tussen leeftijd en het aantal uur dat naar de publieke omroep wordt gekeken. Hoe ouder mensen zijn, hoe vaker ze naar de publieke omroep kijken.*

In de vorige hoofdstukken keken we altijd naar de kruistabel om een indruk van de samenhang te krijgen. Als het gaat om interval- of ratiovariabelen is het verstandig om altijd even naar het spreidingsdiagram te kijken. Daarmee krijg je niet alleen een indruk van de samenhang, maar kun je ook voorkomen dat je op basis van de waarde van r denkt dat er geen samenhang is, terwijl dat wel het geval is. De correlatiecoëfficiënt r geeft namelijk alleen aan in welke mate er sprake is van lineaire samenhang. Als het verband kromlijnig is, is dat niet te zien aan de waarde van r (zie figuur 8.5). Op basis van de waarde van r ($r = 0$) zou je kunnen concluderen dat er geen samenhang bestaat tussen de twee variabelen uit figuur 8.5. Er is inderdaad geen lineair verband, maar uit

het spreidingsdiagram blijkt dat er wel een kromlijnig verband is. Dit zou je gemist hebben als je je alleen op de waarde van r had gebaseerd. De conclusie dat er geen samenhang is, is hier onjuist.



Figuur 8.5 Voorbeeld van een kromlijnig verband

We hebben al gezien in paragraaf 6.5.1 dat een kromlijnig verband ook minder 'krom' kan zijn dan in figuur 8.5 te zien is. Wanneer er veel extreme waarden zijn en daardoor de variabelen niet normaal verdeeld zijn, ontstaat er een sterke kromming die we ook kromlijnig noemen. In dat geval moet je niet de correlatiecoëfficiënt r berekenen maar is het beter Spearmans rho te gebruiken.

8.1.3 Berekening

De basis voor het berekenen van de correlatie is de covariantie en de standaarddeviatie. In hoofdstuk 3 hebben we gezien hoe je de standaarddeviatie berekent. De standaarddeviatie is een soort gemiddelde afstand ten opzichte van het gemiddelde. De variantie is het kwadraat van de standaarddeviatie (oftewel: de standaarddeviatie is de wortel uit de variantie). In hoofdstuk 3 keken we naar de variantie en de standaarddeviatie *binnen* één variabele (bijvoorbeeld in hoeverre de leeftijden van de onderzoekseenheden afweken van het gemiddelde). Het is ook mogelijk om te kijken naar de variantie *tussen* twee variabelen (tussen x en y). Deze vorm van variantie noem je de *covariantie*. De covariantie geeft de mate aan waarin twee variabelen tegelijk variëren. Wanneer er een positieve covariantie is, zullen twee variabelen positief met elkaar correleren. Als de onderzoekseenheden op de ene variabele hoog scoren, zullen ze dat op de andere variabele ook doen. Omgekeerd: wanneer er een negatieve covariantie is, zullen twee variabelen negatief met elkaar correleren. Dit betekent dat onderzoekseenheden die op de ene variabele hoog scoren, op de andere juist laag scoren.

Wanneer je de covariantie tussen de variabelen x en y berekent, noteer je dit als volgt: *Cov (x, y)*.

$$Cov(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

Formule covariantie

Zowel de covariantie tussen x en y als de afzonderlijke standaarddeviaties van x en y komen terug in de formule voor de correlatie.

$$r_{xy}=\frac{Cov(x,y)}{s_x s_y}$$

Formule voor correlatie

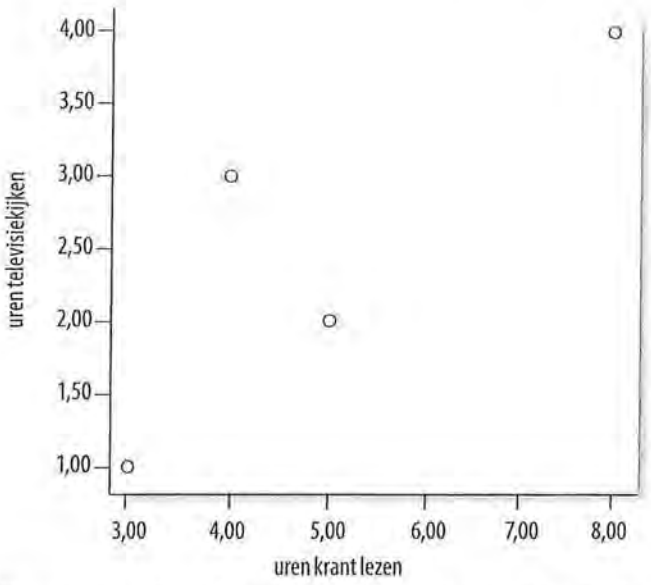
De letter r gebruiken we als symbool voor deze correlatie. De x en de y achter de r geven aan dat het om een correlatie tussen x en y gaat.

De formule van de covariantie lijkt erg op de formule voor variantie, maar nu wordt niet naar één maar naar twee variabelen gekeken (x en y). De basisingrediënten zijn, zoals te zien in de formule, de afwijkingen van x van het gemiddelde van x $(x-\bar{x})$ en de afwijking van y van het gemiddelde van y $(y-\bar{y})$. In de formule van r wordt de covariantie gedeeld door het product van de standaarddeviaties van de twee variabelen. Deze standaarddeviaties zijn zelf ook weer gebaseerd op afwijkingen van het gemiddelde. Dat is dan ook altijd stap 1: het berekenen van de gemiddeldes. Daarna kan de standaarddeviatie van zowel x (s_x) als y (s_y) worden uitgerekend en de covariantie. We zullen dit laten zien aan de hand van een voorbeeld.

We willen nagaan of mensen die vaak de krant lezen ook vaak naar het televisienieuws kijken, en andersom. Dit doen we op basis van de gegevens in tabel 8.2 (krant lezen en televisienieuws kijken is gemeten in uren per week).
Aan het spreidingsdiagram (figuur 8.6) is al te zien dat er een positieve samenhang is tussen het lezen van de krant en televisiekijken.

Tabel 8.2 Datamatrix uren krant en uren tv voor vier respondenten

Respondent	Uren krant (x)	Uren tv (y)
A	3	1
B	4	3
C	5	2
D	8	4



Figuur 8.6 Spreidingsdiagram tussen het lezen van de krant en televisiekijken

Eerst berekenen we het gemiddelde van x en het gemiddelde van y:

$$\overline{x} = \frac{\sum x_i}{n} = \frac{3+4+5+8}{4} = \frac{20}{4} = 5$$

$$\overline{y} = \frac{\sum y_i}{n} = \frac{1+3+2+4}{4} = \frac{10}{4} = 2{,}5$$

Gemiddeld lezen deze vier mensen vijf uur per week de krant, en kijken ze 2,5 uur per week televisie. Vervolgens berekenen we voor zowel het aantal uur krant (x) als het aantal uur televisie (y) de standaarddeviatie (die hebben we immers nodig om de formule van r in te vullen).
Om de standaarddeviatie te berekenen, moeten we eerst de variatie van x (de kwadratensom van het verschil tussen de afzonderlijke x'en en het gemiddelde van x) en de variatie van y (de kwadratensom van het verschil tussen de afzonderlijke y's en het gemiddelde van y) berekenen. Dit doen we op dezelfde manier als al eerder aan bod kwam in hoofdstuk 3. De M wordt in de tabel (tabel 8.3) gebruikt om de gemiddeldes (Mean) aan te geven.

Tabel 8.3 Berekenen van de variatie van x en y

	x	y	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$
A	3	1	$(3-5)=-2$	$-2^2=4$	$(1-2{,}5)=-1{,}5$	$-1{,}5^2=2{,}25$
B	4	3	$(4-5)=-1$	$-1^2=1$	$(3-2{,}5)=0{,}5$	$0{,}5^2=0{,}25$
C	5	2	$(5-5)=0$	$0^2=0$	$(2-2{,}5)=-0{,}5$	$-0{,}5^2=0{,}25$
D	8	4	$(8-5)=3$	$3^2=9$	$(4-2{,}5)=1{,}5$	$1{,}5^2=2{,}25$
Σ	20	10	0	14	0	5
M	5	2,5				

We kunnen nu de standaarddeviaties van beide variabelen berekenen. Daarvoor delen we de variaties door $n-1$. Vervolgens trekken we daaruit de wortel (wanneer we niet zouden worteltrekken zouden we de variantie berekend hebben).

$$s_x=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=\sqrt{\frac{14}{3}}=2{,}160$$

$$s_y=\sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}=\sqrt{\frac{5}{3}}=1{,}291$$

We hebben nu de gegevens voor de noemer van de formule voor r. We hoeven alleen nog maar de covariantie te berekenen. Om deze uit te rekenen hebben we al veel werk verricht. Wat we nog wel moeten berekenen is $\sum(x-\bar{x})(y-\bar{y})$ voor de teller van de formule voor covariantie. We moeten dus voor elke onderzoekseenheid $(x-\bar{x})$ vermenigvuldigen met $(y-\bar{y})$, en deze producten vervolgens bij elkaar optellen.
Aan het sigmateken kunnen we zien dat we dit product voor elke respondent moeten uitrekenen en pas daarna over alle respondenten sommeren:

Hoe je dit op een systematische wijze kunt uitwerken, is te zien in tabel 8.4 (waarvan alleen de laatste kolom nieuwe informatie geeft, de eerdere kolommen bevatten berekeningen die we al in tabel 8.3 hebben gedaan).

Tabel 8.4 Berekenen van de covariantie

	x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
A	3	1	–2	–1,5	–2 * – 1,5 = 3
B	4	3	–1	0,5	–1 * 0,5 = –0,5
C	5	2	0	–0,5	0 * – 0,5 = 0
D	8	4	3	1,5	3 * 1,5 = –4,5
Σ	20	10	0	0	7
M	5	2,5			

Respondent A scoorde 2 uur minder dan het gemiddeld aantal uur krant lezen, 1,5 uur minder dan het gemiddeld aantal uur televisiekijken, en scoort daarom op $(x-\bar{x})(y-\bar{y}) = -2 * -1,5 = 3$. Dit doen we voor elk van de onderzoekseenheden, vervolgens tellen we deze sores bij elkaar op. De som van de producten is 7. We kunnen nu de rest van de formule van de covariantie invullen.

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{7}{3} = 2,333$$

We hebben nu alle gegevens die nodig zijn om de formule van r in te vullen.

$$r_{xy} = \frac{Cov(x,y)}{s_x s_y} = \frac{2,333}{2,160 * 1,291} = 0,837$$

Oftewel: er is een zeer sterke positieve samenhang tussen het aantal uur dat iemand de krant leest en het aantal uur dat hij televisiekijkt. Wanneer het aantal uren dat iemand de krant leest toeneemt, neemt ook het aantal uren dat hij naar de televisie kijkt toe, en andersom.
SPSS laat zien dat onze berekening klopt.

Tabel 8.5 Correlatie tussen uur krant en uur tv (SPSS-output)

Correlations

		uurkrant	uurtv
uurkrant	Pearson Correlation	1	,837
	Sig. (2-tailed)		,163
	N	4	4
uurtv	Pearson Correlation	,837	1
	Sig. (2-tailed)	,163	
	N	4	4

Door de formule voor de covariantie in de formule van r in te vullen kun je r ook meer direct berekenen.

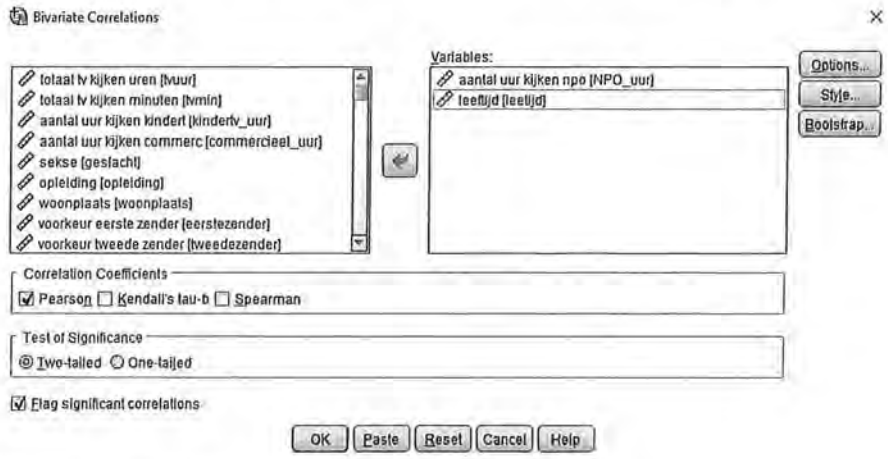
$$r_{xy} = \frac{Cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \div s_x s_y = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Dus je kunt ook de volgende formule voor r gebruiken:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

SPSS Het berekenen van de correlatie

Het berekenen van een correlatie in SPSS gaat via *Analyze → Correlate → Bivariate*. In het vak *Variables* worden de variabelen ingevuld die voor het berekenen van de correlatie nodig zijn. De *Pearson* correlatie staat automatisch aangevinkt.
NB: Via *Bivariate Correlations* kun je SPSS ook Kendalls tau-b en Spearmans rho laten uitrekenen. Je hoeft daarvoor alleen het desbetreffende hokje aan te vinken.



Figuur A Bivariate Correlations-venster: Pearson

Kader 8.2

8.1.4 *Partiële correlaties*

In hoofdstuk 7 zagen we al dat een samenhang tussen twee variabelen beïnvloed kan worden door het toevoegen van een derde variabele. Bij een tabelsplitsing voegde je een derde variabele toe in de *Layers* waardoor partiële associaties werden berekend per categorie van de derde variabele. Bij interval- en ratiovariabelen is het ook mogelijk om de bivariate samenhang te controleren voor een

derde variabele. Anders dan bij de associatiematen op nominaal of ordinaal niveau krijg je nu niet per waarde van de derde variabele een samenhangsmaat (interval- of ratiovariabelen hebben immers vaak erg veel waarden, daardoor zou je door de bomen het bos niet meer kunnen zien), maar wordt een nieuwe correlatie berekend tussen de twee variabelen terwijl je de derde variabele constant houdt.

We leggen dit uit aan de hand van het volgende voorbeeld. Stel je voor dat je onderzoek onder toneelacteurs doet naar het verband tussen plankenkoorts en tekstvastheid. Je vindt tussen de twee variabelen een sterke negatieve correlatie: $r = -0{,}54$ (tabel 8.6). Hoe meer plankenkoorts hoe minder tekstvast de acteurs zijn en andersom (hoe meer tekstvast de acteurs zijn, hoe minder plankenkoorts ze hebben). Vervolgens wil je kijken of een derde variabele een rol speelt in dit verband, en daarom ga je het verband nogmaals onderzoeken, maar nu onder constanthouding van de variabele 'concentratie'. Wanneer je nu eerst tussen deze drie variabelen bivariate correlaties berekent ziet dat er als volgt uit:

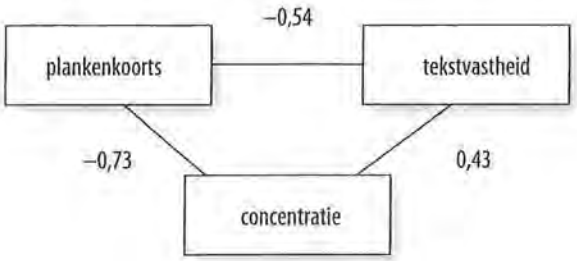
Tabel 8.6 Correlaties van tekstvastheid, plankenkoorts en concentratie (SPSS-output)

Correlations

		Tekstvastheid	Plankenkoorts	Concentratie
Tekstvastheid	Pearson Correlation	1	-,541**	,427**
	Sig. (2-tailed)		,000	,000
	N	103	103	103
Plankenkoorts	Pearson Correlation	-,541**	1	-,729**
	Sig. (2-tailed)	,000		,000
	N	103	103	103
Concentratie	Pearson Correlation	,427**	-,729**	1
	Sig. (2-tailed)	,000	,000	
	N	103	103	103

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

We zien dat concentratie een positieve correlatie heeft met tekstvastheid ($r = 0{,}43$) (dus hoe meer concentratie hoe meer tekstvast de acteur is en omgekeerd) en een negatieve correlatie met plankenkoorts: $r = -0{,}73$ (hoe meer concentratie, hoe minder plankenkoorts en omgekeerd). We kunnen dat nu in een *conceptueel model* tekenen, met concentratie als de mediator:



Figuur 8.7 Conceptueel model met concentratie als mediator

Vervolgens laten we SPSS de correlatie tussen plankenkoorts en tekstvastheid opnieuw berekenen, maar nu controleren we voor de variabele concentratie:

Tabel 8.7 Correlatie plankenkoorts en tekstvastheid waarbij gecontroleerd wordt voor concentratie (SPSS-output)

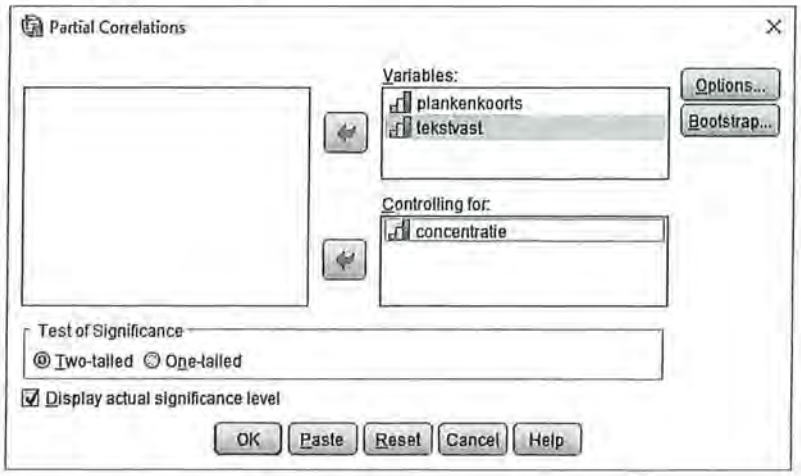
Correlations

Control Variables			Tekstvastheid	Plankenkoorts
Concentratie	Tekstvastheid	Correlation	1,000	-,247
		Significance (2-tailed)	.	,012
		df	0	100
	Plankenkoorts	Correlation	-,247	1,000
		Significance (2-tailed)	,012	.
		df	100	0

Waar de correlatie tussen plankenkoorts en tekstvastheid in eerste instantie −0,54 was, blijkt het verband minder sterk wanneer er gecontroleerd wordt voor de variabele concentratie ($r = -0{,}25$). Dat betekent dat het verband tussen de twee variabelen gedeeltelijk afhangt van de correlatie met concentratie. Er is dus sprake van mediatie. Wanneer het verband geheel zou verdwijnen zou hier sprake zijn van een spurieus verband tussen plankenkoorts en tekstvastheid.

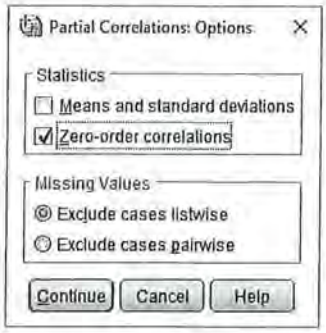
SPSS Het berekenen van partiële correlaties

Om te controleren voor een derde variabele in een correlatieanalyse ga je via *Analyze → Correlate* naar *Partial Correlation*. Bij *Variables* kun je de variabelen invoeren waar je in eerste instantie een correlatie tussen wilt berekenen, bij *Controlling for* zet je de variabele waarvoor je controleert.



Figuur A Partial Correlation-venster

Wanneer je naast de nieuwe, partiële correlatie, de correlaties wilt zien zonder dat gecontroleerd wordt voor een derde variabele, kun je onder *options* het vakje *Zero-order correlations* aanvinken (*Figuur B*).



Figuur B Options-venster

Kader 8.3

8.2 Enkelvoudige regressie

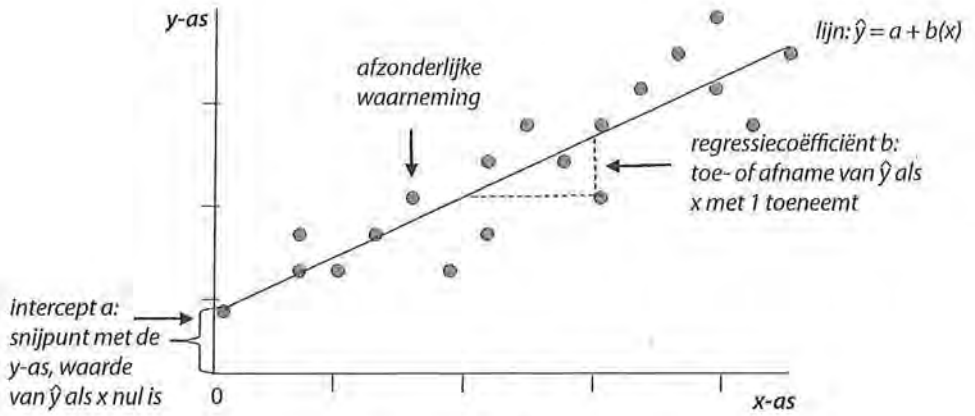
We hebben gezien dat de correlatiecoëfficiënt informatie geeft over de sterkte en de richting van de samenhang. Ook hebben we gezien dat je deze correlatie in een spreidingsdiagram kunt visualiseren. We kunnen op basis van een spreidingsdiagram al een (voorzichtige) conclusie trekken over het verband.
Een regressieanalyse gaat een stapje verder. Bij een regressieanalyse maak je onderscheid tussen afhankelijke en onafhankelijke variabelen. Naast de sterkte en richting van het verband geeft een regressieanalyse een voorspelling van de mate waarin de afhankelijke variabele verandert als gevolg van variatie in de onafhankelijke variabele.
Bij een regressieanalyse wordt dus een relatie tussen een onafhankelijke en afhankelijke variabele verondersteld, en deze is daarmee altijd asymmetrisch. We spreken over een *enkelvoudige regressie* als er één onafhankelijke variabele en één afhankelijke variabele is. De doelstelling van een regressieanalyse is het ontdekken van een patroon in de samenhang tussen x en y, zodat je voorspellingen (in de betekenis van schattingen) kunt doen over y indien x is gegeven. De regressieanalyses die we in dit boek bespreken zijn allemaal lineaire regressies. Met een lineaire regressie bereken je de best passende rechte (= lineaire) lijn door een puntenwolk van een spreidingsdiagram. Dat betekent ook dat, net als bij een Pearsons correlatiecoëfficiënt, er geen sprake mag zijn van een kromlijnig verband, en dat moet je dus altijd eerst nagaan door middel van een spreidingsdiagram. In formule ziet de regressielijn bij een enkelvoudige regressieanalyse er als volgt uit:

$$\hat{y} = a + bx$$

Formule voor enkelvoudige regressie

De lijn is een geschatte lijn. Een regressielijn houdt rekening met alle puntjes in een spreidingsdiagram (de afzonderlijke waarnemingen), en loopt daar zo tussen dat de afstand van alle puntjes tot de lijn minimaal is. Het dakje op de y

geeft aan dat het hier om een schatting of een voorspelling van de waarde van y gaat; $\hat{y}$ is de voorspelde waarde van y.



Figuur 8.8 Grafische weergave van regressielijn

De a in de formule noem je de *intercept* of de *constante*. Dit is het snijpunt van de regressielijn met de y-as, oftewel, a is de voorspelde waarde van y als $x = 0$. De b is de *ongestandaardiseerde regressiecoëfficiënt*. Deze coëfficiënt is bepalend voor de hellingshoek van de lijn. De ongestandaardiseerde regressiecoëfficiënt b geeft aan met hoeveel eenheden de voorspelde waarde van de afhankelijke variabele y verandert als de onafhankelijke variabele x met één eenheid toeneemt.

8.2.1 Berekening

Bij de behandeling van de regressieanalyse bespreken we eerst de wijze van berekenen en dan pas de interpretatie. De interpretatie is namelijk gemakkelijker te begrijpen wanneer je weet hoe je de berekening uitvoert.
Stel, we willen onderzoeken of leeftijd van invloed is op de waardering van het NOS *Journaal*. We vragen drie personen van verschillende leeftijden naar hun waardering van het NOS *Journaal*. Om de waardering van het NOS *Journaal* te meten is een meetinstrument ontwikkeld. De te meten variabele kan de waarde hebben van 0 tot en met 100. De waarde 0 betekent dat de persoon helemaal geen waardering heeft voor het NOS *Journaal* en 100 een zeer hoge waardering. Dit levert de data op die in tabel 8.8 staan.

Tabel 8.8 Datamatrix leeftijd – waardering NOS Journaal

Respondent	Leeftijd	Waardering
A	10	30
B	30	70
C	50	80

Op basis van deze gegevens kun je voorspellen welke waarde y (waardering voor het *NOS Journaal*) zal hebben bij een bepaalde waarde van x (leeftijd). Voor deze voorspelling gebruik je de formule voor de regressielijn: $\hat{y}=a+bx$.
Je hebt de gegevens van de datamatrix nodig om a (de intercept) en b (de ongestandaardiseerde regressiecoëfficiënt) te berekenen. Eén punt van de regressielijn is snel te bepalen, namelijk het punt $(\bar{x},\bar{y})$. Deze twee gemiddeldes liggen op de regressielijn. Van dat gegeven maak je gebruik om de intercept te berekenen. Je kunt de gemiddelden van x en y in de regressievergelijking invullen. Kijken we naar de formule, dan lijkt het logisch te beginnen met het uitrekenen van de intercept (a).

$$a=\bar{y}-b\bar{x}$$

Formule voor de intercept

Maar om de intercept uit te rekenen, heb je de ongestandaardiseerde regressiecoëfficiënt (b) nodig. Je moet daarom wel met de berekening van b beginnen.

$$b=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

Formule voor de ongestandaardiseerde regressiecoëfficiënt

Voor de berekening van b heb je de gemiddeldes van x en y nodig. Je zet alle informatie die je nodig hebt om de formule voor b in te vullen in een tabel. Per onderzoekseenheid bereken je $(x-\bar{x})^2$ (de noemer) en $(x-\bar{x})\,(y-\bar{y})$ (de teller). Pas daarna tel je de gevonden getallen bij elkaar op (zie tabel 8.9). De teller van de ongestandaardiseerde regressiecoëfficiënt wordt hetzelfde berekend als de teller van de covariantie zoals we die in paragraaf 8.1.3 zagen.

Tabel 8.9 Berekenen van de ongestandaardiseerde regressiecoëfficiënt

Respondent	x_i	y_i	$(x-\bar{x})$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(x-\bar{x})\,(y-\bar{y})$
A	10	30	(10 – 30) = –20	400	(30 – 60) = –30	–20 * –30 = 600
B	30	70	(30 – 30) = 0	0	(70 – 60) = 10	0 * 10 = 0
C	50	80	(50 – 30) = 20	400	(80 – 60) = 20	20 * 20 = 400
Σ	90	180		800		1000
M	30	60				

Alle informatie om b te berekenen staat nu in tabel 8.9, zodat je de formule kunt invullen.

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{1000}{800} = 1{,}250$$

Nu kun je ook a berekenen.
$a = \bar{y} - b\bar{x} = 60 - 1{,}25 * 30 = 22{,}500$

De regressievergelijking luidt dus: $\hat{y} = a + bx = 22{,}5 + 1{,}25(x)$

Wat wil dit nu zeggen? De intercept (a) is de voorspelde waarde van y (waardering voor het *NOS Journaal*) wanneer x (leeftijd) de waarde 0 heeft. Letterlijk betekent dit: wanneer iemand nul jaar oud is, zal de waardering voor het *NOS Journaal* 22,5 zijn (want dat was de waarde van de intercept). Een onzinnige voorspelling natuurlijk, wij weten wel beter dan die regressievergelijking. Toch is deze a in de regressievergelijking nodig om voorspellingen te kunnen doen voor elke x (leeftijd). De ongestandaardiseerde regressiecoëfficiënt geeft aan hoeveel de voorspelde waarde van y toeneemt als x met één eenheid toeneemt. Hier zeggen we: het model voorspelt dat wanneer iemand één jaar ouder wordt, de waardering voor het *NOS Journaal* stijgt met 1,25 (op een schaal van 0 tot 100).

Op basis van de intercept en de ongestandaardiseerde regressiecoëfficiënt kun je nu de waardering voor het *NOS Journaal* bij elke willekeurige leeftijd voorspellen door deze leeftijd (x) in de regressievergelijking in te vullen. Je zou bijvoorbeeld kunnen voorspellen hoe een 21-jarige het *NOS Journaal* waardeert:
$\hat{y} = 22{,}5 + 1{,}25 * 21 = 48{,}750$ (op een schaal van 0 tot 100).

We hebben in dit voorbeeld onder onze drie respondenten geen 21-jarige. Als we veel respondenten zouden hebben ondervraagd, met daarin wel een 21-jarige, zouden de scores van een 21-jarige in de puntenwolk die we dan krijgen waarschijnlijk niet precies op de regressielijn liggen. Er blijft variatie rond die regressielijn bestaan, een restvariatie die je niet kunt verklaren met de onafhankelijke variabele. Er zijn namelijk ook andere factoren dan alleen leeftijd die de waardering van het *NOS Journaal* verklaren, maar die hebben we niet met deze regressielijn gemeten. De mate waarin de regressielijn de variatie (of: variantie) verklaart, kunnen we berekenen. Bij regressieanalyse is de *proportie verklaarde variantie* R^2 een belangrijk begrip.[3] Deze lijkt op Goodman en Kruskals tau en lambda (zie hoofdstuk 5). Ook R^2 is gebaseerd op de proportie (of het percentage) voorspellingsverbetering. Ook hier wil je dus bepalen hoe goed de verschillen in de waarden van de onafhankelijke variabele (oftewel: de variantie in de onafhankelijke variabele) de verschillen in de waarden van de afhankelijke variabele verklaren. Het verschil is dat je tau en lambda gebruikt

bij nominale variabelen, en R^2 bij interval- of ratiovariabelen. R^2 is het symbool voor de proportie verklaarde variantie. De formule voor R^2 komt overeen met die van de tau en lambda:

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

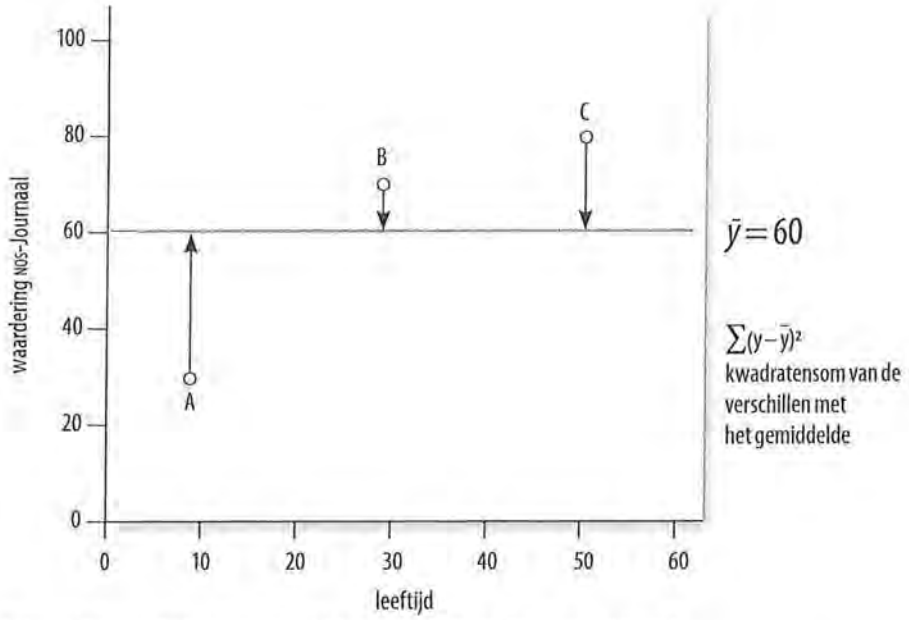
Formule voor proportie verklaarde variantie.

De manier waarop je E_1 en E_2 berekent, is wel anders dan bij tau en bij lambda. Als je bij een interval- of ratiovariabele een voorspelling wilt doen voor y en je daarvoor niet de informatie van x gebruikt maar alleen de gegevens over y, is de beste keuze het gemiddelde van die variabele y. Maar dan maak je voor de afzonderlijke onderzoekseenheden wel een fout die gelijk is aan de waarde van y minus $\bar{y}$. Als je voor alle onderzoekseenheden deze verschillen bij elkaar optelt, is de som 0. Daarom kwadrateren we de verschillen eerst. Die kwadratensom is de *totale variatie*;[4] dit zijn de voorspellingsfouten die je maakt als je het rekenkundig gemiddelde als voorspeller gebruikt.

$$E_1 = \sum (y_i - \bar{y})^2$$

Formule van E_1 bij een regressieanalyse

Een voorbeeld ter verduidelijking. We kijken naar de mate waarin we de waardering van het NOS *Journaal* kunnen voorspellen aan de hand van leeftijd. De waardering voor het NOS *Journaal* is hier de afhankelijke variabele, en bij het berekenen van de E_1 houden we alleen rekening met deze afhankelijke variabele. Van deze variabele kunnen we, omdat het meetniveau hier minimaal interval is, een gemiddelde berekenen. In ons voorbeeld is dat 60. Ongeacht de leeftijd is de gemiddelde waardering voor het Journaal 60 op een schaal van 0 tot en met 100. Persoon A scoorde op waardering van het Journaal 30. De afwijking ten opzichte van het gemiddelde (60) is dus –30. Respondent B scoort met een waardering van 70, 10 punten hoger dan het gemiddelde, en respondent C scoort 20 boven het gemiddelde. Zou je deze afwijkingen bij elkaar optellen, dan kom je uit op nul:
$-30 + 10 + 20 = 0$.
Om die reden kwadrateren we deze afwijkingen afzonderlijk en tellen ze daarna bij elkaar op, zoals we ook bij het berekenen van de variantie (zie hoofdstuk 3) deden. E_1 is dus $-30^2 + 10^2 + 20^2 = 1400$.



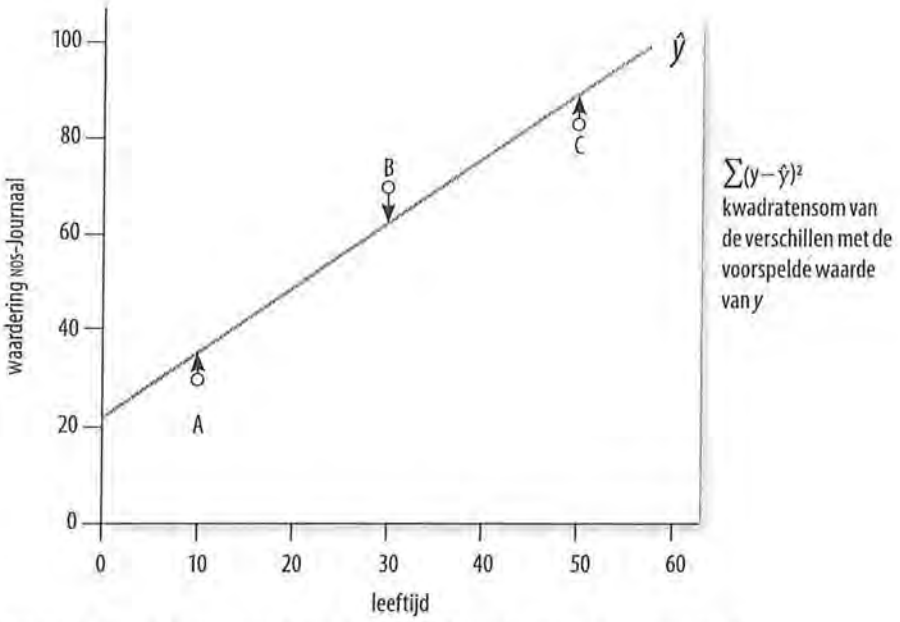
Figuur 8.9 Totale variatie (E_1), afstanden tot het gemiddelde

Bij het berekenen van E_2 maken we wel gebruik van de onafhankelijke variabele. Door de regressievergelijking te gebruiken, pas je informatie over x toe op je voorspelling van y. Ook dan maak je fouten. In dit geval is de voorspellingsfout per onderzoekseenheid y minus $\hat{y}$. De som van de kwadraten van deze verschillen is de *onverklaarde variatie.* De onverklaarde variatie is de voorspellingsfout als je $\hat{y}$ als voorspeller gebruikt. Deze resterende voorspellingsfout, de restvariatie, is gebaseerd op de verschillen met de voorspelde waarden:

$$E_2 = \sum (y - \hat{y})^2$$

Formule van E_2 bij regressieanalyse

De regressielijn in ons voorbeeld was $\hat{y} = 22{,}5 + 1{,}25(x)$. Dat betekent in dit geval dat die lijn voorspelt dat een persoon van 10 jaar oud een waardering heeft van 35 voor het journaal: $\hat{y} = 22{,}5 + 1{,}25 * 10 = 35$. Kijken we in onze datamatrix, dan zien we dat persoon A, die 10 jaar oud is, door de regressielijn wordt overschat, want deze persoon heeft een waardering van 30. Persoon A heeft dus een afwijking van 30 – 35 = -5 ten opzichte van de regressielijn. Voor een 30-jarige voorspelt de regressielijn: $\hat{y} = 22{,}5 + 1{,}25 * 30 = 60$. Deze persoon wordt door de regressielijn dus onderschat, want de score van de 30-jarige in onze datamatrix is 70: 70 – 60 = 10. Tot slot zien we op dezelfde manier dat persoon C als 50-jarige –5 ten opzichte van de regressielijn scoort: $\hat{y} = 22{,}5 + 1{,}25 * 50 = 85$ en 80 – 85 = –5. Ook deze waarden zijn bij elkaar opgeteld nul, dus kwadrateren we ze afzonderlijk voordat we deze bij elkaar optellen om tot E_2te komen: –52 + 102 + –52 = 150.



Figuur 8.10 Onverklaarde variatie (E_2), afstanden tot de regressielijn

We zien in figuur 8.10 dat de afstanden van de punten van de onderzoekseenheden tot de regressielijn kleiner zijn dan de afstanden van deze punten tot het gemiddelde in figuur 8.9. Wanneer alle waarnemingen precies op de lijn zouden liggen, zouden de verschillen in leeftijd perfect de verschillen in waardering verklaren. Nu is er echter een kleine afstand tussen A en C en de lijn, en een iets grotere afstand tussen B en de lijn. Er is dus een gedeelte dat niet wordt verklaard door de lijn: de onverklaarde variatie. Dit noem je ook wel het *residu*.

Wanneer we deze informatie samenvoegen, komen we tot de volgende formule voor de proportie verklaarde variantie.

$$R^2 = \frac{E_1 - E_2}{E_1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{totale variatie} - \text{onverklaarde variatie}}{\text{totale variatie}}$$

Rekenformule voor de proportie verklaarde variantie

Net als tau en lambda heeft R^2 een minimale waarde van 0 en een maximale waarde van 1.

We laten de hele berekening van de R^2 nogmaals zien voor ons voorbeeld (leeftijd en waardering) in een overzichtstabel.

Tabel 8.10 Berekenen van de proportie verklaarde variantie (R^2)

Respondent	x_i	y_i	$(y-\bar{y})$	$(y-\bar{y})^2$	$\hat{y} = a + b\,x$	$(y-\hat{y})$	$(y-\hat{y})^2$
A	10	30	–30	900	22,5 + 1,25 * 10 = 35	(30 – 35) = –5	25
B	30	70	10	100	22,5 + 1,25 * 30 = 60	(70 – 60) = 10	100
C	50	80	20	400	22,5 + 1,25 * 50 = 85	(80 – 85) = –5	25
Σ	90	180	0	1400			150
M	30	60					

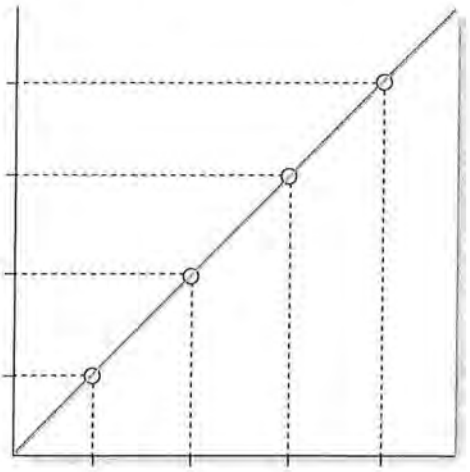
De totale variatie (E_1) is dus 1400, de onverklaarde variatie (E_2) is 150.

Nu we alle informatie hebben, kunnen we de formule invullen.

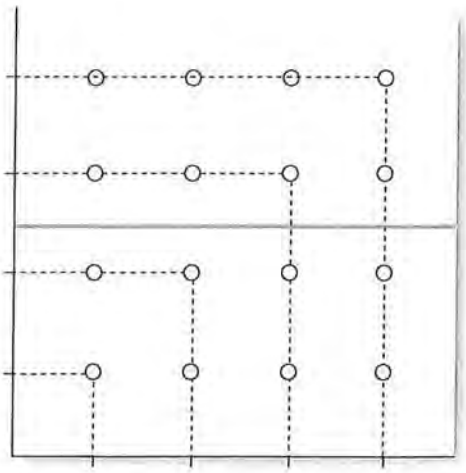
$$R^2 = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = \frac{1400-150}{1400} = 0{,}893$$

We kunnen concluderen dat de varia(n)tie in de waardering voor het *NOS Journaal*, voor 89,3% verklaard wordt door de variantie in de variabele leeftijd. De regressievergelijking is dus een goed verklaringsmodel.

R^2 kan waarden aannemen die liggen tussen de 0 (0% verklaring) en de 1 (100% verklaring). Bij een perfecte verklaring liggen alle punten precies op de regressielijn; er is dan geen restwaarde of residu ($\sum(y-\hat{y})^2 = 0$). Als 0% wordt verklaard, valt de regressielijn samen met de gemiddelde waarde van y en zijn de verschillen met het gemiddelde gelijk aan de restwaarden $\sum(y-\bar{y})^2 = \sum(y-\hat{y})^2$ (zie figuur 8.11 en 8.12).



Figuur 8.11 $R^2 = 1$



Figuur 8.12 $R^2 = 0$

8.2.2 *Interpretatie*

Nu we hebben gezien hoe je de berekeningen van een enkelvoudige regressieanalyse uitvoert, zal de interpretatie van de SPSS-output gemakkelijker zijn. We nemen nog steeds het voorbeeld van de samenhang tussen leeftijd en de waardering van het NOS *Journaal*. We bekijken in kleine stappen de tabellen die SPSS als output van een regressieanalyse geeft (zie kader 8.4 voor de wijze waarop je SPSS een regressieanalyse kunt laten uitvoeren).

Tabel 8.11 Regressieanalyse: intercept (SPSS-output)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	22,500	14,790		1,521	,370
	leeftijd	1,250	,433	,945	2,887	,212

a. Dependent Variable: waardering

De eerste tabel van SPSS die we hier bespreken is de coëfficiëntentabel. Deze zie je in tabel 8.11. De onafhankelijke variabele (in dit geval leeftijd) staat in de linkerkolom van deze coëfficiëntentabel. Wat de afhankelijke variabele (in dit geval waardering) is, wordt onder deze tabel aangegeven. De waarde die meteen onder de B staat, achter (Constant), is de intercept, het snijpunt met de *y*-as (*a*). Deze bedraagt 22,50, zoals we zelf al hadden berekend. De letterlijke betekenis van dit getal is: wanneer iemand nul jaar oud is, is de voorspelde waarde voor de waardering van het NOS *Journaal* 22,50.

Tabel 8.12 Regressieanalyse: ongestandaardiseerde regressiecoëfficiënt (SPSS-output)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	22,500	14,790		1,521	,370
	leeftijd	1,250	,433	,945	2,887	,212

a. Dependent Variable: waardering

De waarde van de ongestandaardiseerde regressiecoëfficiënt *b* wordt onder de intercept (*Constant*) gegeven, achter de onafhankelijke variabele (tabel 8.12). Dit is de waarde waarmee de voorspelling van *y* verandert als *x* met één eenheid toeneemt. Dit is het effect van *x* op *y*. Hier betekent deze waarde dus: wanneer de leeftijd met één jaar toeneemt, stijgt de waardering van het NOS *Journaal* met 1,25.

Tabel 8.13 Regressieanalyse: gestandaardiseerde regressiecoëfficiënt (SPSS-output)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	22,500	14,790		1,521	,370
	leeftijd	1,250	,433	,945	2,887	,212

a. Dependent Variable: waardering

In de kolom *Standardized Coefficients* staat bèta (β). De β is het zuivere effect van de onafhankelijke variabele leeftijd op de afhankelijke variabele waardering. Dit wordt ook wel de *gestandaardiseerde regressiecoëfficiënt* genoemd, en heet zo omdat deze niet afhankelijk is van de meeteenheden van de variabelen. Bèta neemt in de regel alleen waarden aan die tussen –1 en +1 liggen. Bij een enkelvoudige regressieanalyse (regressieanalyse met één onafhankelijke variabele) is bèta altijd gelijk aan de correlatiecoëfficiënt r. We zien dat er een zeer sterke, positieve samenhang is tussen leeftijd en waardering ($\beta = 0{,}95$).

In de tabel *Model Summary* geeft SPSS de proportie verklaarde variantie (R^2) en de multipele correlatie (tabel 8.14). De wortel uit R^2 is R. R is een multipele correlatiecoëfficiënt (zie paragraaf 8.3). Als er één onafhankelijke variabele is, is deze R gelijk aan $|\, r \,|$, de absolute waarde van de correlatie r. Bij meerdere onafhankelijke variabelen is R niet gelijk aan $|\, r \,|$. We zullen hier verder op ingaan in de volgende paragraaf.

Tabel 8.14 Regressieanalyse: proportie verklaarde variantie (SPSS-output)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,945[a]	,893	,786	12,24745

a. Predictors: (Constant), leeftijd

Zoals we eerder al hadden berekend, blijkt de onafhankelijke variabele leeftijd 89,3% van de varia(n)tie in de afhankelijke variabele waardering te verklaren. SPSS geeft ook de waarde van de totale variatie (E_1) en van de onverklaarde variatie (E_2) in een aparte tabel, met als titel ANOVA (tabel 8.15). De totale variatie wordt aangeduid met *Total*, de onverklaarde variatie wordt met *Residual* aangeduid.

Aan de hand van tabel 8.15 kun je ook zelf R^2 uitrekenen.

$$R^2 = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = \frac{E_1 - E_2}{E_1} = \frac{\text{Total} - \text{Residual}}{\text{Total}} = \frac{1400-150}{1400} = 0{,}893$$

Tabel 8.15 Berekening proportie verklaarde variantie (SPSS-output)

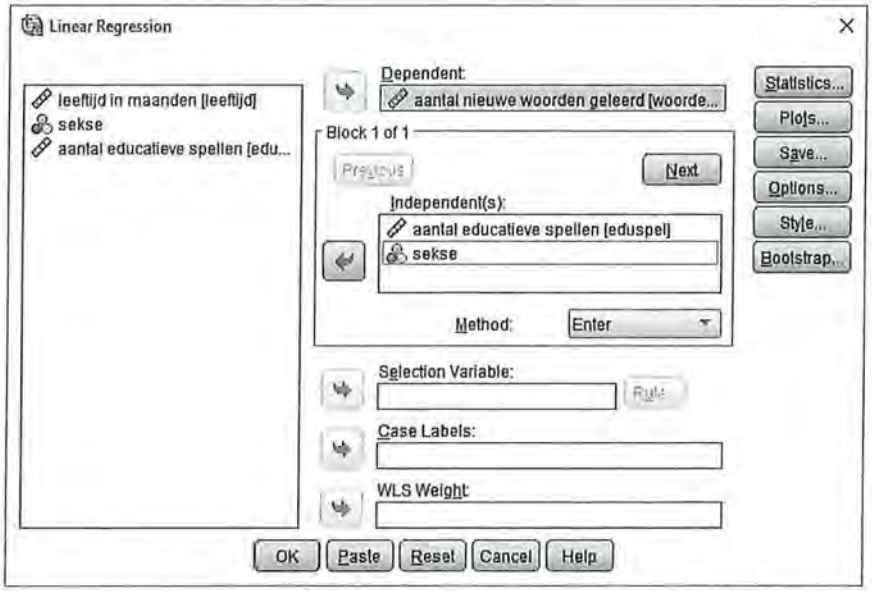
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1250,000	1	1250,000	8,333	,212[b]
	Residual	150,000	1	150,000		
	Total	1400,000	2			

a. Dependent Variable: waardering
b. Predictors: (Constant), leeftijd

SPSS Het uitvoeren van een regressieanalyse

Het uitvoeren van een regressieanalyse in SPSS doe je met *Analyze → Regression → Linear.* In het vakje *Dependent* voer je de afhankelijke variabele in en in het vakje *Independent(s)* één of meerdere onafhankelijke variabelen (voor een meervoudige regressieanalyse zie paragraaf 8.3).



Figuur A Linear Regression-venster

Kader 8.4

We behandelen nog een ander voorbeeld. We kijken weer naar de woordenschat van peuters wanneer ze educatieve spellen op een tablet spelen, maar meten deze variabelen nu allemaal op rationiveau. Een week lang wordt geteld hoe vaak de peuters een educatief spel spelen, en wordt het aantal nieuwe woorden dat zij zeggen in die week geteld. De verwachting is dat peuters die meer educatieve spellen hebben gespeeld (de onafhankelijke variabele) meer nieuwe woorden leren in de week (de afhankelijke variabele). Beide variabelen hebben een ratio meetniveau en we veronderstellen een asymmetrisch verband tussen

de twee variabelen. Een enkelvoudige regressieanalyse is daarom de meest geschikte analysetechniek om dit te onderzoeken. SPSS laat vier tabellen zien, waarvan er twee nodig zijn om een antwoord te geven op de verwachting (tabel 8.16).

Tabel 8.16 Regressieanalyse aantal educatieve spellen en woordenschat (SPSS-output)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,536[a]	,287	,269	1,57891

a. Predictors: (Constant), eduspel aantal educatieve spellen

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7,987	,839		9,516	,000
	eduspel aantal educatieve spellen	,160	,041	,536	3,915	,000

a. Dependent Variable: woordenschat aantal nieuwe woorden geleerd

Je kijkt eerst naar de intercept. Deze is 7,987. Letterlijk betekent deze waarde: wanneer x nul is, is de verwachte waarde van y 7,99. In ons voorbeeld concluderen we: het model voorspelt dat wanneer geen educatieve spellen op de tablet worden gespeeld, de peuter 7,99 nieuwe woorden per week leert.
De ongestandaardiseerde regressiecoëfficiënt (b) bedraagt 0,160. Letterlijk betekent dit: wanneer x met één eenheid stijgt, stijgt de verwachte waarde van y met 0,16. Hier stellen we dus dat het model voorspelt dat wanneer het aantal educatieve spellen spelen toeneemt met één keer, de woordenschat toeneemt met 0,16 woorden. Aan de gestandaardiseerde regressiecoëfficiënt (β) zien we dat dit verband positief en sterk is ($\beta = 0{,}54$). Deze is gelijk aan de R, wat weer de wortel is uit R^2.
Om te kijken of het aantal educatieve spellen spelen een goede verklaring is voor de woordenschat van peuters, kijken we naar de R^2. Deze is 0,287. De variantie in het aantal educatieve spellen spelen, verklaart voor 28,7% de variantie in de woordenschat. Met andere woorden: de mate waarin peuters educatieve spellen spelen op een tablet, verklaart voor 28,7% de mate waarin peuters wekelijks nieuwe woorden leren.

In een onderzoeksverslag zouden we dit als volgt kunnen beschrijven:

> *Uit een enkelvoudige regressieanalyse blijkt dat we vrij goed het aantal nieuwe woorden dat een peuter wekelijks leert kunnen voorspellen aan de hand van het aantal keer educatieve spellen op een tablet spelen ($\beta = 0{,}54$, $n = 40$).[5] De proportie verklaarde variantie is 28,7%. 71,3% van de*

variantie in het leren van nieuwe woorden kan dus niet verklaard worden door het spelen van educatieve spellen op een tablet. Peuters die geen educatieve spellen spelen, leren 7,99 nieuwe woorden per week. Er is een toename van 0,16 nieuwe woorden per week wanneer de frequentie van de educatieve spellen spelen toeneemt met één.[6]

8.3 Meervoudige regressieanalyse

In de vorige paragraaf keken we naar de invloed van één onafhankelijke variabele op één afhankelijke variabele. Bij een meervoudige regressieanalyse kijk je naar het voorspelde effect van meerdere onafhankelijke variabelen. Er is nog steeds maar één afhankelijke variabele. We zullen de meervoudige regressieanalyses niet met de hand berekenen zoals we dat bij de enkelvoudige hebben gedaan, maar ons voornamelijk richten op de interpretatie.
We hebben gezien dat het aantal nieuwe woorden dat geleerd wordt redelijk (voor 28,7%) wordt verklaard door het spelen van educatieve spellen. Daarbij hebben we geen rekening gehouden met andere variabelen. Misschien dat andere factoren ook een rol spelen. Het zou kunnen zijn dat meisjes sneller nieuwe woorden leren dan jongens, of misschien heeft leeftijd wel een grote verklarende invloed en gaat het leren van nieuwe woorden bij drie- en vierjarigen sneller dan bij tweejarigen. Bij een meervoudige regressieanalyse is het mogelijk om ook deze variabelen op te nemen in het verklaringsmodel. Een voorwaarde voor een regressieanalyse was echter wel dat alle variabelen minimaal op intervalniveau gemeten zouden zijn, en 'sekse' is een nominale variabele. Voordat we kijken naar de interpretatie van een regressieanalyse met meerdere onafhankelijke variabelen, bespreken we daarom eerst hoe je nominale variabelen toch in een regressieanalyse op kunt nemen.

8.3.1 Dummyvariabelen

We hebben al gezien dat je met nominale variabelen niet kunt rekenen. We willen echter vaak deze variabelen wel in onze analyses betrekken, zoals de variabele 'geslacht'. Het zou zonde zijn als we met de variabele geslacht alleen kruistabellen kunnen uitvoeren, terwijl het een variabele is die in veel onderzoek meegenomen zal worden. Rekenen met nominale variabelen is mogelijk wanneer we van deze nominale variabelen *dummyvariabelen* maken. Een dummyvariabele is een variabele die dichotoom is, of dichotoom is gemaakt. Een *dichotome variabele* is een variabele die slechts twee mogelijke waarden kan aannemen, bijvoorbeeld man-vrouw, of goed-fout, of ja-nee. In het geval van een dummyvariabele geven we deze waarden altijd de waarden 0 en 1. In het geval van sekse maakt het niet uit of je de man de waarde 1 geeft of de waarde 0. Bij een variabele waarbij je alleen met ja of nee kunt antwoorden ligt het voor de hand om 'nee' de waarde 0 te geven en 'ja' de waarde 1. De waarde

1 geeft dan aan dat de onderzoekseenheid het kenmerk wel heeft en de waarde 0 dat dit niet het geval is. Je zou bijvoorbeeld aan iemand de vraag kunnen stellen: 'ben je een man'? Is het antwoord nee (waarde 0), dan kan het niet anders dan dat deze persoon een vrouw is. We zullen in de volgende paragraaf zien dat wanneer we op deze manier (met nullen en enen) nominale variabelen coderen, we deze toch in een meervoudige regressieanalyse kunnen opnemen.

Sommige nominale variabelen zijn altijd dichotoom, hebben dus altijd slechts twee waarden (zoals sekse), maar een variabele met meerdere waarden kan ook worden gedichotomiseerd tot dummyvariabelen. Wanneer je hebt gevraagd op welke partij iemand stemt, resulteert dat in een nominale variabele met verschillende antwoordcategorieën. Partijvoorkeur is een nominale variabele en daarom niet geschikt voor gebruik in een regressieanalyse. We maken dan in SPSS nieuwe dummyvariabelen (met behulp van *Recode*, zie paragraaf 4.4), waarin we in feite steeds de vraag stellen: 'heeft de respondent op deze partij gestemd?' waarbij steeds geantwoord kan worden met 0 = nee of 1 = ja.

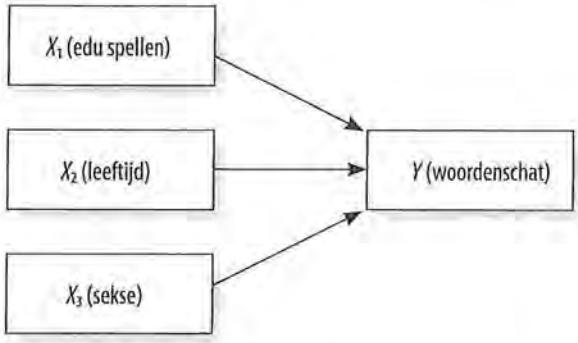
Stel, we hebben respondenten in een enquête de vraag voorgelegd: 'op welke partij zou u bij de komende verkiezingen stemmen?' en ze kunnen (voor de overzichtelijkheid van ons voorbeeld) kiezen uit vier mogelijke antwoorden: 1 = SP, 2 = PvdA, 3 = VVD, 4 = overige partijen. Deze variabele is nominaal, er zit geen rangordening in, er kan niet met de waarden van de antwoorden gerekend worden. We maken nu van deze nominale variabele met vier waarden, drie nieuwe dummyvariabelen. De eerste variabele noemen we 'SP', en heeft de waarden 0 = nee, ik zou niet op de SP stemmen, en 1 = ja, ik zou wel op de SP stemmen. De tweede variabele noemen we 'PvdA', en heeft ook weer twee waarden: 0 = nee, ik zou niet op de PvdA stemmen en 1 = ja, ik zou wel op de PvdA stemmen. En de derde variabele noemen we 'VVD', en heeft ook weer twee waarden: 0 = nee, ik zou niet op de VVD stemmen en 1 = ja, ik zou wel op de VVD stemmen. Voor de laatste categorie, overige partijen, hoeven we geen nieuwe dummyvariabele te maken, omdat wanneer op zowel 'SP' als 'PvdA' als 'VVD' 0 (nee) wordt geantwoord, de respondent automatisch op een van de overige partijen zal stemmen. Bij het maken van een dummyvariabele maak je dus altijd 'het aantal categorieën min 1' aantal dummyvariabelen.

Dummyvariabelen gebruiken we alleen in meervoudige regressieanalyses, en niet in enkelvoudige regressieanalyses. In hoofdstuk 9 zullen we bespreken welke analyse het meest geschikt is wanneer je één nominale onafhankelijke variabele hebt (al dan niet dichotoom) en een interval- of ratiovariabele als afhankelijke variabele. In de volgende paragraaf zullen we laten zien hoe je de dummyvariabelen in een meervoudige regressieanalyse kunt interpreteren.

8.3.2 *Interpretatie meervoudige regressieanalyse*

Een voorwaarde voor een meervoudige regressie is dat de afhankelijke variabele minimaal intervalniveau is, en dat er minimaal één onafhankelijke variabele op minimaal intervalniveau gemeten is. De andere onafhankelijke variabelen kunnen dummyvariabelen of ook interval- of ratiovariabelen zijn.

We gaan verder met het voorbeeld van de mogelijke toename van de woordenschat van peuters bij het spelen van educatieve spellen op een tablet. We voegen nu de variabelen 'leeftijd in maanden' en 'sekse' toe, waarbij we de waarde 0 hebben toegekend aan meisjes, en de waarde 1 aan jongens. Alleen op deze manier mogen we de nominale variabele immers in een meervoudige regressieanalyse gebruiken.



Figuur 8.13 Effect van meerdere onafhankelijke variabelen op een afhankelijke variabele

De eerdere formule voor de regressievergelijking is nu 'verdrievoudigd':

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3$$

Formule voor meervoudige regressie bij drie onafhankelijke variabelen

We hebben drie onafhankelijke variabelen: educatieve spellen (x_1), leeftijd in maanden (x_2) en geslacht (x_3). Elke onafhankelijke variabele heeft zijn eigen ongestandaardiseerde regressiecoëfficiënt (b). Uit de formule blijkt dat er nog steeds maar één intercept is. Dit is het punt waarop de y-as wordt gesneden als alle x'en de waarde nul hebben.
De ongestandaardiseerde regressiecoëfficiënt van een onafhankelijke variabele geeft het effect van die variabele op y weer als de andere onafhankelijke variabelen niet veranderen (constant worden gehouden). Zo is b_1 het effect dat x_1 heeft op y, onder het constant houden van de overige x'en. Wanneer je kijkt naar b_1 om het effect van 'educatieve spellen' op de woordenschat vast te stellen, houd je leeftijd en sekse constant, zoals we dat ook al eerder deden bij tabelsplitsing en partiële correlatie. Hoe een meervoudige regressie er in SPSS-output uitziet, laat tabel 8.17 zien.

Tabel 8.17 Meervoudige regressie met drie onafhankelijke variabelen (SPSS-output)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,729[a]	,531	,475	1,57232

a. Predictors: (Constant), sekse, leeftijd leeftijd in maanden, eduspel aantal educatieve spellen

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6,961	1,136		6,130	,000
	eduspel aantal educatieve spellen	,143	,048	,526	3,261	,002
	leeftijd leeftijd in maanden	2,014	,026	,181	,520	,606
	sekse	-1,687	,531	-,386	-1,293	,204

a. Dependent Variable: woordenschat aantal nieuwe woorden geleerd

De interpretatie begint bij de coëfficiëntentabel (met de titel *Coefficients*). De waarde van de intercept, achter *Constant*, is 6,961. Dat wil zeggen dat wanneer alle onafhankelijke variabelen nul zijn, het voorspelde aantal nieuwe woorden dat wordt geleerd, 6,96 is. Omdat sekse in deze analyse de waarden 0 en 1 heeft (meisjes hadden de waarde 0, jongens de waarde 1), kun je stellen dat meisjes (sekse = 0) van nul maanden (leeftijd = 0) die geen educatieve spellen spelen (eduspel = 0), 6,96 nieuwe woorden per week leren. Dit is dus het punt waarop de regressielijnen de y-as snijden.

We kijken nu naar de eerste ongestandaardiseerde regressiecoëfficiënt, b_1, achter de variabele 'eduspel'. Deze heeft een waarde van 0,143. Wanneer x met één eenheid stijgt, neemt de woordenschat toe met 0,14 als de andere onafhankelijke variabelen ongewijzigd blijven. Het model voorspelt dus dat peuters 0,14 meer nieuwe woorden leren wanneer ze één keer vaker een educatief spel op een tablet spelen, onder het constant houden van leeftijd en sekse. Dit verschilt niet veel met de toename van 0,16 die we zagen bij de enkelvoudige regressieanalyse (tabel 8.16). Het verschil is dat je nu rekening houdt met de andere twee onafhankelijke variabelen. Op basis van de regressievergelijking kun je bijvoorbeeld voorspellen dat een meisje van 25 maanden oud dat vier keer per week een spel speelt, 0,16 nieuwe woorden meer leert dan een meisje van 25 maanden oud dat drie keer per week een spel speelt.

Bij de ongestandaardiseerde regressiecoëfficiënt van leeftijd, b_2, zien we een waarde van 2,014. Dat wil zeggen: wanneer de leeftijd toeneemt met één, dus als een peuter één maand ouder wordt, stijgt het aantal nieuwe woorden met 2,01, onder constanthouding van het 'aantal educatieve spellen' dat gespeeld wordt

en 'sekse'. Een meisje van 25 maanden dat drie keer per week een educatief spel speelt, leert 2,01 nieuwe woorden per week meer dan een meisje van 24 maanden dat drie keer per week een educatief spel speelt.

De laatste ongestandaardiseerde regressiecoëfficiënt, b_3 van sekse, heeft een interpretatie die iets afwijkt van de eerdere twee. Omdat we hier te maken hebben met een dichotome variabele, kunnen we hier iets zeggen over het *gemiddelde* verschil tussen de twee categorieën (meisje-zijn of jongen-zijn). De waarde van b_3 is –1,687. Letterlijk staat er: wanneer sekse met één eenheid toeneemt, neemt het aantal nieuwe woorden dat geleerd wordt af met 1,69, onder het constant houden van het aantal educatieve spellen en leeftijd. Dat is natuurlijk een rare conclusie, want sekse kan niet toenemen (en net zo min afnemen). Omdat meisjes hier de waarde 0 hebben en jongens de waarde 1 hebben, en er geen andere waarden dan dat zijn, kunnen we hier daarom zeggen: jongens leren gemiddeld 1,69 nieuwe woorden minder dan meisjes, waarbij we controleren voor het aantal educatieve spellen dat gespeeld wordt op een tablet en de leeftijd. Een meisje van 22 maanden oud dat vijf keer per week een educatief spel speelt, zal gemiddeld 1,69 nieuwe woorden meer leren in de week dan een jongen van 22 maanden oud die vijf keer per week een educatief spel speelt.

Behalve dat er drie ongestandaardiseerde regressiecoëfficiënten zijn, zijn er nu ook drie gestandaardiseerde regressiecoëfficiënten (bèta's). Deze geven in een meervoudige regressieanalyse de *partiële zuivere effecten* aan. Omdat deze waarden gestandaardiseerd zijn, kun je de effecten van de verschillende onafhankelijke variabelen met elkaar vergelijken. Zonder standaardisatie is dat niet mogelijk, 'Sekse' heeft immers maar twee waarden (meisje en jongen), leeftijd heeft veel meer waarden en is gemeten in maanden (van 0 tot 60 maanden) en het aantal educatieve spellen dat gespeeld wordt kan misschien wel oplopen tot 30 keer per week. De bèta's variëren in de regel van –1 tot +1. In deze analyse zie je dat 'educatieve spellen spelen' de hoogste waarde van bèta heeft, namelijk 0,526. Je kunt daaruit concluderen dat deze onafhankelijke variabele het sterkste effect heeft op het aantal nieuwe woorden dat geleerd wordt en dat leeftijd het minst sterke effect heeft. Het maakt bij de bèta niet uit of de waarde positief of negatief is, een negatieve samenhang kan immers ook zeer sterk zijn.

Omdat er nu drie bèta's zijn, is R, de multipele correlatiecoëfficiënt, niet meer gelijk aan de correlatiecoëfficiënt $| r |$. In de tabel *Model Summary* zie je dat R 0,729 is. De proportie verklaarde variantie (R^2) is 0,531. Aantal educatieve spellen, leeftijd in maanden en sekse, verklaren samen 53,1% van de variantie in het aantal nieuwe woorden dat een peuter per week leert. Dat is vrij veel (er is immers een sterke multipele samenhang). Er blijft wel 46,9% van de variantie niet verklaard, er zijn dus ook nog andere factoren dan die wij hebben gemeten om de woordenschat van peuters te kunnen voorspellen. Door toevoeging van

de variabelen leeftijd en sekse hebben we de verklaarde variantie in het leren van nieuwe woorden kunnen verhogen van 28,7% naar 53,1%.

Op basis van deze gegevens kun je voorspellingen doen door waarden voor de verschillende onafhankelijke variabelen (*x*'en) in te vullen. Eerst stel je de regressievergelijking op:

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 = 6{,}961 + 0{,}143(x_1) + 2{,}014(x_2) - 1{,}687(x_3)$$

Stel, je wilt voorspellen hoeveel nieuwe woorden een jongen van 24 maanden die zeven keer per week een educatief spel op de tablet speelt, leert. Je kunt die gegevens in de regressievergelijking invullen. Je vult voor educatieve spellen (x_1) de waarde 7 in, voor leeftijd de waarde 24, en voor sekse de waarde 1.

Wanneer je dit berekent, kom je op de volgende voorspelling uit:

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 = 6{,}961 + 0{,}143 * 7 + 2{,}014 * 24 - 1{,}687 * 1 = 54{,}611$$

Het model voorspelt dat een jongen van 24 maanden oud die zeven keer per week een educatief spel op een tablet speelt, 54,61 nieuwe woorden per week leert.

8.3.3 *Schijnsamenhang in een meervoudige regressie*

Net als bij tabelsplitsing (hoofdstuk 7) houd je bij het uitvoeren van een meervoudige regressieanalyse andere onafhankelijke variabelen constant. Hierdoor is het mogelijk dat een eerder gevonden verband door toevoeging van een (of meerdere) onafhankelijke variabele(n) verdwijnt.

In een onderzoek onder basisscholieren is aan 129 kinderen gevraagd of zij het Jeugdjournaal leuk vonden (uitgedrukt in een rapportcijfer), en of zij wel eens over het nieuws praatten (gemeten in aantal keer per week). De verwachting daarbij is dat hoe leuker kinderen het Jeugdjournaal vinden (onafhankelijke variabele), hoe vaker zij over het nieuws zullen praten (afhankelijke variabele). Aangezien beide variabelen minimaal intervalniveau zijn, en er sprake is van een asymmetrische relatie, is een enkelvoudige regressieanalyse hier de meest geschikte analyse. Uit de regressieanalyse (tabel 8.18) blijkt een sterk positief verband tussen de twee variabelen ($\beta = 0{,}61$), de verwachting komt dus uit: hoe leuker zij het Jeugdjournaal vinden, hoe meer ze praten over het nieuws.

Tabel 8.18 Enkelvoudige regressieanalyse Jeugdjournaal leuk vinden en praten over het nieuws (SPSS-output)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,154	,041		45,649	,000
	Jeugdjournaal leuk vinden	,654	,016	,614	21,758	,000

a. Dependent Variable: Praten over nieuws

Vervolgens is de variabele geslacht toegevoegd als controlevariabele, waarbij meisjes de waarde 0 kregen en jongens de waarde 1.

Tabel 8.19 Meervoudige regressieanalyse Jeugdjournaal leuk vinden, geslacht en praten over het nieuws (SPSS-output)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,108	,057		36,723	,000
	Jeugdjournaal leuk vinden	,032	,016	,004	20,690	,000
	geslacht (0 = meisje)	-,121	,022	-,629	-5,471	,000

a. Dependent Variable: Praten over nieuws

Aan de gestandaardiseerde regressiecoëfficiënten is nu te zien dat het oorspronkelijke positieve, sterke verband bij Jeugdjournaal verdwijnt: $\beta = 0{,}004$. Bij toevoeging van de variabele geslacht blijkt dat het oorspronkelijke verband een schijnverband is, spurieus is, en dat het praten over het nieuws niet door het al dan niet leuk vinden van het Jeugdjournaal wordt veroorzaakt, maar door of het kind een meisje of een jongen is ($\beta = -0{,}63$). Aan de negatieve waarde van de ongestandaardiseerde regressiecoëfficiënt kunnen we aflezen dat jongens gemiddeld 0,12 keer minder vaak over het nieuws praten dan meisjes, onder constanthouding van of zij het Jeugdjournaal leuk vinden.

8.3.4 *Regressie- en correlatieanalyses in wetenschappelijke tijdschriften*

Regressieanalyses en correlaties worden veelvuldig in wetenschappelijke publicaties gebruikt om hypothesen te toetsen. Hierin wordt meer informatie gegeven dan wij in de voorgaande hoofdstukken hebben besproken, maar we denken dat met de informatie die je nu hebt, je al een heel eind zult komen bij

het kunnen interpreteren van wetenschappelijke resultaten. Als voorbeeld laten we een paar tabellen zien uit onderzoeken die gepubliceerd zijn in het Tijdschrift voor Communicatiewetenschap.

In een experiment van De Leeuw et al.[7] is bijvoorbeeld gekeken naar de invloed van zogenaamd prosociaal televisienieuws op kinderen. Kinderen in groep 7 en 8 van de basisschool werden verdeeld over een controlegroep en een experimentele groep, waarbij kinderen in de experimentele conditie een nieuwsprogramma te zien kregen waarin geld werd ingezameld voor UNICEF, en kinderen in de controlegroep een nieuwsuitzending te zien kregen die ook over UNICEF ging maar waarin het prosociale gedrag (geld inzamelen) niet werd getoond. Voor het experiment maakten de onderzoekers zelf een nieuwsprogramma met de naam *Newz Kids*. De onderzoekers geven eerst de beschrijvende statistieken van de variabelen door percentages, gemiddelden en standaarddeviaties te laten zien:

Tabel 8.20 Beschrijvende statistieken uit artikel van De Leeuw et al. (2014)

	Totaal (N = 372)	Experimentele conditie (n = 183)	Controleconditie (n =189)
Prevalentie			
Jongen	44.1%	43.7%	44.4%
Lijkt een project voor UNICEF leuk	93.7%	93.3%	94.1%
Bereidwilligheid om een project voor UNICEF op te zetten[1]	80.0%	84.8%	76.9%
Gemiddelde (Standaarddeviatie)			
Leeftijd	10.94 (.76)	10.94 (.76)	10.95 (.75)
Waardering van *NewzKids*	7.39 (1.55)	7.36 (1.53)	7.43 (1.57)
Initieel prosociaal gedrag[2]	2.63 (.34)	2.66 (.32)	2.60 (.36)
Mate waarin ouders goede doelen belangrijk vinden	3.57 (.36)	3.55 (.65)	3.59 (.61)
Donatie voor UNICEF[3]	62.16 (30.58)	64.98 (31.50)	59.41 (29.49)

Hieruit kunnen we onder andere aflezen dat kinderen (met een gemiddelde leeftijd van 10,94, SD = 0,76) uit de experimentele conditie, eerder bereid zijn om een project voor UNICEF op te zetten (en die dus meer prosociaal zijn), dan kinderen in de controleconditie, terwijl de waardering voor het fictieve programma in beide groepen ongeveer even hoog is.

In tabel 8.21 is een correlatiematrix te zien tussen alle variabelen die de onderzoekers hebben gemeten. Je ziet hierin dat in wetenschappelijke tijdschriften geen SPSS-tabellen worden gebruikt, maar dat aangepaste tabellen worden gemaakt. In dit boek zullen we niet ingaan op de sterretjes en kruisjes achter de waarden van de associatiematen, maar we kunnen wel de richting en de sterkte van de correlaties aflezen.

Tabel 8.21 Correlatiematrix uit artikel van De Leeuw et al. (2014).

	1	2	3	4	5	6
1. Sekse[1]						
2. Waardering van *NewzKids*	.27 **					
3. Initieel prosociaal gedrag	.28 **	.24 **				
4. Mate waarin ouders goede doelen belangrijk vinden	.22 **	.17 **	.38 **			
5. Conditie[2]	.01	.02	.10 †	-.03		
6.Donatie voor UNICEF	.06	.03	.13 *	.19 **	.09 †	
7. Bereidwilligheid om een project voor UNICEF op te zetten	.32 **	.23 **	.31**	.20 **	.10 †	.06

[1]0 = jongen; 1 = meisje
[2]0 = controleconditie; 1 = experimentele conditie; * $p < .05$, ** $p < .01$, † $p < .10$.

Hoewel wij in dit boek geen bivariate analyses met dummyvariabelen uitvoeren, is wel goed te zien in tabel 8.21 dat je als onderzoeker moet aangeven wat de codering is bij een dummyvariabele (anders weet de lezer niet wat een positieve of negatieve correlatie bij bijvoorbeeld sekse betekent). Zo is te zien dat meisjes op alle variabelen hoger scoren dan jongens (alle correlaties met sekse zijn positief, en meisjes hebben de waarde 1). We kunnen bijvoorbeeld ook aflezen dat hoe hoger de mate is waarin ouders goede doelen belangrijk vinden, hoe hoger het initiële prosociale gedrag is (r = 0,38), en dat dit verband redelijk sterk en positief is.

In een ander onderzoek dat in het Tijdschrift voor Communicatiewetenschap is gepubliceerd vinden we een regressieanalyse met zowel de waarden voor de correlaties als de gestandaardiseerde regressiecoëfficiënten (tabel 8.22). In dit onderzoek van Slot et al. (2014)[8] is onder andere onderzocht of kinderen in de leeftijd van 9 tot 12 jaar reclame in online sociale netwerken (met de naam *Habbo*) begrijpen en in welke mate zij gevoelig zijn voor de mening van leeftijdsgenoten met betrekking tot merknamen die in deze netwerken gebruikt worden, en of er een verlangen was naar het geadverteerde merk.

Ook bij deze tabel zullen we niet op alle statistieken ingaan (zoals de standaardfouten en de sterretjes achter de waarden), maar we kunnen al veel van de waarden interpreteren. We zien bijvoorbeeld dat meisjes minder verlangen hebben naar de geadverteerde merken dan jongens ($\beta = -0.11$) onder constanthouding van de overige onafhankelijke variabelen. Wanneer gekeken wordt naar het kunnen begrijpen van de persuasieve intentie, valt op dat onder constanthouding van de overige variabelen er bijna geen invloed is op het verlangen (β = 0,09), waar wel een matige samenhang bestond (r = 0,20) wanneer we de andere onafhankelijke variabelen niet in de analyse betrekken. We kunnen ook zien dat alle onafhankelijke variabelen samen voor 36% de variantie in het verlangen naar geadverteerde merken verklaren, en dat de gevoeligheid voor de mening van *peers* met betrekking tot de merken in *Habbo* het sterkste

effect heeft op het verlangen naar merken waar reclame voor wordt gemaakt (β = 0,37), gevolgd door de kritische houding ten opzichte van reclame (hoe kritischer de houding, hoe minder het verlangen, β = –0,34).

Tabel 8.22 Regressieanalyse in het artikel van Slot et al. (2013)

	Verlangen naar geadverteerde merken		
	β	*SE*	*r*
Controlevariabelen			
Leeftijd	-.08	(.07)	-.15
Geslacht (1 = meisjes)	-.11	(.13)	-.10
Speelfrequentie *Habbo*	.04	(.08)	.02
Reclamewijsheid			
Reclameherkenning	-.06	(.09)	-.09
Begrip commerciële bron	-.10	(.12)	-.13
Begrip persuasieve intentie	.09	(.09)	.20*
Kritische houding t.o.v reclame	-.34***	(.08)	-.46***
Gevoeligheid voor *peer*invloed			
*Peer*invloed merken i.h. algemeen	-.03	.11	.08
*Peer*invloed merken in *Habbo*	.37***	(.08)	.50***
N	148		
Totaal R^2 (aangepast)	.36		

β = genormaliseerde bètaregressiecoëfficiënten; *SE* = standaardfouten; *r* = correlaties verlangen naar geadverteerde merk

8.4 Samenvatting

De associatiematen die je gebruikt op interval- en rationiveau zijn de correlatiecoëfficiënt (*r*), de proportie verklaarde variantie (R^2) en de gestandaardiseerde regressiecoëfficiënt (β). Een correlatie geeft aan wat de sterkte en richting is van de samenhang tussen twee variabelen. De proportie verklaarde variantie gebruik je bij een regressieanalyse en geeft aan in welke mate de onafhankelijke variabele(n) de varia(n)tie in de afhankelijke variabele verklaart/verklaren. Bèta's geven het zuivere effect van de onafhankelijke variabele(n) op de afhankelijke variabele aan.

Een regressievergelijking geeft een voorspelling van de afhankelijke variabele *y* op basis van de onafhankelijke variabele(n) *x* (of meerdere *x'en*). Bij een meervoudige regressieanalyse kijk je naar het partiële effect van een onafhankelijke variabele, waarbij je de andere onafhankelijke variabelen constant houdt. Het constant houden van (controleren voor) een derde variabele is ook al eerder aan de orde geweest bij tabelsplitsing (zie hoofdstuk 7).

In meervoudige regressieanalyses kunnen ook dummyvariabelen worden gebruikt, waarbij een nominale variabele (indien nodig) wordt omgezet naar een dichotome variabele met de waarden nul en één. Bij een dichotome variabele

geeft de ongestandaardiseerde regressiecoëfficiënt het gemiddelde verschil tussen de nul- en de één-categorie aan.

Tabel 8.23 Overzicht associatiematen

	Nominaal	Ordinaal	Interval en ratio
Symmetrisch	Cramers V phi	Gamma Kendalls tau-b Spearmans rho	Correlatie (r)
Asymmetrisch	Goodman en Kruskals tau lambda	Somers' d	Proportie verklaarde variantie (R^2) Gestandaardiseerde regressiecoëfficiënt (β)

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

Noten

1 Leeftijd is hier de onafhankelijke variabele, want kijktijd kan nooit leeftijd beïnvloeden. Een asymmetrische relatie mag echter ook beantwoord worden met een symmetrische associatiemaat, hoewel deze misschien niet altijd het *meest geschikt* zal zijn.
2 We hadden ook kunnen kiezen voor uren tv als *x* en uren krant als *y*.
3 De begrippen variatie en variantie worden hier beide gebruikt om hetzelfde aan te duiden. In sommige publicaties zul je misschien het woord variatie in plaats van variantie zien staan, wij kiezen hier voor de term variantie.
4 Ook bij deze term worden variatie en variantie door elkaar gebruikt, ze duiden hier hetzelfde aan.
5 De *n* kan niet uit bovenstaande tabellen worden afgelezen, deze informatie moet je uit de datamatrix zelf halen.
6 Een regressievergelijking beschrijven kan op verschillende manieren en is minder eenduidig dan bij de vorige associatiematen. Zorg in ieder geval dat de onderdelen intercept, (on)gestandaardiseerde regressiecoëfficiënt, proportie verklaarde variantie, (aantal) onderzoekseenheden en de variabelen worden besproken.
7 De Leeuw, N.H., Rozendaal, E., Kleemans, M., Anschütz, D.J. & Buijzen, M. (2014). 'Prosociaal nieuws en prosociaal gedrag in kinderen', *Tijdschrift voor Communicatiewetenschap (42)*4, 342-357.
8 Slot, N., Rozendaal, E., Van Reijmersdal E.A., & Buijzen, M. (2013). 'Hoe kinderen reageren op reclame in online sociale netwerken: reclamewijsheid en de invloed van leeftijdsgenoten', *Tijdschrift voor Communicatiewetenschap (41)* 1, 19-40.

Associatiematen: tot slot 9

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

9.1 Eta en eta-kwadraat

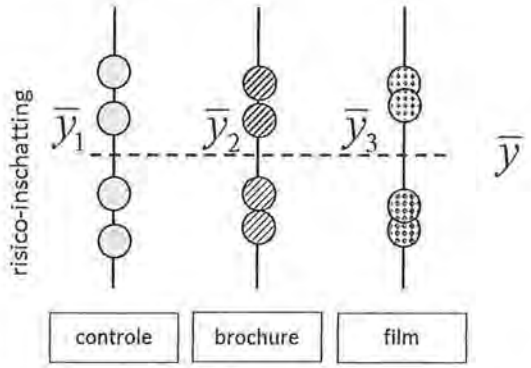
Eta is een associatiemaat die geschikt is voor een asymmetrische relatie, waarbij de onafhankelijke variabele een nominaal meetniveau heeft en de afhankelijke variabele een interval of ratio meetniveau. Eta hoort bij de *variantieanalyse.*[1] Bij het uitvoeren van een variantieanalyse worden de gemiddelden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

9.1.1 Interpretatie

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

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

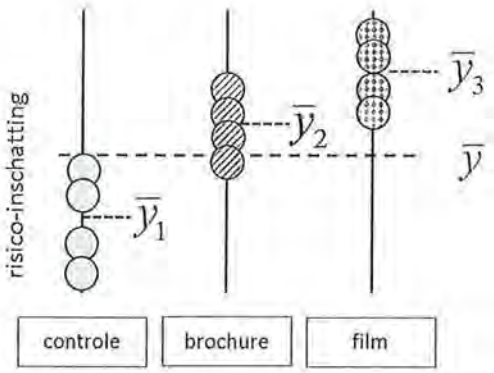
We schetsen hieronder twee mogelijke scenario's. In figuur 9.1 zie je de mogelijkheid dat de proefpersonen onderling veel van elkaar verschillen, maar dat tussen de groepen niet veel verschil is in de gemiddelde risico-inschatting. Zowel in de controlegroep, als in de groep die de brochure heeft gelezen,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



Figuur 9.1 Geen variantie tussen de groepen: gemiddelden gelijk

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

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



Figuur 9.2 Wel variantie tussen de groepen: verschillende gemiddelden

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

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

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

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

Formule voor eta²

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

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

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

Formule voor E_1 bij eta²

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

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

Formule voor E_2 bij eta²

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

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

Formule voor eta²

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

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

Report

risico-inschatting

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

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

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

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

ANOVA Table

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

Zou je nu eta of eta^2 berekenen, dan komt deze dus uit op 0:

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

SPSS bevestigt dit uiteraard:

Tabel 9.3 Associatiemaat eta en eta² (SPSS-output)

Measures of Association

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

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

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

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

Report

risico-inschatting

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

ANOVA Table

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

Measures of Association

	Eta	Eta Squared
risicoinschatting * conditie	,796	,634

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

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

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

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

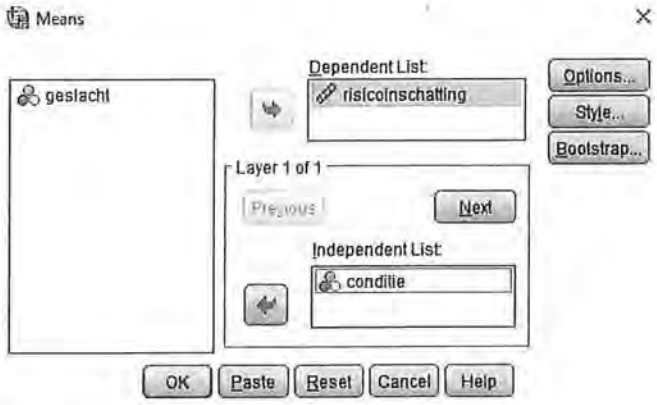
$$\eta = \sqrt{0{,}634} = 0{,}796$$

Onze conclusie aan de hand van bovenstaande output is:

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

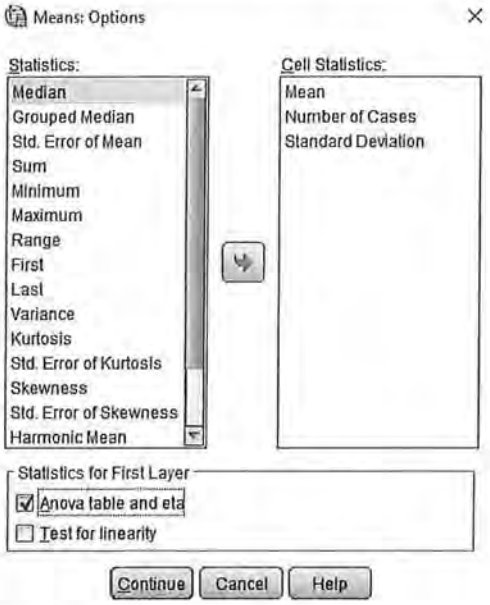
SPSS Vergelijken van gemiddelden tussen afzonderlijke groepen

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



Figuur A Means-venster

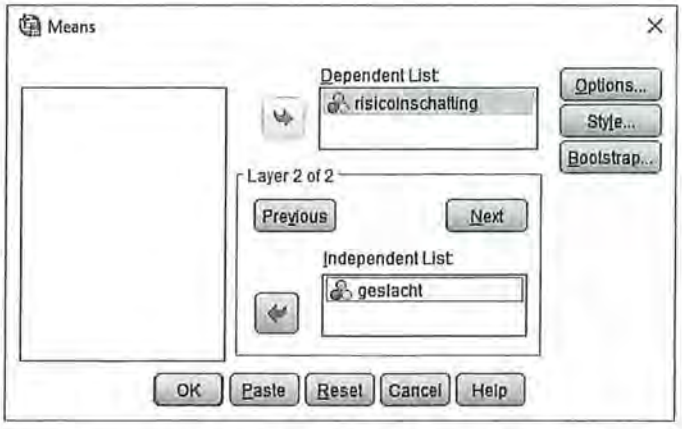
Via *Options* kun je aangeven welke informatie SPSS moet vermelden. Standaard staan hier het gemiddelde (*Mean*), het totaal aantal onderzoekseenheden (*Number of Cases*) en de standaarddeviatie (*Standard Deviation*) weergegeven. Bij *Statistiscs for First Layer* kan worden aangegeven dat eta (en daarmee ook eta²) berekend moet worden (figuur B).



Figuur B Options-venster

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

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



Figuur C Means-venster Layer 2 of 2

9.1.2 Berekening

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

Tabel 9.5 Datamatrix van sekse en leestijd

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

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

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

Tabel 9.6 Waarde van E_1 berekenen

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

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

Tabel 9.7 Waarde van E_2 berekenen

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

Nu kun je eta² berekenen en daarna eta:

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

en

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

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

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

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

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

Measures of Association

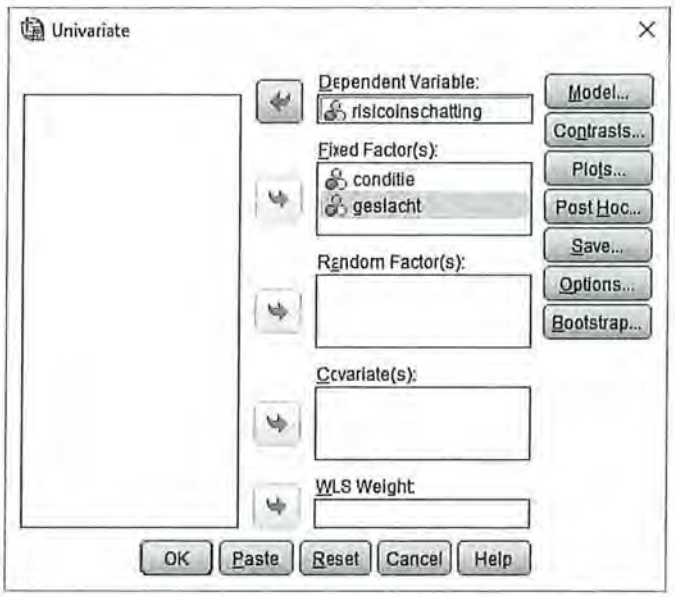
	Eta	Eta Squared
leestijd * geslacht	,814	,663

9.1.3 *Interactie-effecten bij variantieanalyse*

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

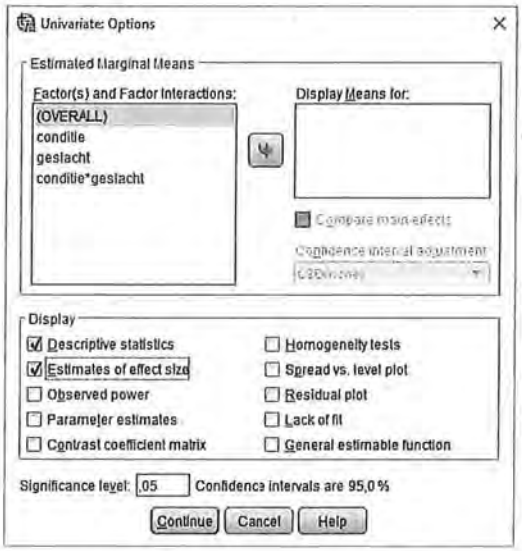
SPSS Berekenen van interactie-effecten via GLM

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



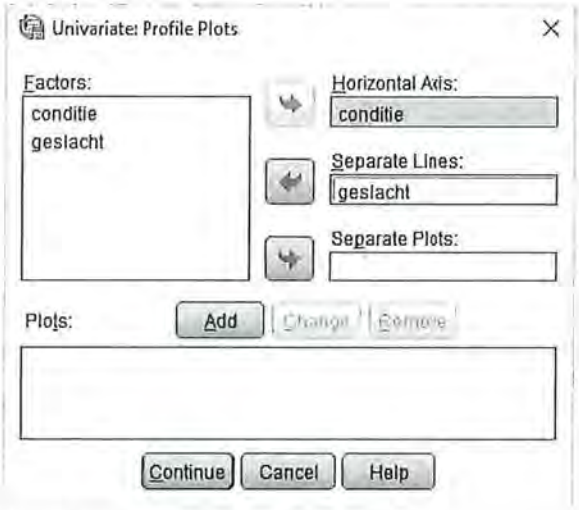
Figuur A Univariate-venster

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



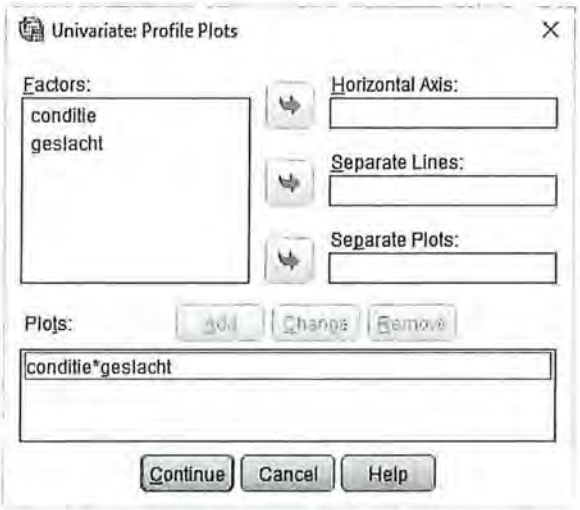
Figuur B Options-venster

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



Figuur C Profile Plots-venster

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



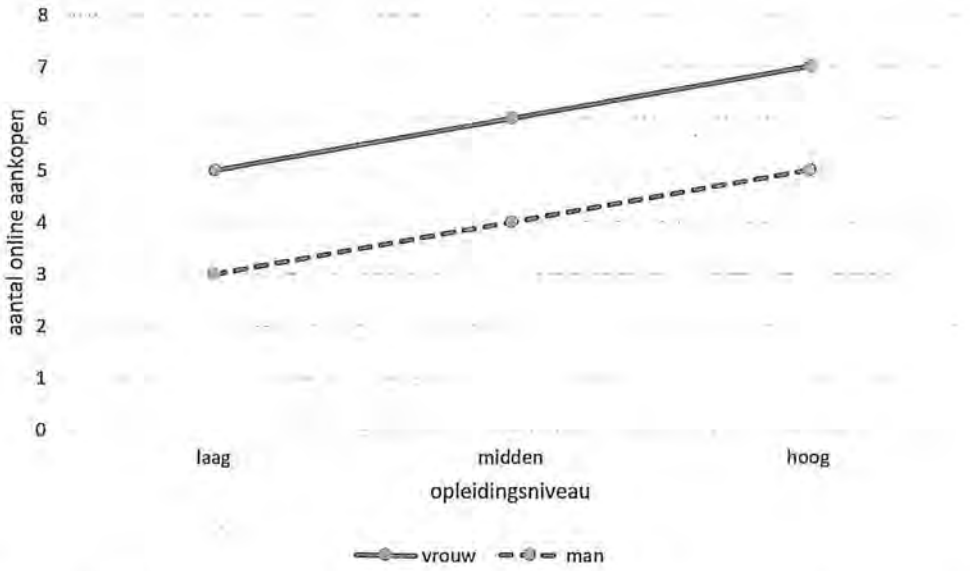
Figuur D Profile Plots-venster na Add

Kader 9.2

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

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

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

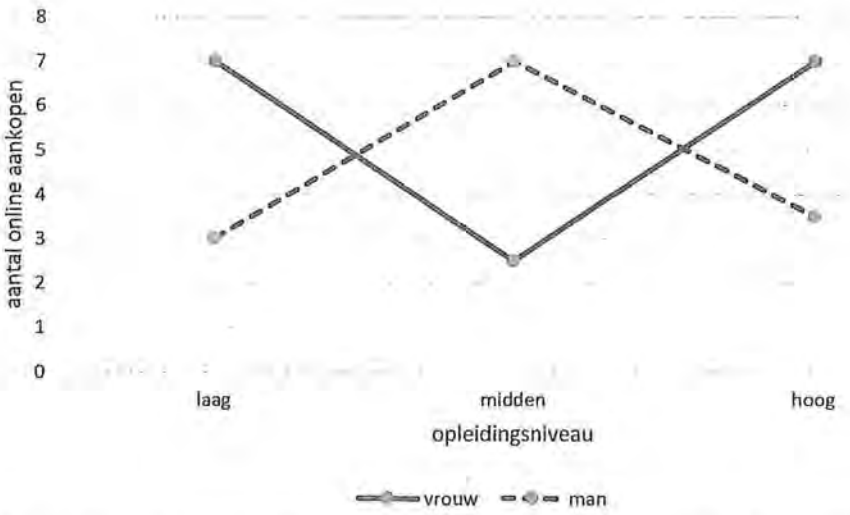


Figuur 9.3 Geen interactie-effect van opleiding en geslacht op aantal online aankopen

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

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

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



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

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

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

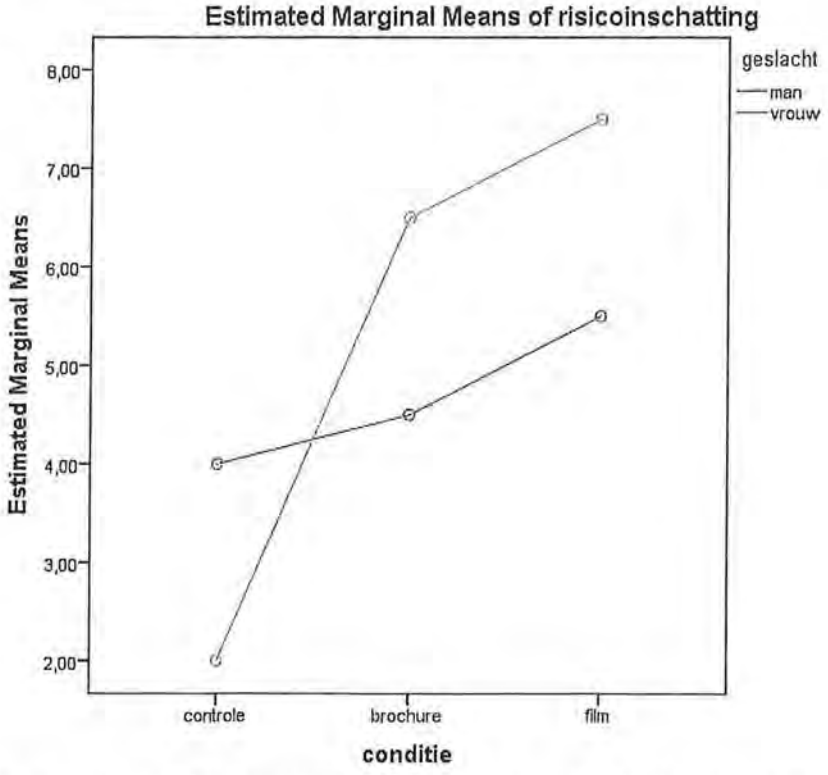
Tabel 9.9 Gemiddelden naar conditie en geslacht voor risico-inschatting (SPSS-output)

Report

risico-inschatting

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

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



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

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

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

Tests of Between-Subjects Effects

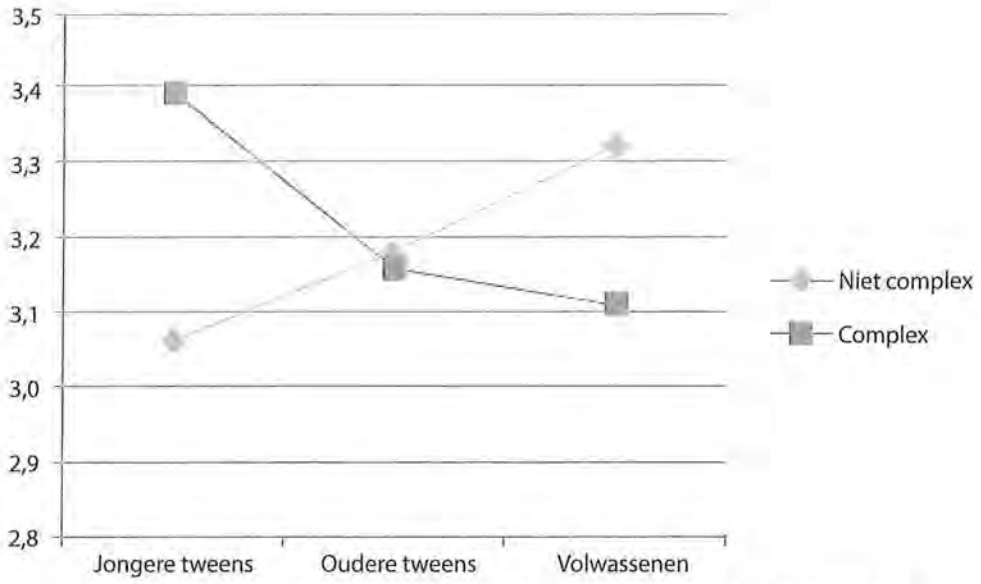
Dependent Variable: risico-inschatting

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

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

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

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



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

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

9.2 Het kiezen van een associatiemaat

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

9.2.1 *Formulering van uitspraken op basis van je onderzoek*

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

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

Tabel 9.11 Kruistabel televisiekijken naar leeftijdsgroepen

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

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

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

9.2.2 Kenmerken van de associatiematen

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

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

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

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

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

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

γ = 1

tau-b = .67

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

γ = .37

tau-b = .19

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

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

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

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

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

Cramers V = 0,62

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

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

Cramers V=0,27

9.3 Samenvatting

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

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

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

Noten

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

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

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

Schaalconstructie 10

In dit laatste hoofdstuk staan we stil bij het construeren van schalen. In de voorgaande hoofdstukken bedoelden we met 'schaal' veelal een antwoordschaal (een onderzoekseenheid kan op een schaal van 1 tot en met 5 (of 7 of 9 of ...) aangeven in hoeverre hij het eens is met een stelling) of een indexscore (een samengestelde schaal waarbij verschillende concrete variabelen bij elkaar werden opgeteld). Hier zullen we stilstaan bij een gemiddelde schaal. Met *schaal* bedoelen we hier de combinatie van een aantal variabelen (items) waarop onderzoekseenheden een score krijgen op een abstract, complex kenmerk. Een abstract complex kenmerk wordt ook wel een *concept* of *latente variabele* genoemd. Dit zijn kenmerken die je niet direct kunt meten of observeren, zoals bijvoorbeeld met een vraag naar een 'mening' of 'gedrag'. Je kunt wel door middel van verschillende *manifeste variabelen*, variabelen die je wel direct kunt meten, proberen zo goed mogelijk in de buurt te komen van het meten van dit concept. Je moet dan echter wel weten of je de juiste manifeste variabelen hebt gebruikt, en of je geen systematische en/of toevallige fouten hebt gemaakt bij het meten van het construct. Hoe je dit kunt nagaan zal in dit hoofdstuk beschreven worden.

10.1 Validiteit van een meting

Wanneer we kenmerken van onderzoekseenheden gaan meten kan het zijn dat we daar fouten bij maken. Deze fouten hebben invloed op de *betrouwbaarheid* en de *validiteit* van onze metingen. De betrouwbaarheid van een meting hangt af van het aantal toevallige fouten dat gemaakt wordt, en zal besproken worden in paragrafen 10.2, 10.3.2 en 10.6. Wanneer er *systematische* fouten worden gemaakt, hebben we problemen met de *validiteit* van de meting. Bij validiteit gaat het om de inhoudelijke betekenis van een meting; meet je inderdaad wat je had willen meten?

Sommige concepten zijn moeilijker te meten dan andere. De variabele 'leeftijd' bijvoorbeeld is vrij simpel te meten. Je kunt aan iemand vragen: 'Hoe oud ben je?' (... jaar), of je kunt vragen naar iemands geboortejaar. Welke vraagstelling ook gekozen wordt, bij het meten van leeftijd zul je weinig systematische fouten maken: je meet wat je wilt meten, namelijk hoe oud iemand is.

Bij complexe of abstracte begrippen zijn de problemen met betrekking tot validiteit groter, omdat de manier waarop het concept gemeten wordt minder voor

de hand ligt of omdat de metingen op wezenlijk verschillende manieren kunnen worden uitgevoerd. Meestal worden voor het meten van dergelijke concepten meerdere metingen verricht die elk afzonderlijk slechts een aspect van het abstracte begrip meten. Deze metingen zijn in de sociale wetenschappen meestal uitspraken of vragen in een vragenlijst. Door middel van een *factoranalyse* kan worden nagegaan of een gemeenschappelijke dimensie ten grondslag ligt aan de vragen die gekozen zijn om een abstract begrip te meten. Als uit de factoranalyse blijkt dat de gebruikte vragen een gemeenschappelijke onderliggende dimensie hebben, is dat een indicatie voor de validiteit van de meting van het abstracte begrip. Met een factoranalyse kun je aantonen dat verschillende variabelen (aspecten van) hetzelfde begrip meten.

10.1.1 Latente en manifeste variabelen

Stel dat je in je onderzoek wilt meten hoe 'media-afhankelijk' je respondenten zijn. De meest simpele optie zou zijn je respondenten op een antwoordschaal van bijvoorbeeld 1 tot 7 te laten aangeven hoe afhankelijk ze van de media zijn. Helaas valt op deze simpele optie heel wat af te dingen. Media-afhankelijkheid is een abstract en complex begrip. Je respondenten weten niet dat jij met media-afhankelijkheid doelt op 'de mate waarin een individu (informatie uit) de media gebruikt om zijn of haar doelen te bereiken'. Als je respondenten direct naar hun media-afhankelijkheid vraagt, zullen ze zelf de betekenis van dit begrip invullen en deze invulling zal sterk variëren onder je respondenten. Op een schaal van 1 tot 7 zullen zij aangeven hoe media-afhankelijk zij zichzelf vinden of hoe media-afhankelijk zij willen zijn. De kans dat twee personen die in feite even media-afhankelijk zijn, op deze schaal hetzelfde antwoord geven, is erg klein. Kortom, met deze simpele vraag meet je niet wat je wilt meten: de meting is dus niet valide. Beter is na te gaan hoe media-afhankelijkheid zich bij een individu manifesteert en daar vragen of uitspraken over te formuleren.
Als onderzoeker begin je eerst met een *theoretische definitie* van het verschijnsel. In dit geval zou de theoretische definitie van 'afhankelijkheid van media' zijn 'de mate waarin een individu (informatie uit) de media gebruikt om zijn of haar doelen te bereiken'. Daarmee is het begrip echter nog niet geoperationaliseerd, dat wil zeggen meetbaar gemaakt voor je onderzoek. De *operationele definitie* van media-afhankelijkheid zou kunnen zijn: 'de mate waarin mensen vinden dat ze zich met de media vermaken, ze meer te weten komen over normen en waarden in de maatschappij, ze zichzelf en de wereld om zich heen beter begrijpen, ze weten wat ze in allerlei situaties het best kunnen doen en ze zichzelf met anderen kunnen vermaken'. In deze operationele definitie is een aantal manifeste variabelen opgesomd waarvan je als onderzoeker vindt dat het aspecten van het begrip 'media-afhankelijkheid' zijn.
Je gaat vervolgens de manifeste uitingen van het latente verschijnsel 'media-afhankelijkheid' meten. Je zou de respondenten bijvoorbeeld de volgende zes uitspraken (zes direct te meten uitspraken, zes *manifeste variabelen*) kunnen

voorleggen waarbij ze op een schaal van 1 tot 7 kunnen aangeven in hoeverre ze het ermee eens zijn (1 = zeer mee oneens, 7 = zeer mee eens):

Door de media te gebruiken ...
1. ... kan ik mezelf goed vermaken.
2. ... kom ik meer te weten over normen en waarden in de maatschappij.
3. ... begrijp ik meer van de wereld om mij heen.
4. ... ga ik mijzelf beter begrijpen.
5. ... weet ik wat ik in allerlei situaties het best kan doen.
6. ... kan ik me met anderen vermaken.

Als in deze manifeste variabelen de latente variabele media-afhankelijkheid inderdaad tot uitdrukking komt, zullen respondenten die sterk media-afhankelijk zijn op deze stellingen hoog scoren en zullen respondenten die nauwelijks afhankelijk van de media zijn, lage scores hebben. De correlaties tussen deze zes variabelen zullen dus hoog moeten zijn als er een gemeenschappelijke dimensie (latente variabele) aan ten grondslag ligt.

Manifeste variabelen (ook wel: *items*) zijn dus kenmerken die direct gemeten kunnen worden en latente variabelen (ook wel: *concepten*) zijn abstracte kenmerken die niet direct gemeten kunnen worden, maar waarvoor door het meten van een aantal manifeste variabelen een indicatie kan worden gevonden.

Tabel 10.1 Latente en manifeste variabelen

	Manifest (uiterlijk kenmerk)	Latent (abstract, innerlijk kenmerk)
Enkelvoudig (één variabele)	Bijv. sekse Bijv. leeftijd	Niet mogelijk
Samengesteld / complex (meer variabelen)	**Index** Bijv. kijktijd gemeten door totaal aantal minuten televisiekijken per week samen te voegen met het aantal uur per week	**Schaal** Bijv. media-afhankelijkheid gemeten door zes verschillende items

Tabel 10.1 laat zien dat manifeste variabelen zowel enkelvoudig als samengesteld kunnen worden gemeten, en dat latente variabelen (die zijn samengesteld door manifeste variabelen) altijd complex zijn. In de volgende paragraaf laten we zien hoe je door middel van een factoranalyse vast kunt stellen of een latente variabele valide gemeten is.
Wanneer blijkt dat de manifeste variabelen inderdaad hetzelfde verschijnsel meten (we laten in paragraaf 10.3.1 zien hoe je dat kunt vaststellen), kun je concluderen dat je een valide meting hebt uitgevoerd.[1]

10.2 Betrouwbaarheid van een meting

Naast een valide meting is het ook van belang dat je een betrouwbare meting hebt uitgevoerd. De *betrouwbaarheid* van een meting is de mate waarin die meting vrij is van toevallige fouten. Je wilt graag dat je meting valide en betrouwbaar is, dus vrij van zowel systematische als toevallige fouten. Nadat je hebt vastgesteld of een meting valide is (zie paragraaf 10.3.1) ga je daarom na of de meting betrouwbaar is (zie ook paragraaf 10.3.2).

We hadden al gezien dat als een verschijnsel of kenmerk iets gecompliceerder is, zoals het eerder genoemde kenmerk 'afhankelijkheid van de media', het vaak niet mogelijk is om dit kenmerk slechts met één variabele te meten. Doordat het verschijnsel of kenmerk verschillende aspecten omvat, zijn meer variabelen nodig om het gehele begrip te dekken. We verwachten dat de variabelen (de *items*) die media-afhankelijkheid moeten gaan meten sterk met elkaar samenhangen. Respondenten met een hoge media-afhankelijkheid scoren op alle items hoog en respondenten met een lage media-afhankelijkheid scoren op alle items laag. Als de correlaties tussen al deze items hoog zijn, kunnen we stellen dat de schaal om media-afhankelijkheid te meten *intern consistent* is. Het is echter onhandig om uitspraken te doen over de interne consistentie op basis van een groot aantal correlatiecoëfficiënten (in bovenstaand voorbeeld zou je zes correlatiecoëfficiënten krijgen). We gebruiken daarom *Cronbachs alfa*, waarmee we met één kengetal de mate van interne consistentie aangeven. Deze maat voor interne consistentie geeft aan hoe betrouwbaar je schaal is. Of het meetinstrument voor het meten van media-afhankelijkheid betrouwbaar is, stel je dus vast door na te gaan of respondenten consistent scoren op de afzonderlijke variabelen. Als dat het geval is, kun je een gemiddelde schaal maken die de mate van afhankelijkheid van de media aangeeft.

10.3 Schaalconstructie

Met schaalconstructie bedoelen we het maken (construeren) van een nieuwe schaal op basis van bestaande variabelen. Zoals we al besproken hebben, is die nieuwe schaal een latente variabele (een construct, een abstract begrip) die niet in één keer is gemeten, maar op basis van een aantal variabelen (manifeste variabelen, items) is samengesteld. Een voorwaarde voor schaalconstructie is dat alle items op dezelfde manier gemeten zijn (dat wil zeggen: alle vragen moeten met dezelfde antwoordschaal beantwoord zijn)[2], en zij moeten allemaal minimaal een ordinaal meetniveau hebben. De uiteindelijke schaal, waar de gemiddelde scores van de respondenten op de verschillende items worden weergegeven, heeft altijd een interval meetniveau (het is immers een gemiddelde schaal).

Schaalconstructie heeft drie doelen. Ten eerste willen we een *valide meting* van een concept door meting van verschillende items. We willen de verschillende

aspecten van een latente variabele meten door middel van verschillende variabelen (items) en we willen nagaan of deze items inderdaad in voldoende mate hetzelfde meten.
Ten tweede wil je een *preciezere meting*. Wanneer je scores op verschillende items met elkaar combineert tot één schaal, krijg je meer genuanceerde verschillen tussen de scores van de onderzoekseenheden. Bij het maken van een schaal is het ook mogelijk om ordinale items, bijvoorbeeld variabelen met ordinale *Likertschalen*, samen te voegen. Een Likertschaal is een veelgebruikte antwoordschaal in de sociale wetenschappen om houdingen en meningen te meten, waarbij respondenten kunnen antwoorden op een schaal van het ene uiterste tot het andere uiterste, bijvoorbeeld van 1 = nooit tot en met 5 = zeer vaak, of van 1 = mee oneens tot en met 5 = mee eens. Veelal worden vijfpuntsschalen gebruikt, maar ook zeven- en negenpuntsschalen kunnen een Likertschaal zijn. Formeel zijn dit ordinale schalen. Maar bij een ordinale schaal van 1 (mee oneens) tot 5 (mee eens) *lijken* de intervallen tussen de getallen min of meer gelijk. Daarom kunnen we met dergelijke ordinale variabelen toch rekenen (zoals optellen, aftrekken en gemiddeldes berekenen). Wanneer je nu een gemiddelde schaal maakt, is de schaal niet langer discreet maar continu, en zijn ook scores mogelijk van 2,3 of 4,1 enzovoorts.
Tot slot heeft schaalconstructie tot doel *data te reduceren*. Als je de zes items die media-afhankelijkheid meten samenvoegt, hoef je in je onderzoek niet zes verschillende analyses uit te voeren om te kijken naar bijvoorbeeld de verschillen tussen mannen en vrouwen en hun media-afhankelijkheid (uitgaande van ons voorbeeld waar we zes vragen hadden gesteld om de mate van media-afhankelijkheid te meten), maar slechts één (namelijk één analyse met geslacht en media-afhankelijkheid als variabelen).

Bij het maken van een schaal worden altijd de volgende stappen gevolgd:
1. factoranalyse (vaststellen van de validiteit van de schaal);
2. betrouwbaarheidsanalyse (vaststellen van de interne consistentie van de schaal);
3. maken van de schaal;
4. beschrijven van de schaal.

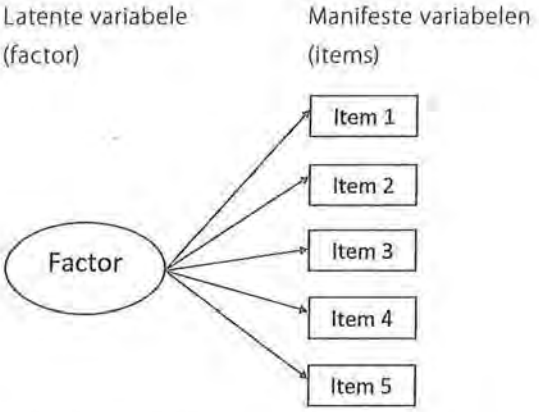
10.3.1 Factoranalyse

We gaan kijken of we een schaal kunnen construeren voor de latente variabele 'smartphoneverslaving'. In een onderzoek is gekeken naar de mate waarin studenten aan een universiteit verslaafd zijn aan hun smartphone.[3]
Om smartphoneverslaving te meten is een enquête afgenomen onder studenten waarbij op een zevenpunts Likertschaal variërend van 1 = geheel mee oneens tot en met 7 = geheel mee eens geantwoord kon worden:
1. Ik raak geïrriteerd wanneer mensen mij storen terwijl ik mijn smartphone gebruik.

2. Ik check voortdurend mijn smartphone om te kijken of ik niets heb gemist op sociale media.
3. Ik neem mijn telefoon mee naar het toilet, zelfs als ik heel nodig moet.
4. Er is niets leuker dan bezig te zijn met mijn smartphone.
5. Ik zou niet meer zonder smartphone kunnen.

Of dit vermoeden klopt, en of we inderdaad met deze vijf variabelen één latente variabele meten, zal moeten blijken uit de resultaten van de factoranalyse. Op basis van de gemeenschappelijke samenhang (correlaties) tussen een groep variabelen wordt een *factor* gevormd, de latente variabele. Die factor representeert wat de groep variabelen gemeenschappelijk heeft, een onderliggende dimensie. De onderliggende dimensie bij onze vijf vragen zou dan 'smartphoneverslaving' zijn.

Schematisch ziet het er als volgt uit:



Figuur 10.1 Latente en manifeste variabelen

In figuur 10.1 is te zien dat de pijlen vanuit de factor naar de items wijzen. Het idee van een factoranalyse is namelijk dat *omdat* de latente eigenschap aanwezig is, er op een bepaalde manier op de verschillende items wordt gescoord, en niet andersom. Kijken we bijvoorbeeld naar de eerste vraag die gesteld is om smartphoneverslaving te meten, dan stellen we bij een factoranalyse dat *omdat* een respondent verslaafd is aan zijn of haar smartphone, diegene hoog zal scoren op de vraag 'ik raak geïrriteerd wanneer mensen mij storen terwijl ik mijn smartphone gebruik'. Het is niet zo dat 'irritatie bij storen' de oorzaak is van smartphoneverslaving.

Als we de scores van de respondenten op de voornoemde vijf variabelen bij elkaar op zouden tellen, zou de resulterende variabele aangeven in welke mate de studenten verslaafd zijn aan hun smartphone. Maar mogen we dat wel doen? Dragen alle vijf variabelen bij aan het meten van hetzelfde begrip? We kunnen voor de eerste impressie een correlatiematrix uitdraaien (tabel 10.2) om te kijken of er een samenhang is tussen de variabelen. We vermoeden immers dat

een hoge score op de ene variabele gepaard zal gaan met een hoge score op de andere variabele.

Tabel 10.2 Correlatiematrix voor de zes variabelen m.b.t. smartphoneverslaving (SPSS-output)

Correlations

		v1 irritatie bij storen	v2 checken sociale media	v3 mee-nemen naar toilet	v4 niets leuker	v5 niet zonder kunnen
v1 irritatie bij storen	Pearson Correlation	1	,372**	,463**	,570**	,557**
	Sig. (2-tailed)		,000	,000	,000	,000
	N	239	239	239	239	239
v2 checken sociale media	Pearson Correlation	,372**	1	,584**	,549**	,618**
	Sig. (2-tailed)	,000		,000	,000	,000
	N	239	239	239	239	239
v3 meenemen naar toilet	Pearson Correlation	,463**	,584**	1	,565**	,580**
	Sig. (2-tailed)	,000	,000		,000	,000
	N	239	239	239	239	239
v4 niets leuker	Pearson Correlation	,570**	,549**	,565**	1	,714**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	239	239	239	239	239
v5 niet zonder kunnen	Pearson Correlation	,557**	,618**	,580**	,714**	1
	Sig. (2-tailed)	,000	,000	,000	,000	
	N	239	239	239	239	239

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

We zien dat er inderdaad een positieve samenhang is tussen alle items. Je zou kunnen redeneren dat studenten die de hoog scoren op 'ik neem mijn telefoon mee naar het toilet, zelfs als ik heel nodig moet' (v3) ook hoog scoren op 'er is niets leuker dan bezig te zijn met mijn smartphone' (v4).

Hoewel we aan de correlatiematrix een impressie kunnen ontlenen van een construct, blijft het echter een overzicht van allemaal afzonderlijke bivariate correlaties. In een factoranalyse wordt de samenhang berekend van de afzonderlijke items met de onderliggende factor. Deze correlatie tussen een manifeste variabele (bijvoorbeeld 'Ik zou niet meer zonder smartphone kunnen') en een factor (hier: 'smartphoneverslaving') wordt de *factorlading* genoemd. Deze factorladingen variëren tussen –1 (perfecte negatieve samenhang met de factor) en +1 (perfecte positieve samenhang met de factor). Een factorlading van 0 betekent dat er geen relatie met de factor is. Er kan gesproken worden van een betekenisvolle samenhang wanneer de factorlading groter is dan |0,45|.[4]

Weer een andere naam voor factor, latente variabele of construct is *component*. In tabel 10.3 zie je de factorladingen die horen bij onze vijf items. Er wordt in dit geval één component gevonden waarbij alle items, alle manifeste variabelen, een factorlading hebben die groter is dan 0,45. Dit wijst erop dat deze vijf items

inderdaad een valide meting zijn voor de latente variabele 'smartphoneverslaving'.[5]

Tabel 10.3 Componentenmatrix met factorladingen (SPSS-output)

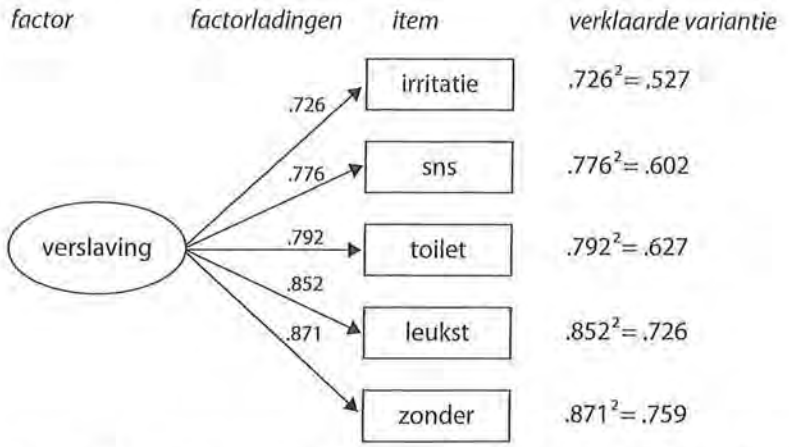
Component Matrix[a]

	Component
	1
v1 irritatie bij storen	,726
V2 checken sociale media	,776
v3 meenemen naar toilet	,792
v4 niets leuker	,852
v5 niet zonder kunnen	,871

Extraction Method: Principal Component Analysis.
a. 1 components extracted.

Bij een factoranalyse worden coëfficiënten berekend, zodanig dat de factor de items zo goed mogelijk voorspelt. Je krijgt dus bij elke pijl vanuit de latente variabele een waarde (figuur 10.2), en die waarde geeft aan hoe goed de latente variabele de scores op dat item kan voorspellen. Die waarden zijn de factorladingen, en net als bij de gestandaardiseerde regressiecoëfficiënten en associatiematen liggen deze tussen –1 en +1 en hebben ze een richting. Je kunt deze factorladingen dus net zo interpreteren als de correlatiecoëfficiënt en de gestandaardiseerde regressiecoëfficiënt. De factorladingen geven de samenhang weer tussen de latente variabele en de manifeste items die we gemeten hebben. Een factorlading van 0 betekent dat er helemaal geen relatie is tussen de factor en het item, en een factorlading van +1 zou betekenen dat er een perfecte positieve samenhang zou zijn. Dat betekent dat een factorlading ook negatief kan zijn. Dat betekent niet dat de manifeste variabele dan niet bijdraagt aan een valide schaal. Wanneer de factorlading hoog genoeg is (dat wil zeggen hoger dan |0,45|) is het waarschijnlijk zo dat de manifeste variabele met een negatieve correlatie met de factor in dat geval 'omgekeerd' is gecodeerd. Wanneer we bijvoorbeeld de eerste vraag hadden geformuleerd als 'ik raak nooit geïrriteerd wanneer mensen mij storen terwijl ik mijn smartphone gebruik', zouden we nog steeds met die vraag 'smartphoneverslaving' meten, maar dan is juist een lage score op deze variabele een indicatie van de verslaving. In dat geval moet je dit item hercoderen (zie hoofdstuk 4) voordat je je schaal maakt.

Net als bij een regressieanalyse kunnen we ook een uitspraak doen over de proportie verklaarde variantie. Bij een factoranalyse kijken we dan hoe goed de variantie in de factor, de variantie in de afzonderlijke items verklaart.



Figuur 10.2 Factorladingen en verklaarde variantie per item voor 'smartphoneverslaving'

De verklaarde variantie kun je per item uitrekenen door het kwadraat te nemen van de factorlading. Uit figuur 10.2 kunnen we bijvoorbeeld opmaken dat de verslaving aan een smartphone sterk en positief samenhangt met de vraag of studenten zich geïrriteerd voelen als ze gestoord worden tijdens het telefoongebruik (*factorlading* = 0,73). De variantie in de factor (oftewel: de spreiding in de scores binnen de schaal voor smartphoneverslaving) verklaren voor 52,7% de variantie in hoe op het item van irritatie is gescoord. Zo kunnen we ook aflezen dat de variantie in smartphoneverslaving voor 75,9% de variantie in het antwoord 'ik zou niet meer zonder smartphone kunnen' verklaart.
Deze itemvarianties (de varianties per afzonderlijk item) worden niet door SPSS getoond, maar wel de *totale verklaarde variantie*. Dit is de gemiddelde verklaarde variantie van de afzonderlijke itemvarianties. In ons voorbeeld is de totale verklaarde variantie dus:

$$\text{totale verklaarde variantie} = \frac{0{,}527+0{,}602+0{,}627+0{,}726+0{,}759}{5} = \frac{3{,}241}{5} = 0{,}649$$

Dat wil zeggen dat de variantie in de factor voor 64,9% de variantie in de vijf items verklaart. Deze informatie vinden we ook in tabel 10.4, waar we onder andere in de laatste kolom (*Extraction Sums of Squared Loadings*) het percentage van 64,80 zien (de decimalen achter de komma wijken af wegens afrondingsverschillen). Dit betekent ook dat je een percentage informatieverlies hebt van 35,1% (100 – 64,90). De mate van smartphoneverslaving verklaart de variantie in de vijf items niet perfect. Ook andere factoren dan smartphoneverslaving zullen dus een rol spelen bij het beantwoorden van de vijf vragen. Er is geen standaardrichtlijn voor wat voldoende verklaarde variantie is en wat niet. Meestal beargumenteert de onderzoeker dat zelf bij zijn of haar resultaten.

Tabel 10.4 Totale verklaarde variantie en eigenwaarde in factoranalyse (SPSS-output)

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,240	64,802	64,802	3,240	64,802	64,802
2	,657	13,142	77,944			
3	,452	9,039	86,984			
4	,375	7,503	94,486			
5	,276	5,514	100,000			

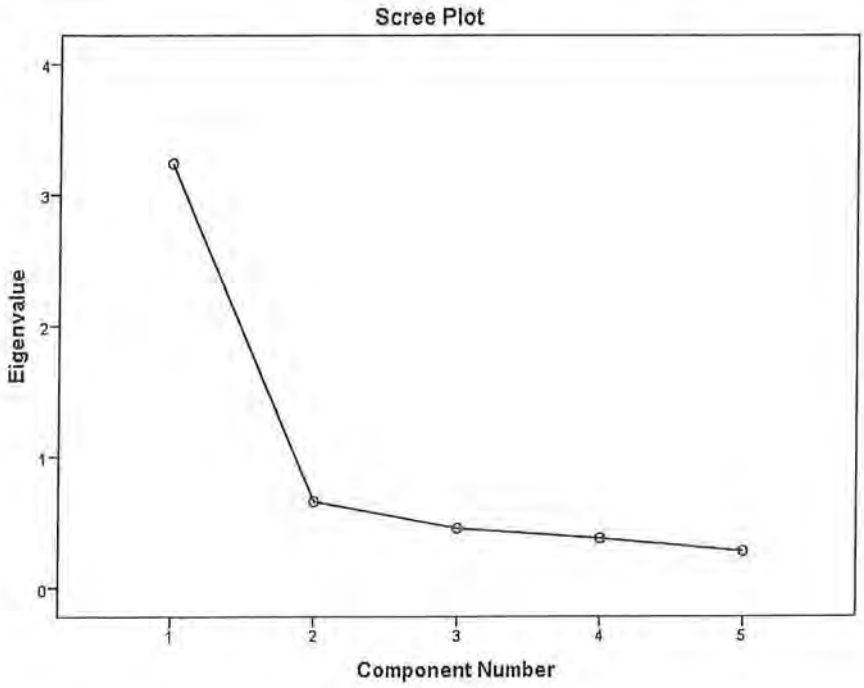
Extraction Method: Principal Component Analysis.

Als de verklaarde variantie laag is, meten we met onze items niet wat we willen meten en kunnen we geen valide schaal maken van de gebruikte items. Het kan ook zijn dat sommige items wel, en andere items niet bij de schaal horen (hier komen we op terug in de volgende paragraaf). Hoe meer de items hetzelfde meten, hoe hoger de *gemeenschappelijke variantie* zal zijn. Dit is de variantie die de latente variabele (component) gemeen heeft met de manifeste variabele. Wanneer alle items iets anders zouden meten dan in dit geval smartphoneverslaving, is de gemeenschappelijke variantie laag en kan er niet één factor of component worden gevormd. We willen dus dat de gemeenschappelijke variantie zo hoog mogelijk is, en de *specifieke variantie* en/of de *foutenvariantie* zo laag mogelijk zijn. De specifieke variantie is de variantie die alleen door die specifieke variabele gemeten wordt en die dus uniek is voor die variabele. Foutenvariantie is de variantie die geheel willekeurig is en die geen systematisch verband heeft met enige andere bron van variantie. Dit zou kunnen komen door meetfouten die worden veroorzaakt door onzorgvuldig invullen van een vragenlijst, gokken bij multiplechoicevragen enzovoort. Welk deel precies specifieke variantie en welk deel foutenvariantie is, is niet uit een factoranalyse af te leiden.

Hoe hoger de gemeenschappelijke variantie, hoe sterker de items onderling met elkaar en met de factor correleren, en hoe 'sterker' en informatiever onze factor is. Deze informativiteit wordt uitgedrukt in de *eigenwaarde* van de component. In tabel 10.4 zien we ook de *eigenwaarde* (in het Engels *Eigenvalue*) staan, in de eerste kolom. De eigenwaarde geeft aan hoe informatief de factor is. Wanneer je de items samenneemt tot één factor, is deze dan informatiever dan de losse items apart? Vaak besluiten onderzoekers om alleen factoren of componenten te gebruiken waarvoor de eigenwaarde groter is dan 1 (*criterium van Kaiser*). Dit is een arbitraire grens.

Ook de eigenwaarde wordt berekend aan de hand van de individuele itemvarianties. De *eigenwaarde* van een factor wordt berekend door de factorladingen (correlaties tussen variabelen en de factor) te kwadrateren en daarna te sommeren. Het is dus de teller van onze som bij het berekenen van de totale variantie, en

die is hier 3,241, zoals ook in tabel 10.4 te zien is in de eerste kolom onder *Initial Eigenvalues – Total*. De hoogst mogelijke eigenwaarde is in ons voorbeeld 5. We hebben immers vijf items gemeten, en wanneer deze precies hetzelfde zouden meten (zonder enige specifieke variantie of foutenvariantie) hebben we een perfecte schaal geconstrueerd zonder enig informatieverlies.
In tabel 10.4 zien we dat SPSS vijf componenten berekent als we vijf manifeste variabelen hebben. Alleen de eerste component heeft een eigenwaarde groter dan 1 en is de factor die we kunnen gebruiken. Dat komt goed uit want het was ook de bedoeling dat deze items één concept (namelijk smartphoneverslaving) zouden meten.
Een alternatief voor het gebruik van de grens van de waarde 1 voor de eigenwaarde is een *screeplot*. Ook daarmee kun je kijken naar de componenten die door een factoranalyse worden onderscheiden. De eigenwaarden van de componenten of factoren zijn hierin op volgorde gezet en vaak is duidelijk een 'elleboog' te zien in de lijn die deze eigenwaarden met elkaar verbindt (zie figuur 10.3). De onderzoeker beperkt zijn factoren dan tot het punt van de elleboog (de 'knik'). Ook op basis van deze screeplot kunnen we concluderen dat er één component ten grondslag ligt aan de vijf manifeste items. Een screeplot is voornamelijk een handig grafische hulpmiddel wanneer er meer dan één component wordt onderscheiden, waar we in de volgende paragraaf bij stil zullen staan.



Figuur 10.3 Screeplot met één component

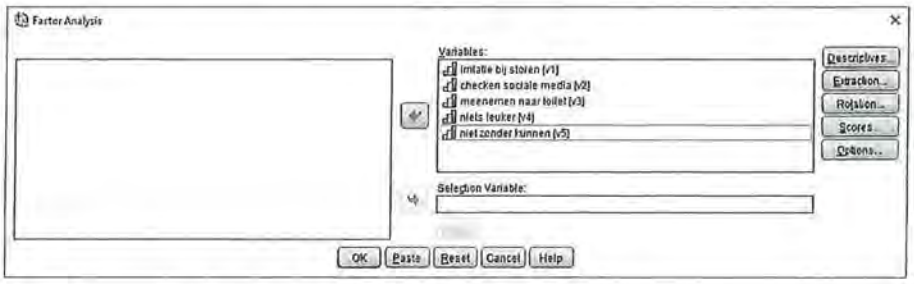
Een factoranalyse dient dus om te kijken in welke mate een aantal manifeste variabelen dezelfde latente variabele meten. Dit is een vorm van *constructvaliditeit*. Constructvaliditeit geeft aan in hoeverre het meetinstrument meet wat je wilt meten. Als de manifeste variabelen unidimensioneel zijn, dat wil zeggen

als er slechts één dimensie/factor aan ten grondslag ligt, kun je concluderen dat je variabelen één latente variabele meten.

Concluderend kunnen we aan de hand van de factoranalyse stellen dat onze vijf items een valide schaal vormen voor de latente variabele 'smartphoneverslaving'. Er werd namelijk één component gevormd met een eigenwaarde hoger dan 1. Bovendien was de totale verklaarde variantie voldoende hoog, en waren alle factorladingen hoger dan |0,45|.

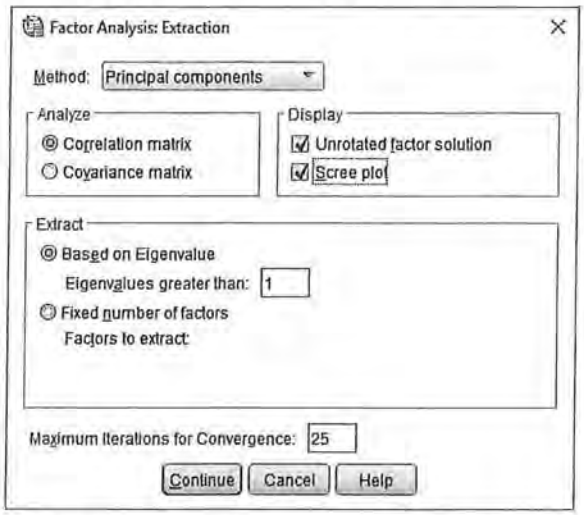
SPSS Uitvoeren van een factoranalyse

Wanneer we een factoranalyse in SPSS laten uitvoeren (via *Analyze → Dimension Reduction → Factor*), hebben we tal van opties om aan te geven welke informatie we willen.
In het vak *Variables* plaatsen we alle variabelen waar we een factoranalyse over willen uitdraaien. In figuur A is dat gedaan voor de vijf variabelen waarvan we denken dat ze een valide schaal voor smartphoneverslaving zullen meten.



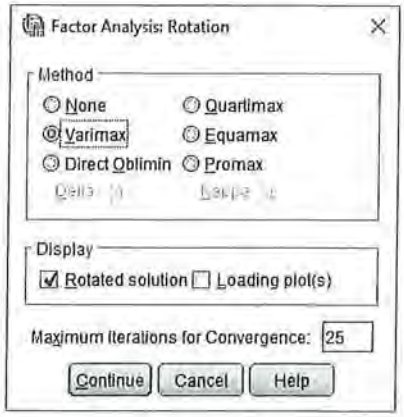
Figuur A Factor Analysis-venster

Wanneer je vervolgens op *Extraction* klikt, kun je onder *Display* aangeven dat je een screeplot wilt uitdraaien (figuur B).



Figuur B Extraction-venster: Scree plot aanvinken

Tot slot moet je bij Rotation (figuur C) aangeven dat je een *Varimaxrotatie* wilt laten uitvoeren. Het doel en nut daarvan wordt besproken in paragraaf 10.4.



Figuur C Rotation-venster: Varimax aanvinken

Kader 10.1

10.3.2 Betrouwbaarheidsanalyse: interne consistentie

Nadat we ons hebben verzekerd van de validiteit van onze meting, gaan we na of de schaal ook betrouwbaar is. Is het niet op basis van toeval dat de items met elkaar correleren en daardoor een valide schaal vormen? Zoals we in paragraaf 10.2 reeds aankondigden, wordt de betrouwbaarheid van de schaal, de interne consistentie van de schaal, gemeten aan de hand van Cronbachs alfa. SPSS geeft ons hiervoor twee tabellen, één met de waarde van alfa, en één waarin staat of we de mogelijkheid hebben om te schaal te verbeteren (dat wil zeggen: betrouwbaarder te maken).

Tabel 10.5 Betrouwbaarheidsanalyse door Cronbachs alfa (SPSS-output)

Reliability Statistics

Cronbach's Alpha	N of Items
,856	5

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
v1 irritatie bij storen	20,34	24,494	,577	,848
v2 checken sociale media	20,87	21,640	,650	,832
v3 meenemen naar toilet	21,15	20,002	,672	,831
v4 niets leuker	20,31	23,149	,738	,813
v5 niet zonder kunnen	20,40	22,250	,766	,803

De richtlijnen bij Cronbachs alfa zijn anders dan bij de interpretatie van factorladingen. We hanteren een ondergrens van 0,60 voordat we kunnen spreken van een redelijk betrouwbare schaal. Vanaf 0,80 vinden we de schaal betrouwbaar. In tabel 10.5 is in de tabel *Reliability Statistics* af te lezen dat onze vijf items een betrouwbare schaal kunnen vormen voor de latente variabele smartphoneverslaving: $\alpha = 0{,}86$. In de tabel daaronder, *Item-Total Statistics,* wordt per variabele aangegeven wat de betrouwbaarheid van de schaal zou zijn wanneer we dat item niet in onze schaal zouden opnemen (in de laatste kolom onder *Cronbach's Alpha if Item Deleted*). In dit geval kunnen we de schaal niet verbeteren: geen van de waarden in deze kolom is hoger dan de oorspronkelijke alfa van 0,856. We hoeven overigens de schaal ook niet te verbeteren, want de betrouwbaarheid is al hoog genoeg.

Mocht je een negatieve (maar voldoende hoge) factorlading bij het interpreteren van de factoranalyse over het hoofd hebben gezien, dan zul je in de kolom *Corrected Item-Total Correlation* dit nog kunnen nagaan door te kijken of hier een negatieve waarde in voorkomt. Is dat het geval, dan is het waarschijnlijk zo dat het item nog gehercodeerd moet worden.

Ook Cronbachs alfa kunnen we met de hand berekenen, en dat doen we aan de hand van een correlatiematrix. Aangezien alfa in de formule gebruikmaakt van correlaties op minimaal intervalniveau, dus van Pearsons correlaties, gebruiken wij deze ook, hoewel officieel het meetniveau van de variabelen ordinaal is. De informatie die we hiervoor gebruiken is de correlatiematrix zoals weergegeven in tabel 10.2.

Als verschillende variabelen hetzelfde verschijnsel (moeten) meten, is het te verwachten dat de correlatie tussen die variabelen hoog is. Zou één van de variabelen een lage of een qua richting tegengestelde correlatie met de andere variabelen vertonen, dan is deze variabele niet consistent. Die variabele wijkt dan af van het algemene patroon. Cronbachs alfa geeft een indicatie van de interne consistentie van de schaal die is samengesteld op basis van een aantal items. Bij de berekening van Cronbachs alfa maak je gebruik van de onderlinge correlaties. Het is een soort gemiddelde correlatie. Daarnaast is ook het aantal items van invloed op de waarde van Cronbachs alfa.

$$\alpha = \frac{k\bar{r}}{1+(k-1)\bar{r}}$$

Formule voor Cronbachs alfa

In deze formule staat k voor het aantal items en $\bar{r}$ voor het gemiddelde van de correlaties.

Met behulp van de correlaties tussen de vijf items die tezamen de schaal smartphoneverslaving vormen (tabel 10.2) kun je Cronbachs alfa berekenen. We hebben tien correlatiecoëfficiënten. De gemiddelde correlatie is de som van deze correlaties, gedeeld door het aantal correlaties:

$$\bar{r} = \frac{0{,}372+0{,}463+0{,}570+0{,}557+0{,}584+0{,}549+0{,}618+0{,}565+0{,}580+0{,}714}{10}$$

$$= \frac{5{,}572}{10} = 0{,}557$$

We hebben vijf items, dus voor k vullen we 5 in.

$$\alpha = \frac{k\bar{r}}{1+(k-1)\bar{r}} = \frac{5*0{,}557}{1+(5-1)*0{,}557} = \frac{2{,}785}{3{,}228} = 0{,}863$$

De met de hand berekende alfa valt iets hoger uit dan de alfa zoals berekend door SPSS omdat wij met drie decimalen achter de komma rekenen en SPSS met meer decimalen. De interpretatie blijft hetzelfde: α is groter dan 0,80 en we hebben een goede betrouwbare schaal. De vijf variabelen vormen een goede schaal die intern consistent is. We kunnen nu ook ons databestand uitbreiden door een nieuwe variabele 'smartphoneverslaving' te maken.

10.3.3 Maken en beschrijven van de schaal

Nu we hebben vastgesteld dat we een valide en betrouwbare schaal kunnen construeren voor smartphoneverslaving, kunnen we de schaal daadwerkelijk gaan maken. Dit doen we door gebruik te maken van het commando 'MEAN' zoals beschreven in hoofdstuk 4 (paragraaf 4.3, en kader 4.2). Onze nieuwe schaal heeft als meetniveau interval: het is een gemiddelde schaal waarin scores op vijf items bij elkaar worden gevoegd en gedeeld door het aantal items. We kunnen voor het beschrijven van de schaal gebruikmaken van *Explore* (zie paragraaf 3.4 en kader 3.1).

Op deze schaal van 1 tot 7, waarbij hoe hoger gescoord wordt, hoe meer de onderzoekseenheden smartphoneverslaafd zijn, scoren studenten gemiddeld 5,15 met een standaarddeviatie van 1,16. Dat wil zeggen dat de studenten behoorlijk verslaafd zijn aan hun smartphone. Uit deze tabel kan ook worden opgemaakt dat de schaal niet te scheef verdeeld is (*Skewness* = –0,62).

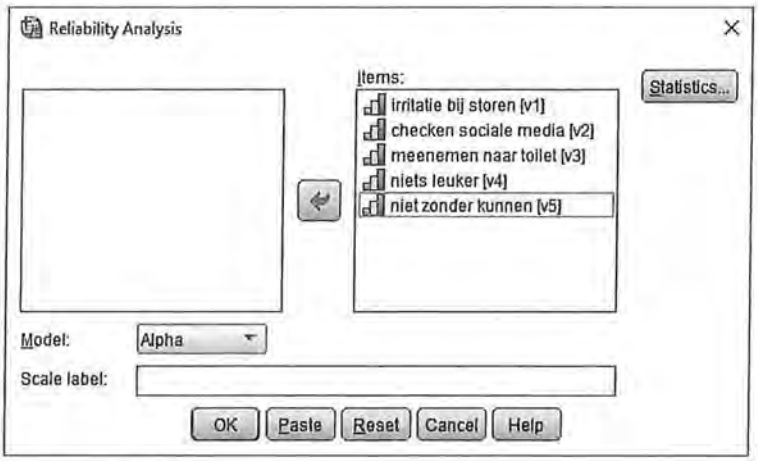
Tabel 10.6 Explore om schaal te beschrijven (SPSS-output)

Descriptives

			Statistics	Std. Error
smartphoneverslaving	Mean		5,1540	,07503
	95% Confidence Interval for Mean	Lower Bound	5,0062	
		Upper Bound	5,3018	
	5% Trimmed Mean		5,2055	
	Median		5,2000	
	Variance		1,346	
	Std. Deviation		1,16000	
	Minimum		1,20	
	Maximum		7,00	
	Range		5,80	
	Interquartile Range		1,80	
	Skewness		-,624	,157
	Kurtosis		,201	,314

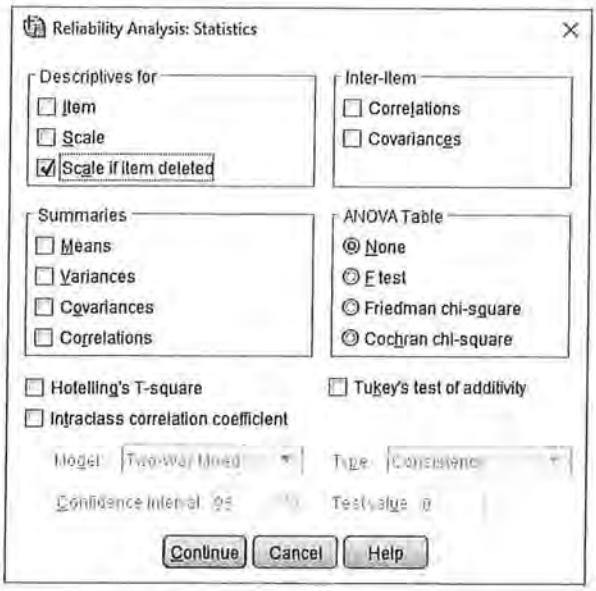
SPSS Berekenen van Cronbachs alfa

Het berekenen van Cronbachs alfa in SPSS gaat via *Analyze → Scale → Reliability Analysis*. In het vak *Items* vul je vervolgens de variabelen in die samen een schaal zouden moeten vormen (zie figuur A).



Figuur A Reliability Analysis-venster

Via *Statistics* kun je aan SPSS extra informatie vragen (zie figuur B), waarbij het aanvinken van 'Scale if Item deleted' de belangrijkste is.



Figuur B Statistics-venster

Kader 10.2

We hebben nu alle stappen van een schaalconstructie doorlopen en komen tot de volgende conclusie:

> *Uit een factoranalyse blijkt dat we één factor kunnen vormen voor de latente variabele 'smartphoneverslaving' met vijf manifeste items. Er is één component met een eigenwaarde hoger dan 1 (EV = 3,42) en de totale verklaarde variantie is 64,80%. Dit is een valide schaal waarbij geldt: hoe hoger wordt gescoord, hoe meer verslaafd de studenten zijn. De schaal is betrouwbaar (α = 0,86). Op de schaal, die loopt van 1 tot en met 7, scoren studenten gemiddeld hoog: M = 5,15, SD = 1,16.*

10.4 Meerdere factoren

Met een factoranalyse bereken je op basis van onderlinge correlaties een nieuwe variabele (een factor of component), die het best de onderliggende dimensie weergeeft. Het resultaat van een factoranalyse kan ook zijn dat er meer dan één dimensie ten grondslag ligt aan de variabelen die je gebruikt. We leggen dit uit aan de hand van een fictief voorbeeld.

Stel dat je wilt weten of leeftijd samenhangt met de vraag of mensen graag uitgaan. Je verwacht dat naarmate mensen ouder zijn, ze minder graag uitgaan. Je formuleert verschillende uitspraken over uitgaan, waarbij respondenten op een schaal van 1 tot 5 kunnen aangeven in welke mate ze het eens zijn met die

uitspraken (1 = zeer mee oneens, 5 = zeer mee eens). De volgende uitspraken leggen we ter beantwoording aan respondenten voor:

Ik ga graag naar het café.
Ik ga graag naar een popconcert.
Ik ga graag naar een (dans)club.
Ik ga graag naar een klassiek concert.
Ik ga graag naar het theater.
Ik ga graag naar het museum.

Als de zes stellingen over uitgaan, de manifeste variabelen, inderdaad 'het uitgaan', de latente variabele, meten, moeten ze voldoende informatie gemeenschappelijk hebben en elk apart iets toevoegen. Maar als je naar de stellingen kijkt, is het voorstelbaar dat die 'gemeenschappelijkheid' ontbreekt. Wanneer iemand bijvoorbeeld graag naar het theater gaat, graag een klassiek concert bezoekt, maar liever niet naar een café, popconcert of dansclub gaat, zou zijn score op de som van de zes items hetzelfde zijn als de score van iemand die juist graag naar café, popconcert en dansclub gaat en liever niet naar theater, concert of museum. Dat dit inderdaad het geval is, blijkt uit de correlatiematrix (tabel 10.7). Tussen de variabelen café, popconcert en club zien we sterke samenhangen en ook tussen de variabelen klassiek concert, theater en museum, terwijl de overige correlaties zwak tot nihil zijn. In deze correlatiematrix zijn duidelijk twee clusters te onderscheiden. Het eerste cluster wordt gevormd door de variabelen V1, V2 en V3 (café, popconcert en club) en het tweede cluster door de variabelen V4, V5 en V6 (klassiek concert, theater en museum).

Tabel 10.7 Correlatiematrix verschillende uitgaansbestemmingen

	V1	V2	V3	V4	V5	V6
V1 Café	1,00					
V2 Popconcert	0,680	1,00				
V3 Club	0,721	0,720	1,00			
V4 Klassiek concert	0,127	0,165	0,117	1,00		
V5 Theater	0,149	0,204	0,157	0,668	1,00	
V6 Museum	0,206	0,230	0,199	0,641	0,674	1,00

Je zou kunnen redeneren dat diegenen die hoog scoren op de variabelen klassiek concert, theater en museum hun vrije tijd liever aan 'kunst en cultuur' besteden en diegenen die hoog scoren op café, popconcert en club hun vrije tijd liever aan 'sociaal vermaak' besteden. De manifeste variabelen 'bezoek theater', 'museumbezoek' en 'bezoek klassieke concerten' zouden dan samen de latente variabele 'uitgaan: kunst en cultuur' meten. Café-, popconcert- en dansclubbezoek zijn dan de manifeste variabelen die de latente variabele 'uitgaan: sociaal vermaak' meten.

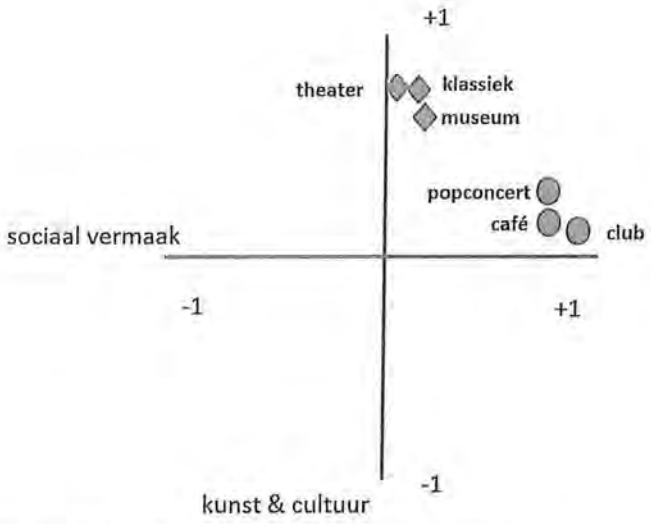
De namen van de factoren ('uitgaan: kunst en cultuur' en 'uitgaan: sociaal vermaak) hebben we in dit geval gekozen op basis van theoretische overwegingen. Een onderzoeker kan de naamgeving van de factoren ook pas naderhand bepalen op basis van de variabelen die het sterkst aan een factor bijdragen (de hoogste factorladingen hebben). Als onderzoekers een factoranalyse uitvoeren zonder dat ze van tevoren een idee hebben van mogelijke latente variabelen, en vervolgens op basis van de sterkst ladende manifeste variabelen een naam aan de gevonden factoren geven, spreken we van een *explorerende* factoranalyse. De naamgeving aan de factoren is afhankelijk van de keuze van de onderzoeker zelf. Bij een explorerende factoranalyse speelt de eigen interpretatie een grote rol bij de duiding van de factoren/latente variabelen. De onderzoeker gebruikt de variabelen die op een factor een hoge lading hebben om tot een interpretatie van de factor te komen. Als onderzoekers op theoretische gronden verwachten met een aantal manifeste variabelen bepaalde latente verschijnselen te meten, is sprake van een *confirmatieve* factoranalyse. De factoranalyse dient dan als bevestiging van de verwachtingen van de onderzoeker met betrekking tot de onderliggende structuur, dimensies van een verschijnsel. Met een confirmatieve factoranalyse hoopt de onderzoeker vooraf benoemde latente variabelen in de factoren te herkennen.

In de correlatiematrix van tabel 10.7 konden we zelf al twee clusters onderscheiden. Als we in SPSS een factoranalyse uitvoeren, wordt dit bevestigd. Er zijn inderdaad twee factoren. In de zogenoemde factorladingmatrix (tabel 10.8) zien we voor elke variabele de correlaties – de factorladingen – met die twee latente variabelen.

Tabel 10.8 Factorladingen voor twee componenten

	Factor 1	Factor 2
Café	0,888	0,083
Popconcert	0,880	0,136
Club	0,907	0,076
Klassiek	0,093	0,887
Theater	0,049	0,877
Museum	0,150	0,864

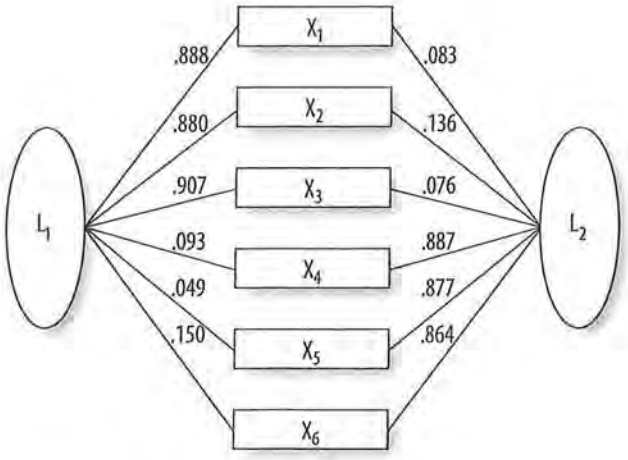
Als we deze informatie van de factorladingen grafisch zouden weergeven op een assenstelsel (figuur 10.4), wordt nogmaals duidelijk dat het hier om twee onderscheiden clusters gaat. Het idee van een factoranalyse is dat een item óf bij de ene factor, óf bij de andere factor wordt ingedeeld.



Figuur 10.4 Factorladingen van uitgaan items

De variabelen 'café', 'popconcert' en 'club' scoren allemaal hoog op factor 1, maar laag op factor 2. Bij de 'kunst en cultuur'-variabelen is dat juist omgekeerd. Hieruit is af te leiden dat 'café', 'popconcert' en 'club' gezamenlijk één factor, of latente variabele, vormen, namelijk 'uitgaan: sociaal vermaak', en 'klassiek concert', 'theater' en 'museum' de latente variabele 'uitgaan: kunst en cultuur' vormen.

Deze factorladingen staan bij de lijnen in de schematische voorstelling van de latente en manifeste variabelen van figuur 10.5.

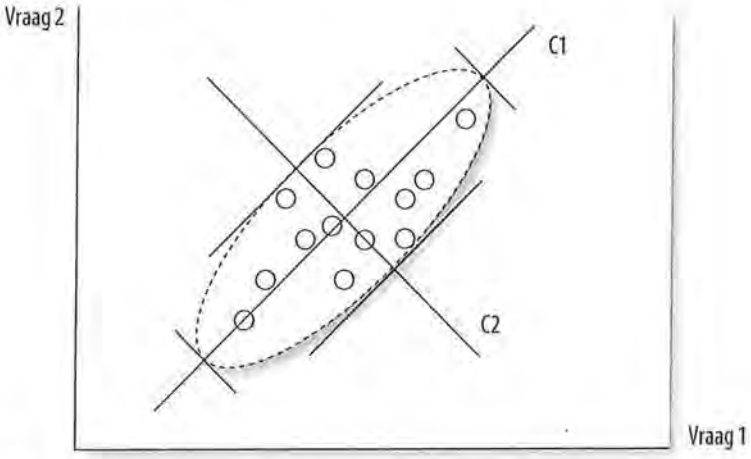


Figuur 10.5 Factorladingen schematisch

10.4.1 Het vinden van de factoren

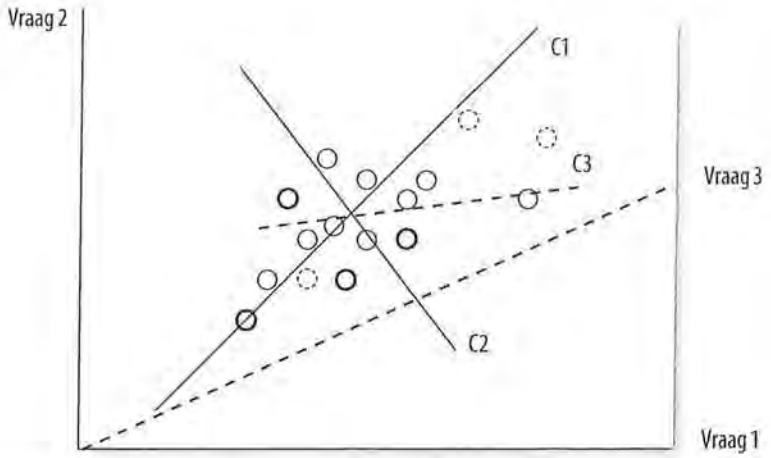
Er zijn verschillende methoden waarop onderliggende factoren uit de correlaties tussen de manifeste variabelen kunnen worden geëxtraheerd. *Principale componentenanalyse* (PCA) is een van deze methoden. Bij principale componentenanalyse worden de onderliggende dimensies die uit de samenhang tussen de variabelen worden geëxtraheerd, componenten[1] genoemd.

Om te verduidelijken hoe de onderliggende dimensies gevonden worden, starten we eerst met maar twee (manifeste) variabelen. In het spreidingsdiagram van figuur 10.6 zijn de scores van de respondenten op twee variabelen weergegeven. Aan de puntenwolk in dit spreidingsdiagram is te zien dat er samenhang is tussen de twee variabelen. Er kan een lijn worden getrokken die de variantie binnen de puntenwolk het best verklaart. Dit is de eerste component (C1). De tweede component (C2) wordt hier loodrecht op gezet. Door de tweede component loodrecht op de eerste te plaatsen is er geen samenhang tussen de twee componenten. Deze tweede lijn verklaart ook nog enige variantie in de puntenwolk, maar veel minder. Deze twee componenten zijn de onderliggende dimensies die de variantie in de puntenwolk verklaren, zij vertonen onderling geen samenhang.



Figuur 10.6 Spreidingsdiagram van twee variabelen

Als we hetzelfde doen voor drie manifeste variabelen, ontstaat een driedimensionale puntenwolk en vinden we een derde component die loodrecht op de eerste twee kan worden getekend. We zien dan een driedimensionaal assenstelsel en dat kunnen we ons nog wel voorstellen, maar bij vier manifeste variabelen is het visualiseren van de componenten al onmogelijk.



Figuur 10.7 Driedimensionaal spreidingsdiagram van drie variabelen

Bij de principale componentenanalyse worden op deze manier altijd evenveel componenten gevonden als er manifeste variabelen zijn en deze componenten verklaren tezamen 100% van de variantie. Bij factoranalyse wordt in een puntenwolk gezocht naar een lijn, een factor, die zo veel mogelijk van de gemeenschappelijke variantie verklaart. Een volgende factor moet daarna zo veel mogelijk van de nog resterende gemeenschappelijke variantie verklaren.

10.4.2 Het interpreteren van de factoren en de noodzaak van rotatie

Bij de principale componentenanalyse zijn er evenveel componenten als manifeste variabelen. Een onderzoeker wil het aantal variabelen graag zo ver mogelijk reduceren tot die componenten of factoren die er werkelijk toe doen. Hoe bepaal je nu hoeveel factoren er echt toe doen? Dat gebeurt op basis van de verklaarde variantie en de eigenwaarden (zie ook paragraaf 10.3.1). De eerste component is altijd de component die de meeste variantie verklaart, bij elke volgende is de verklaarde variantie lager en die is dus ook minder belangrijk. Naast een percentage verklaarde variantie wordt de informativiteit van een factor uitgedrukt in de eigenwaarde. Deze werd per component berekend door de factorladingen te kwadrateren en bij elkaar op te tellen. Deel je de uitkomst daarvan door het aantal items, dan krijg je de proportie verklaarde variantie per component.

Om te illustreren hoe je als onderzoeker probeert je factoren te interpreteren, gaan we verder met de zes stellingen die in paragraaf 10.1.1 zijn geformuleerd om de media-afhankelijkheid van respondenten te meten. Anders dan in het fictieve voorbeeld van de uitgaansvariabelen zijn in de correlatiematrix van de media-afhankelijkheidsvariabelen niet direct clusters te onderscheiden (tabel 10.9). De correlaties variëren tussen 0,02 en 0,58.

Tabel 10.9 Correlatiematrix media-afhankelijkheidsvariabelen (SPSS-output)

Correlations

		v1 ver-maken	v2 normen en waarden	v3 begrip wereld	v4 weten wat te doen	v5 mijzelf begrijpen	v6 anderen vermaken
v1 verma-ken	Pearson Correlation	1	,015	,122	,277**	,352**	,584**
	Sig. (2-tailed)		,818	,059	,000	,000	,000
	N	239	238	239	239	239	239
v2 normen en waar-den	Pearson Correlation	,015	1	,330**	,312**	,357**	,151*
	Sig. (2-tailed)	,818		,000	,000	,000	,019
	N	238	238	238	238	238	238
v3 begrip wereld	Pearson Correlation	,122	,330**	1	,399**	,403**	,240**
	Sig. (2-tailed)	,059	,000		,000	,000	,000
	N	239	238	239	239	239	239
v4 weten wat te doen	Pearson Correlation	,277**	,312**	,399**	1	,451**	,318**
	Sig. (2-tailed)	,000	,000	,000		,000	,000
	N	239	238	239	239	239	239
v5 mijzelf begrijpen	Pearson Correlation	,352**	,357**	,403**	,451**	1	,440**
	Sig. (2-tailed)	,000	,000	,000	,000		,000
	N	239	238	239	239	239	239
v6 anderen vermaken	Pearson Correlation	,584**	,151*	,240**	,318**	,440**	1
	Sig. (2-tailed)	,000	,019	,000	,000	,000	
	N	239	238	239	239	239	239

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

We gebruiken SPSS om door middel van een factoranalyse hier de onderliggende dimensies (factoren) uit te halen. In de output (tabel 10.10) vinden we een tabel met de geëxtraheerde componenten met de waarden van de eigenwaarden. De eigenwaarde van de eerste component is 2,63, deze component verklaart 43,8% van de variantie in de zes variabelen. De eigenwaarde van de tweede component is 1,23, deze component verklaart 20,5% van de variantie in de zes variabelen. De overige componenten hebben een eigenwaarde kleiner dan 1. Op basis daarvan kan de onderzoeker besluiten het onderzoek te beperken tot de eerste twee componenten. Gezamenlijk verklaren deze twee componenten 64,3% van de variantie. Daarmee accepteert de onderzoeker informatieverlies: de zes manifeste variabelen geven meer informatie over de variatie onder de respondenten dan de twee latente variabelen waartoe hij zich gaat beperken. Daar staat tegenover dat hij nu twee variabelen kan maken waarmee de mate van media-afhankelijkheid wordt gekwantificeerd.

Tabel 10.10 Verklaarde variantie bij factoranalyse met media-afhankelijkheidsvariabelen (SPSS-output)

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,628	43,803	43,803	2,628	43,803	43,803	2,001	33,343	33,343
2	1,230	20,493	64,296	1,230	20,493	64,296	1,857	30,953	64,296
3	,674	11,227	75,524						
4	,582	9,700	85,224						
5	,494	8,233	93,458						
6	,393	6,542	100,000						

Extraction Method: Principal Component Analysis.

SPSS geeft ons ook een componentenmatrix (tabel 10.11) waarin de factorladingen op de twee componenten met een eigenwaarde groter dan 1 zijn gegeven. Deze tabel is te vergelijken met tabel 10.8, waar we de factorladingen op 'uitgaan' lieten zien. Op basis van de factorladingen in deze matrix kan de onderzoeker proberen de componenten te benoemen. Dat wordt in dit geval bemoeilijkt doordat een aantal variabelen op beide componenten hoog laden. Bij bijvoorbeeld de variabele 'mijzelf begrijpen' is de factorlading op de eerste component 0,786 en op de tweede component 0,062. Deze variabele 'hoort' dus duidelijk bij component 1. Maar bij de variabele 'normen en waarden' is de factorlading op de eerste component 0,507 en op de tweede component 0,587. Beide ladingen zijn in ieder geval hoger dan |0,45|, wat betekent dat de factorlading hoog genoeg is. Nu zouden we gewoon kunnen zeggen: 0,587 (factorlading van de variabele 'normen en waarden' op component 2) is hoger dan 0,507 (factorlading van de variabele 'normen en waarden' op component 1), maar zo eenvoudig werkt het niet.

Tabel 10.11 Componentenmatrix (SPSS-output)

Component Matrix[a]

	Component	
	1	2
v1 vermaken	,598	-,660
V2 normen en waarden	,507	,587
v3 begrip wereld	,622	,422
v4 weten wat te doen	,711	,181
v5 mijzelf begrijpen	,786	,062
v6 anderen vermaken	,709	-,485

Extraction Method: Principal Component Analysis.
a. 2 components extracted.

We kunnen ook dit probleem gelukkig door SPSS laten oplossen.

De twee componenten vormen een assenstelsel waar de zes variabelen mee samenhangen. Als we met dat assenstelsel gaan draaien (roteren), veranderen de factorladingen (de correlaties van de variabelen met de twee componenten). SPSS probeert een stand te vinden waarbij het aantal hoge factorladingen op een component wordt gemaximaliseerd en het aantal lage factorladingen wordt geminimaliseerd. Zo kan een beter onderscheid worden gemaakt tussen de twee componenten. Het resultaat vinden we in de geroteerde componentenmatrix (tabel 10.12). Het vinden van de optimale stand van het assenstelsel gebeurt volgens een 'trial and error'-procedure. Factorladingen worden steeds opnieuw berekend tot een optimale stand is bereikt (we noemen dit een iteratief proces). In dit geval vindt SPSS de optimale factorladingen na drie pogingen. Dit is onder de geroteerde componentenmatrix aangegeven ('Rotation converged in 3 iterations'). Er bestaan meerdere vormen van roteren. Wij gebruiken de methode *Varimaxrotatie.*

Tabel 10.12 Componentenmatrix na rotatie (SPSS-output)

Rotated Component Matrix[a]

	Component	
	1	2
v1 vermaken	,002	,890
V2 normen en waarden	,770	-,097
v3 begrip wereld	,745	,104
v4 weten wat te doen	,649	,342
v5 mijzelf begrijpen	,626	,481
v6 anderen vermaken	,202	,835

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

Nu is interpretatie van de componenten/factoren eenvoudiger. Vooral de tweede factor is direct duidelijk. De variabelen 'vermaken' en 'met anderen vermaken' hebben op deze component de hoogste factorladingen. Hier gaat het om afhankelijkheid van de media voor entertainment, voor het vermaak. De eerste factor zou surveillance kunnen worden genoemd, de afhankelijkheid van de media om te kunnen functioneren in je sociale omgeving. Dit interpreteren van de factoren blijft een subjectief proces en het is mogelijk dat een andere onderzoeker op basis van dezelfde componentenmatrix tot andere naamgevingen besluit. Na rotatie hebben we twee nieuwe componenten die niet meer overeenkomen met de initiële componenten en daardoor is ook de verklaarde variantie van de twee componenten veranderd. In tabel 10.10 zien we deze cijfers in de een-na-laatste kolom. Na rotatie verklaart component 1 (variabelen v2 tot en met v5) 33,3% van de variantie in de afzonderlijke items en component 2 (variabelen v1 en v6) 31,0%.

Nu we weten dat we twee valide schalen kunnen maken die beide een ander aspect van media-afhankelijkheid meten, moeten we ook tweemaal een betrouwbaarheidsanalyse uitvoeren, en tweemaal een schaal maken en beschrijven.

10.5 Gebruik en presentatie van de resultaten

Als je een factoranalyse in SPSS draait, kun je SPSS vragen de gevonden factoren als variabelen aan je databestand toe te voegen. Die 'factorvariabelen' kun je vervolgens gebruiken in je verdere analyses, maar je kunt ook besluiten de factoranalyse te gebruiken als rechtvaardiging om de variabelen die hoog op een factor laden, zelf op te tellen tot een nieuwe variabele. Welke keuze je ook maakt, het is wel noodzakelijk om op een correcte wijze verslag te doen van je factoranalyse. Naast de factoranalyse bespreek je de betrouwbaarheid van de schaal of schalen, het gemiddelde en de standaarddeviatie van de schalen, en de antwoordschaal om de waarden in perspectief te zetten. Op basis van de factoranalyse van de media-afhankelijkheidsvariabelen schrijf je bijvoorbeeld in je onderzoeksverslag:

> *'Media-afhankelijkheid' is gemeten door de respondenten te vragen in welke mate ze de media gebruiken om zichzelf te vermaken, meer te weten te komen over de normen en waarden in de maatschappij, meer begrip te krijgen van de wereld om hen heen, zichzelf beter te begrijpen, te weten wat ze in allerlei situaties het best kunnen doen en zich met anderen te vermaken. Uit de factoranalyse met Varimaxrotatie bleek dat er twee dimensies zijn te onderscheiden (eigenwaarde > 1), te weten 'surveillance' en 'entertainment'. Samen verklaren deze factoren 64,3% van de variantie. Beide schalen zijn redelijk betrouwbaar ('surveillance' $\alpha = 0{,}74$, 'entertainment' $\alpha = 0{,}71$) (tabel 10.13). Respondenten zijn redelijk afhankelijk van media om in hun sociale omgeving te functioneren ($M = 4{,}86$, $SD = 0{,}99$) en redelijk afhankelijk van de media voor entertainment ($M = 4{,}76$, $SD = 1{,}48$), beide gemeten op een schaal van 1 tot en met 7.*

Tabel 10.13 Factoranalyse (met rotatie) media-afhankelijkheid

Door de media te gebruiken ...	Surveillance	Entertainment
– kan ik mezelf goed vermaken		**0,89**
– kom ik meer te weten over normen en waarden in de maatschappij	**0,77**	
– begrijp ik meer van de wereld om mij heen	**0,75**	
– ga ik mijzelf beter begrijpen	**0,65**	0,48
– weet ik wat ik in allerlei situaties het best kan doen	**0,63**	
– kan ik me met anderen vermaken		**0,84**
Eigenwaarde	2,00	1,86
Verklaarde variantie	33,3%	31,0%
Cronbachs alfa	0,74	0,71

Factorladingen na varimaxrotatie; factorladingen < |0,45| zijn niet weergegeven.[6]

10.6 Overige vormen van betrouwbaarheid

In voorgaande paragrafen hebben we naar de interne consistentie gekeken om de betrouwbaarheid van een schaal te bepalen. We willen tot slot nog stilstaan bij andere vormen van betrouwbaarheid. Het uitrekenen van statistieken op basis van de variabelen in je databestand is alleen maar zinvol als je waarnemingen op een betrouwbare wijze zijn gemeten. De betrouwbaarheid van een meting is de mate waarin die meting vrij is van toevallige fouten. Dat betekent dat als je een meting nog een keer uitvoert op dezelfde manier bij dezelfde onderzoekseenheden, je dezelfde waarde moet vinden. Als dat niet het geval is, is je meting niet betrouwbaar.

De manier waarop je een kenmerk van je onderzoekseenheden meet, verdient daarom de nodige aandacht. Sommige kenmerken zijn gemakkelijk te meten, zoals geslacht en leeftijd. De kans op toevallige fouten is voor deze kenmerken niet groot, maar ook niet uitgesloten. Andere kenmerken zijn minder eenvoudig te meten, bijvoorbeeld een concept als 'vertrouwen in de media' of 'media-afhankelijkheid'. De kans op toevallige fouten wordt dan groter.
De betrouwbaarheid van metingen kunnen we met behulp van verschillende procedures en statistieken vaststellen. De te onderscheiden procedures leiden tot het vaststellen van verschillende aspecten van betrouwbaarheid, namelijk betrouwbaarheid in de zin van *stabiliteit*, *equivalentie* en *interne consistentie* die we al eerder in dit hoofdstuk hebben we besproken. Om na te gaan of de metingen *intern consistent* zijn gebruiken we een schaalanalyse. In deze paragraaf gaan we dieper in op de stabiliteit en equivalentie van metingen. Of een meting stabiel is, stel je vast door herhaalde metingen, de test-hertestprocedure.

Je herhaalt dan je meting op dezelfde manier. Als we eenzelfde verschijnsel op een andere manier kunnen meten (parallelle tests), kunnen we nagaan of de resultaten van die metingen tot dezelfde (*equivalente*) resultaten leiden. Als herhaalde metingen of parallelle metingen tot dezelfde resultaten leiden, is de meting betrouwbaar. Voor het testen van stabiliteit en equivalentie kunnen we voor een deel de in de vorige hoofdstukken besproken associatiematen gebruiken, maar daarnaast worden in dit hoofdstuk ook andere statistieken genoemd.

10.6.1 Stabiliteit en equivalentie

Als je metingen betrouwbaar zijn, zou je elke keer dat je dezelfde meting bij dezelfde onderzoekseenheden opnieuw uitvoert dezelfde waarden moeten vinden. Er is dan sprake van *stabiliteit*. Als het mogelijk is eenzelfde meting te herhalen, is dat een test-hertestbetrouwbaarheidsmeting. Wanneer je een vraag in een vragenlijst twee keer aan dezelfde respondenten voorlegt, krijg je in je databestand twee variabelen waartussen je de samenhang kunt berekenen. Aangezien beide metingen hetzelfde verschijnsel meten, zou het verband, de samenhang, zeer sterk/perfect moeten zijn. Om dat vast te stellen kun je de eerder besproken associatiematen gebruiken. De waarde die je dan vindt, is een indicatie voor de betrouwbaarheid van je metingen. Je interpreteert de waarden van de associatiematen dan wel anders dan wanneer we de samenhang tussen twee verschillende variabelen berekenen. Als je twee keer de respondenten naar hun inkomen vraagt en je daarna de correlatiecoëfficiënt berekent (omdat het hier om variabelen op rationiveau gaat), ben je niet erg tevreden met een r van 0,80. Als er helemaal geen toevallige fouten in je metingen zitten, zouden de twee variabelen identiek moeten zijn. Je verwacht dan dat r gelijk is aan 1.
Het herhalen van een meting is niet altijd mogelijk. Respondenten zullen aardig geïrriteerd raken als je vragen twee keer stelt. Bij het coderen van categorieën voor een inhoudsanalyse en bij observaties is dat gemakkelijker te verwezenlijken. Voor het vaststellen van de betrouwbaarheid van een meting kun je de meting door dezelfde codeur of observator laten herhalen (intracodeur- of intraobservatorbetrouwbaarheid) of een tweede codeur of observator inschakelen die dezelfde onderzoekseenheden codeert of observeert (intercodeur- of interobservatorbetrouwbaarheid). De eerste vorm waarborgt de stabiliteit van de meting, de tweede vorm de *equivalentie* van een meting. Equivalentie betekent dat bij dezelfde onderzoekseenheden op (minimaal) twee manieren hetzelfde verschijnsel wordt gemeten. Wanneer de resultaten (nagenoeg) identiek zijn, is er sprake van een betrouwbare meting. In een vragenlijst zou je dat kunnen doen door een parallelle test, waarbij je op een andere manier dezelfde informatie bij dezelfde onderzoekseenheden meet. Dit kun je doen door bijvoorbeeld niet alleen de leeftijd in jaren te vragen, maar ook het geboortejaar. Een hoge correlatie tussen die twee variabelen duidt op betrouwbare metingen.

De meeste simpele manier voor het berekenen van de inter- of intracodeurbetrouwbaarheid bij een inhoudsanalyse is het berekenen van het *overeenstemmingspercentage*. Het overeenstemmingspercentage is een heel grove maat die aangeeft welk percentage van de coderingen door twee codeurs identiek zijn uitgevoerd. Je berekent dit overeenstemmingspercentage op basis van een kruistabel. In deze kruistabel worden de codering door codeur 1 en de codering door codeur 2 van eenzelfde variabele tegen elkaar afgezet.
In tabel 10.14 staan de coderingen van de gevolgen van klimaatverandering (1 = geen, 2 = droogte, 3 = overstromingen) die in nieuwsberichten genoemd zijn. Er zijn tien nieuwsberichten (de onderzoekseenheden) die door zowel codeur 1 als door codeur 2 zijn gecodeerd. De coderingen van codeur 1 staan in de kolommen en de coderingen van codeur 2 staan in de rijen van de kruistabel. In de diagonaal van deze tabel vinden we de berichten die van beide codeurs dezelfde waarde hebben gekregen. Van drie berichten vinden ze allebei dat er 'geen gevolgen' in worden genoemd. In drie berichten hebben ze allebei geconstateerd dat hierin 'de droogte' is vermeld en van twee berichten vinden ze beide dat 'de overstromingen' genoemd worden. Die 3 + 3 + 2 = 8 berichten hebben ze identiek gecodeerd. Over de overige twee gecodeerde berichten verschillen ze van mening. Het overeenstemmingspercentage is 8÷10 = 80%. Met 80% overeenstemming zul je als onderzoeker over het algemeen niet zo heel erg tevreden zijn. Het betekent immers ook dat de ene codeur in 20% van de gevallen de operationalisatie van de begrippen anders heeft opgevat dan de andere codeur. Met een klein aantal onderzoekseenheden zoals in ons voorbeeld, hebben twee fouten een groter effect op het overeenstemmingspercentage dan wanneer je bijvoorbeeld op honderd artikelen twee keer geen overeenstemming bereikt. Deze betrouwbaarheidsstatistiek gebruik je dan ook meestal als proefcoderingen om je codeersysteem nog te kunnen verbeteren.

Tabel 10.14 Coderingen gevolgen klimaatverandering door twee codeurs:

Codeur 1 / Codeur 2	(1) Geen gevolgen	(2) Droogte	(3) Over-Stromingen	Totaal
(1) Geen gevolgen	3	0	0	3
(2) Droogte	1	3	1	5
(3) Overstromingen	0	0	2	2
Totaal	4	3	3	10

Verschillende statistische maten die gebruikt worden bij inhoudsanalyse zijn gebaseerd op de mate van overeenstemming, zoals Scotts pi, Krippendorfs alfa en Cohens kappa. Deze maten houden ook nog rekening met de toevalskans. We zullen deze maten niet bespreken in dit boek. Associatiematen als Kendalls tau-b en Spearmans rho (ordinaal) en Pearsons productmomentcorrelatie *r* (interval, ratio) kunnen ook worden gebruikt voor het vaststellen van de stabiliteit of equivalentie van een meting.

10.7 Samenvatting

In dit hoofdstuk hebben we laten zien welke stappen genomen moeten worden in het construeren van een schaal. Daartoe ga je na of je valide en betrouwbare metingen hebt uitgevoerd. We hebben laten zien hoe je door middel van een factoranalyse latente variabelen kunt vinden die ten grondslag liggen aan de (manifeste) variabelen die je hebt gemeten. Je kunt factoranalyse gebruiken om iets te zeggen over de validiteit van je metingen.
Als de factoranalyse aantoont dat je manifeste variabelen unidimensioneel zijn en dus hetzelfde verschijnsel meten, rechtvaardigt dit het samenvoegen van de manifeste variabelen tot één schaalvariabele voor het te meten verschijnsel. Op die manier kan een groot aantal variabelen worden teruggebracht tot hanteerbare hoeveelheden (datareductie), waardoor de analyses voor je onderzoek vereenvoudigd worden.
Daarnaast moet de meting betrouwbaar zijn. Betrouwbaarheid is de mate waarin je waarnemingen vrij zijn van toevallige fouten. Daarbij onderscheid je drie aspecten: stabiliteit, equivalentie en interne consistentie.
Bij een test-hertest kun je op basis van associatiematen iets zeggen over de stabiliteit van je metingen. Een parallelle test geeft informatie over de equivalentie. De interne consistentie van een schaal waarvoor je meerdere items gebruikt, stel je vast door Cronbachs alfa uit te rekenen.

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

Noten

1 Er zijn veel soorten validiteit, die we niet allemaal zullen bespreken in dit boek. Bij het uitvoeren van een factoranalyse om na te gaan of je een valide meting hebt uitgevoerd, gaat het in de regel om *construct- of begripsvaliditeit*, waarbij je nagaat of de items die je hebt gemeten inderdaad een indicator zijn van datgene wat je beoogde te meten.
2 Er zijn manieren om schalen te maken met wegingen van scores bij variabelen die niet op dezelfde manier gemeten zijn. Deze zullen wij echter niet in dit boek behandelen.
3 Dit voorbeeld is gebaseerd op het volgende onderzoek: Kwon, M., Kim, D.J., Cho, H. & Yang, S. (2013). *The Smartphone Addiction Scale: Development and Validation of a Short Version for Adolescents.* PLoS ONE 8(12): e83558. doi:10.1371/journal.pone.0083558.
4 Net als bij de richtlijnen bij de interpretatie van een associatiemaat is dit geen universele regel.
5 Het onderscheiden van meerdere componenten wordt besproken in de volgende paragraaf.
6 Bij principale componentenanalyse zouden we consequent de term component moeten gebruiken, maar omdat de interpretatie van factoren en componenten identiek is, worden in dit hoofdstuk beide termen gebruikt.

Formuleblad Beschrijvende statistiek

Centrum- en spreiding, z-score

Gemiddelde

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ of } \bar{x} = \frac{\sum_{j=1}^{k} x_j f_j}{n}$$

Variatie

$$\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Variantie

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Standaarddeviatie

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Z-score

$$z = \frac{x_i - \bar{x}}{s}$$

Waarin:

n = totaal aantal waarnemingen

k = totaal aantal groepen

x_i = score van waarneming i op variabele x

x_j = score van waarneming j op variabele x

f_j = frequentie van j

i = index die loopt van 1 tot n

j = groepsscore die loopt van 1 tot k

$\sum_{i=1}^{n} x_i$ = som van de waarden van variabele x van de i-de tot en met de n-de waarneming

Associatiematen op minimaal nominaal niveau

Chi-kwadraat

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

Waarin:
f_o = geobserveerde frequentie
f_e = verwachte frequentie
r = aantal rijen
k = aantal kolommen

Cramers V

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

Phi

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

Lambda

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

Met $E_1 = n - fMo(y)$
en $E_2 = n - \sum fMo(y)_{kx}$

Waarin:
E_1 = aantal voorspellingsfouten zonder x
E_2 = aantal voorspellingsfouten met x
$fMo(y)$ = frequentie van de modus van y
$fMo(y)_{kx}$ = frequentie van de modus van y per kolom van x

Goodman & Kruskals tau

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

Met $E_1 = \sum_i \left(\frac{n - R_i}{n} R_i \right)$
en $E_2 = \sum_j E_{2j}$
waarbij
$E_{2j} = \sum_j \left(\frac{C_j - O_{ij}}{C_j} O_{ij} \right)$

R_i = totaal van Rij i
C_j = totaal van Kolom j
O_{ij} = aantal waarnemingen in rij i en kolom j

Associatiematen op minimaal ordinaal niveau

Gamma

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

Somers' d

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

Kendalls tau-b

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

Spearmans rho

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

Waarin:
Nc = aantal concordante paren
Nd = aantal discordante paren
Ty = aantal geknoopte paren op y
Tx = aantal geknoopte paren op x

Waarin:
d = verschil tussen de rangnummers van de waarden van x en y

Associatiematen op minimaal interval niveau

Covariantie

$$Cov(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

Waarin:
x_i = score van waarneming i op variabele x
y_i = score van waarneming i op variabele y

Correlatie

$$r_{xy}=\frac{Cov(x,y)}{s_x s_y}$$

s_x = standaarddeviatie van x
s_y = standaarddeviatie van y

$$\text{of } r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y}$$

Regressielijn

Enkelvoudig:

$$\hat{y}=a+bx$$

Waarin:
a = intercept, constante
b = ongestandaardiseerde regressiecoëfficiënt

Intercept

$$a=\bar{y}-b\bar{x}$$

Ongestandaardiseerde regressiecoëfficiënt

$$b=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

Proportie verklaarde variantie

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

E_1 = totale variantie
E_2 = onverklaarde variantie

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

Meervoudige regressievergelijking

$$\hat{y}=a+b_1x_1+b_2x_2+\ldots+b_kx_k$$

b_k = ongestandaardiseerde regressiecoëfficiënt van variabele x_k
x_k = score op de k-de onafhankelijke variabele

Variantieanalyse

Eta²

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

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

Waarin:
y = gemiddelde van alle onderzoekseenheden
y_j = waarde van y per groep van j
$\bar{y}_j$ = gemiddelde van groep j
k = totaal aantal groepen

Eta

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

Betrouwbaarheidsanalyse

Cronbachs alfa

$$\alpha = \frac{k\bar{r}}{1+(k-1)\bar{r}}$$

Waarin:
k = aantal variabelen
$\bar{r}$ = gemiddelde correlatie

Rechter overschrijdingskansen in de standaardnormale verdeling

z	$P_R(z)$	z	$P_R(z)$	z	$P_R(z)$	z	$P_R(z)$
0,00	0,5000	0,38	0,3520	0,76	0,2236	1,14	0,1271
0,01	0,4960	0,39	0,3483	0,77	0,2206	1,15	0,1251
0,02	0,4920	0,40	0,3446	0,78	0,2177	1,16	0,1230
0,03	0,4880	0,41	0,3409	0,79	0,2148	1,17	0,1210
0,04	0,4840	0,42	0,3372	0,80	0,2119	1,18	0,1190
0,05	0,4801	0,43	0,3336	0,81	0,2090	1,19	0,1170
0,06	0,4761	0,44	0,3300	0,82	0,2061	1,20	0,1151
0,07	0,4721	0,45	0,3264	0,83	0,2033	1,21	0,1131
0,08	0,4681	0,46	0,3228	0,84	0,2005	1,22	0,1112
0,09	0,4641	0,47	0,3192	0,85	0,1977	1,23	0,1093
0,10	0,4602	0,48	0,3156	0,86	0,1949	1,24	0,1075
0,11	0,4562	0,49	0,3121	0,87	0,1922	1,25	0,1056
0,12	0,4522	0,50	0,3085	0,88	0,1894	1,26	0,1038
0,13	0,4483	0,51	0,3050	0,89	0,1867	1,27	0,1020
0,14	0,4443	0,52	0,3015	0,90	0,1841	1,28	0,1003
0,15	0,4404	0,53	0,2981	0,91	0,1814	1,29	0,0985
0,16	0,4364	0,54	0,2946	0,92	0,1788	1,30	0,0968
0,17	0,4325	0,55	0,2912	0,93	0,1762	1,31	0,0951
0,18	0,4286	0,56	0,2877	0,94	0,1736	1,32	0,0934
0,19	0,4247	0,57	0,2843	0,95	0,1711	1,33	0,0918
0,20	0,4207	0,58	0,2810	0,96	0,1685	1,34	0,0901
0,21	0,4168	0,59	0,2776	0,97	0,1660	1,35	0,0885
0,22	0,4129	0,60	0,2743	0,98	0,1635	1,36	0,0869
0,23	0,4090	0,61	0,2709	0,99	0,1611	1,37	0,0853
0,24	0,4052	0,62	0,2676	1,00	0,1587	1,38	0,0838
0,25	0,4013	0,63	0,2643	1,01	0,1562	1,39	0,0823
0,26	0,3974	0,64	0,2611	1,02	0,1539	1,40	0,0808
0,27	0,3936	0,65	0,2578	1,03	0,1515	1,41	0,0793
0,28	0,3897	0,66	0,2546	1,04	0,1492	1,42	0,0778
0,29	0,3859	0,67	0,2514	1,05	0,1469	1,43	0,0764
0,30	0,3821	0,68	0,2483	1,06	0,1446	1,44	0,0749
0,31	0,3783	0,69	0,2451	1,07	0,1423	1,45	0,0735
0,32	0,3745	0,70	0,2420	1,08	0,1401	1,46	0,0721
0,33	0,3707	0,71	0,2389	1,09	0,1379	1,47	0,0708
0,34	0,3669	0,72	0,2358	1,10	0,1357	1,48	0,0694
0,35	0,3632	0,73	0,2327	1,11	0,1335	1,49	0,0681
0,36	0,3594	0,74	0,2296	1,12	0,1314	1,50	0,0668
0,37	0,3557	0,75	0,2266	1,13	0,1292	1,51	0,0655

z	$P_R(z)$	z	$P_R(z)$	z	$P_R(z)$	z	$P_R(z)$	z	$P_R(z)$
1,52	0,0643	1,90	0,0287	2,28	0,0113	2,66	0,0039	3,04	0,0012
1,53	0,0630	1,91	0,0281	2,29	0,0110	2,67	0,0038	3,05	0,0011
1,54	0,0618	1,92	0,0274	2,30	0,0107	2,68	0,0037	3,06	0,0011
1,55	0,0606	1,93	0,0268	2,31	0,0104	2,69	0,0036	3,07	0,0011
1,56	0,0594	1,94	0,0262	2,32	0,0102	2,70	0,0035	3,08	0,0010
1,57	0,0582	1,95	0,0256	2,33	0,0099	2,71	0,0034	3,09	0,0010
1,58	0,0571	1,96	0,0250	2,34	0,0096	2,72	0,0033	3,10	0,0010
1,59	0,0559	1,97	0,0244	2,35	0,0094	2,73	0,0032	3,11	0,0009
1,60	0,0548	1,98	0,0239	2,36	0,0091	2,74	0,0031	3,12	0,0009
1,61	0,0537	1,99	0,0233	2,37	0,0089	2,75	0,0030	3,13	0,0009
1,62	0,0526	2,00	0,0228	2,38	0,0087	2,76	0,0029	3,14	0,0008
1,63	0,0516	2,01	0,0222	2,39	0,0084	2,77	0,0028	3,15	0,0008
1,64	0,0505	2,02	0,0217	2,40	0,0082	2,78	0,0027	3,16	0,0008
1,65	0,0495	2,03	0,0212	2,41	0,0080	2,79	0,0026	3,17	0,0008
1,66	0,0485	2,04	0,0207	2,42	0,0078	2,80	0,0026	3,18	0,0007
1,67	0,0475	2,05	0,0202	2,43	0,0075	2,81	0,0025	3,19	0,0007
1,68	0,0465	2,06	0,0197	2,44	0,0073	2,82	0,0024	3,20	0,0007
1,69	0,0455	2,07	0,0192	2,45	0,0071	2,83	0,0023	3,21	0,0007
1,70	0,0446	2,08	0,0188	2,46	0,0069	2,84	0,0023	3,22	0,0006
1,71	0,0436	2,09	0,0183	2,47	0,0068	2,85	0,0022	3,23	0,0006
1,72	0,0427	2,10	0,0179	2,48	0,0066	2,86	0,0021	3,24	0,0006
1,73	0,0418	2,11	0,0174	2,49	0,0064	2,87	0,0021	3,25	0,0006
1,74	0,0409	2,12	0,0170	2,50	0,0062	2,88	0,0020	3,30	0,0005
1,75	0,0401	2,13	0,0166	2,51	0,0060	2,89	0,0019	3,35	0,0004
1,76	0,0392	2,14	0,0162	2,52	0,0059	2,90	0,0019	3,40	0,0003
1,77	0,0384	2,15	0,0158	2,53	0,0057	2,91	0,0018	3,45	0,0003
1,78	0,0375	2,16	0,0154	2,54	0,0055	2,92	0,0018	3,50	0,0002
1,79	0,0367	2,17	0,0150	2,55	0,0054	2,93	0,0017	3,60	0,0002
1,80	0,0359	2,18	0,0146	2,56	0,0052	2,94	0,0016	3,70	0,0001
1,81	0,0351	2,19	0,0143	2,57	0,0051	2,95	0,0016	3,80	0,0001
1,82	0,0344	2,20	0,0139	2,58	0,0049	2,96	0,0015	3,90	0,00005
1,83	0,0336	2,21	0,0136	2,59	0,0048	2,97	0,0015	4,00	0,00003
1,84	0,0329	2,22	0,0132	2,60	0,0047	2,98	0,0014		
1,85	0,0322	2,23	0,0129	2,61	0,0045	2,99	0,0014		
1,86	0,0314	2,24	0,0125	2,62	0,0044	3,00	0,0013		
1,87	0,0307	2,25	0,0122	2,63	0,0043	3,01	0,0013		
1,88	0,0301	2,26	0,0119	2,64	0,0041	3,02	0,0013		
1,89	0,0294	2,27	0,0116	2,65	0,0040	3,03	0,0012		

Literatuur

Aron, A., Aron, E.N. & Coups, E.J. (2005). *Statistics for the behavioral and social sciences; a brief course* (derde editie). Upper Saddle River, NJ: Pearson.

Brink, W.P. van den & Koele, P. (1985). *Statistiek Deel 1,2 en 3.* Meppel: Boom.

DeVellis, R.F. (1991). *Scale development: Theory and applications.* Newbury Park, CA: Sage.

Fielding, J. & Gilbert, N. (2000). *Understanding social statistics.* Londen: Sage.

Field, A.P. (2009). *Discovering statistics using SPSS* (third edition). London, England: SAGE.

Gibbons, J.D. (1993). *Nonparametric measures of association.* Newbury Park, CA: Sage.

Kerlinger, F.N. (1973). *Foundations of behavioral research* (tweede editie). Londen: Holt, Rinehart and Winston.

Kim, J.-O. & Mueller, C.W. (1978). *Factor analysis: Statistical methods and practical issues.* Beverly Hills, CA: Sage.

Levin, J. & Fox, J.A. (2003). *Elementary statistics in social research* (negende editie). Boston, MA: Allyn and Bacon.

McClave, J.T., Benson, P.G. & Sincich, T. (2003). *Statistiek; een inleiding voor het hoger onderwijs* (achtste editie). Amsterdam: Pearson Education Benelux.

Norušis, M.J. (1991). *SPSS/PC + Studentware plus.* Chicago, il: SPSS inc.

Sapsford, R. (2007). *Survey research.* (Second edition). London, UK: Sage.

Segers, J.H.G. (1983). *Sociologische onderzoeksmethoden deel 1.* Assen: Van Gorcum.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York, NY: McGraw-Hill.

Spitz, J.C. (1971). *Statistiek voor psychologen, pedagogen, sociologen.* Amsterdam: Agon Elsevier.

Stempel, G.H. III & Westley, B.H. (red.) (1981). *Research methods in mass communication.* Englewood Cliffs, NJ: Prentice-Hall.

Tacq, J. (1991). *Van probleem naar analyse; de keuze van een gepaste multivariate analysetechniek bij een sociaal-wetenschappelijke probleemstelling.* De Lier: Academisch Boeken Centrum.

Triola, M.F. (2004). *Elementary statistics* (negende internationale editie). Boston: Pearson Addison Wesley.

Viswanathan, M. (2005). *Measurement error and research design.* Thousand Oaks, CA: Sage.

Wonnacott, R.J. & Wonnacott, T.H. (1985). *Introductory statistics* (vierde editie). New York, NY: John Wiley & Sons.

Register

Over de auteurs

Bregje van Groningen studeerde wijsbegeerte aan de Universiteit van Leiden en communicatiewetenschap aan de Universiteit van Amsterdam. Sinds haar afstuderen werkt zij als docent bij de afdeling communicatiewetenschap aan de Universiteit van Amsterdam. Zij houdt zich als docent vooral bezig met onderzoeksgerelateerde cursussen, zoals Methoden van Communicatie Onderzoek en Beschrijvende Statistiek en onderzoekspractica. Daarnaast geeft zij de colleges voor Wetenschapsfilosofie en Methodologie. Binnen de afdeling Communicatiewetenschap is zij werkzaam bij de programmagroep Media Entertainment.

Connie de Boer studeerde sociale geografie aan de Universiteit Utrecht. Na haar afstuderen werkte zij bij het Baschwitz Instituut voor massapsychologie en openbare mening. Zij is gepromoveerd op het proefschrift *Peilingen in de pers*, en ze is co-auteur van de boeken *Media en publiek* en *Publieke opinie*. Thans is zij als hoofddocent verbonden aan de afdeling Communicatiewetenschap van de Universiteit van Amsterdam.