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

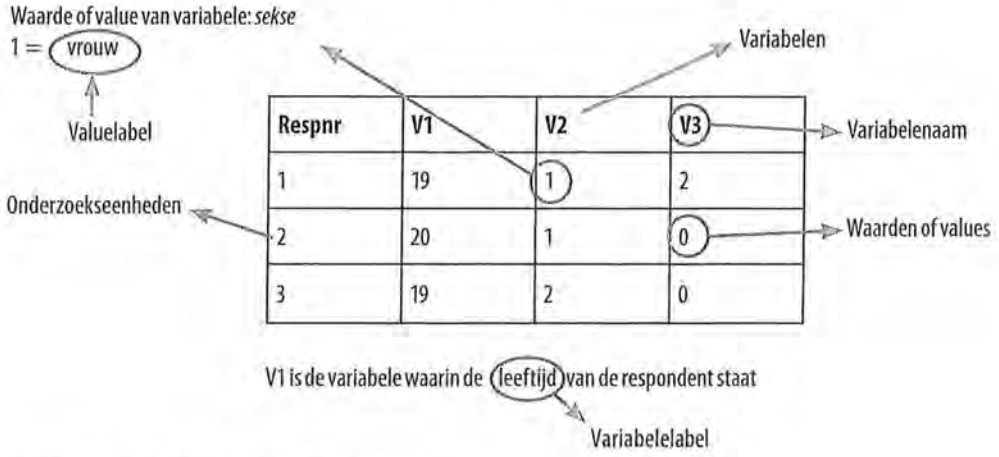
Persoon	Leeftijd	Sekse	NPO1	NPO3	RTL4	RTL5	Net5	BBC
1	19	1	2	0	2	2	4	0
2	20	1	0	0	1	2	2	1
3	19	2	0	1	1	1	0	3
4	22	1	3	2	2	3	3	2
5	24	2	1	0	3	1	1	1
6	21	2	0	3	2	0	1	1
7	20	1	2	2	1	2	3	2

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)

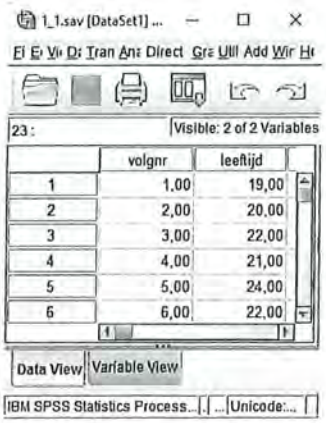
Leeftijd

		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.

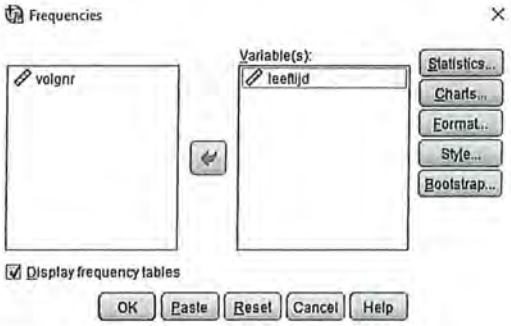
SPSS Invoeren van gegevens en het maken van een frequentietabel

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).



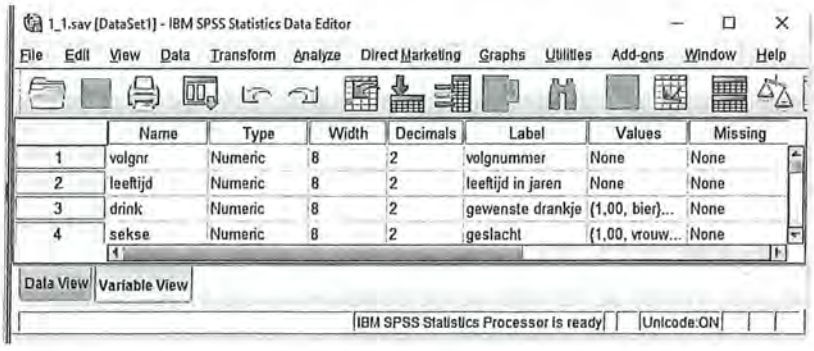
Figuur A Datamatrix

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

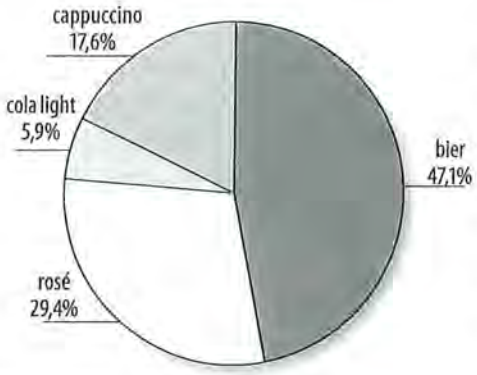
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	19	4	20,0	23,5	23,5
	20	5	25,0	29,4	52,9
	21	3	15,0	17,6	70,6
	22	4	20,0	23,5	94,1
	24	1	5,0	5,9	100,0
	Total	17	85,0	100,0	
Missing	geen antwoord	3	15,0		
Total		20	100,0		

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.

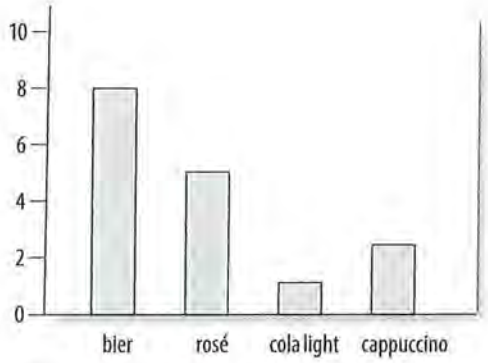
1.2.3 Grafieken

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)

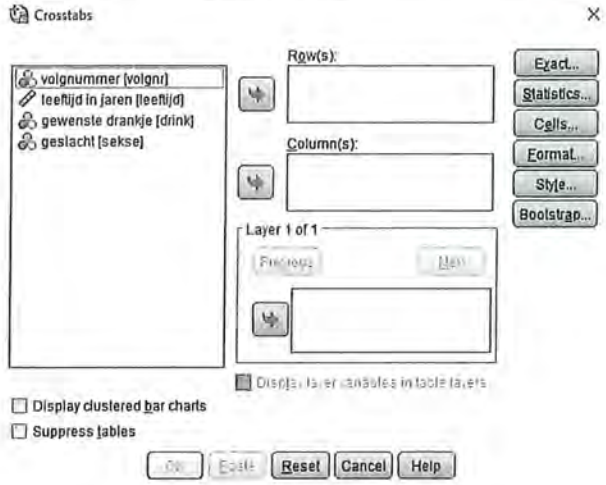
V57 Interesse in kunst *V49 sekse Crosstabulation

			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%

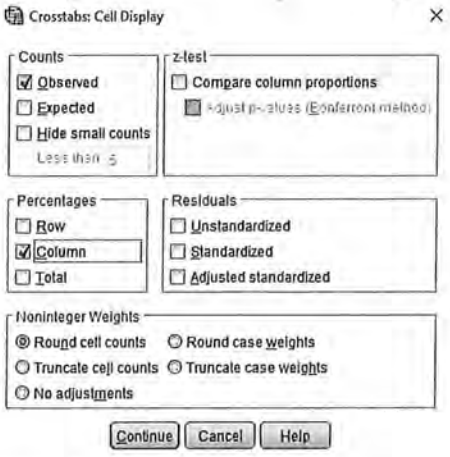
Bovenstaande kruistabel, die gepercenteerd is op de kolommen, laat bijvoorbeeld zien dat van de 896 mannen die aan deze enquête hebben deelgenomen 6,9% sterk in kunst is geïnteresseerd, tegenover 10,3% van de 943 vrouwen.

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.

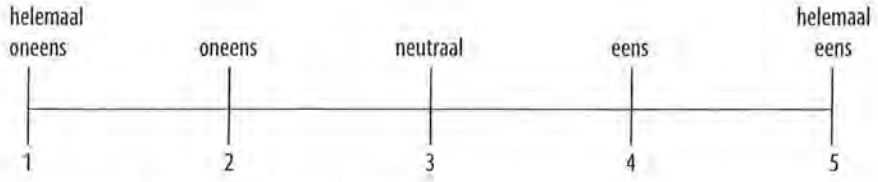
1.5.2 Ordinaal meetniveau

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

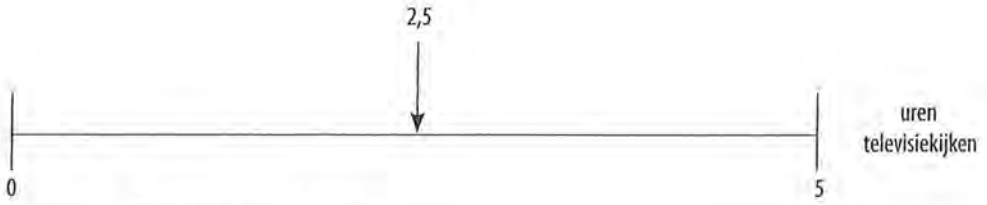
1.6 Waarden van variabelen

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.