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.

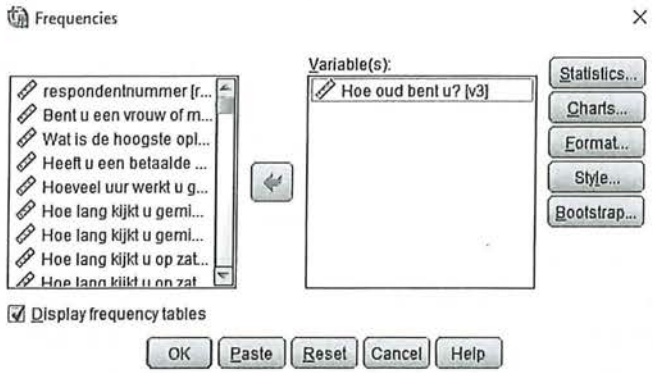
4.1 Syntax

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.

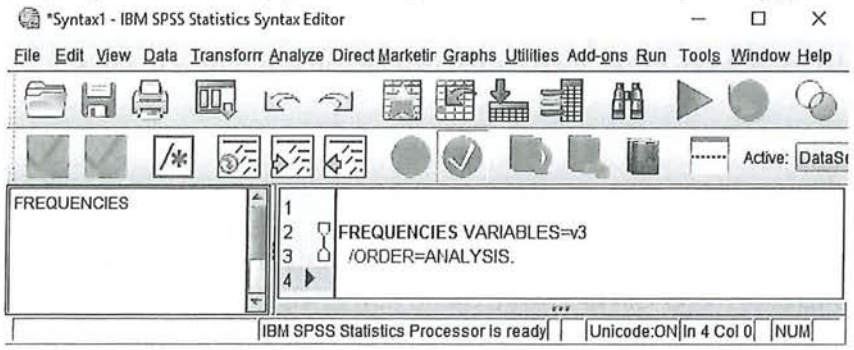
```
FREQUENCIES
VARIABLES=v3
/STATISTICS=STDDEV MEAN
/ORDER= ANALYSIS.
```

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:

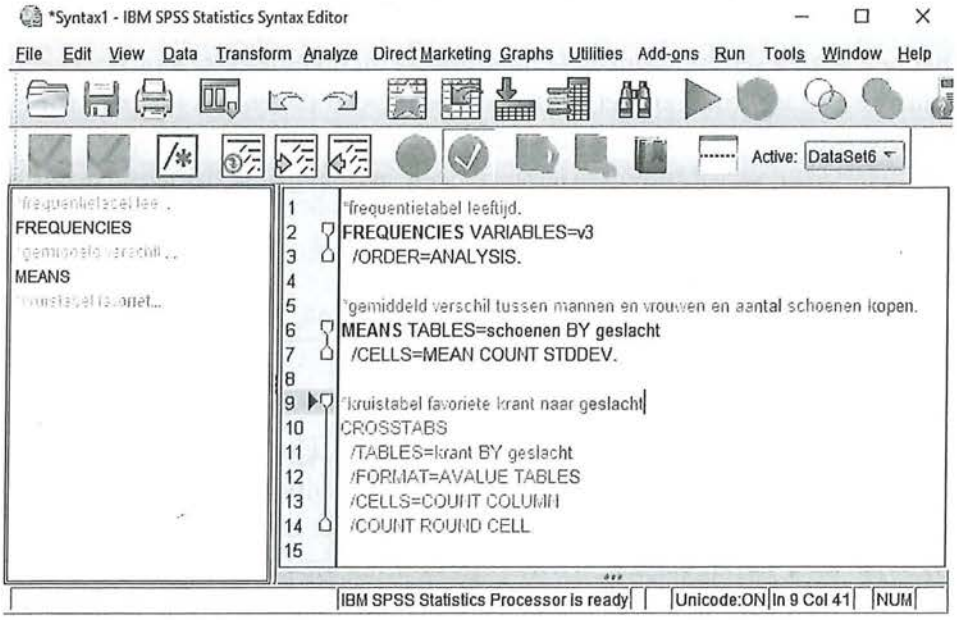


Figuur 4.2 Syntax in syntaxscherm

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

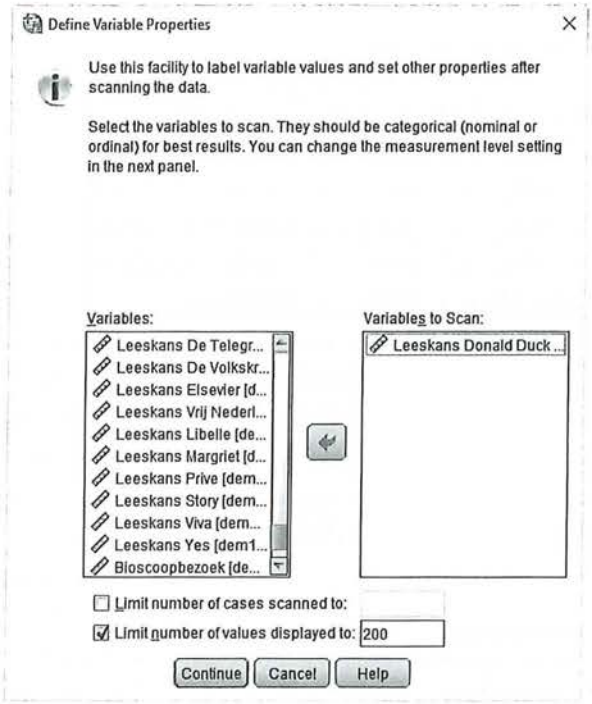
int_politiek interesse in politiek

		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		

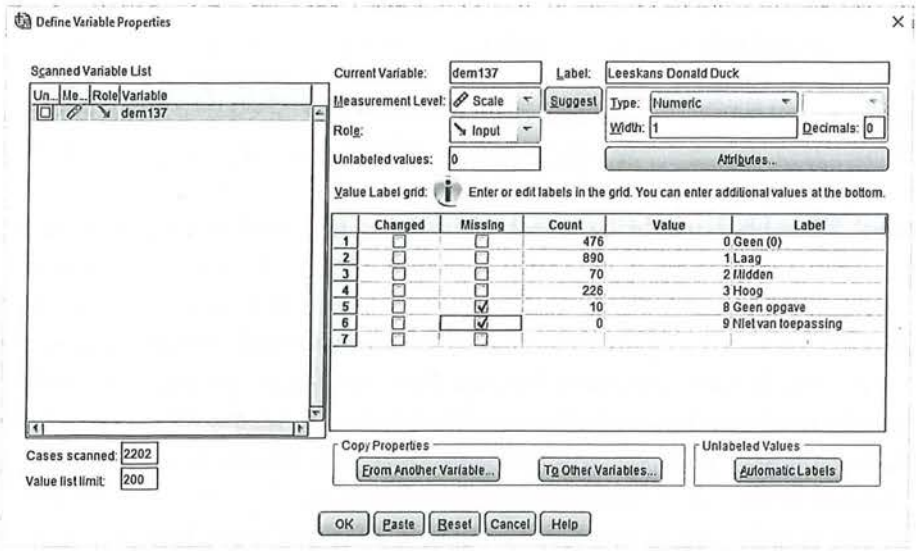
In tabel 4.2 hebben we de waarde 9 wél missing gemaakt (Zie kader 4.1 voor de handeling in SPSS). In deze tabel is te zien dat 361 respondenten de vraag hebben beantwoord met 'weet ik niet', en dat deze mensen niet in de analyse van deze variabele zijn meegenomen. Er is nu ook een verschil tussen de 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.

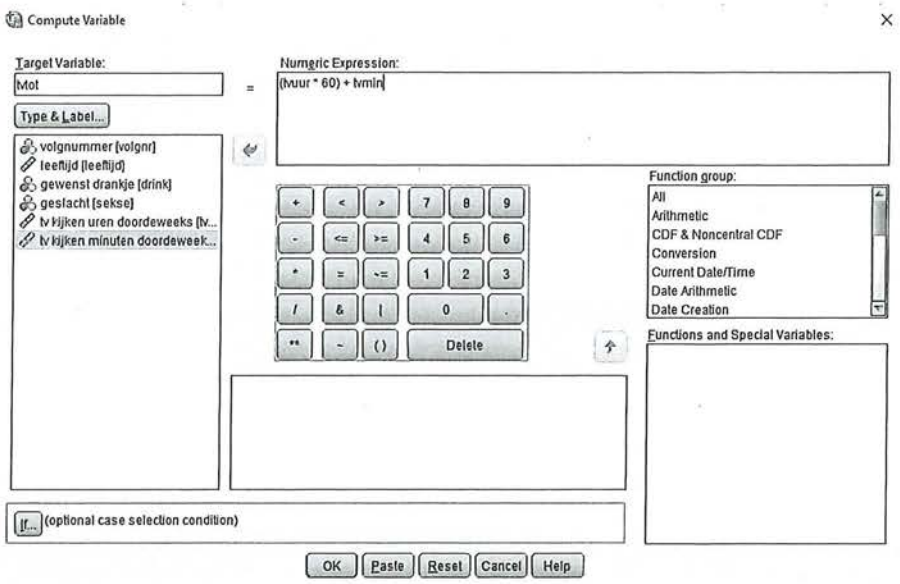
SPSS Nieuwe variabele maken door middel van Compute

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

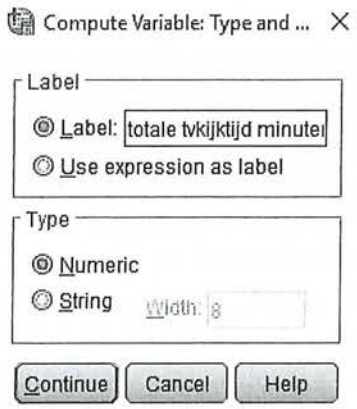
In het *Compute Variable*-venster (*figuur A*) dat dan verschijnt, kies je een nieuwe naam voor je variabele (onder *Target Variable:*). Let er daarbij op dat je geen spaties of leestekens gebruikt (op de *underscore* na herkent SPSS deze namelijk niet). Vervolgens kun je SPSS vertellen hoe die nieuwe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

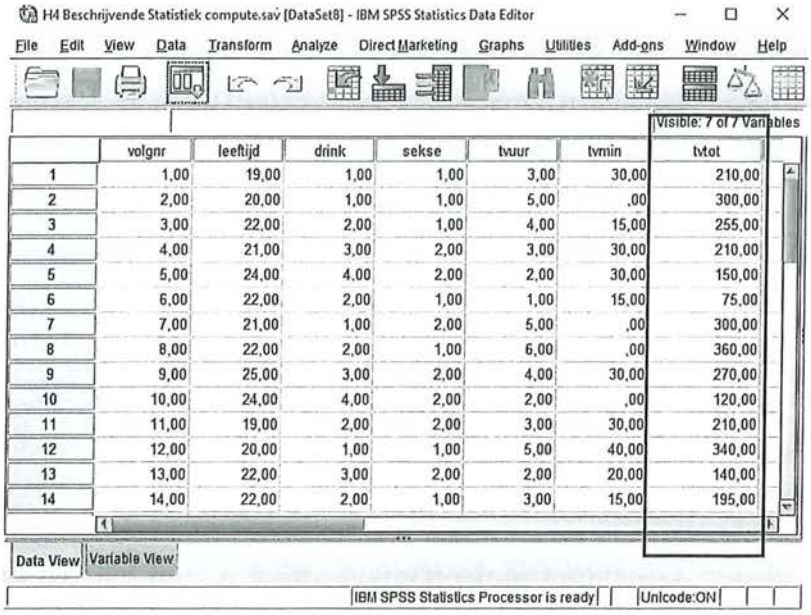


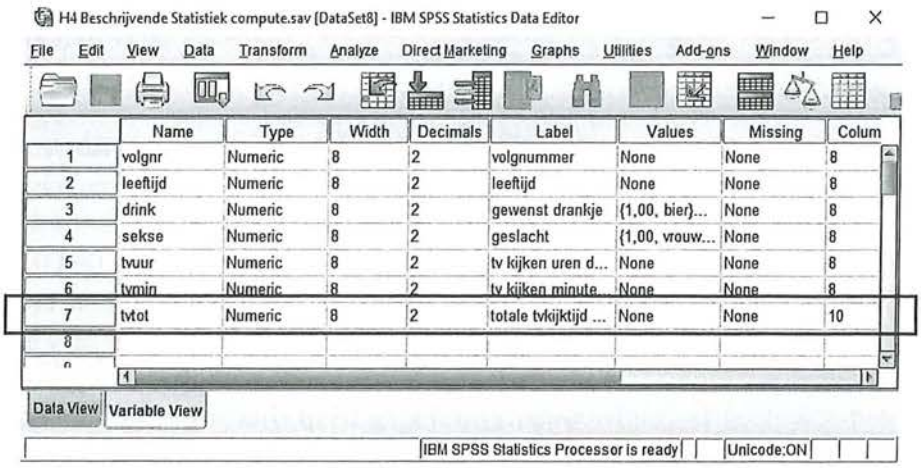
Figuur A Compute Variable-venster



Figuur B Type & Label-venster

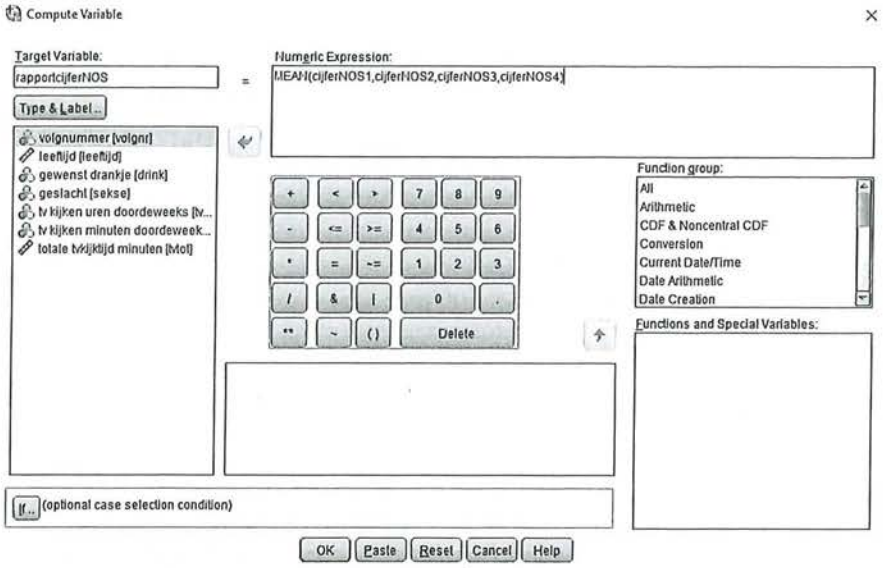
De nieuwe variabele verschijnt, na het runnen van de syntax, achteraan in je datamatrix in de *Data View*, en onderaan in de *Variable View* (zie Figuur C).





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)

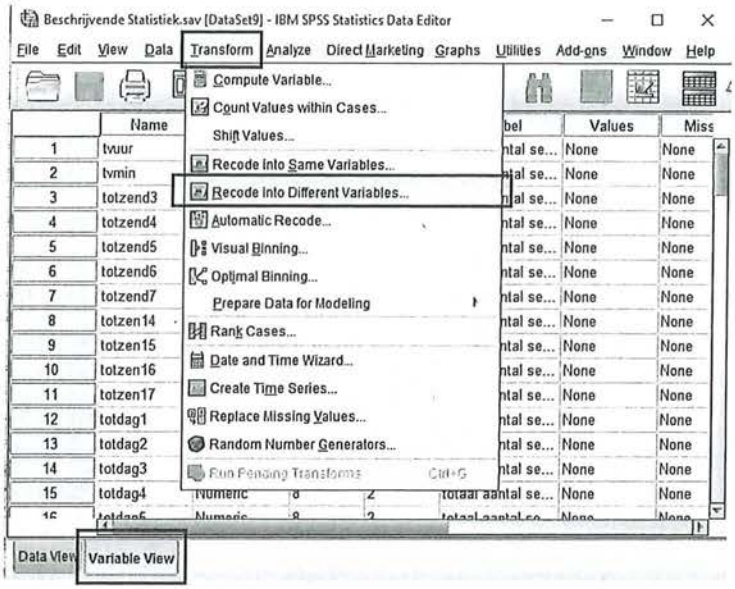
minpoHER minuten pub omr klassen

		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!

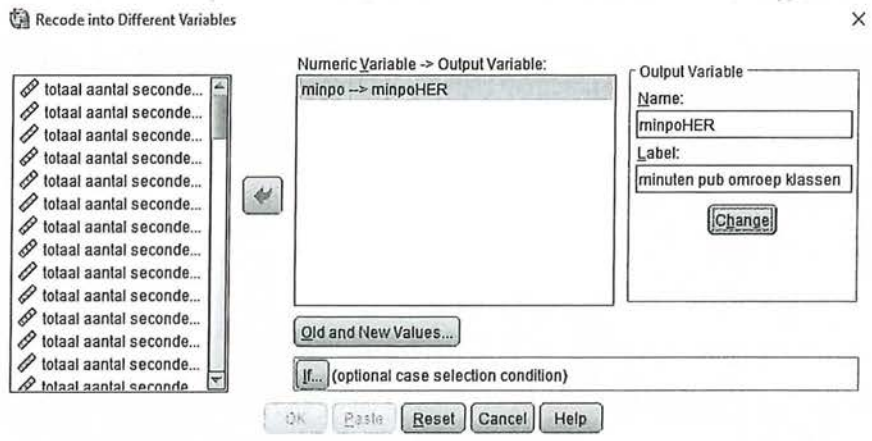
SPSS Hercoderen van variabelen

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.



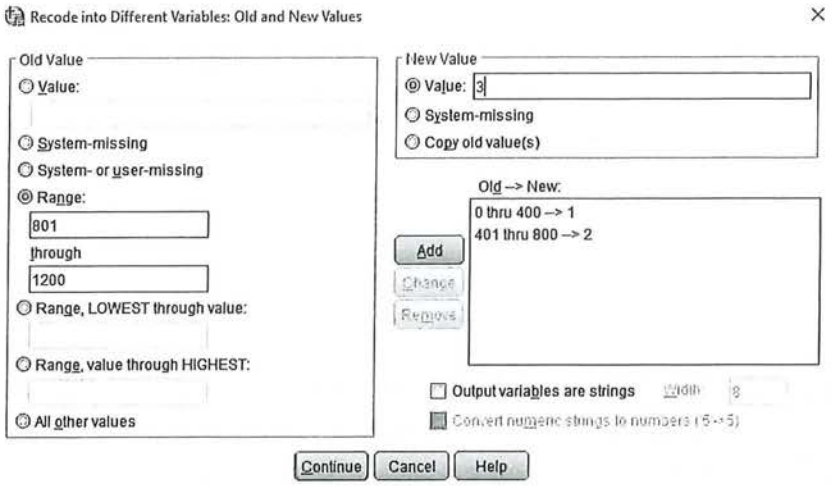
Figuur A Hercoderen van een variabele

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.



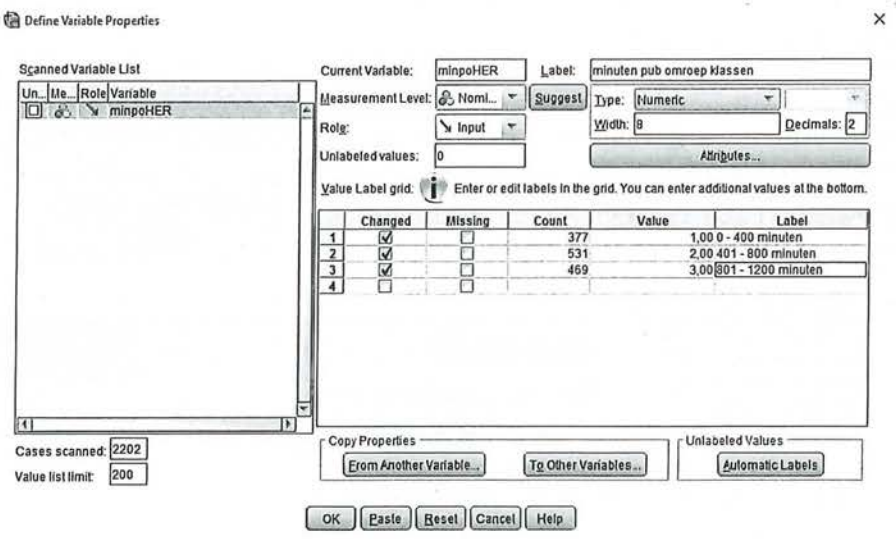
Figuur B Recode into Different Variables-venster

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

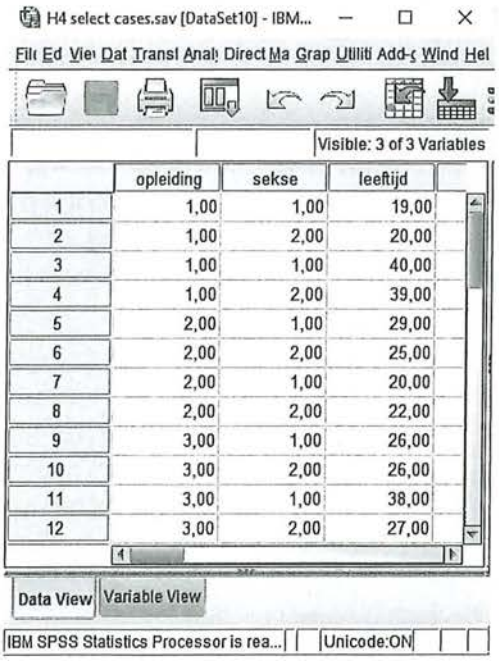
Het is ook mogelijk om met *LOWEST through value* te werken (vanaf de laagste waarde *tot en met* een bepaalde waarde, in bovenstaand voorbeeld *Range, Lowest through value: 500*) en *value through HIGHEST* (vanaf een bepaalde waarde *tot en met* de hoogste mogelijke waarde die deze variabele aanneemt, 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)

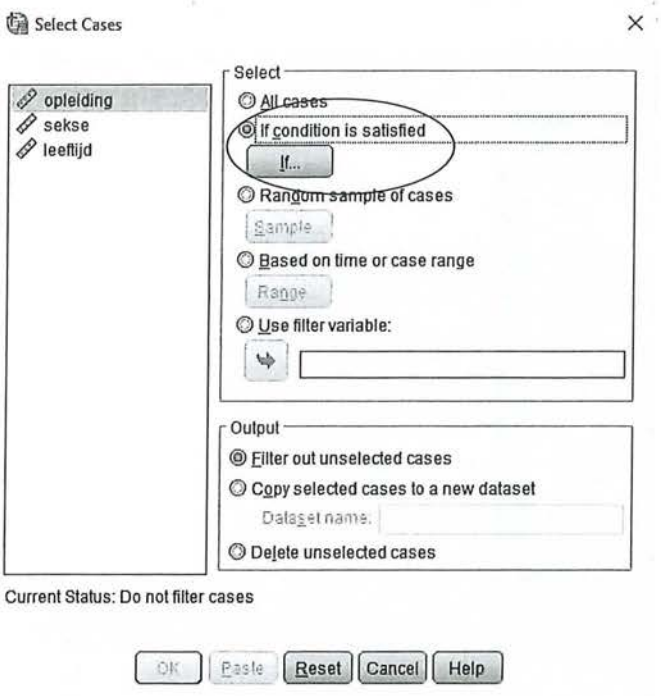
De mediaan van opleidingsniveau is 2 (midden opgeleid), de modus van sekse is 2 (er zijn meer mannen dan vrouwen) en de gemiddelde leeftijd is 27,58 (SD = 7,55).
Wanneer we nu een frequentieverdeling van bijvoorbeeld opleidingsniveau uit zouden draaien, krijgen we daar de informatie te zien van alle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

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

opleiding

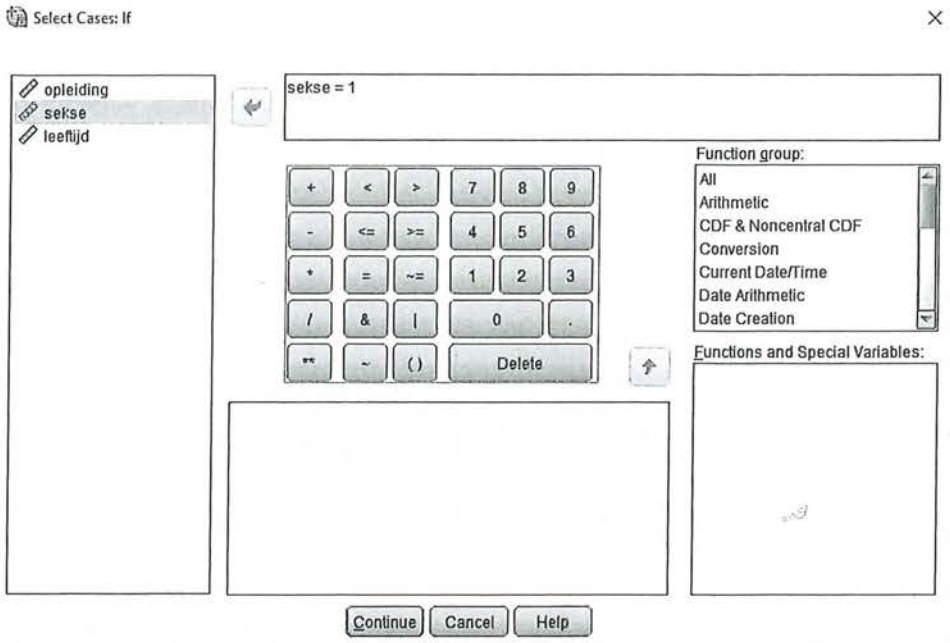
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,00 laag	4	33,3	33,3	33,3
	2,00 midden	4	33,3	33,3	66,7
	3,00 hoog	4	33,3	33,3	100,0
	Total	12	100,0	100,0	

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*).



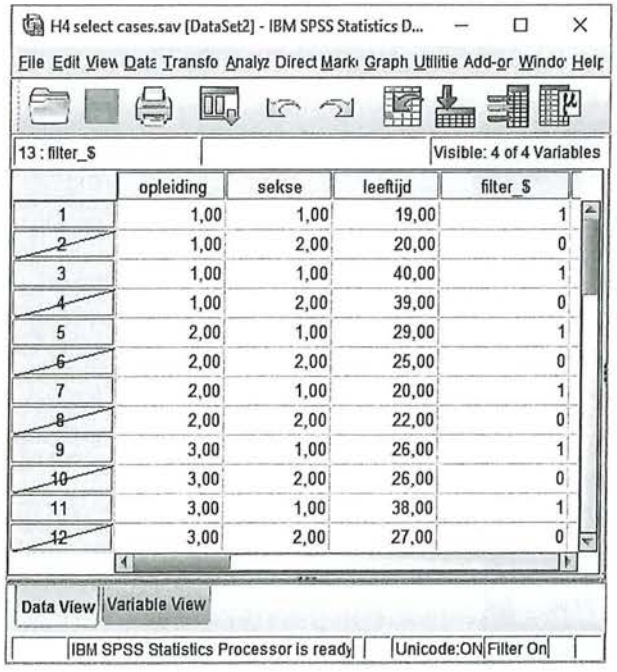
Figuur 4.5 Select Cases-venster

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.



	opleiding	sekse	leeftijd	filter_$
1	1,00	1,00	19,00	1
2	1,00	2,00	20,00	0
3	1,00	1,00	40,00	1
4	1,00	2,00	39,00	0
5	2,00	1,00	29,00	1
6	2,00	2,00	25,00	0
7	2,00	1,00	20,00	1
8	2,00	2,00	22,00	0
9	3,00	1,00	26,00	1
10	3,00	2,00	26,00	0
11	3,00	1,00	38,00	1
12	3,00	2,00	27,00	0

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

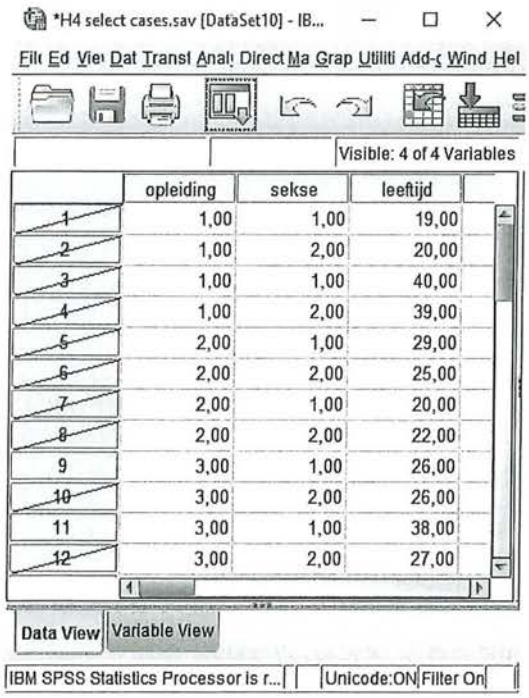
variabele die aangeeft of de respondent wel (waarde 1) of niet (waarde 0) geselecteerd is na het selecteren van onze cases.
Alle analyses die we vanaf dit moment met de variabelen uitvoeren, gaan alleen nog maar over vrouwen. Wanneer je nu een frequentieverdeling maakt van opleidingsniveau, is de mediaan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

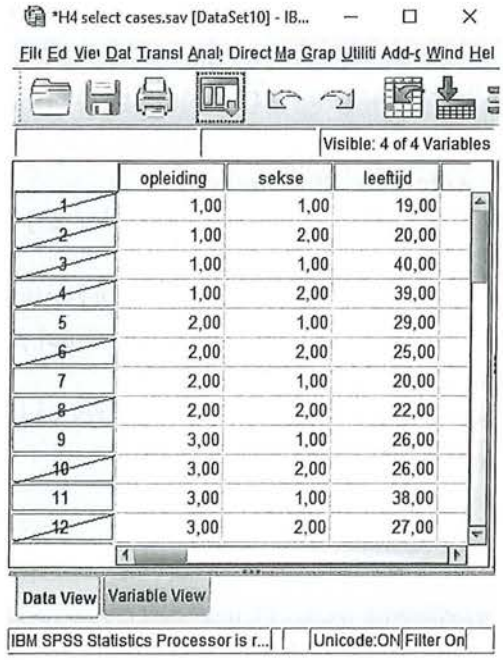
Wanneer je ook het gemiddelde zou berekenen, zou je zien dat waar eerst de gemiddelde leeftijd van alle respondenten 27,58 was, de gemiddelde leeftijd van de vrouwen 30,6 (*SD* = 8,35) is.
In een volgende stap willen we niet alleen vrouwen selecteren, maar willen we alleen die vrouwen selecteren die 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.