



Het bewaren van verschillende versies van je databestand tijdens databeheer

Tijdens het schoonmaken en bewerken van je dataset in SPSS moet je even gestructureerd te werk gaan als bij het uitvoeren van analyses. Zomaar aanpassingen maken in je data zonder notities hiervan bij te houden, of steeds in het originele databestand blijven werken en geen kopieën opslaan, zal voor problemen zorgen. Zodra je namelijk zaken hebt veranderd in je databestand en die aanpassingen opgeslagen hebt, kun je niet meer terug en zullen waardevolle data die per ongeluk aangepast of verwijderd worden, niet meer in originele staat beschikbaar zijn.

Om dit te vermijden, zijn er twee zaken die je kunt ondernemen. Ten eerste kun je notities bijhouden in de *Syntax Editor* met betrekking tot de wijzigingen die je doorgevoerd hebt. Heb je bijvoorbeeld een aantal cases verwijderd omdat ze te veel ontbrekende waarden hadden, schrijf dan neer hoeveel je er verwijderd hebt, met welke reden en welke nummers die cases hadden. Heb je systematische of eenmalige fouten aangepakt, schrijf dan neer via welke commando's je dit gedaan hebt en welke nieuwe waarden ze eventueel gekregen hebben. Ten tweede kun je steeds via kopieën van de originele dataset werken en de verschillende versies van die dataset tussendoor opslaan, om te vermijden dat je fouten niet meer kunt terugdraaien. Probeer hierbij ook in de benamingen van die verschillende versies gestructureerd te werk te gaan. Begin alvast telkens met de datum waarop je die versie van het databestand het laatst bewerkt hebt en duid bijvoorbeeld aan hoeveel cases er in elke versie zitten (bv. '2022_03_29_Naam Databestand_512 cases.sav'). Wanneer je helemaal klaar bent met data cleaning en andere bewerkingen van je data (zie paragraaf 4.4), kun je de laatste versie een benaming geven zoals '2022_03_31_Naam Databestand_Finaal.sav'. Het kan dus bijvoorbeeld zijn dat je vijf of zes verschillende versies van je dataset staan heft tussen je originele ruwe databestand en je finale databestand. Als je later beseft dat je tijdens het databeheer een fout gemaakt hebt, kun je via de *Syntax* opsporen waar je precies die fout gemaakt hebt en kun je vanaf die versie van je databestand opnieuw beginnen om de fout te herstellen, zonder alle wijzigingen die je ervoor gemaakt hebt opnieuw te moeten doorvoeren.

4.4 DATABESTANDEN SAMENVOEGEN, FILTEREN OF SPLITSEN

Naast data cleaning moet je vaak nog andere bewerkingen uitvoeren op je databestand vooraleer je een analyse uitvoert. In deze paragraaf overlopen we hoe je twee databestanden kunt samenvoegen, hoe je kunt werken met een filter om sommige cases uit analyses te weren en hoe je een databestand kunt splitsen in verschillende groepen onderzoeksseenheden, waardoor elke analyse apart uitgevoerd wordt voor elke groep, in plaats van voor alle cases samen. Dit zijn vaak voorkomende databewerkingen waar je als onderzoeker ongetwijfeld mee te maken zult krijgen.

4.4.1 SAMENVOEGEN VAN DATABESTANDEN VIA MERGE

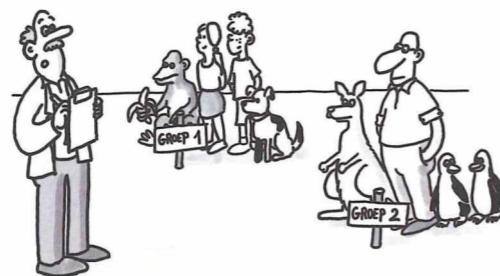
Een eerste bewerking die vaak voorkomt, is het samenvoegen of *mergen* van verschillende databestanden. Het **samenvoegen van databestanden** kun je doen in SPSS, maar deze functie is beperkt tot het samenvoegen van maximaal twee bestanden in een keer. Heb je dus bijvoorbeeld vier databestanden die samengevoegd moeten worden, dan zul je dit in drie stappen moeten doen. Databestanden samenvoegen kan op twee verschillende manieren gebeuren, afhankelijk van de reden waarom je ze wilt samenvoegen.

Een eerste optie is dat je cases toevoegt aan een databestand (*Add cases*). Het **toevoegen van cases** kan gebeuren wanneer je bijvoorbeeld een vragenlijst afgenumen hebt en verwerkt hebt, maar er na de einddatum toch nog enkele ingevulde vragenlijsten binnenkomen. Deze laattijdige cases kun je dan toevoegen aan je databestand met alle andere cases. Het kan ook zijn dat de dataverzameling verspreid gebeurde, bijvoorbeeld via verschillende instanties of zowel op papier als digitaal. In die situaties heb je ook meerdere databestanden die je moet samenvoegen om alle data samen te kunnen analyseren. Belangrijk bij het toevoegen van cases is dat de databestanden dezelfde variabelen bevatten. Het nieuw gevormde databestand bestaat dan ook uit dezelfde variabelen, maar het aantal cases van het originele databestand is verhoogd met het aantal cases uit het bestand dat je eraan toegevoegd hebt. Figuur 4.25 illustreert het toevoegen van cases op een visuele manier.



Figuur 4.25

De respondenten die besloten om mee te werken aan de eerste wave van de C&W studie kunnen kiezen of ze de vragenlijst online of op papier invullen. Stel dat je hierbij afgesproken hebt dat je collega-onderzoeker de papieren vragenlijsten manueel zal invoeren en dat jij de digitale vragenlijsten zal verwerken. Nu wil jij het databestand met de papieren vragenlijsten (DATA_WAVE1_CenW_papier.sav met n = 10) van je collega samenvoegen of *mergen* met jouw databestand (DATA_WAVE1_CenW_digitaal.sav met n = 2010). Het toevoegen van cases kun je doen aan de hand van de onderstaande acties. Je kunt deze cases uiteraard ook manueel toevoegen, maar omdat dit bij een groot aantal cases zeer tijdrovend is en risico's inhoudt om (invul)fouten te maken, is het veel beter ervoor te kiezen de databestanden automatisch samen te voegen.



"Op basis van welke karakteristieken moet ik deze twee groepen samenvoegen?"

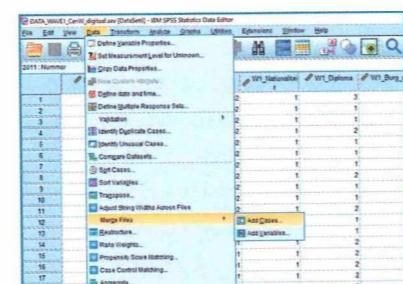
Samenvoegen van databestanden via Add Cases

ACTIE 1

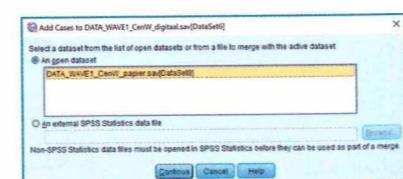
Open de twee databestanden in SPSS die je wilt samenvoegen en navigeer vanuit het databestand waarin je cases wilt toevoegen (hier: DATA_WAVE1_CenW_digitaal.sav) naar **Data > Merge Files > Add Cases**.

ACTIE 2

In het **Add Cases**-dialoogvenster dat nu geopend is, selecteer je onder **An open dataset** het tweede databestand dat je geopend had (hier: DATA_WAVE1_CenW_papier.sav) en dat je met het eerste bestand wilt samenvoegen. Via **Browse** kun je indien nodig ook naar andere bestanden zoeken. Klik daarna op **Continue**.



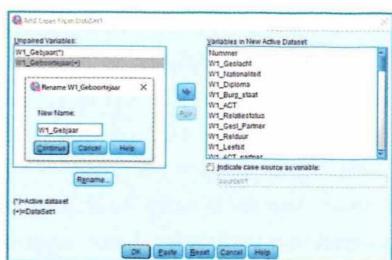
Figuur 4.26 ACTIE 1



Figuur 4.27 ACTIE 2

ACTIE 3

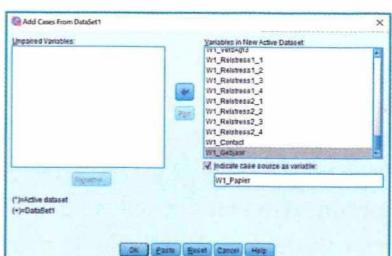
In dit tweede dialoogvenster staan rechts de variabelen die perfect overeenkomen in beide datasets en links de variabelen die SPSS niet met zekerheid kan matchen. Het is belangrijk om er bij het toevoegen van cases voor te zorgen dat alle variabelen exact dezelfde naam hebben. Hier zien we dat in de dataset van de papieren vragenlijsten (+) de variabele geboortejaar anders benoemd is dan in de digitale vragenlijst (*), waardoor het **Unpaired Variables** zijn. Om dit op te lossen, selecteer eerst 'W1_Geboortejaar', klik daarna op **Rename** en pas de naam aan naar 'W1_Gebaar'. Klik daarna op **Continue**. Om de variabelen te paren, gebruik je de control toets om beide variabelen te selecteren en klik je op **Pair**. Je ziet nu 'W1_Gebaar' in het rechterkader verschijnen.



Figuur 4.28 ACTIE 3

ACTIE 4

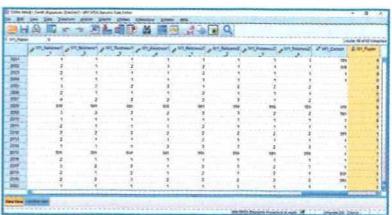
Vink ten slotte nog de optie **Indicate case source as variable** aan. Hiermee geef je de opdracht om een variabele te maken die aanduidt welke cases al in het originele bestand zaten (= 0) en welke nieuw toegevoegd werden (= 1). Kies voor die variabele eventueel nog een naam (hier: W1_Papier) en klik dan op **Paste** indien je het SPSS-commando in je **Syntax Editor** wilt plakken of klik op **OK** als je het SPSS-commando wilt uitvoeren.



Figuur 4.29 ACTIE 4

ACTIE 5

Als je het samenvoegen uitgevoerd hebt, controleer dan even in de *Data View* of alle cases toegevoegd werden. Je ziet hier bij de variabele 'W1_Papier' dat de tien laatste cases nieuw toegevoegd werden. Let op: sla dit samengevoegde databestand onder een volledig nieuwe naam op (bijvoorbeeld DATA_WAVE1_CenW_digitaal+papier.sav), want zo ben je het originele digitale databestand niet kwijt.



Figuur 4.30 ACTIE 5

ADD FILES /FILE=*

```
/FILE='DataSet1'  
/RENAME W1_Geboortejaar=W1_Gebaar  
/IN=W1_Papier.
```

VARIABLE LABELS W1_Papier
'Case source is DataSet1'.
EXECUTE.

Naast het toevoegen van cases is er ook een tweede optie in het samenvoegen van bestanden, waarbij je extra variabelen toevoegt aan een databestand (*Add variables*). Het **toevoegen van variabelen** komt bijvoorbeeld vaak voor bij onderzoek dat een vertrouwelijke component kent. Zo is het vanuit ethisch oogpunt noodzakelijk om de identiteit van respondenten te beschermen, wanneer ze vertrouwelijke informatie delen. Hier kun je dan werken met twee aparte databestanden: in het ene bestand verzamel je de persoonlijke informatie van je respondenten waarmee ze geïdentificeerd kunnen worden, en in het andere databestand sta je de vertrouwelijke informatie op die ze gedeeld hebben. Soms zal je echter beide databestanden nodig hebben voor een analyse, waardoor je ze tijdelijk moet kunnen samenvoegen. Ook wanneer je gegevens uit verschillende bronnen moet combineren voor je onderzoek, moet je variabelen toevoegen aan een dataset. Stel dat je bijvoorbeeld bij de lokale politie van een stad informatie verzameld hebt over het aantal woninginbraken per stadsbuurt en dat je die wilt vergelijken met een aantal andere kenmerken van die stadsbuurten (bv. het aantal officieel geregistreerde woningen). Die andere buurtgegevens zul je niet bij de lokale politie, maar bij andere administratieve diensten van de stad moeten opvragen. Deze verschillende databestanden zul je nadien moeten samenvoegen om je analyses te kunnen uitvoeren. Tot slot worden bestanden ook samengevoegd bij longitudinaal onderzoek. Onderzoeksseenheden daarbij op verschillende tijdstippen bekeken, gemeten of bevrageerd. Hierbij moet je dan je databestand telkens updaten met de resultaten van het nieuwe meetmoment om longitudinale analyses te kunnen uitvoeren. Figuur 4.31 illustreert het toevoegen van variabelen op een visuele manier.



Figuur 4.31

Wanneer je twee databestanden met dezelfde cases, maar verschillende variabelen, wilt samenvoegen, heb je een gemeenschappelijke link of match nodig tussen die twee databestanden. Dit matchingscriterium noemen we een *key variable* of **sleutelvariabele** en zorgt ervoor dat elke case een uniek nummer heeft in beide databestanden, waardoor de waarden voor de verschillende variabelen onder de juiste case gekoppeld worden in het samengevoegde bestand. De kenmerken van die sleutelvariabele moeten in elk SPSS-bestand hetzelfde zijn op het vlak van *name*, *type* en *width* (zie paragraaf 4.2.1) en beide bestanden moeten opeen gesorteerd staan op die sleutelvariabele (dit wordt automatisch door SPSS gedaan bij

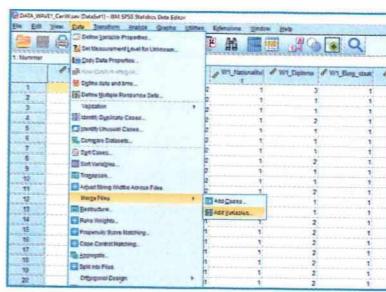
het samenvoegen). Merk hier echter op dat niet per se alle cases in beide bestanden moeten voorkomen en dat het oké is als de bestanden ook een aantal dezelfde variabelen bevatten. Het samenvoegen zal dan nog steeds lukken, maar de cases die niet in beide bestanden voorkomen, zullen dan een aantal ontbrekende waarden hebben en de variabelen die dubbel voorkomen, zullen niet toegevoegd worden.

Laten we nu eens kijken hoe we variabelen kunnen toevoegen aan een databestand in SPSS. In paragraaf 4.3.2 controleerden we het databestand afkomstig uit de tweede wave (2021) van het C&W onderzoek. Tijdens de tweede wave namen 511 respondenten deel, waarvan 333 respondenten ook deelgenomen hadden aan de eerste wave (2020) van het onderzoek. Er zijn dus in totaal 178 compleet nieuwe deelnemers in Wave 2, die tijdens de eerste wave niet deelgenomen hadden. Deze longitudinale afname laat ons toe na te gaan wat er veranderd is op het vlak van welzijn na één jaar corona bij die 333 respondenten. Om dit te kunnen nagaan, moet je eerst de twee databestanden samenvoegen. De onderstaande acties leggen uit hoe dit in zijn werk gaat in SPSS.

Samenvoegen van databestanden via Add Variables

ACTIE 1

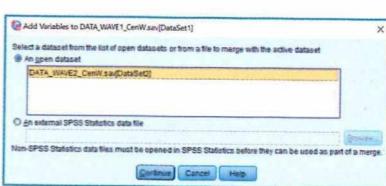
Open de twee databestanden in SPSS die je wilt samenvoegen en navigeer vanuit het databestand waarin je variabelen wilt toevoegen (hier: DATA_WAVE1_CenW.sav) naar **Data > Merge Files > Add Variables**.



Figuur 4.32 ACTIE 1

ACTIE 2

In het **Add Variables**-dialoogvenster dat nu geopend is, selecteer je onder **An open dataset** het tweede databestand dat je geopend had (hier: DATA_WAVE2_CenW.sav) en dat je met het eerste bestand wilt samenvoegen. Via **Browse** kun je indien nodig ook naar andere bestanden zoeken. Klik daarna op **Continue**.



Figuur 4.33 ACTIE 2

ACTIE 3

In dit tweede dialoogvenster onder tabblad **Merge Method** selecteer je **One-to-one merge based on key values** en zorg je dat **Sort files by key values before merging** aangevinkt staat. SPSS kiest hier automatisch al de variabele 'Nummer' als **key variable** of sleutelvariabele. Om dit aan te passen, ga je naar het tabblad **Variables**.

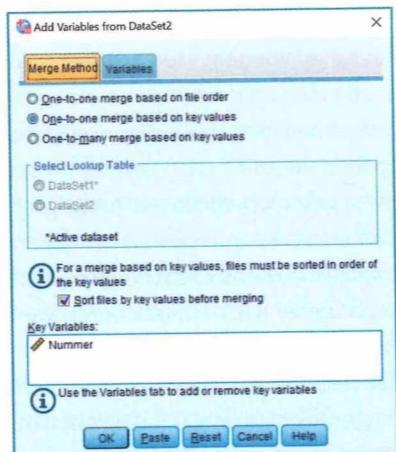
ACTIE 4

In het tabblad **Variables** kun je indien nodig specifieke variabelen uitsluiten tijdens het samenvoegen. Je zou bijvoorbeeld 'W2_Gebjaar' kunnen uitsluiten, want die is voor beide waves normaal gezien hetzelfde. Hier doen we dit niet, omdat sommige respondenten uit Wave 2 niet aan Wave 1 deelgenomen hebben. Je kunt hier ook de **key variable** wijzigen indien nodig. Klik ten slotte op **Paste** indien je het SPSS-commando in je **Syntax Editor** wilt plakken of klik op **OK** als je het SPSS-commando wilt uitvoeren.

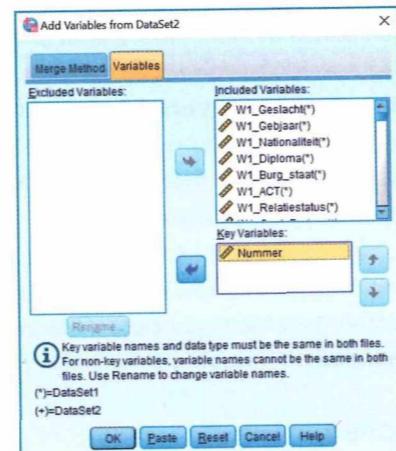
ACTIE 5

Als je het samenvoegen uitgevoerd hebt, controleer dan even of het totale aantal respondenten klopt (hier: $2020 + 178 = 2198$). Indien je geen nieuwe onderzoekseenheden in je tweede databestand hebt, blijft het aantal respondenten in het nieuw samengevoegde bestand gelijk aan het aantal van het eerste databestand. Sla dit samengevoegde databestand het best onder een volledig nieuwe naam op (bv. DATA_WAVE1_WAVE2_CenW.sav), want zo ben je het originele databestand niet kwijt.

```
DATASET ACTIVATE DataSet1.  
SORT CASES BY Nummer.  
DATASET ACTIVATE DataSet2.  
SORT CASES BY Nummer.  
DATASET ACTIVATE DataSet1.  
MATCH FILES /FILE=*<  
/FILE='DataSet2'  
/BY Nummer.  
EXECUTE.
```



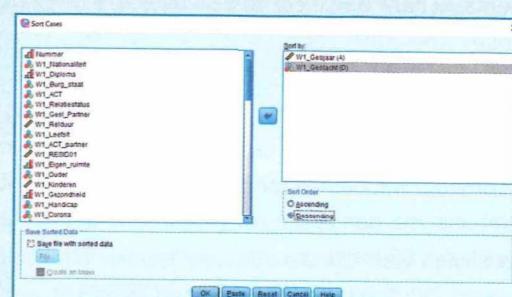
Figuur 4.34 ACTIE 3



Figuur 4.35 ACTIE 4

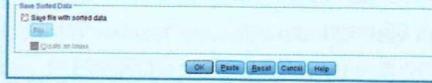
De cases in je dataset sorteren

We zagen in deze paragraaf dat het samenvoegen van bestanden enkel mogelijk is wanneer beide bestanden oplopend gesorteerd staan op de sleutelvariabele die je gekozen hebt. Ook naast het samenvoegen van databestanden kan het **sorteren van cases** nuttig zijn. Meestal staan de cases of rijen in de *Data View* gesorteerd in de volgorde waarop ze ingegeven werden in de dataset: de cases die het meest recent toegevoegd werden, staan onderaan de dataset. De C&W databestanden die we doorheen dit boek gebruiken, staan gesorteerd op het unieke nummer dat ze gekregen hebben onder de variabele 'Nummer', van laag naar hoog (oplopend). Maar je zou evengoed in plaats van op dat unieke nummer de dataset 'DATA_WAVE1_CenW.sav' kunnen sorteren op geboortejaar, te beginnen met de oudste respondenten bovenaan. Je kunt dit manueel doen in de *Data View* door met de rechtermuisknop te klikken op de hoofding van de kolom 'W1_Gebjaar' en te klikken op *Sort Ascending*, waardoor de cases oplopend gesorteerd zullen staan op geboortejaar. We raden echter aan om het sorteren van data via het commando *Sort Cases* te doen, want zo kun je in je *Syntax* bijhouden op welke variabele je het databestand gesorteerd hebt. Bovendien kun je via dit commando de gegevens met de nieuwe sortering als een apart databestand opslaan, zodat je de originele volgorde van cases niet kwijt bent: het sorteren van data kan na het uitvoeren namelijk niet ongedaan gemaakt worden via *Undo*.



Figuur 4.36

Als je onder het tabblad *Data* op *Sort Cases* klikt, opent er een dialoogvenster (zie figuur 4.36). Hier kun je de variabele(n) waarop je de gegevens wilt sorteren in het vakje *Sort by* slepen. Als je de cases op meerdere variabelen wilt sorteren, is de volgorde belangrijk. Staat er bijvoorbeeld eerst 'W1_Gebaar' en daarna 'W1_Geslacht', dan sorteert SPSS de cases eerst op geboortejaar en worden de respondenten die hetzelfde geboortejaar hebben daarna gesorteerd op geslacht. De keuze tussen *Ascending* (oplopend) en *Descending* (aflopend) die je voor elke variabele maakt onder *Sort Order*, verschijnt tussen haakjes naast elke variabele. Vink ten slotte *Save file with sorted data* aan als je de nieuw gesorteerde gegevens als een apart bestand wilt opslaan. Klik in dat geval ook op de knop *File...* om de opslaglocatie en de naam van het nieuwe bestand te kiezen. Als je dan op *Paste* klikt, zal het onderstaande commando in de *Syntax* verschijnen. Voer je dit commando uit, dan zullen de cases in het databestand oplopend op geboortejaar gesorteerd staan en daarna aflopend op geslacht.



SORT CASES BY W1 Gebiaar(A) W1 Geslacht(B)

4.4.2 FILTEREN OP BEPAALDE CASES VIA SELECT CASES

Als onderzoeker streef je meestal naar een zo rijk mogelijke dataset met heel veel informatie en een groot aantal verschillende cases. Je gaat bijvoorbeeld zowel studenten als werken-den bevragen, jongeren en ouderen, en zowel mensen die in een relatie zitten als alleenstaanden. Toch betekent dat niet dat je altijd alle cases van je databestand in je analyses wilt betrekken. Soms wil je uitsluitend focussen op een subgroep van je dataset en ga je enkel kijken naar de vrouwelijke respondenten of naar de respondenten die kinderen hebben. Je kunt hiervoor aan SPSS de opdracht geven om alleen die specifieke cases te selecteren bij het uitvoeren van analyses en alle andere cases even uit de dataset te filteren. Het **selecteren van cases** doe je door de dataset die je wilt filteren, te openen en te navigeren naar *Data* » *Select Cases*. In dit dialoogvenster zie je links onderaan of er een filter staat of niet (*Current Status: Do not filter cases*) en zijn er verschillende opties mogelijk om cases te selecteren in het kader *Select* (zie figuur 4.37):

All cases:

Deze optie staat standaard aangeduid: er wordt geen selectie gemaakt, alle cases van de dataset worden gebruikt bij het uitvoeren van analyses. Indien je eerder een selectie gemaakt hebt via *Select cases* en je die filter opnieuw wilt uitzetten, moet je de opties *All cases* weer aanduiden.

If condition is satisfied:

Hier kun je cases selecteren op basis van een voorwaarde die je zelf opstelt. Als je bijvoorbeeld wilt dat respondenten die ouder dan 65 jaar zijn niet meegenomen worden in een analyse, dan stel je als voorwaarde dat de variabele 'Leeftijd' een waarde kleiner of gelijk aan 65 heeft.

Random sample of cases:

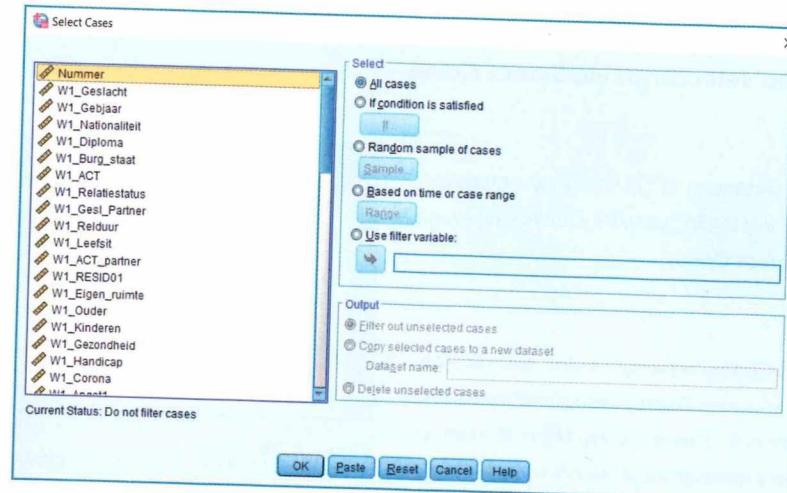
Bij deze optie trekt SPSS een aselecte of willekeurige steekproef aan respondenten uit je dataset. Je kunt hierbij aanduiden hoe groot die selectie relatief gezien moet zijn (*Approximately xx % of all cases*) of je duidt aan hoeveel cases SPSS at random moet selecteren uit een zelfgekozen aantal cases, bijvoorbeeld de eerste 600 cases of alle cases (*Exactly xx cases from the first xx cases*).

Based on time or case range:

Hier kun je cases selecteren op basis van hun rijnummer in de *Dataview*. Je geeft hierbij een bepaald bereik op door een rijnummer in te vullen als bovengrens (*First Case*) en een rijnummer als ondergrens (*Last Case*). Alle cases die voor het eerste of na het laatste opgegeven rijnummer komen, worden niet geselecteerd.

Use filter variable:

Bij deze laatste optie kun je een dichotome (of binaire) variabele met waarden 0 of 1 (zie hoofdstuk 2) opgeven als basis om te selecteren, bijvoorbeeld een variabele relatiestatus waarbij respondenten waarde 0 krijgen als ze in een relatie zaten tijdens de periode van dataverzameling en waarde 1 als dat niet het geval was. Indien je deze variabele ingeeft in het vakje onder *Use filter variable*, zal SPSS alle cases selecteren die waarde 1 hebben en alle andere cases (waarde 0 of missing) niet.



Figuur 4.37

We passen dit toe op een voorbeeld uit de C&W dataset, waarbij we ons afvragen hoe het gesteld is met de algemene gezondheid van de Belgische bevolking in maart 2020. Tijdens de eerste wave van het onderzoek (DATA_WAVE1_CenW.sav) werd deze vraag gesteld, waarbij de respondenten een antwoord konden aanduiden gaande van '1 = Slecht' tot '5 = Heel erg goed'. Dit levert in de dataset de variabele 'W1_Gezondheid' op. Indien je een frequentietabel opvraagt van deze variabele, ziet die er als volgt uit (zie figuur 4.38):

Hoe is het gesteld met je algemene gezondheid, op dit moment?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Slecht	32	1,6	1,6	1,6
	2 Redelijk	346	17,1	17,1	18,7
	3 Goed	802	39,7	39,7	58,4
	4 Erg goed	655	32,4	32,4	90,8
	5 Heel erg goed	185	9,2	9,2	100,0
	Total	2.020	100,0	100,0	

Figuur 4.38

Stel dat je echter in het kader van een deelonderzoek specifiek geïnteresseerd bent in de algemene gezondheid van studenten. Op figuur 4.38 kun je enkel aflezen hoe het gesteld is met de algemene gezondheid van alle respondenten. Je moet dus eerst een selectie maken van de respondenten die student zijn, om te kunnen zien hoe hun antwoorden verdeeld zijn op de vraag rond algemene gezondheid. Je kunt de variabele 'W1_ACT' gebruiken uit de dataset (waarbij ACT een afkorting is van Activiteit) om een onderscheid te maken tussen studenten ($W1_ACT = 1$) en niet-studenten ($W1_ACT = 2$ of 3) en alleen cases met studenten te selecteren in SPSS, via de onderstaande acties:

Een subgroep selecteren via Select Cases

ACTIE 1

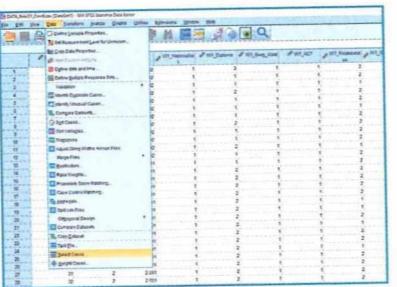
Open het databestand in SPSS waarin je een selectie wilt maken (hier: DATA_WAVE1_CenW.sav) en ga naar **Data** » **Select Cases**.

ACTIE 2

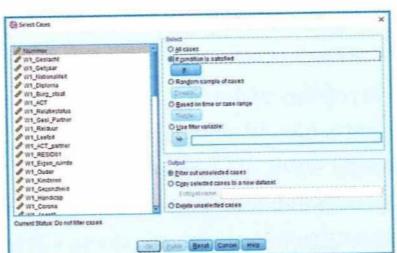
In het **Select Cases**-dialoogvenster dat nu geopend is, duid je onder **Select** de optie **If condition is satisfied** aan. Klik daarna op de knop **If...** om voorwaarde vast te leggen waarop geselecteerd moet worden.

ACTIE 3

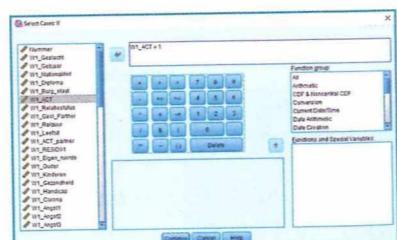
In dit tweede dialoogvenster tref je bovenaan een leeg kader aan waarin je de selectievoorwaarde kunt vastleggen. De blauwe toetsen in het midden kun je gebruiken als operatoren om de voorwaarde te formuleren. Links kun je op zoek gaan naar de variabelen die je nodig hebt, rechts kun je eventueel gebruikmaken van speciale functies om je voorwaarde te formuleren. Als je voorwaarde afgewerkt is, druk dan op **Continue**.



Figuur 4.39 ACTIE 1



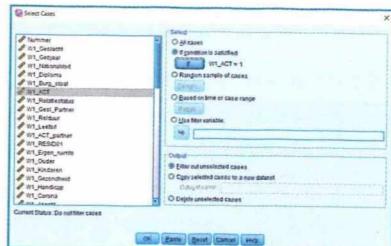
Figuur 4.40 ACTIE 2



Figuur 4.41 ACTIE 3

ACTIE 4

Terug in het eerste dialoogvenster zie je nu de geformuleerde voorwaarde staan naast de knop **If....**. Selecteer in het **Output**-kader **Filter out unselected cases** en druk dan op **Paste** indien je het SPSS-commando in je **Syntax Editor** wilt plakken of klik op **OK** als je het SPSS-commando wilt uitvoeren.



Figuur 4.42 ACTIE 4

USE ALL.

COMPUTE filter_\\$=(W1_ACT = 1).

VARIABLE LABELS filter_\\$ 'W1_ACT = 1 (FILTER)'.

VALUE LABELS filter_\\$ 0 'Not Selected' 1 'Selected'.

FORMATS filter_\\$ (f1.0).

FILTER BY filter_\\$.

EXECUTE.

Indien je het bovenstaande commando in de Syntax uitvoert en daarna opnieuw een frequentietabel opvraagt van de variabele 'W1_Gezondheid', krijg je een andere frequentietabel te zien (zie figuur 4.43). Door aan SPSS de opdracht te geven om enkel de subgroep studenten mee te nemen in analyses, zie je nu dat het totaal aantal cases van deze variabele op 724 staat, terwijl er in totaal 2020 respondenten in de dataset zitten. Op basis van deze output zie je echter niet dat het om studenten gaat: je moet zelf goed op de hoogte zijn welke filter er van kracht is. Rechts onderaan in de *Data View* staat de aanduiding *Filter On* als er een filter aan staat in de dataset en wanneer dat het geval is, zul je zien dat de rijnummers van cases die niet geselecteerd werden, doorgestreept zijn (indien je dit laatste niet ziet, ga dan naar *Edit* en vink *Hide excluded cases* uit). Daarnaast kun je opnieuw even navigeren naar **Data** » **Select Cases** om te zien welke filter er precies aan staat. Als je de selectie niet langer nodig hebt, kun je in het dialoogvenster van **Select Cases** terug *All cases* aanduiden en op **OK** klikken. Zo worden opnieuw alle cases meegenomen in de analyse en is er geen filter meer actief. De aanduiding *Filter On* verdwijnt dan ook in de *Data View*.

Hoe is het gesteld met je algemene gezondheid, op dit moment?

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid				
1 Slecht	21	2,9	2,9	2,9
2 Redelijk	146	20,2	20,2	23,1
3 Goed	256	35,4	35,4	58,4
4 Erg goed	242	33,4	33,4	91,9
5 Heel erg goed	59	8,1	8,1	100,0
Total	724	100,0	100,0	

Figuur 4.43



Het verwijderen van cases via Select Cases

Op het einde van paragraaf 4.3.2 zagen we hoe je manueel cases kunt verwijderen uit je dataset. Via het commando **Select Cases** kunnen we echter ook op een systematische manier **meerdere cases tegelijk verwijderen**. Stel dat je bijvoorbeeld onderzoek wilt uitvoeren naar smartphoneverslaving onder jongeren en met die doelgroep voor ogen een vragenlijst verspreidt. De doelgroep die je wilt bereiken, zijn jongeren van minimaal 15 jaar en maximaal 25 jaar. Je vermeldt daarom in het begin van de vragenlijst dat je specifiek op zoek bent naar respondenten uit die doelgroep, maar toch vullen sommige respondenten je vragenlijst gewoon in zonder deze informatie te lezen. Als gevolg krijg je een dataset met een paar cases die niet binnen de doelgroep passen en dus respondenten die jonger dan 15 jaar of ouder dan 25 jaar zijn. Indien je in de vragenlijst de leeftijd van de respondenten gevraagd hebt, kun je deze ongewenste cases opsporen en uit het databestand verwijderen. Je kunt dit manueel doen, maar de kans is groot dat je toch hier en daar een case over het hoofd zult zien die eigenlijk verwijderd moet worden. Je kunt deze cases wel systematisch en in één keer verwijderen door de acties van het **Select Cases** commando te volgen. Tijdens stap 3 formuleer je een voorwaarde waarbij je alle cases selecteert die tot je doelgroep behoren en die je dus wilt behouden, in dit geval alle cases waarbij de variabele leeftijd groter of gelijk aan 15 is en kleiner of gelijk aan 25 is. In stap 4 duid je dan aan dat alle permanent verwijderd mogen worden via *Delete unselected cases* (in plaats van tijdelijk uit te filteren). Als je dit commando uitvoert, worden alle respondenten die jonger dan 15 jaar of ouder dan 25 jaar in één keer verwijderd uit je dataset. Let op, dergelijke bewerkingen zijn onomkeerbaar en doe je niet zomaar. Zorg om die reden dat je op een kopie werkt van je originele dataset, want als je de voorwaarde verkeerd geformuleerd hebt, kun je sommige cases die wel bruikbaar waren voor je onderzoek kwijtraken.

4.4.3 VERGELIJKEN VAN GROEPEN VIA SPLIT FILE

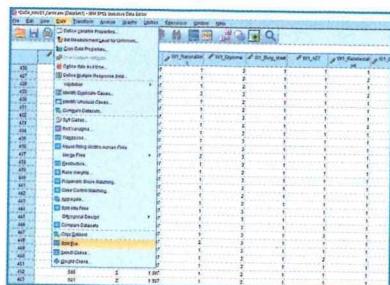
In sommige gevallen ben je niet alleen geïnteresseerd in een subgroep van de dataset (bv. mensen die in een relatie verkeren), maar wil je ook de vergelijking maken tussen die subgroep en de andere cases van de dataset (bv. mensen die niet in een relatie verkeren). In plaats van enkel de resultaten te zien van een analyse voor één specifieke subgroep (wat je via **Select Cases** kunt verwezenlijken), wil je dus graag de aparte resultaten van een analyse zien voor meerdere subgroepen tegelijk, zodat je die een op een kunt vergelijken. Dit kan handig zijn wanneer je vermoedt dat bepaalde subgroepen van onderzoekenheden (sterk) verschillen van elkaar op bepaalde kenmerken. Zo zou je bijvoorbeeld het aantal gepleegde misdrijven kunnen vergelijken tussen druggebruikers en niet-druggebruikers, aan de hand van een frequentietabel van elke groep. In SPSS kun je dit doen door het databestand op te splitsen in groepen van cases aan de hand van een categorische variabele. Het **splitsen van databestanden** wil niet zeggen dat je een databestand effectief in stukken gaat knippen, maar wel dat de uitvoer van alle analyses steeds per groep getoond wordt in de *Output Viewer*.

We laten dit zien aan de hand van het voorbeeld over de algemene gezondheid uit de C&W studie. Stel dat je de gezondheid van de studenten wilt vergelijken met de andere subgroepen, namelijk '2 = Ik werk thuis' en '3 = Ik werk op mijn werkplek'. Dit kun je doen door de onderstaande acties te volgen:

Subgroepen vergelijken via Split File

ACTIE 1

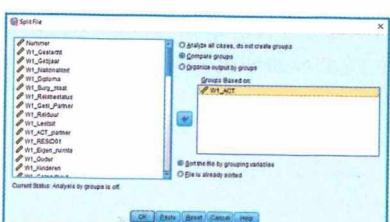
Open het databestand in SPSS dat je wilt opsplitsen in subgroepen (hier: DATA_WAVE1_CenW.sav) en ga naar **Data > Split File**.



Figuur 4.44 ACTIE 1

ACTIE 2

In het **Split File**-dialoogvenster dat nu geopend is, duid je de optie **Compare groups** aan als je wilt dat de resultaten van alle subgroepen apart in één uitvoertabel komen of de optie **Organize output by groups** als je voor elke subgroep een aparte uitvoertabel wilt (het verschil hiertussen is echter miniem). Sleep daarna de variabele(n) die je voor de subgroepen wilt vergelijken naar het kader **Groups Based on**. Laat **Sort the file by grouping variables** aangeduid staan en druk dan op **Paste**



Figuur 4.45 ACTIE 2



4.5 OEFENREEKS HOOFDSTUK 4

indien je het SPSS-commando in je **Syntax Editor** wilt plakken of klik op **OK** als je het SPSS-commando wilt uitvoeren.

SORT CASES BY W1_ACT.
SPLIT FILE LAYERED BY W1_ACT

Als je het bovenstaande commando selecteert en uitvoert in de Syntax en daarna een frequentietabel opvraagt van de variabele 'W1_Gezondheid', dan zie je dat de frequentietabel in drie subgroepen is opgesplitst (zie figuur 4.46):

Hoe is het gesteld met je algemene gezondheid, op dit moment?					
Welke situatie is het meest op jou van toepassing gedurende de laatste twee weken?		Frequency	Percent	Valid Percent	Cumulative Percent
1 Ik studeer thuis	Valid	1 Slecht	21	2,9	2,9
		2 Redelijk	146	20,2	20,2
		3 Goed	256	35,4	35,4
		4 Erg goed	242	33,4	91,9
		5 Heel erg goed	59	8,1	100,0
		Total	724	100,0	100,0
2 Ik werk thuis	Valid	1 Slecht	10	1,0	1,0
		2 Redelijk	155	15,3	15,3
		3 Goed	445	43,8	43,8
		4 Erg goed	320	31,5	91,5
		5 Heel erg goed	86	8,5	100,0
		Total	1.016	100,0	100,0
3 Ik werk op mijn werkplek	Valid	1 Slecht	1	,4	,4
		2 Redelijk	45	16,1	16,1
		3 Goed	101	36,1	36,1
		4 Erg goed	93	33,2	33,2
		5 Heel erg goed	40	14,3	14,3
		Total	280	100,0	100,0

Figuur 4.46

Alle analyses die je zult uitvoeren in SPSS zullen na deze bewerking telkens voor elke subgroep apart getoond worden. Rechts onderaan in de *Data View* staat de aanduiding *Split by* als het bestand momenteel gesplitst wordt. Deze opsplitsing blijft net zoals bij *Select Cases* altijd van kracht tot je ze zelf weer uitzet, door opnieuw naar *Data* » *Split File* te navigeren, *Analyze all cases, do not create groups* aan te duiden en te bevestigen met *OK*. Je kunt ook het onderstaande commando in de *Syntax* uitvoeren:

SPLIT FILE OFF.

Oefening H4.1. Databestand aanmaken. Als je surft naar pelckmans.be/toegepaste-statistiek, vind je de eerste pagina van een ingevulde potlood en papier-vragenlijst die overeenkomt met de vragen die gesteld werden in de eerste wave van het Corona & Welzijn onderzoek. Open een leeg databestand, maak de variabelen aan, voeg de *variable labels* en *value labels* toe, en vul de antwoorden in van de respondent. Ga na of er verschillen zijn tussen jouw ingegeven waarden en deze van respondent nummer 184 uit 'DATA_WAVE1_CenW.sav'.

Oefening H4.2. Frequenties. Open 'DATA_WAVE1_CenW.sav'. Hoeveel procent van de respondenten rapporteerde coronasymptomen (W1_Corona) in de eerste wave?

Antwoord: % van de respondenten rapporteerde coronasymptomen.

Oefening H4.3. Cases toevoegen. Open 'DATA_WAVE1_CenW.sav' en vraag een frequentietabel op van 'W1_Relatiestatus'. Open 'Data_oefening413_add_cases.sav' en voeg beide bestanden samen. Vraag opnieuw een frequentietabel op van 'W1_Relatiestatus'. Hoeveel procent van de respondenten heeft een relatie in het oorspronkelijke en in het nieuwe databestand?

Antwoord:% heeft een relatie in het oorspronkelijke databestand.% heeft een relatie in het nieuwe databestand.

Oefening H4.4. Variabelen toevoegen. Open ‘DATA_WAVE1_CenW.sav’ en vraag een frequentietabel op van het geslacht (W1_Geslacht). Open ‘Data_oefening415_add_variables.sav’ en voeg beide bestanden samen. Vraag opnieuw een frequentietabel op van het geslacht (W1_Geslacht). Wat is het percentage mannen en vrouwen in de twee frequentietabellen? Hoe komt dit?

Antwoord:% zijn mannen en% zijn vrouwen in het oorspronkelijke bestand.% zijn mannen en% zijn vrouwen in het nieuwe bestand. De reden is