Appendix B
Computeranalyse van kwantitatieve data

Ten behoeve van de analyse van kwantitatieve data worden verschillende programmapakketten gebruikt, zoals SAS (Statistical Analyzing System) en SPSS, naast allerlei programma's voor speciale doeleinden. Deze appendix geeft een korte introductie in het gebruik van een van de bekendste programmapakketten, SPSS. Het bedrijf SPSS is in 2009 overgenomen door IBM. De laatste versie – SPSS 17.0 – is opgevolgd onder de nieuwe naam PASW[1] Statistics 18.0. Wij gebruiken in het hiernavolgende de versie SPSS 11.0. De tekst bevat enkele oefeningen.

B.1 Werken met SPSS

Het startvenster van SPSS vraagt je onder meer of je een bestaand databestand wilt openen. Wij selecteren de 1991 U.S. General Social Survey.sav (GSS91).

> *Het is natuurlijk ook mogelijk om je eigen data te gebruiken. Dit onderwerp wordt behandeld in sectie B.2.*

Vervolgens krijg je de keuze tussen twee hoofdvensters (Data View en Variable View). Gebruikers switchen vaak van de een naar de ander.
Klik, om te beginnen, eerst op Data View. We zien nu de scores van 1517 achtereenvolgende surveyrespondenten op de variabelen; dit is een echte datamatrix. De data zijn van een Amerikaans onderzoek waaruit een deelverzameling van eenheden en variabelen is getrokken.
Klik nu op Variable View. Voor elke variabele worden tien eigenschappen genoemd. *Width* slaat op het aantal posities per variabele (met een maximum van acht), 'Decimals' op het aantal plaatsen achter de komma. *Label* staat voor de volledige naam van de variabele zoals je die in je rapport zou willen vermelden. De namen verschijnen volledig op het scherm als je, boven in het scherm,

1 De nieuwe naam PASW komt van Predictive Analytics Software.

klikt op de scheidslijn tussen de woorden Label en Values, en je je hand naar rechts beweegt met ingedrukte muisknop. Je kunt tijdelijk de ruimte voor 'Values' op dezelfde manier vergroten. Maar als je op het kleine grijze vlakje rechts (nadat je op 'Values' voor een van de variabelen hebt geklikt) klikt, verschijnt er een klein scherm met de naam 'Value Labels'. Het bevat, bijvoorbeeld, de informatie dat voor de variabele SEX code 1 is toegewezen aan de mannen, en code 2 aan de vrouwen. Klik op OK.

Herhaal de instructie voor enkele andere variabelen.

De kolom 'Missing' slaat op de zogenoemde 'system-missing values'. Dit betreft de codes die toegewezen worden voor 'geen antwoord', 'weet niet' en 'niet van toepassing'. Bij variabele 9 (Education) worden de codes 97, 98 en 99 voor dit doel gebruikt.
Op dit moment zijn 'Columns' en 'Align' niet van belang. 'Measure' of meetniveau is natuurlijk wel van belang. SPSS onderscheidt 'nominal', 'ordinal' en 'scale'. 'Scale' is identiek met 'kwantitatief meten', dat wil zeggen, meten op interval- en rationiveau.

Om vertrouwd te raken met een paar belangrijke procedures kun je vervolgens het best wat met het programma spelen. Klik bijvoorbeeld in het menu op ANALYZE, en daarna op DESCRIPTIVE STATISTICS. Als je nu op FREQUENCIES klikt, verschijnt het dialoogvenster FREQUENCIES. In het gedeelte aan de linkerkant kun je alle variabelen zien (de bronnenlijst). Je kunt bijvoorbeeld een overzicht van de verdeling op de variabele SEX krijgen door met de muis naar SEX te gaan en te klikken) en de geselecteerde variabele naar de rechterkant te verplaatsen (klik op de pijl). Ten slotte klik je op OK. Wacht een moment tot de tabel verschijnt. Aan de linkerkant is alles wat rechts staat netjes aangegeven. Door op Statistics of Charts te klikken verschijnt een klein rood pijltje voor de tekst of de tabel die je in je rapport wilt opnemen.

Herhaal deze procedure met de variabelen RACE end GENERAL HAPPINESS. Maak eerst het dialoogvenster FREQUENCIES leeg door op RESET te klikken, anders verschijnt alles wat je al eerder hebt gezien.

Nadat we dit alles hebben geoefend en hebben gezien hoe eenvoudig dat is, vragen we ons af wat er gebeurt 'als je op een paar andere knoppen drukt'. Wat zien we, bijvoorbeeld, als we onder DESCRIPTIVE STATISTICS niet FREQUENCIES, maar DESCRIPTIVES aanklikken?

Kies de variabele AGE; breng deze over van de linkerkant naar de rechterkant, en klik op OK. Nu zie je een tabel met het aantal respondenten (n = 1514). Je ziet ook dat de jongste respondent 18 jaar oud was, en de oudste 89, en dat de gemiddelde leeftijd van de steekproef 45.63, was, met een standaardafwijking (zie hoofdstuk 11) van 17.81. Het venster 'DESCRIPTIVES' geeft ons dus enkele eenvoudige beschrijvende maten voor de betreffende variabele. Door onder DESCRIPTIVE STATISTICS op EXPLORE te klikken krijgen we een nog veel groter aantal statistisch beschrijvende maten te zien.
Als je op CROSSTABS klikt, vraagt het dialoogvenster je twee variabelen in te voeren, de ene als rijen, de andere als kolommen in de tabel die je wilt maken. Voer ze na elkaar in en klik op OK.

> Herhaal de procedure door de variabelen van plaats te laten wisselen. Je ziet nu dat de tabel 90° is geroteerd, maar geheel dezelfde informatie geeft. Wat in de rijen en wat in de kolommen moet staan, wordt uitgelegd in hoofdstuk 12.

Als afsluiting van deze eerste ronde om de smaak te pakken te krijgen, klikken we niet ANALYSE in het menu aan, maar GRAPHS. Daarna klikken we op BAR. Het dialoogvenster BAR CHARTS verschijnt. Klik op het SIMPLE-icoontje, en daarna op DEFINE. We kiezen nu een variabele, bijvoorbeeld REGION, en brengen deze over naar het kleine vlakje door op CATEGORY AXIS te klikken, en daarna op OK. Het resultaat spreekt voor zichzelf.

> Als je in GRAPHS bent, klik dan eens niet op BAR, maar op PIE, en herhaal dezelfde procedure. Om een fraai kleurenplaatje te krijgen, kies dan eens AGE als variabele (vergeet niet eerst te resetten). Het is natuurlijk niet een adequaat plaatje voor zo'n fijn verdeelde variabele als deze.

Vanuit ANALYSE kun je altijd terug naar de datamatrix door het outputvenster te sluiten.

Je hebt ongetwijfeld gezien dat er nog veel meer mogelijkheden zijn voor analyse dan de paar die we hiervoor hebben toegelicht. Je krijgt een idee van de vele technieken in de sociale en gedragswetenschappen (en ook in andere disciplines) als je wat langer met SPSS 'speelt'. Je zult ook wel gezien hebben dat de vele dialoogvensters waarbinnen je kunt kiezen, vergelijkbaar zijn met zelfbedieningszaken voor vele producten: de mogelijke analyseprocedures. De praktische aspecten daarvan leer je al doende; in de hoofdstukken 10-12 van dit boek wordt de basistheorie van enkele zaken uitgelegd.

B.2 Eigen data invoeren

Om je data in te voeren begin je met een schoon scherm. Switch, indien nodig, naar Variable View. Gebruik het menu om 'type in data' te kiezen. Eerst moeten de variabelen worden gedefinieerd. Ga met je muis naar het kleine vlakje linksboven en typ de naam van de eerste variabele. Het is misschien handig om het identificatienummer dat elke respondent heeft, als eerste variabele te kiezen. Afhankelijk van de grootte van N kies je daarvoor een *width* van drie of vier posities.

> Denk eraan: de nummers die links vermeld zijn, zijn niet de respondentnummers, maar eenvoudige op elkaar volgende nummers. Als je een respondent met volgnummer 0003 hebt verwijderd, schuiven de nummers gewoon op.

De volgende, inhoudelijke variabele is bijvoorbeeld SEX. Typ SEX. Klik daarna het vlakje rechts aan, onder TYPE. In het dialoogvenster kun je verschillende zaken aangeven. Kies bij Variable Type voor 'numerical', en geef ook de breedte en het aantal decimalen aan. Klik daarna op OK. Voor 'SEX' heb je geen decimalen nodig, dus geven we bij Decimal Place voor deze variabele een 0. Als we dat niet zouden doen, zou een code 2 altijd worden geschreven als 2,00, maar we hebben hier geen komma's en decimalen nodig. Het definiëren van de relevante VALUES en LABELS is echter noodzakelijk. Dit maakt het in een later stadium gemakkelijker om te zien welke variabelen in je bestand zitten, met de betekenis van elke antwoordcategorie. Voor SEX ben je vrij om 'sekse', 'gender' of wat dan ook dat je zinvol lijkt, als label toe te kennen. Voor die toewijzing zijn acht posities beschikbaar (acht in letter, getal, ander symbolenstelsel of wat je maar wilt gebruiken, maar je moet altijd met een letter beginnen). Vervolgens kun je bij VALUE aangeven dat de code 1 gebruikt wordt voor de mannen. Klik op ADD, en voer code 2 voor vrouwen in. Klik opnieuw op ADD, en bevestig dat met OK. Op die manier sla je de toegewezen codenummers op samen met hun betekenis. Klik op OK.

> Er staan nog vier andere kolommen: Missing, Columns, Align en Measure. We zullen verderop de eerste kolom toelichten. De resterende drie kun je laten zoals ze staan. Nadat je alle eigenschappen van de variabele SEX hebt ingevoerd, ga je naar de volgende variabele door de cel onder SEX (links) aan te klikken. Variabelen zoals AGE, hebben categorieën die direct kunnen worden ingevoerd; zij worden niet gedefinieerd (behalve natuurlijk de missing values).

Als alle variabelen zijn ingevoerd, kies je het venster DATA VIEW.
De scores van de eerste respondent kunnen nu worden ingevoerd. Daarna kun je met de tweede respondent verder gaan, enzovoort. Als je een nummer intypt, verschijnt dit in de cel linksboven. Als je het pijltje met de cursor naar rechts laat gaan, kom je bij de plaats voor de score van die respondent op de tweede variabele. Nadat de score op de laatste variabele is ingevuld, ga je met de cursor naar de eerste variabele voor de tweede respondent. Het nu getypte getal verschijnt in de cel linksboven. Met de pijltjestoets kun je altijd naar links of rechts, boven of onder om foutjes te herstellen.

> Je kunt natuurlijk de datamatrix ook verticaal vullen door eerst de eerste variabele voor alle respondenten in te vullen, en dan naar de volgende variabele te gaan, door de cursor naar boven en naar beneden te bewegen. Maar dit is niet erg gebruikelijk.

Hoe voer je een ontbrekende waarneming in? Sla deze eenvoudig over, typ dus niets, maar beweeg je cursor niet één stap, maar twee stappen naar rechts. De datamatrix toont een komma of een punt in de overgeslagen cel. In de berekeningen wordt de score van de respondent op deze variabele als 'missing' beschouwd.
Het wordt een beetje ingewikkelder als je onderscheid wilt maken tussen 'geen antwoord' en 'niet van toepassing'. Als ze allebei met een komma of een 'blank' zijn genoteerd, kunnen ze niet worden onderscheiden: beide zijn 'missing values'. En door het toewijzen van een of ander getal aan 'niet van toepassing' worden deze getallen ten onrechte meegeteld in berekeningen van bijvoorbeeld het gemiddelde of percentages. Wat je wel kunt doen, is het toewijzen van een bepaalde unieke code aan weigeringen respectievelijk aan 'niet van toepassing' en 'weet niet'. Je wijst bijvoorbeeld bij 'niet van toepassing' een code 8 toe bij een vraag met zes of minder 'echte' antwoordcategorieën. Bij een variabele die twee posities beslaat, gebruik je code 98. Op dezelfde manier gebruik je de codes 9 en 99 voor 'weet niet' en een 7 en 97 voor 'weigering' (een en ander is ook te vinden in hoofdstuk 6). Gebruik daarna de definitie van de variabele om aan te geven dat deze antwoordcategorieën als 'missing value' beschouwd moeten worden. Klik onder Variable View op de tab en voor de betrokken variabele op de kolom 'missing'. Klik het grijze vlakje aan en geef in het dialoogvenster aan dat 7, 8 and 9 (respectievelijk 97, 98 en 99) 'Discrete missing values' zijn. Klik op OK. Door te vragen naar een frequentieverdeling zie je hoeveel personen in elk van die speciale categorieën zitten. Maar zodra een berekening wordt uitgevoerd, worden personen met deze waarden buiten beschouwing gelaten.

Bestanden worden bewaard door in het menu FILE aan te klikken, daarna SAVE AS, en een of andere informatiedrager te selecteren en een naam aan het bestand te geven, bijvoorbeeld a:\owndata1, of: c:\exercise\data1, of als een SPSS-bestand te bewaren.

Oefening: voer de data in van oefening 1.4 en bewaar het bestand. Ga naar het startvenster, activeer het bestand, en maak een paar frequentieverdelingen en kruistabellen.

B.3 Foutencontrole

Bij de variabele SEX worden maar twee waarden gebruikt, 1 en 2. Alle andere waarden zijn onmogelijk. Zou je per ongeluk een 3 of een 5 intypen, dan komen deze keurig in de frequentieverdeling te staan. Je kunt een check op deze soort fouten doen door de frequentieverdelingen van alle variabelen in één job via ANALYSE en FREQUENCIES te bekijken. Je doet dit door van de eerste naar de laatste variabele te gaan met de linker muisknop ingedrukt. Laat dan de knop los, en klik de pijlpunt. De variabelen komen aan de rechterkant van het venster te staan. Klik daarna op OK. Als in een kleine datamatrix zich ergens zo'n fout voordoet, is het gemakkelijk de betreffende respondent te vinden en de score te corrigeren (cursor op de fout en de juiste code intypen). In een grotere datamatrix kan dit laatste niettemin veel tijd kosten, maar dan kun je in Data View klikken op de naam van de variabele, en vervolgens in het menu op Edit klikken. In het dialoogvenster van Find kun je de fout waar je naar op zoek bent, vinden. Door op Find Next te klikken vind je de respondent met de foutieve score, die dan gecorrigeerd kan worden.

B.4 Hercoderen

Een onderzoeker kan verschillende redenen hebben om een variabele te hercoderen, dat wil zeggen op een andere manier te coderen dan de oorspronkelijke codering. Een van die redenen kan zijn dat je een lange lijst van antwoordalternatieven wilt samenvatten in een korte opsomming. Zo ligt het voor de hand om bijvoorbeeld in het GSS91-bestand de frequentieverdeling van AGE te vereenvoudigen (ieder geboortejaar heeft oorspronkelijk een eigen code!) tot een verdeling met vijf leeftijdsgroepen. Het is verstandig om elke nieuw gecodeerde variabele een eigen naam te geven, hier bijvoorbeeld AGENEW.

Een andere reden doet zich voor bij Likertschalen. In de meeste schalen komen zowel positief als negatief geformuleerde items voor. Om de correlaties tussen de items goed te beoordelen is het gemakkelijker om de negatieve items eerst om te scoren (of de positieve, dat maakt niets uit). Dat wil zeggen: 1 wordt 5, 2 wordt 4, 3 blijft 3, 4 wordt 2 en 5 wordt 1, als je met vijf antwoordcategorieën werkt. Als een aantal items op deze wijze gehercodeerd wordt, moeten alle correlaties in de matrix positief zijn.
Hoe ga je te werk in SPSS als je de variabele 'RACE', met drie waarden, wilt vervangen door een dichotome variabele?

Klik TRANSFORM in het menu
Klik RECODE
Klik INTO DIFFERENT VARIABLES (omdat je een nieuwe naam wilt toekennen)
Klik RACE
Klik op het pijltje dat naar rechts wijst
Klik in het vakje OUTPUT VARIABLE, en typ een nieuwe naam in, bijvoorbeeld RACENEW
Klik CHANGE
Klik OLD AND NEW VALUES
Typ in het vakje OLD VALUE: 1
Typ in het vakje NEW VALUE: 1
Klik ADD
Typ in het vakje OLD VALUE: 2
Typ in het vakje NEW VALUE: 2
Klik ADD
Typ in het vakje OLD VALUE: 3
Typ in het vakje NEW VALUE: 2

Als alle waarden van de oorspronkelijke variabele zijn veranderd (vergeet ook niet de onveranderde waarden, zoals wanneer een 3 een 3 blijft, aan te geven), klik dan op OK.
Als er meer dan één variabele moet worden gehercodeerd, klik dan niet op OK na de eerste variabele, maar op CONTINUE. Pas na de laatste variabele klik je op OK.

De procedure is een beetje anders als we een kwantitatieve variabele met veel categorieën willen vereenvoudigen, bijvoorbeeld een leeftijdsverdeling willen vervangen door een eenvoudig 'jonger dan 25' en '25 en ouder'. Het begin van de procedure is hetzelfde, maar als je bij 'OLD AND NEW VALUES' komt, volgt:

Klik OLD AND NEW VALUE
Klik RANGE LOWEST THROUGH (dit betekent: alle leeftijden te beginnen met de laagste tot aan ...)
Typ: 24
Klik NEW VALUE
Typ: 1
Klik ADD
Klik RANGE ... THROUGH HIGHEST (dit betekent: alle leeftijden beginnend met ... tot aan de hoogste)
Typ: 25
Klik NEW VALUE
Typ: 2
Klik ADD
Als nog meer van deze jobs moeten worden gedaan, klik op CONTINUE, en herhaal de procedure
Ten slotte: klik op OK.

Een nieuw gemaakte variabele wordt automatisch toegevoegd aan de datamatrix als laatste kolom.

B.5 Het optellen van scores over enkele variabelen

Bij samengestelde meetinstrumenten moeten vaak de scores over enkele variabelen worden opgeteld. In het GSS91-bestand, bijvoorbeeld, verwijzen de variabelen hlth1 tot en met hlth9 naar het al dan niet voorgekomen zijn van trieste gebeurtenissen in het leven die invloed kunnen hebben op iemands lichamelijke en geestelijke gezondheid. Je kunt je voorstellen dat niet per iedere afzonderlijke gebeurtenis verbanden worden gelegd met iemands eigen levensgeluk of iemands carrière, maar dat eerst een 'optelsom' wordt gemaakt over alle gebeurtenissen. Nu moet je eerst weten hoe de variabelen gescoord zijn. Daar kom je het snelst achter door vanuit 'variable view' op 'values' van de betreffende variabele te klikken. Je krijgt dan een 'value labels'-venstertje te zien. Het blijkt dan dat een ja-antwoord score 1, een nee-antwoord score 2 heeft gekregen. Over negen vragen is het theoretische minimum dus 9, en het maximum 18. Iemand die géén van de gebeurtenissen heeft meegemaakt, krijgt dus de hoogste score, 18. Hoe maken we een nieuwe variabele, die we 'healthtotal' (HLTHTOT) noemen?

Het is niet zo handig dat hier de codes 1 en 2 zijn gebruikt. Veel handiger voor variabelen waarbij moet worden aangegeven of iets aanwezig is of niet, is: code 1 = ja en code 0 = nee. Als je dan optelt, krijg je precies het aantal keren ja. Ook is het gemiddelde te beschouwen als een percentage. Dus: liever eerst hercoderen!

Klik op TRANSFORM
Klik op COMPUTE
Typ in het vakje onder TARGET VARIABLE: HLTHTOT
Typ onder NUMERIC EXPRESSION: hlth1+hlth2+hlth3+hlth4+hlth5+hlth6+hlth7+hlth8+hlth9
Klik op OK.

De nieuwe variabele hlthtot wordt nu als laatste aan de datamatrix toegevoegd. Je kunt nu op de gewone manier de frequentieverdeling en enkele beschrijvende statistische gegevens oproepen. Je ziet als resultaat dat de laagst voorkomende waarde 12 is, en de hoogst voorkomende waarde 18. Het gemiddelde is ongeveer gelijk aan 17. De mediaan is wellicht een betere maat.
Bij een ander type samengestelde variabelen kun je onder NUMERIC EXPRESSION andere algebraïsche bewerkingen doen. Als je bijvoorbeeld zou beschikken over een datamatrix met gegevens van landen, waaronder OPP (oppervlakte) en INW (aantal inwoners), zou je een nieuwe variabele bevolkingsdichtheid (aantal inwoners per vierkante kilometer) kunnen berekenen door de opdracht INW : OPP te typen.

Sommige meetinstrumenten bestaan uit verschillende – soms wel tien tot vijftien – deelinstrumentjes. Elk deelinstrumentje is een vraag of een bewering waarop een standpunt van de respondent wordt verwacht. Al die vragen verwijzen naar hetzelfde, bijvoorbeeld naar de houding van de respondent ten opzichte van allochtonen, of naar de houding ten opzichte van het gebruik van het openbaar vervoer. De achtergrond van zo'n samengesteld meetinstrument is dat er niet één enkele perfecte vraag denkbaar is. Daarom stelt de onderzoeker verschillende vragen en als we de scores van een respondent over al die vragen optellen (of het gemiddelde daarvan berekenen), hebben we een meer betrouwbare benadering van de houding van de respondent. De basisgedachte van zo'n Likertschaal bespraken we al in hoofdstuk 7.
Maar dat idee veronderstelt wel dat al die vragen inderdaad hetzelfde meten, in dezelfde richting wijzen.