Inleiding SPSS 24

Inleiding SPSS 24

voor IBM SPSS Statistics

Eelko Huizingh

Boom

Opmaak binnenwerk: Holland Graphics, Amsterdam
Omslagontwerp: Carlito's Design, Amsterdam
Basisontwerp omslag: Studio Bassa, Culemborg

ISBN 978 90 2440 696 8
ISBN 978 90 2440 697 5 (e-book)
NUR 123/916

www.boomhogeronderwijs.nl

Voorwoord

Statistische analyses zijn het wetenschappelijke wereldje allang ontgroeid. Sla maar een willekeurige krant open en u vindt er de resultaten van statistisch onderzoek. Actualiteitenrubrieken presenteren eigen opiniepeilingen en in verkiezingstijden worden dagelijks nieuwe prognoses bekendgemaakt. De opkomst van statistische software heeft in niet geringe mate bijgedragen aan deze ontwikkeling.
SPSS is al enkele decennia een van de belangrijkste statistische softwarepakketten. In de loop der tijd is het pakket steeds opnieuw aangepast aan de veranderende wensen van gebruikers en de veranderende mogelijkheden van de technologie. Het belangrijkste verschil tussen *SPSS voor Windows* en alle voorgaande SPSS-versies is het bedieningsgemak. Het wordt steeds eenvoudiger om analyses uit te voeren.
Het grote bedieningsgemak kent ook een keerzijde: een t-toets met SPSS uitvoeren lukt echt iedereen. Maar weet ook iedereen welke conclusies uit de resultaten mogen worden getrokken? En: of in dit geval eigenlijk wel een t-toets mocht worden uitgevoerd? Deze vragen hebben te maken met het interpreteren van de analyseresultaten en de veronderstellingen van de analysetechniek. Met het toenemende bedieningsgemak verschuift de aandacht naar deze twee onderwerpen. Het 'domme' rekenwerk kan aan de computer worden overgelaten.

Doelstelling boek

Een goed boek over SPSS dient dus niet alleen te gaan over 'het indrukken van knoppen'. Het dient het gebruik van SPSS in een breder perspectief te plaatsen. De doelstelling van dit boek is dan ook:

> *Het leren benutten van de mogelijkheden van SPSS bij het verantwoord uitvoeren van statistisch onderzoek.*

Om deze doelstelling te realiseren is dit boek in twee delen ingedeeld. Deel I is getiteld 'Leren werken met SPSS'. In dit deel worden de mogelijkheden van SPSS geschetst per fase van het onderzoeksproces. Daarnaast bevat deel I een drietal sessies. De drie sessies zijn zo geschreven dat u ze zittend achter het beeldscherm kunt doornemen. Alle te geven muis- en toetsaanslagen worden genoemd, voorzien van de nodige tekst en uitleg. Dit maakt het boek uitstekend geschikt voor zelfstudie. Tijdens de drie sessies wordt het uitvoeren van een klein onderzoek nagebootst, zodat u precies die dingen doet die u later bij uw eigen onderzoek ook zult doen.
Deel II, 'Werken met SPSS', is bedoeld als naslagdeel. U verzamelt gegevens en wilt hiermee bepaalde bewerkingen of analyses uitvoeren en staat voor de vraag hoe u dit met SPSS kunt doen. Deel II is gewijd aan de vele analyses die SPSS kent. Bij elke analyse worden de dialoogkaders afgebeeld en de door SPSS gemaakte uitvoer getoond, besproken en geïnterpreteerd. Om u behulpzaam te zijn bij het verantwoord uitvoeren van statistisch onderzoek, worden in elk hoofdstuk ook de te maken veronderstellingen en verwante analysetechnieken genoemd. Om de toegankelijkheid van deel II te vergroten, is achterin

het boek een overzicht opgenomen van de verschillende soorten analyses afgezet tegen het meetniveau van de gegevens.

Gebruik boek

Zelf geef ik aan de Rijksuniversiteit Groningen al vele jaren cursussen waarbij we SPSS gebruiken. Gebaseerd op mijn ervaringen heb ik destijds het boek *Inleiding SPSS* geschreven. Van dit boek zijn inmiddels meer dan vijftien versies verschenen, waarbij het boek steeds is aangepast aan de nieuwe mogelijkheden van SPSS en ideeën opgedaan tijdens het gebruik van mijn boek.
Mijn ervaring is dat het leren werken met SPSS en het gebruiken van SPSS bij statistisch onderzoek twee verschillende zaken zijn. Vandaar ook de tweedeling in dit boek. Het uitgangspunt bij het leren werken met SPSS is dat een softwarepakket het beste vanachter het beeldscherm geleerd kan worden. Over een softwarepakket moet niet uitgebreid worden verteld of gelezen – zelf de knoppen indrukken is de beste leerschool. Daarom bevat deel I een drietal sessies waarin alle belangrijke onderdelen van SPSS aan de orde komen.
Voor het goed uitvoeren van statistische analyses is een makkelijk toegankelijk naslagwerk nodig. Hierin moeten analyses stap voor stap worden uitgelegd en de uitkomsten op een begrijpelijke manier worden toegelicht. Daarnaast dient duidelijk te zijn welke conclusies getrokken mogen worden uit welke uitkomsten. Deze ideeën bepalen de opzet van de hoofdstukken in deel II.

Dankwoord

Een aantal mensen heeft een voor mij belangrijke bijdrage geleverd bij de totstandkoming van dit boek. Allereerst wil ik IBM SPSS dank zeggen voor de uitstekende samenwerking die we nu al meer dan twintig jaar hebben. Het contact met de verschillende medewerkers blijkt telkens opnieuw een plezierige en efficiënte ervaring en vormt voor mij een aangename stimulans om dit boek steeds weer bij te werken.
Daarnaast bedank ik Ilse en Evelien. Zij hebben de groei van dit boek in vele stadia meegemaakt en zijn inmiddels zelfs uitgegroeid tot gebruikers ervan, iets wat ik destijds nooit had kunnen bedenken. Tot slot gaat mijn dank uit naar Agnes. Haar rustige en stimulerende aanwezigheid heeft mijn leven weer een nieuwe impuls gegeven. Heerlijk om genieten en productief bezig zijn zo goed met elkaar te kunnen combineren!

Assen, februari 2017

Eelko Huizingh

Inhoud

Deel I
Leren werken met SPSS

Dit eerste deel bestaat uit drie inleidende en drie praktische hoofdstukken. In de eerste drie hoofdstukken worden de achtergronden en mogelijkheden van SPSS beschreven en in de andere drie hoofdstukken wordt in de vorm van sessies de werking van SPSS toegelicht.

Hoofdstuk 1 behandelt de geschiedenis en de mogelijkheden van het SPSS-gegevens-analysesysteem en in hoofdstuk 2 staat de vraag centraal hoe u SPSS kunt gebruiken binnen een onderzoek. Het onderzoeksproces wordt daartoe in zeven fasen opgedeeld en per fase wordt aangegeven welke ondersteuning SPSS kan bieden. In hoofdstuk 2 worden ook steeds de bijbehorende SPSS-opdrachten genoemd.
Voordat u met SPSS kunt gaan analyseren, moet eerst een gegevensbestand worden gemaakt. Hiertoe moeten metingen of enquêtevragen naar variabelen worden 'vertaald' en, na het bepalen van de codering, de gegevens in een computerbestand worden ingevoerd. Van dit alles wordt een overzicht opgesteld, dit is het in hoofdstuk 3 besproken codeboek.

De laatste drie hoofdstukken van deel I zijn praktisch van aard. Elk hoofdstuk bestaat uit een sessie waarin aan de hand van een eenvoudig voorbeeld alle onderwerpen aan de orde komen die van belang zijn om te kunnen werken met SPSS. De drie sessies zijn zo geschreven dat u ze zittend achter het beeldscherm kunt doornemen. Alle te geven muis- en toetsaanslagen worden genoemd, voorzien van de nodige tekst en uitleg. Omdat tijdens de drie sessies het uitvoeren van een klein onderzoek wordt nagebootst, doet u precies die dingen die u later bij uw eigen onderzoek ook zult gaan doen.

Wat is het doel van de sessies?
Voor ervaren SPSS-gebruikers is het een snelle kennismaking met de nieuwste versie van SPSS. Voor degenen die nu voor het eerst kennismaken met SPSS (in SPSS-jargon zijn zij de *new friends,* in tegenstelling tot de hierboven besproken *old friends*), geven de sessies in vogelvlucht een indruk van de opbouw, de werkwijze en de mogelijkheden van SPSS.

1 Achtergronden van SPSS voor Windows

Inleiding

In een wereld waarin veranderingen elkaar steeds sneller opvolgen, is SPSS een van de oudste en nog steeds veelgebruikte softwarepakketten. De eerste versie van SPSS verscheen al in 1968! Paragraaf 1.1 beschrijft de ontwikkeling die het pakket sindsdien heeft doorgemaakt. De programmatuur van SPSS bestaat uit een basismodule, die u altijd nodig hebt om te kunnen werken met SPSS, en daarnaast een aantal uitbreidingsmodules. De mogelijkheden hiervan worden in paragraaf 1.2 kort besproken.

1.1 SPSS: historie en ontwikkeling

Statistiek is voor een groot deel niets anders dan het vele malen herhaald uitvoeren van eenvoudige rekenkundige bewerkingen. Neem het berekenen van een veelgebruikt kengetal als de standaarddeviatie (de standaardafwijking). Dit gaat als volgt:

- Tel het aantal waarnemingen.
- Bereken het totaal van alle waarnemingen.
- Deel het totaal door het aantal waarnemingen, dit is het gemiddelde.
- Trek van iedere waarneming het gemiddelde af.
- Kwadrateer deze verschillen.
- Tel alle kwadraten bij elkaar op.
- Deel dit totaal door het aantal waarnemingen minus 1.
- Trek de wortel uit de uitkomst van de deling.

Hoewel het bovenstaande een hele mond vol is, ziet u dat de kennis die nodig is voor deze berekeningen nauwelijks het niveau van de basisschool ontstijgt. Deze berekening een keer met de hand uitvoeren is nuttig om inzicht te krijgen in statistiek. Maar wat een werk zou het zijn om dit voor honderd waarnemingen te moeten doen, nog afgezien van de kans op fouten! Het is dan ook niet verwonderlijk dat al vanaf het eerste begin van het computertijdperk dit rekenapparaat dankbaar gebruikt wordt door statistici. Het is evenmin verwonderlijk dat pas met de ontwikkeling van computers de toegepaste statistiek een hoge vlucht heeft kunnen nemen.
Al sinds 1968 draagt SPSS hieraan zijn steentje bij. SPSS is destijds ontwikkeld als analyseprogramma voor sociale wetenschappers. De letters SPSS vormden in die tijd de afkorting van Statistical Package for the Social Sciences. Tegenwoordig is SPSS veel meer: het programma wordt nu verkocht als een modulair totaalpakket voor gegevensinvoer, gegevensverwerking en gegevenspresentatie. De doelgroep bestaat dan ook allang niet meer uit alleen sociale wetenschappers, SPSS is overal bruikbaar waar gegevens worden verzameld, geanalyseerd en gepresenteerd in tabellen en grafieken.

Een sterk punt van SPSS is dat het nagenoeg alle veelgebruikte analysetechnieken kan uitvoeren. Dit maakt het pakket onder andere bijzonder geschikt voor het analyseren van vragenlijsten. Een van de kenmerken van vragenlijsten is namelijk dat variabelen op verschillende meetniveaus voorkomen. Dus niet alleen ratio of interval, maar ook nominaal of ordinaal.

Veranderende statistiek en technologie
Gedurende de vele jaren die zijn verstreken sinds 1968 is er flink gesleuteld aan SPSS om het programma aan te passen aan de voortschrijdende ontwikkelingen op het terrein van statistiek en technologie. Veel statistische methoden die in de laatste vier decennia zijn ontwikkeld of verbeterd, werden in SPSS opgenomen. Daarnaast heeft SPSS de ontwikkeling doorgemaakt van grote *mainframes,* opgesteld in speciaal hiervoor ingerichte rekencentra, tot *laptops* met de afmetingen van een schrijfblok. De vroegere SPSS-gebruikers moesten een stapeltje met opdrachten klaarmaken en inleveren, eerst in de vorm van ponskaarten bij de balie van een rekencentrum en later in de vorm van een computerbestand via een terminal. Na een tijdje wachten konden ze de geprinte uitvoer ophalen. Latere versies van SPSS werden steeds gebruiksvriendelijker. Naast vele kleine stapjes voorwaarts kende SPSS in deze ontwikkeling twee grote stappen. De eerste werd in 1983 gezet met het verschijnen van SPSS/PC, de eerste SPSS-versie voor personal computers. Negen jaar later, in 1992, volgde de tweede grote stap met de komst van SPSS voor Windows, een versie die nog veel gemakkelijker te gebruiken was. SPSS voor Windows was namelijk de eerste versie van SPSS waarbij kennis van de speciale SPSS-opdrachtentaal niet meer nodig is. SPSS is sindsdien nog gemakkelijker en sneller te (leren) gebruiken.

De gebruikers
Niet alleen SPSS, maar ook de gebruikers zijn sterk veranderd in de loop der tijd. De 'eerste' SPSS-gebruikers hebben veel van de gebruikte technieken nog met de hand moeten uitvoeren en kenden deze daarom van haver tot gort: veronderstellingen, berekeningswijze en betekenis van de uitkomsten waren voor hen gesneden koek. Dit geldt voor de hedendaagse SPSS-gebruikers veel minder, met het gevaar dat analyses verkeerd uitgevoerd of geïnterpreteerd worden. Dit gevaar wordt nog vergroot door het toegenomen bedieningsgemak: iedere leek kan een willekeurige geavanceerde analyse met SPSS uitvoeren, maar begrijpen is vers twee. SPSS kan hiervoor weinig bescherming bieden, dus een goede statistische basiskennis blijft voor het toepassen van veel technieken onontbeerlijk.

1.2 Het SPSS-gegevensanalysesysteem

SPSS biedt een zeer brede verzameling van statistische methoden. Het nadeel hiervan is dat veel gebruikers mogelijkheden worden geboden die ze nooit zullen gebruiken. Daarom is SPSS opgedeeld in modules. Naast de basismodule kunnen gebruikers kiezen uit een reeks speciale modules, die ook afzonderlijk worden verkocht.
In dit boek worden alleen de mogelijkheden van de basismodule besproken. Deze module bevat opdrachten om gegevensbestanden te maken en te bewerken, en daarnaast de meest gebruikte analysemethoden. Voorbeelden hiervan zijn frequentietabellen, kruistabellen, vele soorten grafieken, t-toetsen, variantie-, correlatie- en regressieanalyse (zie paragraaf 2.5 voor een uitgebreider overzicht).

Het bedrijf SPSS is in 2009 overgenomen door IBM. Toen is ook de naam van de software veranderd in 'IBM SPSS'. In dit boek wordt versie 24.0.0 van SPSS besproken. Deze versie is op de markt gebracht in 2016 en heet officieel 'IBM SPSS Statistics 24'. Voor het gemak spreken we in dit boek uitsluitend van SPSS. Voor oudere versies van SPSS is het boek 'Inleiding SPSS 22.0' meer geschikt.

2 Het gebruik van SPSS bij statistisch onderzoek

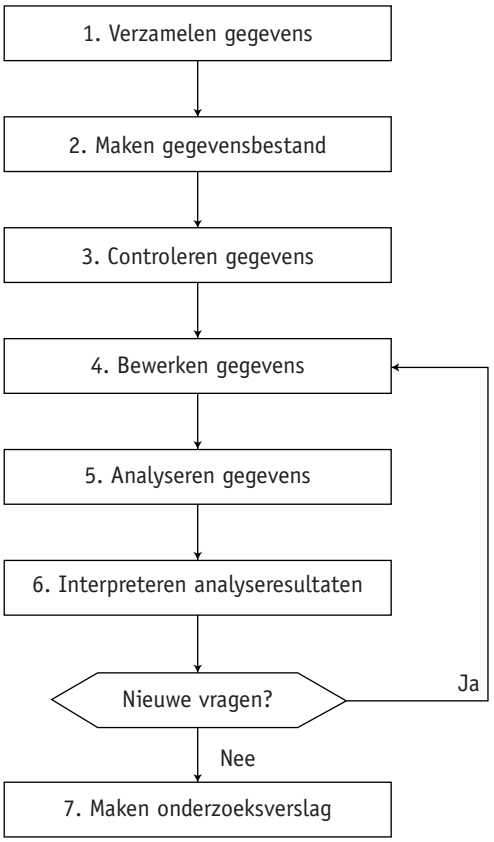
Inleiding

Het doel van dit hoofdstuk is om een overzicht te geven van de mogelijkheden van SPSS. We doen dit aan de hand van de verschillende fasen in het onderzoeksproces. De onderzoeksdraad wordt opgepakt bij het verzamelen van de gegevens en weer losgelaten nadat alle analyses zijn uitgevoerd en het onderzoeksverslag wordt geschreven. Voor elke fase wordt besproken welke ondersteuning SPSS kan bieden. Tevens worden de namen van de te gebruiken opdrachten gegeven. Doordat ook telkens de paragrafen worden genoemd waarin deze opdrachten worden behandeld, is dit hoofdstuk ook tijdens een onderzoek goed te gebruiken als referentiepunt.
De zeven fasen die in het onderzoeksproces worden onderscheiden, worden in paragraaf 2.1 kort beschreven. In de volgende zes paragrafen komt telkens een fase aan de orde. Het maken van het gegevensbestand wordt beschreven in paragraaf 2.2, daarna volgt het controleren van de gegevens (zie paragraaf 2.3). De volgende stap, het bewerken van de gegevens, is nodig als een analyse niet met alle originele gegevens moet worden uitgevoerd (paragraaf 2.4). Het analyseren van de gegevens is de *raison d'être* van SPSS, paragraaf 2.5 is daarom ook de langste paragraaf. Na het analyseren volgt het interpreteren van de analyseresultaten, zie paragraaf 2.6. In de laatste fase wordt een onderzoeksverslag gemaakt (paragraaf 2.7).

2.1 Het onderzoeksproces

Hoewel vele wetenschappers faseringen hebben bedacht om het proces van statistisch onderzoek beter te structureren, zien de meeste studies er in hoofdlijnen hetzelfde uit. Aan het begin van het onderzoek wordt een probleemstelling geformuleerd (een algemene vraagstelling) waaruit een aantal toetsbare hypothesen worden afgeleid (concrete vraagstellingen). Om het toetsen van de hypothesen mogelijk te maken, worden gegevens verzameld, waarna de hypothesen met behulp van statistische analyses worden getoetst. Het onderzoeksproces eindigt met het opstellen van het onderzoeksverslag waarin conclusies en aanbevelingen hun plaats vinden. Het gedeelte van het onderzoeksproces waarbij SPSS wordt gebruikt, bestaat uit zeven fasen (zie figuur 2.1). Elke fase wordt kort besproken.

1. Het verzamelen van gegevens – Er zijn veel verschillende manieren om gegevens te verzamelen, zoals via vragenlijsten, invulformulieren of meetapparatuur (bijvoorbeeld scanning). Deze eerste fase is alleen nodig als speciaal ten behoeve van het onderzoek gegevens moeten worden verzameld (primaire gegevens). Als gebruik wordt gemaakt van al verzamelde gegevens (secundaire gegevens) kan deze stap uiteraard worden overgeslagen.



Figuur 2.1 De fasen in het onderzoeksproces waarbij SPSS voor ondersteuning kan zorgen.

2. Het maken van een gegevensbestand – Na het verzamelen van de gegevens is de volgende stap het maken van een gegevensbestand dat SPSS kan analyseren. In deze fase worden variabelen gedefinieerd en de verzamelde gegevens gecodeerd, ingetypt en opgeslagen in een bestand. Het is ook mogelijk dat de gegevens zich al in een bestand bevinden, in dat geval kunnen de gegevens direct in SPSS worden ingelezen.

3. Het controleren van de gegevens – De volgende stap is het controleren van de gegevens. Pas als blijkt dat de gegevens foutloos ingevoerd en ingelezen zijn, kan het eigenlijke analysewerk beginnen. Als de ingelezen gegevens fouten bevatten, dan vindt terugkoppeling naar de vorige fase plaats: de foutieve gegevens worden gewijzigd, waarna de gegevens opnieuw worden gecontroleerd.

4. Het bewerken van de gegevens – Vaak is het nodig om, voordat de analyses worden uitgevoerd, de gegevens eerst te bewerken. Voorbeelden van bewerkingen zijn het maken van klasse-indelingen en het selecteren van een groep waarnemingen (bijvoorbeeld alleen mannen of alleen mensen ouder dan zestig jaar). Deze fase is facultatief; als u alle ori-

ginele gegevens in een analyse wilt gebruiken, hoeven de gegevens ook niet bewerkt te worden.

5. Het analyseren van de gegevens – Pas de vijfde fase bestaat uit het daadwerkelijk analyseren van de gegevens. In deze fase moet voor het beantwoorden van elke onderzoeksvraag de meest geschikte analysemethode worden gekozen. Deze keuze is afhankelijk van het doel van de analyse en de mate waarin de gegevens voldoen aan de veronderstellingen van een analysemethode. Voor het uitvoeren van de analyse geeft u alle benodigde specificaties op, zoals de namen van variabelen en de gewenste kengetallen.

6. Het interpreteren van de analyseresultaten – Het resultaat van een analyse is een tabel of grafiek met allerlei kengetallen zoals gemiddelden, overschrijdingskansen en correlatiecoëfficiënten. Deze informatie moet worden geïnterpreteerd om vast te stellen of de analyse een bevredigend antwoord op een onderzoeksvraag heeft geleverd. Vaak leidt interpretatie van de analyseresultaten weer tot nieuwe vragen, zodat een volgende bewerkings- en analyseronde begint.

7. Het maken van een onderzoeksverslag – Als de gegevens uitputtend zijn geanalyseerd, wordt een onderzoeksverslag opgesteld. Dit rapport bevat onder meer de interpretatie van de uitkomsten tezamen met de relevante SPSS-uitvoer (grafieken en tabellen). Voor een deel zal de SPSS-uitvoer in de lopende tekst worden opgenomen, de meer gedetailleerde uitkomsten krijgen vaak een plaats in bijlagen.

In het vervolg van dit hoofdstuk wordt elke fase, met uitzondering van de eerste, in een afzonderlijke paragraaf besproken. Voor elke fase wordt aangegeven welke ondersteuning SPSS u te bieden heeft.

2.2 Het maken van het gegevensbestand

Na het verzamelen van de gegevens moeten deze worden opgenomen in een bestand dat SPSS kan analyseren. Dit wordt het *gegevensbestand* genoemd. De verzamelde gegevens representeren kenmerken, meningen of aspecten van personen, bedrijven, huizen, productieseries, subsidies, enzovoort. SPSS gebruikt de term *waarneming* (in het Engels 'case') voor het object of subject waar iets van wordt gemeten. De gemeten kenmerken worden aangeduid als *variabelen*. Het SPSS-gegevensbestand is een matrix waarin de waarnemingen in de rijen staan en de variabelen in de kolommen.
Het maken van een gegevensbestand bestaat uit twee stappen, namelijk het omzetten van de metingen in variabelen en het invoeren van de gegevens in het gegevensbestand. Beide stappen zullen we kort bespreken. Het is ook mogelijk dat de gegevens zich al in een bestand bevinden, bijvoorbeeld een Excel-bestand. De mogelijkheden van SPSS voor het inlezen van andere bestanden worden aan het einde van deze paragraaf besproken.

Van metingen tot variabelen
Omdat gegevensbestanden variabelen bevatten, moet u de gemeten kenmerken 'vertalen' naar variabelen en voor elke variabele vaststellen welke waarden hiervoor mogen worden opgegeven. Bij een enquête betekent dit dat elke vraag moet worden omgezet in een of

meer variabelen en dat voor elk antwoord een code moet worden bepaald (bijvoorbeeld eens zijn met een stelling is code 1 en oneens code 0, zie paragraaf 3.1). De omzetting van metingen in variabelen wordt in een overzicht vastgelegd, dit wordt het *codeboek* genoemd. Het codeboek bevat onder meer de naam en omschrijving van elke variabele en de gebruikte codering. Het wordt gebruikt voor het definiëren van de variabelen en het coderen van de metingen (zie paragraaf 3.2). Na het invoeren van alle variabelendefinities kunt u in SPSS een codeboek opvragen, zie paragraaf 7.2.
Het *opgeven van de variabelen* gebeurt bij SPSS in het variabelenblad van de Data Editor. Dit is een spreadsheet waarin u de variabelen en hun kenmerken kunt invoeren. In feite geeft u aan SPSS de inhoud van het codeboek op (zie paragraaf 4.4).
Het codeboek wordt ook gebruikt voor het omzetten van de metingen in waarden van een variabele. Dit wordt het *coderen van de metingen* genoemd. Bij een enquête stelt u voor elk antwoord dat iemand heeft gegeven vast welke waarde hiervoor in het gegevensbestand moet worden ingevoerd.

Het invoeren van de gegevens
De volgende stap is het vullen van het gegevensbestand met gegevens. De gemakkelijkste en veiligste manier om gegevens in te voeren is door gebruik te maken van *de gegevensverzamelingmodule SPSS Data Collection Author*. Als u niet over Author beschikt, kunt u de gegevens ook intypen in het gegevensblad (van de Data Editor). Het gegevensblad is een spreadsheet waarbij de rijen worden gevormd door waarnemingen en de kolommen door variabelen. Het invoeren van de gegevens in het gegevensblad wordt besproken in de paragrafen 3.3 en 4.5. Paragraaf 7.3 behandelt een aantal nuttige functies van de Data Editor voor het invoegen en verplaatsen van variabelen en waarnemingen.

Het inlezen van gegevens uit een ander pakket
Soms zijn de gegevens al in een softwarepakket vastgelegd, in dat geval moet SPSS het gegevensbestand van dat pakket inlezen. SPSS kan bestanden van verschillende andere programma's lezen, zoals databasepakketten (bijvoorbeeld Access) en spreadsheets (bijvoorbeeld Excel). Daarnaast kan SPSS ASCII-bestanden lezen en uiteraard ook gegevensbestanden die met andere versies van SPSS zijn gemaakt, zie paragraaf 7.5 tot en met 7.7.

2.3 Het controleren van de gegevens

Nadat de gegevens zijn ingevoerd (of ingelezen), moet worden gecontroleerd of hierbij geen fouten zijn gemaakt. Als u de gegevens via het gegevensblad hebt ingevoerd, moet u de gegevens handmatig controleren. Vaak is dit een tijdrovende stap, maar hiermee voorkomt u wel dat later analyses moeten worden overgedaan omdat het gegevensbestand nog fouten bevat. U kunt de gegevens controleren door in de Data Editor een voor een de cellen af te lopen. Een andere mogelijkheid is om de inhoud van dit venster uit te printen. Dit overzicht kunt u dan vergelijken met de bron waaruit de gegevens zijn ingevoerd, bijvoorbeeld de ingevulde vragenlijsten.
Verder is het raadzaam om, nog voor het begin van de echte analyses, van alle variabelen een frequentietabel op te vragen (met de opdracht Frequencies, zie de paragrafen 4.7 en 11.1). De tabellen maken duidelijk of voor een variabele niet-bestaande codes zijn ingevoerd (bijvoorbeeld een 2 bij een variabele met de codering 0 en 1). Daarnaast geven de frequentietabellen u alvast een goed inzicht in het gegevensmateriaal.

Als het gegevensbestand nog fouten bevat, kent de Data Editor een aantal handige opdrachten om de foutieve gegevens snel in de spreadsheet op te sporen (zie paragraaf 7.4).

2.4 Het bewerken van de gegevens

De vierde fase in het onderzoeksproces is het bewerken van de gegevens. Deze stap is nodig als u bij een analyse niet alle originele gegevens nodig hebt. Als u klasse-indelingen wilt gebruiken, alleen een specifieke groep wilt analyseren of gegevensbestanden wilt samenvoegen, dan volgt eerst de stap van het bewerken van de gegevens en daarna pas de analyse van de gegevens.
De bewerking heeft betrekking op een van de volgende drie elementen:

- *Variabelen,* zoals het maken van klasse-indelingen of het berekenen van nieuwe variabelen.
- *Waarnemingen,* zoals het selecteren van een groep waarnemingen of het sorteren van de waarnemingen.
- *Het hele gegevensbestand,* zoals het samenvoegen of kantelen van gegevensbestanden.

Elke soort bewerkingen wordt in een afzonderlijke subparagraaf besproken, tussen haakjes staat steeds de betreffende paragraaf in deel II.

2.4.1 Het bewerken van variabelen

Met deze categorie bewerkingsopdrachten kunt u:

1. Berekeningen met variabelen uitvoeren.
2. Tellen hoe vaak een bepaalde waarde bij verschillende variabelen voorkomt.
3. Gegevens van vorige of volgende waarnemingen gebruiken
4. Klasse-indelingen maken.
5. Rangnummers bepalen.
6. Tekstvariabelen omzetten in numerieke variabelen.
7. Voor tijdreeksgegevens ontbrekende waarden laten invullen.

1. Berekeningen met variabelen uitvoeren: Compute Variable (8.1) – Met de opdracht Compute Variable wordt een nieuwe variabele gemaakt die het resultaat is van een berekening met een of meerdere bestaande variabelen. In de berekening kunt u functies gebruiken, onder andere voor afronden, worteltrekken en logaritmen. De opdracht kan ook voor alleen een specifieke groep waarnemingen worden uitgevoerd.
Als u weet hoeveel iemand heeft betaald voor baanhuur, tenniskleding en een tennisracket, kunt u met Compute Variable een nieuwe variabele tennisuitgaven maken, zijnde de som van de andere drie variabelen.

2. Tellen hoe vaak een bepaalde waarde voorkomt: Count (8.2) – Met de opdracht Count wordt een nieuwe variabele gemaakt die weergeeft hoe vaak een bepaalde waarde bij een aantal variabelen voorkomt. De opdracht kan ook voor alleen een specifieke groep waarnemingen worden uitgevoerd.
Stel dat u in vijf stellingen het belang van bepaalde aspecten van het milieu hebt benadrukt en voor elke stelling hebt gevraagd of respondenten het met die stelling eens zijn.

Dit levert u dus vijf antwoorden (vijf variabelen) per persoon op. Met Count kunt u een nieuwe variabele Milieu maken die aangeeft met hoeveel stellingen iemand het eens is (Milieu heeft dus waarden tussen nul en vijf).

3. Gegevens van vorige of volgende waarnemingen gebruiken: Shift Values (8.3) – SPSS gebruikt bij berekeningen of analyses alleen gegevens van dezelfde waarneming. Soms is dit niet handig, bijvoorbeeld als een actie in de ene periode invloed heeft in een volgende periode. Voor de analyse hebt u dan zowel gegevens van de huidige waarneming als van een voorgaande (of volgende) waarneming nodig. Met Shift Values verschuift u gegevens van een vorige (of volgende) waarneming naar de huidige waarneming.
Stel dat u de reclame-uitgaven in de vorige maand wilt relateren aan de verkopen in de huidige maand. Met de opdracht Shift Values maakt u dan eerst een nieuwe variabele die de reclame-uitgaven in de vorige maand weergeeft. Daarna kunt u de correlatie berekenen tussen deze nieuwe variabele en de verkopen in de huidige maand.

4. Klasse-indelingen maken: Visual Binning en Recode (8.4) – Klasse-indelingen worden vaak gebruikt om een continue variabele met veel verschillende waarden in een frequentie- of kruistabel op te kunnen nemen. Door de variabele te hercoderen worden de vele verschillende waarden gereduceerd tot enkele klassen.
Als u de leeftijd in jaren hebt gemeten en een frequentietabel wilt maken, dan moeten eerst leeftijdsgroepen worden gevormd met de opdracht Visual Binning (of Recode), bijvoorbeeld jonger dan 20 jaar, tussen 20 en 40 jaar, en ouder dan 40 jaar.

5. Tekstvariabelen omzetten in numerieke variabelen: Automatic Recode (8.5) – De opdracht Automatic Recode zet een tekstvariabele automatisch om in een numerieke variabele.
Stel dat een aantal mensen gevraagd is naar hun favoriete tennisspeler en in het gegevensbestand de spelersnamen voluit zijn opgenomen, bijvoorbeeld 'Agassi', 'Federer' en 'Roddick'. Met de opdracht Automatic Recode kunt u deze variabele dan omzetten in een numerieke variabele met de waarden 1, 2 en 3. SPSS zorgt er automatisch voor dat 'Agassi' de omschrijving van code 1 wordt, enzovoort.

6. Rangnummers bepalen: Rank Cases (8.6) – Met de opdracht Rank Cases maakt u een variabele die de rangorde van de waarnemingen op basis van een of meer variabelen weergeeft.
Als u van honderd personen de leeftijd weet, dan kunt u met de opdracht Rank Cases een nieuwe variabele maken die voor de jongste persoon de waarde 1 heeft en voor de oudste de waarde 100 (de rangnummers kunnen ook andersom worden toegekend).

7. Voor tijdreeksgegevens ontbrekende waarden laten invullen: Replace Missing Values (8.7) – Met de opdracht Replace Missing Values kunt u voor tijdreeksgegevens de waarschijnlijke waarde van ontbrekende waarden laten berekenen en deze hiervoor laten invullen.
Stel dat van een kind gedurende een jaar elke maand de lengte wordt gemeten en dat de waarde voor de maand mei ontbreekt. Door voor de maand mei het gemiddelde van de maanden april en juni in te vullen, krijgt u voor mei een waarde die doorgaans acceptabel zal zijn.

2.4.2 Het bewerken van waarnemingen

Met de bewerkingsopdrachten voor waarnemingen kunt u:
1. Een opsplitsing in groepen opgeven voor identieke analyses per groep.
2. Waarnemingen selecteren.
3. Waarnemingen een verschillend gewicht toekennen.
4. Waarnemingen sorteren.

1. Een opsplitsing in groepen opgeven voor identieke analyses per groep: Split File (9.1) – Met de opdracht Split File kunt u een bepaalde groepsindeling opgeven, waarna SPSS elke volgende analyse automatisch voor elke groep afzonderlijk uitvoert.
Stel dat u voor de groep mannen en de groep vrouwen dezelfde kruistabellen, staafdiagrammen en correlatieanalyses wenst. U geeft dan eerst met de opdracht Split File de groepsindeling op en als u daarna de analyses uitvoert, wordt elke analyse voor elke groep afzonderlijk uitgevoerd.

2. Waarnemingen selecteren: Select Cases (9.2) – De opdracht Select Cases selecteert een bepaalde groep waarnemingen waarna de volgende analyses alleen voor deze groep waarnemingen worden uitgevoerd. U kunt waarnemingen selecteren op basis van een bepaalde voorwaarde, het toeval of het waarnemingnummer.
Als het geslacht van elke respondent bekend is, kunnen met Select Cases alle vrouwen worden geselecteerd. Als het gegevensbestand een groot aantal namen en adressen bevat, kunt u met Select Cases hieruit een aselecte steekproef trekken. Ook als u een analyse met alleen de eerste vijftig respondenten wilt uitvoeren, kunt u dit met Select Cases opgeven.

3. Waarnemingen een verschillend gewicht toekennen: Weight Cases (9.3) – SPSS kent elke waarneming een gelijk gewicht toe, maar met de opdracht Weight Cases kunt u hierin verandering aanbrengen. Weight Cases is ook zeer handig wanneer u niet beschikt over de oorspronkelijke waarnemingen, maar alleen over de verdeling daarvan (zie het voorbeeld in paragraaf 9.3).
Stel dat in het onderzoek mannen ten opzichte van vrouwen oververtegenwoordigd zijn. Dit betekent dat verhoudingsgewijs er in de steekproef meer mannen zijn dan in de hele populatie. Om de verhouding ten opzichte van vrouwen te herstellen, zou u mannen een lager gewicht kunnen geven.

4. Waarnemingen sorteren: Sort Cases (9.4) – Normaal gesproken staan de waarnemingen in het gegevensbestand in de volgorde waarin u ze hebt ingevoerd. Met de opdracht Sort Cases kunt u sorteren en de waarnemingen dus in een andere volgorde zetten. Sort Cases is handig in samenhang met de opdracht om te selecteren op basis van het waarnemingnummer (zie Select Cases).
Als u een analyse alleen met de jongste vijftig respondenten wilt uitvoeren, dan sorteert u eerst de waarnemingen op basis van de leeftijd van de respondent en daarna selecteert u de eerste vijftig waarnemingen.

2.4.3 Het bewerken van het hele gegevensbestand

In tegenstelling tot de andere bewerkingsopdrachten maken de bewerkingsopdrachten die betrekking hebben op het hele gegevensbestand een *nieuw gegevensbestand*. Met deze categorie opdrachten kunt u:
1. Gegevensbestanden samenvoegen.
2. Waarnemingen samenvoegen (aggregeren).
3. Het gegevensbestand herstructureren (kantelen).

1. Gegevensbestanden samenvoegen: Merge Files (10.1) – Soms staan de gegevens verdeeld over verschillende gegevensbestanden. Met de opdracht Merge Files kunt u deze bestanden samenvoegen door aan een gegevensbestand waarnemingen of variabelen toe te voegen.
Stel dat twee personen elk bij een eigen PC de antwoorden op een vragenlijst invoeren. Er ontstaan dan twee gegevensbestanden met dezelfde variabelen maar met andere respondenten. In dit geval moeten de waarnemingen uit het ene bestand worden toegevoegd aan het andere bestand.
Stel dat u mensen op twee momenten vragen hebt gesteld over politieke partijen (voor de verkiezingen en na de verkiezingen) en dat hun antwoorden in twee verschillende gegevensbestanden zijn vastgelegd (een bestand 'voor' en een bestand 'na' de verkiezingen). Beide bestanden bevatten dezelfde waarnemingen maar verschillende variabelen. Nu moeten de variabelen uit het ene bestand worden toegevoegd aan het andere bestand.

2. Waarnemingen samenvoegen: Aggregate (10.2) – Met de opdracht Aggregate kunt u het detailniveau van uw analyses veranderen door waarnemingen samen te voegen. Voor elke variabele kunt u opgeven hoe SPSS die moet aggregeren (bijvoorbeeld door de waarnemingen op te tellen, het gemiddelde te bepalen of de waarde van de eerste waarneming in de groep te nemen).
Stel dat in een winkel voor elk merk wasmachine per dag is geregistreerd: het aantal verkochte apparaten en de die dag geldende prijs. U kunt dan aggregeren naar weekcijfers door voor elke week het aantal per dag verkochte wasmachines bij elkaar op te tellen en de weekprijs te berekenen als het gemiddelde van de dagprijzen.

3. Het gegevensbestand herstructureren: Restructure Data Wizard (10.3) – Het gegevensbestand is een matrix bestaande uit rijen en kolommen. In de rijen horen de waarnemingen en in de kolommen de variabelen. In de praktijk is dat niet altijd het geval en voor die situaties kent SPSS de Restructure Data Wizard.
Stel dat het gegevensbestand de verkopen per maand bevat van een aantal soorten jam (zoals abrikozen-, aardbeien- en kersenjam). Om te toetsen of er gemiddeld per maand meer abrikozen- dan kersenjam wordt verkocht, moeten de jamsoorten de variabelen vormen en de maanden de waarnemingen. Om te toetsen of er gemiddeld in januari meer jam wordt verkocht dan in juni, moeten de waarnemingen en variabelen van plaats verwisselen. De maanden worden dan de variabelen en de jamsoorten de waarnemingen.

2.5 Het analyseren van de gegevens

De vijfde stap in het onderzoeksproces is het uitvoeren van de statistische analyses. In het vorige hoofdstuk is aangegeven dat SPSS bestaat uit een basismodule en een aantal uitbreidingsmodules. In dit boek worden alleen analysetechnieken beschreven uit de basismodule. Andere modules bevatten een keur aan meer geavanceerde technieken.
De analyses in de basismodule die in dit boek worden beschreven zijn in de volgende zes groepen te verdelen:

1. Het beschrijven van een variabele.
2. Het beschrijven van groepen waarnemingen.
3. Het toetsen van verschillen tussen onafhankelijke groepen.
4. Het toetsen van verschillen tussen gerelateerde groepen.
5. Het bepalen van de samenhang tussen twee variabelen.
6. Het verklaren van een variabele door een of meer andere variabelen.

Elke groep wordt in een afzonderlijke subparagraaf besproken. Een schematisch overzicht van de uit te voeren analyses afgezet tegen het meetniveau van de variabelen vindt u aan de binnenzijde van het kaft achter in dit boek.
Omdat elke analysemethode eisen stelt aan het meetniveau van de variabelen, behandelen we in de volgende subparagraaf eerst de vier verschillende meetniveaus. Bent u al bekend met de betekenis van meetniveaus, dan kunt u deze bespreking overslaan.

2.5.1 Het meetniveau van een variabele

In het vervolg van deze paragraaf wordt een groot aantal analyses beschreven. Voor elk analysedoel, bijvoorbeeld het beschrijven van groepen of het bepalen van de samenhang tussen variabelen, bestaan meerdere analysemethoden. U moet dus steeds bepalen welke methode voor u de meest geschikte is. Een belangrijk criterium bij deze keuze is het *meetniveau* van de variabelen (ook wel schalingsniveau genoemd). De verschillende methoden stellen namelijk eisen aan het meetniveau van de variabelen. Alleen als uw variabelen voldoen aan het vereiste meetniveau, mag u de betreffende methode toepassen. Overigens maken veel analysemethoden ook nog andere veronderstellingen, zodat alleen het voldoen aan het meetniveau niet voldoende is. In deel II, waarin elke analysemethode wordt beschreven, wordt uitgebreider ingegaan op de andere veronderstellingen.
In een onderzoek analyseert u de waargenomen eigenschappen van bepaalde verschijnselen. De mate waarin deze eigenschappen gemeten kunnen worden, kan verschillen. In oplopende meetbaarheid worden de volgende vier typen schalen onderscheiden: nominaal, ordinaal, interval en ratio. Elk van deze vier meetniveaus wordt kort beschreven en toegelicht met enkele voorbeelden.

1. Nominale schaal – Bij een nominale schaal krijgen de eigenschappen een willekeurige waarde. Een nominale schaal wordt gebruikt als een eigenschap eigenlijk niet meetbaar is, maar alleen identificeerbaar. Voorbeelden zijn kleur haar, merk tennisracket, geslacht en bloedgroep. Voor elke eigenschap worden categorieën onderscheiden en aan elke categorie wordt een getal toegekend. Dat getal dient als etiket en niet om de omvang van een eigenschap weer te geven. De haarkleuren rood, blond en zwart kunnen worden weergegeven met de cijfers 1, 2, en 3, maar elke andere volgorde van waarden was ook bruikbaar

geweest. Een hogere haarkleur betekent niet dat die persoon donkerder haar, langer haar of meer haar heeft.
In het bovenstaande voorbeeld zijn getallen toegekend aan een eigenschap, het komt ook voor dat de eigenschap al zelf in getallen wordt uitgedrukt maar het meetniveau toch nominaal is. Voorbeelden van dit soort eigenschappen zijn telefoonnummers, de rugnummers in een voetbalelftal en de nummers van bankrekeningen. Ook dan geeft de hoogte van een nummer geen informatie over de waarde van een eigenschap.

2. Ordinale schaal – Bij een ordinale schaal krijgen de eigenschappen niet meer een willekeurige waarde, maar geeft de schaal een rangorde weer. Een hogere waarde op de schaal geeft aan dat een eigenschap groter is, of langer, hoger, belangrijker, beter of aantrekkelijker. Echter, er wordt slechts gemeten of een eigenschap meer of minder voorkomt en niet in welke mate de eigenschap meer of minder voorkomt.
Zo kunt u met behulp van een 5-puntsschaal meten hoe lekker een snoepje wordt gevonden, waarbij 1 is 'beslist niet lekker' en 5 'heel erg lekker'. Een snoepje met score 4 is lekkerder dan een snoepje met score 2, maar we kunnen niet zeggen dat het ene snoepje twee keer zo lekker is als het andere. Ook hoeft het verschil tussen score 1 en 2 niet even groot te zijn als het verschil tussen score 4 en 5. Een ander voorbeeld zijn de rangen in het leger. Een hogere rang is belangrijker, maar we kunnen niet zeggen hoeveel belangrijker. Ordinale schalen worden vaak gebruikt om meningen (percepties of attitudes) van mensen te meten.
De waarde van een eigenschap geeft dus informatie over die eigenschap. In principe bent u vrij te kiezen welk uiterste van een eigenschap de hoogste score en de laagste score krijgt. Als u op een 5-puntsschaal meet in hoeverre men het eens is met een bepaalde stelling, kunt u voor 'volledig eens' in principe zowel de score 1 als de score 5 gebruiken. Het is echter gebruikelijk *oplopende schalen* te gebruiken. Dus als de schaal weergeeft 'de mate van eens zijn' met een stelling, moet een lagere score aangeven dat men het minder eens is met de stelling. We kiezen dan voor 1 is 'volledig oneens' en 5 'volledig eens'.

3. Intervalschaal – De intervalschaal geeft ook een rangordening weer, maar nu heeft het verschil tussen de waarden wel een betekenis. Eén eenheid verschil verwijst altijd naar hetzelfde verschil. Het nulpunt van de schaal is echter arbitrair. Een voorbeeld van een intervalschaal is de temperatuur gemeten in graden Celsius. Het temperatuurverschil tussen vijf en tien graden Celsius is even groot als het verschil tussen dertig en vijfendertig graden. Het nulpunt is echter arbitrair gekozen, dat wil zeggen dat nul graden Celsius niet de laagst mogelijke temperatuur is. Daarom kunnen we ook niet zeggen dat twintig graden Celsius twee keer zo warm is als tien graden. Dit is gemakkelijk in te zien als we ons realiseren dat de temperatuur even goed in graden Fahrenheit kan worden gemeten als in graden Celsius.
Een ander voorbeeld betreft kalenderjaren. De periode tussen het jaar 800 en het jaar 1000 duurde even lang als die tussen 1800 en 2000. Het nulpunt is echter ook nu arbitrair, het jaar 2000 kwam dus niet twee keer zo laat als het jaar 1000.

4. Ratioschaal – De ratioschaal heeft alle eigenschappen van de intervalschaal èn er is sprake van een natuurlijk nulpunt. Dit betekent dat de verschillen tussen de getallen op de schaal een reële en gelijke betekenis hebben, evenals de verhouding tussen twee getallen. Voorbeelden van dit soort maatstaven zijn lengte, gewicht, afstand, geldbedragen en

aantallen. Stel dat plaats A op 20 kilometer afstand van plaats X ligt, en plaats B op 100 kilometer. We kunnen dan niet alleen zeggen dat B 80 kilometer verder van X verwijderd is dan A, maar ook dat B vijf keer zo ver verwijderd is van X als A.

Veel methoden die gebruikt kunnen worden voor ratiovariabelen mogen ook worden toegepast voor intervalvariabelen. Vandaar dat beide groepen ook wel met één term worden aangeduid als *continue variabelen*.

2.5.2 Het beschrijven van een variabele

Opdrachten om inzicht te krijgen in de verdeling van een variabele zijn doorgaans de eerste analyse-opdrachten in een reeks analyses. Met dergelijke analyses wordt namelijk inzicht verkregen in de samenstelling van het gegevensmateriaal, waarvan bij volgende analyses gebruik kan worden gemaakt. U kunt een variabele beschrijven door het opvragen van:
1. De frequentie van elke waarde.
2. De centrale tendentie.
3. Kengetallen voor de spreiding.
4. De overeenkomst met een theoretische verdeling.
5. De trendmatige ontwikkeling.

1. De frequentie van elke waarde – Het bepalen van de frequentie van elke waarde betekent niets anders dan het turven van elke waarde ('een rechte telling'). De frequenties kunnen worden getoond in een tabel of grafiek. In het eerste geval wordt gesproken van een *frequentietabel*. SPSS kent hiervoor de opdracht Frequencies (4.7 en 11.1). Aan het meetniveau van de variabele wordt geen eis gesteld, wel is het zo dat voor continue variabelen met een groot aantal verschillende waarden de tabel al snel heel groot en onoverzichtelijk wordt. Dan is het handig de variabele eerst in klassen in te delen (met de opdracht Visual Binning, 8.4). Bij een groot aantal soortgelijke nominale of ordinale variabelen kan *meervoudige-antwoordenanalyse* een duidelijk overzicht van de frequenties geven (zie hoofdstuk 14).
Voor een grafische weergave van frequenties worden *staafdiagrammen* gebruikt bij nominale en ordinale variabelen (5.2 en 12.2). Bij continue variabelen worden de waarden per staaf gegroepeerd in een *histogram* (12.7) of samengevat in de vorm van een *stamdiagram* (11.3) of een *boxdiagram* (12.9).

2. De centrale tendentie – De centrale tendentie verwijst naar het gemiddelde van een groep waarnemingen. De volgende drie maatstaven worden gehanteerd:
1. De *modus*: de waarde die het meest voorkomt. Vooral gebruikt bij nominale variabelen.
2. De *mediaan*: de waarde van de middelste waarneming. Vooral gebruikt bij ordinale variabelen.
3. Het *rekenkundig gemiddelde*: de som van alle waarnemingen gedeeld door het aantal waarnemingen. Vooral gebruikt bij interval- en ratiovariabelen. Omdat het rekenkundig gemiddelde de meest gebruikte centrale tendentiemaatstaf is, wordt deze maatstaf vaak kortweg aangeduid als 'het gemiddelde'.

Al deze drie centrale tendentiemaatstaven worden berekend met de opdracht Frequencies (drukknop Statistics, 11.1.1).
U kunt ook toetsen of het steekproefgemiddelde overeenkomt met een ander gemiddelde (bijvoorbeeld het nationale gemiddelde of een norm):

- Voor dichotome variabelen (variabelen met slechts twee waarden) wordt hiervoor de *binomiale toets* gebruikt (18.2.1).
- Voor ordinale variabelen kan de mediaan worden getoetst met de *Wilcoxon signed-rank toets* (18.2.4).
- Voor interval- en ratiovariabelen kan een *t-toets voor één steekproef* worden uitgevoerd met de opdracht One-Sample T Test (15.2).

3. Kengetallen voor de spreiding – U kunt een groot aantal kengetallen opvragen die inzicht geven in de spreiding van een variabele. Voorbeelden zijn voor variabelen op minimaal ordinaal niveau *percentielen,* het *minimum* en het *maximum*. Voor interval- en ratiovariabelen bestaan veel meer spreidingsmaatstaven, zoals het *bereik* (het verschil tussen minimum en maximum), de *standaarddeviatie* en de *variantie*.
SPSS kent drie opdrachten met vergelijkbare mogelijkheden voor het berekenen van deze spreidingsmaatstaven: de drukknop Statistics bij de opdrachten Frequencies (11.1) en Explore (11.3) en de drukknop Options bij de opdracht Descriptives (11.2).

4. De overeenkomst met een theoretische verdeling – Er zijn verschillende manieren om inzicht te krijgen in de mate waarin de waargenomen verdeling van een variabele overeenkomt met een theoretische verdeling, bijvoorbeeld de normale verdeling. U kunt dit doen door het opvragen van kengetallen, grafieken of het uitvoeren van een toets.
Kengetallen die inzicht geven in de overeenkomst met een normale verdeling zijn de scheefheid en de kurtosis. De *scheefheid* geeft weer in welke mate een verdeling symmetrisch is. Bij de normale verdeling zijn de beide helften aan weerszijden van het gemiddelde (de modus) elkaars spiegelbeeld en de verdeling is dus volledig symmetrisch. Bij een positieve scheefheid heeft de verdeling 'een staart' naar rechts en zijn er relatief gezien meer waarnemingen groter dan de modus. De *welving of kurtosis* is een maatstaf voor de relatieve platheid van de verdeling ten opzichte van de normale verdeling. De kurtosis is positiever als de top hoger ligt en negatiever als de top lager ligt. De normale verdeling heeft een kurtosis van nul. Beide kengetallen vereisen ten minste een ordinale schaal en kunnen zowel met de opdracht Frequencies, Descriptives als Explore worden opgevraagd (zie hoofdstuk 11).
Verschillende grafieken geven een goed inzicht in de overeenkomsten tussen een waargenomen verdeling en een normale verdeling. Voorbeelden zijn *boxdiagrammen* (12.9) en *histogrammen* met een normale curve (12.7). Meer in het bijzonder geldt dit voor een *normaal-kwantielplot* en een *afwijkingengrafiek* (11.3.2).
U kunt ook toetsen in hoeverre de waargenomen verdeling overeenkomt met een theoretische verdeling, bijvoorbeeld de normale verdeling. Dit kan met de *Kolmogorov-Smirnov toets* (18.2.3). De variabele dient dan ten minste ordinaal geschaald te zijn. De verdeling van een nominale variabele kunt u toetsen aan een theoretische verdeling met behulp van de *chi-kwadraattoets* (18.2.2). De verdeling waaraan moet worden getoetst, kunt u zelf opgeven. U kunt bijvoorbeeld onderzoeken of er evenveel mannen als vrouwen zijn ondervraagd door een theoretische verdeling van 1:1 op te geven. Het is ook mogelijk dat uit een andere bron, bijvoorbeeld het CBS, bekend is dat de onderzochte populatie twee

keer zoveel mannen bevat als vrouwen. Met de chi-kwadraattoets kunt u dan nagaan of de steekproef representatief is ten aanzien van het kenmerk geslacht. Dit is het geval als ook in de steekproef de verhouding tussen mannen en vrouwen (ongeveer) 2:1 is.

5. De trendmatige ontwikkeling – Als de waarnemingen elkaar in de tijd opeenvolgende metingen weergeven, kan het zinvol zijn de trendmatige ontwikkeling te onderzoeken. Dit kan met behulp van *staafdiagrammen* (5.2 en 12.2), *lijngrafieken* (12.3) of *oppervlaktegrafieken* (12.4).

2.5.3 Het beschrijven van groepen waarnemingen

Voor het beschrijven van groepen waarnemingen kan, afhankelijk van het meetniveau van de variabele, een van de volgende analyse-opdrachten worden gebruikt:
- Een *kruistabel* bevat informatie over het aantal waarnemingen per groep en wordt opgesteld met de opdracht Crosstabs (13). Kruistabellen worden vooral gebruikt bij nominale of ordinale variabelen.
- Voor interval- of ratiovariabelen berekent de opdracht Means (15.1) *voor elke groep kengetallen* zoals het gemiddelde, de spreiding en het aantal waarnemingen.

Grafieken voor het beschrijven van groepen zijn:
- *Gegroepeerde en gestapelde staafdiagrammen* (12.2) voor nominale en ordinale variabelen.
- *Lijn- en oppervlaktegrafieken* (12.3 en 12.4) worden gebruikt voor het analyseren van de trendmatige ontwikkeling van verschillende groepen.

Een andere mogelijkheid is om met de opdracht *Split File* (9.1) de waarnemingen in verschillende groepen te splitsen en daarna analyses voor het beschrijven van één variabele uit te voeren (zoals het opvragen van een frequentietabel, kengetallen of een grafiek). Deze analyse wordt dan voor elke onderscheiden groep afzonderlijk uitgevoerd.

2.5.4 Het toetsen van verschillen tussen onafhankelijke groepen

Om te toetsen of onafhankelijke groepen significant van elkaar verschillen, kent SPSS een aantal toetsen. Het is gebruikelijk om een onderscheid te maken tussen situaties met twee groepen en met meer dan twee groepen.
Bij twee onafhankelijke groepen kunt u een van de volgende toetsen uitvoeren:
- De *chi-kwadraattoets* voor nominale variabelen, met de drukknop Statistics bij de opdracht Crosstabs (13.3).
- De *Mann-Whitney toets* (18.3.1) of de *Kolmogorov-Smirnov toets* (18.3.2) voor ordinale variabelen.
- De *t-toets* voor interval- en ratiovariabelen met de opdracht Independent Samples T Test (15.3).

Bij meer dan twee onafhankelijke groepen worden de volgende toetsen gebruikt:
- De *chi-kwadraattoets* voor nominale variabelen, met de drukknop Statistics bij de opdracht Crosstabs (13.3).
- De *mediaantoets* (18.3.3) en de *Kruskal-Wallis toets* (18.3.4) voor ordinale variabelen.

- *Variantieanalyse* (de F-toets) voor interval- en ratiovariabelen. SPSS kent twee opdrachten voor variantieanalyse: de opdracht One-Way ANOVA (16.2) als u op basis van één variabele groepen onderscheidt en de opdracht GLM Univariate (16.3) als de groepen op basis van twee of meer variabelen worden onderscheiden.

2.5.5 Het toetsen van verschillen tussen gerelateerde groepen

Soms zijn de waarnemingen in de onderscheiden groepen niet onafhankelijk van elkaar maar aan elkaar gerelateerd. Dit is het geval als u niet zomaar mannen en vrouwen ondervraagt, maar echtparen. Of als u dezelfde personen voor en na een bepaalde gebeurtenis ondervraagt. Om te toetsen of gerelateerde groepen significant van elkaar verschillen, kent SPSS de volgende toetsen:

- De *tekentoets* (18.4.1) en de *Wilcoxon matched-pair signed-rank-toets* (18.4.2) voor ordinale variabelen bij twee gerelateerde groepen.
- De *Friedman-toets* (18.4.3) voor ordinale variabelen bij meer dan twee gerelateerde groepen.
- De *gepaarde t-toets* voor interval- en ratiovariabelen met de opdracht Paired-Samples T Test (15.4).

2.5.6 Het bepalen van de samenhang tussen twee variabelen

Om inzicht te krijgen in de mogelijke samenhang tussen twee variabelen kunt u opvragen:

- Een *kruistabel* voor nominale en ordinale variabelen met de opdracht Crosstabs (13).
- Een *spreidingsdiagram* waarbij de ene variabele op de horizontale as en de andere op de verticale as staat, terwijl de waarnemingen in de vorm van een puntenwolk worden afgebeeld. Spreidingsdiagrammen worden gemaakt met de opdracht Graphs-Chart Builder (6.2 en 6.5) en vereisen minimaal ordinale variabelen. Spreidingsdiagrammen geven niet alleen goed inzicht in de mate waarin twee variabelen samenhangen maar ook in de vorm van de samenhang (bijvoorbeeld rechtlijnig of kromlijnig).

Of en de mate waarin twee variabelen met elkaar samenhangen, kan ook in de vorm van een kengetal worden weergeven, namelijk:

- De *chi-kwadraattoets* om voor nominale variabelen te bepalen of twee variabelen al dan niet onafhankelijk van elkaar zijn (13.3).
- De *phi-coëfficiënt* om voor nominale variabelen de mate van samenhang te bepalen (13.4).
- De *Spearman correlatiecoëfficiënt* en de *Kendall's tau* voor ordinale variabelen (beide 13.4 en 17.1).
- De *Pearson correlatiecoëfficiënt* voor interval- en ratiovariabelen. SPSS kent hiervoor de opdracht Bivariate Correlations (17.1) en als u wilt corrigeren voor de invloed van een derde variabele de opdracht Partial Correlations (17.2).

2.5.7 Het verklaren van een variabele door een of meer andere variabelen

Om te toetsen of er een lineair verband bestaat tussen een afhankelijke variabele en een of meer onafhankelijke variabelen, wordt *regressieanalyse* gebruikt. Regressieanalyse levert een vergelijking op waarmee de afhankelijke variabele numeriek kan wor-

den verklaard. Belangrijke veronderstellingen zijn dat er een oorzakelijk verband moet bestaan tussen de afhankelijke variabele en de onafhankelijke variabelen (causaliteit), en de variabelen interval- of ratiogeschaald zijn. Het verband tussen de variabelen moet lineair (rechtlijnig) zijn, of rechtlijnig gemaakt kunnen worden. Regressieanalyse wordt uitgevoerd met de opdracht Linear Regression (17.3). Als het verband kromlijnig is en via een wiskundige formule kan worden beschreven, kunt u de opdracht Curve Estimation gebruiken voor het uitvoeren van de regressieanalyse (zie paragraaf 17.4).

2.6 Het interpreteren van de analyseresultaten

De uitvoer van een analyse bestaat uit een tabel of grafiek met aantallen, percentages, gemiddelden, overschrijdingskansen, coëfficiënten, enzovoort. Dit zijn slechts getallen, louter cijfertjes. Om een antwoord op een onderzoeksvraag te krijgen, moet u deze getallen interpreteren. Over het algemeen biedt SPSS hiervoor slechts beperkte hulp: interpreteren is nog altijd mensenwerk.
SPSS vergemakkelijkt het interpreteren wel op een aantal manieren. Zo bewaart het programma automatisch alle uitvoer in de Viewer. Hierin kunt u de uitvoer bekijken, zonodig van commentaar voorzien, wijzigen en bewaren. SPSS markeert interessante uitkomsten door bijvoorbeeld sterretjes achter een significante correlatiecoëfficiënt te zetten of opvallende cellen te arceren. Ook berekent SPSS steeds de overschrijdingskans (p-waarde) van een toets (zoals chi-kwadraat-, t-, of F-toets). U kunt SPSS ook regels opgeven om cellen in een tabel met opvallende uitkomsten, bijvoorbeeld een zeer hoog of laag percentage, te laten markeren, zie paragraaf 6.8. Een enkele keer kunt u SPSS al bij het uitvoeren van een analyse interpretatiemaatstaven laten toepassen. Een voorbeeld is regressieanalyse, waar kan worden opgegeven aan welke eisen een variabele moet voldoen om opgenomen te worden in de vergelijking. SPSS stelt dan zelf vast of een variabele voldoet aan de door u opgegeven voorwaarden.
Om het interpreteren van de analyseresultaten te vereenvoudigen, is van elke techniek in deel II een voorbeeld opgenomen. Hierbij wordt niet alleen de exacte opdracht weergegeven en besproken, maar ook wordt de resulterende SPSS-uitvoer afgedrukt en uitgebreid besproken en geïnterpreteerd.

2.7 Het maken van het onderzoeksverslag

Het onderzoeksverslag wordt vrijwel altijd in een tekstverwerker opgesteld en niet binnen SPSS. Toch bevat de Viewer wel de mogelijkheden om een eenvoudig rapport te maken. In de Viewer kunt u uitvoer selecteren, van commentaar voorzien, bewaren en printen. Daarnaast kunt u SPSS uitvoer in diverse formaten opslaan, onder meer als pdf-, Word- of Powerpoint-bestand of als web report, voor het publiceren van resultaten op internet, zie paragraaf 6.9.
Als u het onderzoeksrapport in een tekstverwerker opstelt, kunt u delen van de uitvoer in de Viewer markeren en kopiëren naar een tekstverwerker. Op deze wijze kunt u zowel tekst, tabellen als grafieken in het onderzoeksverslag opnemen (zie paragraaf 6.10 voor een voorbeeld).

3 Van gegevensbron tot gegevensbestand

Inleiding

SPSS wordt veel gebruikt voor het analyseren van gegevens die verzameld zijn met behulp van vragenformulieren. Voorbeelden van vragenformulieren zijn enquêtes, garantieformulieren, bezoekverslagen en standaardbeoordelingen. In dit hoofdstuk zal worden beschreven hoe de overstap van een vragenformulier naar een SPSS-gegevensbestand kan worden gemaakt. De eerste stap is het 'vertalen' van de vragen op het formulier naar variabelen die een plaats krijgen in het gegevensbestand en het bepalen van de codering (zie paragraaf 3.1).
Om de overgang van gegevensbron naar gegevensbestand zo goed mogelijk te laten verlopen, wordt een overzicht opgesteld waarin staat welke variabelen bij welke vragen horen en welke codering is aangebracht. Dit overzicht wordt het codeboek genoemd (zie paragraaf 3.2). De laatste stap in de overgang van gegevensbron naar gegevensbestand is het daadwerkelijk invoeren van de gegevens. Dit gebeurt met een invoerformulier ontworpen in Author of met de Data Editor. Paragraaf 3.3 geeft enkele tips om dit zo snel mogelijk en met de kleinste kans op fouten te doen.
In dit boek zullen we gebruikmaken van een fictief onderzoek onder tennisspelers. De in dit onderzoek gebruikte vragenlijst en codeboek staan in paragraaf 3.4.

3.1 Van vragen naar variabelen

Voordat u aan de slag kunt met SPSS, moeten de via vragenformulieren verzamelde gegevens een plaats krijgen in een SPSS-gegevensbestand. Omdat SPSS met variabelen werkt, is een *vertaling* van de vragen naar variabelen noodzakelijk. Voor elke vraag moet worden bepaald hoeveel en welke variabelen met welke codering hiervoor worden gebruikt. Deze paragraaf bespreekt drie mogelijke situaties:

- gesloten vragen,
- open vragen, en
- vragen met meerdere antwoorden per respondent.

Elke soort vragen wordt in een afzonderlijke subparagraaf besproken.

3.1.1 Gesloten vragen

Gesloten vragen zijn vragen waarop maar een beperkt aantal, van tevoren bekende antwoorden mogelijk is. Vaak is het aantal antwoorden beperkt, eenvoudigweg omdat er niet meer antwoorden mogelijk zijn, bijvoorbeeld alleen maar 'ja' en 'nee' (en eventueel ook nog 'weet niet'). Gesloten vragen zijn ook vragen naar continue eigenschappen die in een discrete vraagstelling worden verwoord, bijvoorbeeld een vraag naar iemands inkomen

waarbij vooraf inkomensgroepen zijn aangegeven. Bij gesloten vragen is meestal zowel de vertaling als het bepalen van de codering niet moeilijk: voor één vraag wordt één variabele gemaakt, terwijl de codering al in de vraag opgesloten ligt.
Een voorbeeld uit een onderzoek onder tennisspelers kan als illustratie dienen (zie tabel 3.1). De vraag naar nieuwe tennisschoenen is een gesloten vraag waarbij elke respondent slechts één antwoord kan geven. We kunnen daarom volstaan met één variabele. Voor de codering zijn er twee mogelijkheden:

1. *Een tekstvariabele met de codering J = Ja en N = Nee*. Dit heeft als voordeel dat u de antwoorden niet eerst hoeft te coderen, ze kunnen direct van het vragenformulier worden overgetypt. Tekstvariabelen kennen echter ook belangrijke nadelen, waardoor ze over het algemeen zijn af te raden (zie paragraaf 3.2.3).
2. *Een numerieke variabele met de codering 0 = Nee en 1 = Ja*. Omdat oplopende schalen altijd de voorkeur verdienen (zie paragraaf 2.5.1), wordt de omgekeerde codering (0 = Ja en 1 = Nee) niet gebruikt.

Tabel 3.1 Een voorbeeld van een gesloten vraag waarbij slechts één antwoord per respondent mogelijk is.

Heeft u in het afgelopen jaar nieuwe tennisschoenen gekocht?
☐ Ja
☐ Nee

3.1.2 Open vragen

Open vragen zijn vragen waarbij van tevoren de antwoordmogelijkheden niet zijn aangegeven. We maken hierbij onderscheid tussen kwantitatieve en kwalitatieve vragen. Een *kwantitatieve vraag* meet een eigenschap in de vorm van een getal. De vertaling van vragen naar variabelen is dan heel eenvoudig: voor elke vraag één variabele waarbij het ingevulde aantal de waarde is van de variabele. Dit geldt voor verschijnselen als geldbedragen (aantal euro's), afstanden (aantal kilometers), of temperatuur (aantal graden Celsius). In deze gevallen is codering niet nodig, zie het eerste voorbeeld in tabel 3.2, waar de leeftijd wordt uitgedrukt in het aantal jaren.

Tabel 3.2 Twee voorbeelden van een open vraag.

Kwantitatieve vraag
Wat is uw leeftijd? _____ (aantal jaren)
Kwalitatieve vraag
Wie is uw favoriete tennisspeler? _____ (naam)

Naast kwantitatieve vragen staan *kwalitatieve vragen*, die een niet-numerieke eigenschap van een verschijnsel meten. Voorbeelden zijn kleuren, steden, merken en titels van boeken. De tweede vraag in tabel 3.2 is een kwalitatieve vraag. Omdat elke respondent slechts één tennisspeler mag noemen, kunnen we ook nu volstaan met één variabele. Voor de codering zijn er twee mogelijkheden:

1. *Een numerieke variabele* waarbij de codering loopt van 1 tot en met het aantal genoemde spelers. Het nadeel van deze oplossing is dat coderen dan alleen achteraf kan, na het afnemen van alle enquêtes. Bij schriftelijk onderzoek met een klein aantal respondenten is dat geen bezwaar, maar bij veel respondenten of bij telefonische en persoonlijke interviews kan dit echter niet. Eventueel zou u de codering ook gedurende het invoeren van de gegevens kunnen bepalen, namelijk door elke nieuwe spelersnaam een volgend cijfer toe te kennen. Deze methode is echter rommelig met een grote kans op fouten en is alleen mogelijk als één persoon alle gegevens invoert.
2. *Een tekstvariabele* bestaande uit de hele spelersnaam of een gedeelte hiervan, bijvoorbeeld alleen de eerste drie letters. Het voordeel van deze oplossing is dat niet met coderen hoeft te worden gewacht totdat alle enquêtes zijn afgenomen. Tekstvariabelen kennen echter ook belangrijke nadelen (zie paragraaf 3.2.3). Bij het gebruik van een gedeelte van een naam als codering kunnen namen met dezelfde beginletters voor problemen zorgen.

Gesloten vragen of open vragen? – Als u de keuze hebt tussen een gesloten of een open vraag gaat de voorkeur uit naar een open kwantitatieve vraag omdat de variabele dan op het hoogst mogelijke meetniveau wordt gemeten (interval of ratio). Veel verschijnselen zijn echter niet in getallen uit te drukken en in dat geval verdienen gesloten vragen de voorkeur. De reden is dat het coderen van open vragen lastiger is. Een belangrijke stimulans voor het toenemende gebruik van gesloten vragen is de opkomst van computerondersteund enquêteren, waarbij het noodzakelijk is van tevoren de codering te bepalen.
Toch worden ook open kwalitatieve vragen vaak gebruikt. Soms omdat van tevoren alle mogelijke antwoorden niet bekend zijn en soms omdat het onwenselijk is de antwoorden te noemen. Een voorbeeld van dit laatste is een onderzoek naar de bekendheid van een merk wasmiddel: bij een gesloten vraag hoeft de respondent een merk alleen maar te herkennen en bij een open vraag moet de respondent de merknamen zelf bedenken. Gesloten vragen leiden dan ook tot een aanzienlijk hogere merkbekendheid.

Gedeeltelijk open vragen – Als van tevoren niet alle mogelijke antwoorden bekend zijn wordt, om het coderen te vergemakkelijken, nogal eens gebruikgemaakt van *gedeeltelijk open vragen*. Tabel 3.3 bevat de vraag naar de favoriete tennisspeler in de vorm van een gedeeltelijk open vraag. De van tevoren bekende antwoorden worden genoemd (gesloten vraag) en tevens wordt de respondent de mogelijkheid geboden een ander antwoord in te vullen (open vraag).

Tabel 3.3 Een voorbeeld van een gedeeltelijk open vraag.

Wie is uw favoriete tennisspeler?
☐ Novak Djokovic
☐ Andy Murray
☐ Rafael Nadal
☐ Iemand anders, namelijk: _____

Ook in dit voorbeeld zijn er twee mogelijkheden voor de codering. Als het onderzoek schriftelijk wordt afgenomen en het aantal vragenlijsten beperkt is, kunnen de bij

'Iemand anders' ingevulde namen worden geïnventariseerd en opgenomen worden als code 4, 5, enzovoort. In dat geval wordt de vraag vertaald in één variabele. Gebruikelijker is om twee variabelen te maken, een numerieke variabele voor het gesloten gedeelte van de vraag en een tekstvariabele (of eventueel een numerieke variabele) voor het open gedeelte van de vraag. In dit geval wordt de vraag dus vertaald in twee variabelen. Eventueel kunnen achteraf de beide variabelen worden samengevoegd met de opdracht Compute en de zelden genoemde namen samengevoegd tot de categorie 'Overige' met de opdracht Recode (zie hoofdstuk 8).

3.1.3 Vragen met meerdere antwoorden per respondent

De vertaling van vragen naar variabelen wordt moeilijker als een respondent meerdere antwoorden op één vraag kan geven. In dit geval wordt gesproken van *meervoudige-antwoorden*. In hoofdstuk 14 wordt besproken hoe u deze antwoorden kunt analyseren, in deze paragraaf richten we ons op de vertaling naar variabelen.

Tabel 3.4 Een voorbeeld van een gesloten vraag waarbij meerdere antwoorden per respondent mogelijk zijn.

Welke spelsoort beoefent u?
☐ Enkelspel
☐ Dubbelspel

Tabel 3.4 bevat een voorbeeld van een gesloten vraag waarbij meerdere antwoorden per respondent mogelijk zijn. Het is mogelijk dat iemand zowel enkel- als dubbelspel speelt. De vraag kan op twee manieren worden vertaald naar variabelen:

1. *Eén variabele met alle mogelijke antwoordcombinaties*. In dit geval zijn vier antwoordcombinaties mogelijk: geen van beide, alleen enkelspel, alleen dubbelspel, en enkel- en dubbelspel. We kunnen dan één variabele maken met de codes 1 tot en met 4 voor de vier verschillende antwoordcombinaties.
2. *Een afzonderlijke variabele voor elke deelvraag*. De vraag kan worden opgevat als bestaande uit twee deelvragen, namelijk 'Speelt u enkelspel?' en 'Speelt u dubbelspel?'. Beide deelvragen kennen Ja en Nee als mogelijke antwoorden. We maken dan twee variabelen met elk als codering 0 = Nee en 1 = Ja (of N en J, zie paragraaf 3.1.1).

De eerste oplossing lijkt voor de hand te liggen omdat dan kan worden volstaan met één variabele en dus het minst ingetypt hoeft te worden. Het is echter lang niet altijd de handigste oplossing. U kunt voor de eerste oplossing kiezen als de vier onderscheiden groepen ook als vier verschillende groepen worden gezien. Als u echter de groep enkelspelers wilt analyseren en het niet uitmaakt of men naast enkelspel ook nog dubbelspel speelt, verdient de tweede oplossing de voorkeur. Dit geldt ook als er meerdere deelvragen zijn, omdat dan het aantal antwoordcombinaties snel oploopt. Bij twee deelvragen zijn er vier antwoordcombinaties, bij drie deelvragen acht en bij vier deelvragen al zestien. Dergelijke aantallen antwoordcombinaties zijn niet meer hanteerbaar. Het coderen van de antwoorden is gemakkelijker (en dus sneller en met minder kans op fouten) als u voor elke deelvraag een afzonderlijke variabele gebruikt. U kunt dan nog steeds de antwoorden

als een groep analyseren door de betreffende variabelen te definëren als een multipele response set (zie hoofdstuk 14). Overigens, als later mocht blijken dat u niet de juiste keuze hebt gemaakt, dan kunt u met de opdrachten Compute en Recode (zie hoofdstuk 8) de ene codering altijd nog omzetten in de andere.

3.2 Het codeboek

Het codeboek bevat een overzicht van welke variabelen bij welke vragen horen en met welke codering zij in het gegevensbestand staan. Codeboeken zijn nuttig zowel bij het invoeren van gegevens als tijdens het analyseren om te kunnen opzoeken naar welke vraag een variabele verwijst ("hoe luidt de exacte formulering van die vraag?") of wat de betekenis van een code is. Een codeboek bestaat uit vier kolommen (zie tabel 3.5). In deze paragraaf worden alle kolommen, met uitzondering van de eerste, in een afzonderlijke subparagraaf besproken. Paragraaf 3.4 bevat een voorbeeld van een codeboek. U kunt ook in SPSS een codeboek opvragen, zie paragraaf 7.2.

Tabel 3.5 De structuur van een codeboek.

Kolom	**Inhoud**
1	Nummer van de vraag op het vragenformulier
2	Naam van de variabele
3	Omschrijving van de variabele
4	Gebruikte codering

3.2.1 De naam van de variabele

Elke variabele moet een naam krijgen (zie paragraaf 7.1 voor een overzicht van de restricties die voor variabelenamen gelden). Vaak valt het echter niet mee om een goede en korte naam te bedenken en daarom is een omschrijving van de naam onontbeerlijk.
Het is aan te raden als eerste variabele het nummer van de respondent op te nemen. Als de waarnemingen (enquêteformulieren) niet genummerd zijn, nummer ze dan zelf doorlopend vanaf 1. Het voordeel hiervan is dat een typefout of een onwaarschijnlijk lijkende waarde aan de hand van dit nummer snel is terug te vinden in de originele vragenlijsten. Overigens nummert SPSS ook zelf uw waarnemingen door ze een rijnummer toe te kennen. Dit nummer, dat in de kaderrand van het gegevensblad staat, verandert echter als u waarnemingen sorteert of selecteert. Daarom kunt u beter het nummer van de respondent in een afzonderlijke variabele vastleggen.

3.2.2 De omschrijving van de variabele

De derde kolom in het codeboek bevat een (korte) omschrijving van de variabele. SPSS noemt de omschrijving van de naam van de variabele een *variable label*. Zeker als u veel variabelen hebt of variabelen die moeilijk in één kort kernwoord zijn samen te vatten, zullen de gekozen variabelenamen al snel cryptisch overkomen. Bovendien hebben afkortingen de neiging voor de geestelijke vader (of moeder) veel duidelijker te zijn dan voor anderen. Vandaar dat een aparte kolom voor de omschrijving wordt opgenomen.

3.2.3 De codering

De wijze waarop de antwoorden zijn gecodeerd, staat in de vierde kolom van het codeboek. Voor de vraag of iemand in het afgelopen jaar nieuwe tennisschoenen heeft gekocht, wordt ingevuld: 0 = Nee en 1 = Ja (of J = Ja en N = Nee). Bij variabelen die ongecodeerd worden opgenomen (zoals leeftijd, prijs en aantal werknemers) worden de eenheden vermeld, bijvoorbeeld de leeftijd in jaren en het aantal werknemers in honderdtallen. Soms wordt een speciale code gebruikt als een vraag niet (correct) is ingevuld, bijvoorbeeld een 0 (nul) als de leeftijd niet is ingevuld. In dat geval moet u bij het definiëren van variabelen aan SPSS opgeven dat deze code verwijst naar een ontbrekende waarde (zie verder paragraaf 6.4).
Het is af te raden om al bij het coderen groepsindelingen te maken; voer zoveel mogelijk de originele antwoorden in. Groepsindelingen hebben een aantal nadelen. Behalve dat andere klasse-indelingen niet meer mogelijk zijn, wordt door het indelen in groepen het meetniveau van de variabele verlaagd. De leeftijd in jaren is een ratiovariabele, de leeftijd ingedeeld in klassen is ordinaal. Door het verlagen van het meetniveau zijn verschillende analyses niet meer mogelijk omdat dan niet meer wordt voldaan aan het vereiste meetniveau (zie paragraaf 2.5). Zo kan de leeftijd ingedeeld in klassen niet meer worden gebruikt voor onder meer correlatie- en regressieanalyse.

Tekstvariabelen versus numerieke variabelen – De gebruikte codering bepaalt ook het type van de variabele: numeriek of tekst. Over het algemeen is het gebruik van tekstvariabelen af te raden. Nadelen van tekstvariabelen zijn:

- *Het verschil tussen hoofd- en kleine letters.* SPSS maakt, voor wat betreft de waarde van tekstvariabelen, onderscheid tussen hoofdletters en kleine letters. Voor SPSS zijn j en J dus twee verschillende antwoorden. U dient daarom of alleen hoofdletters of alleen kleine letters te gebruiken.
- *Bij sommige opdrachten kunnen alleen numerieke variabelen worden gebruikt.* Een voorbeeld is het maken van een groepsindeling bij variantieanalyse (One-Way ANOVA). In dat geval dient u de tekstvariabele eerst met de opdracht Automatic Recode om te zetten in een numerieke variabele (zie paragraaf 8.5).
- *Het gebruik van apostroffen.* Bij verwijzingen naar de waarde van een tekstvariabele moeten apostroffen om de waarde staan, vergelijk NewShoe = 'J' met NewShoe = 1. Dit betekent extra typewerk en een grotere kans op fouten.
- *Minder tekens in te voeren.* Als u numerieke codes gebruikt, hoeft u over het algemeen minder in te typen dan als u voluit teksten moet invoeren. Dit betekent tijdwinst en minder kans op fouten.

Vanwege deze nadelen van tekstvariabelen worden doorgaans, waar mogelijk, numerieke variabelen gebruikt. Het eventuele nadeel van numerieke codes dat niet duidelijk is wat code 1 of code 2 betekent, wordt ondervangen door omschrijvingen toe te kennen aan de waarden van een variabele. SPSS noemt dit *value labels* (zie paragraaf 5.3).

3.3 Het intypen van de gegevens

De laatste stap in de overgang van gegevensbron naar gegevensbestand is het daadwerkelijk invoeren van de gegevens. Dit gebeurt in het gegevensblad (zie paragraaf 4.5) of

met een invulformulier of vragenlijst ontworpen in IBM SPSS Data Collection Author. Het intypen van gegevens is misschien niet zo'n boeiende bezigheid, het is wel de basis voor de kwaliteit van alle volgende analyses. Het intypen moet dus niet alleen zo snel mogelijk, maar vooral ook foutloos gebeuren.
Als u de gegevens vanaf ingevulde vragenformulieren intypt, is het verstandig eerst de codering op alle formulieren aan te brengen en dan pas met invoeren te beginnen. Tegelijkertijd een code bedenken en de code intypen leidt gemakkelijk tot fouten.
Trek ook ruimschoots de tijd uit voor het controleren van het gegevensbestand. Bedenk daarbij dat één invoerfout soms verkeerde waarden voor meerdere variabelen betekent. Een veel gemaakte fout is namelijk dat een variabele wordt overgeslagen waardoor ook de volgende variabelen een verkeerde waarde krijgen.

Tot slot zullen we nog een open deur intrappen die niet vaak genoeg kan worden ingetrapt: wanneer u een gegevensbestand hebt gemaakt, dient u meteen een kopie te maken. Niets is vervelender dan het opnieuw moeten intypen van een brij cijfertjes....

3.4 Het tennisonderzoek

In dit boek zullen we gebruikmaken van een fictief onderzoek onder tennisspelers. In dit onderzoek zijn de tien vragen gesteld die in tabel 3.6 zijn afgebeeld. In het volgende hoofdstuk zullen we een deel van de bijbehorende antwoorden in een gegevensbestand invoeren.
De vragenlijst is eenvoudig van opzet. Op alle vragen is per respondent slechts één antwoord mogelijk, zodat elke vraag wordt vertaald naar één variabele. Tabel 3.7 bevat het bijbehorende codeboek.
Zoals u ziet, begint het codeboek met een variabele voor het nummer van de respondent. Verder zijn twee vragen vertaald als tekstvariabelen (vraag 5 en vraag 8). Dit is met opzet gedaan omdat we deze vragenlijst ook in de volgende hoofdstukken gebruiken en we dan over tekstvariabelen willen kunnen beschikken. Zoals eerder gesteld verdient bij een echt onderzoek een numerieke codering doorgaans de voorkeur.

Tabel 3.6 De vragenlijst uit het tennisonderzoek.

1	Hoe vaak tennist u gemiddeld per maand? ... aantal keren
2	Welke spelsoort beoefent u? 1 Alleen enkelspel 2 Alleen dubbelspel 3 Zowel enkelspel als dubbelspel
3	Hoeveel heeft u in het afgelopen jaar uitgeven aan baanhuur? ... aantal euro's
4	Hoeveel heeft u in het afgelopen jaar uitgeven aan tenniskleding? ... aantal euro's
5	Heeft u in het afgelopen jaar nieuwe tennisschoenen gekocht? Ja / Nee
6	Zo ja: Hoe duur waren deze? 1 Minder dan 50 euro 2 Tussen 50 en 100 euro 3 Meer dan 100 euro
7	Bezoekt u tennistoernooien? 1 Zelden 2 Soms 3 Vaak
8	Wat is uw geslacht? Man / Vrouw
9	Wat is uw leeftijd? ... aantal jaren
10	Hoeveel bedraagt uw netto maandinkomen? ... aantal euro's

Tabel 3.7 Het codeboek voor de vragenlijst van het tennisonderzoek.

Vraagnr.	**Naam variabele**	**Omschrijving**	**Codering**
–	Respondentnr	Respondentnummer	Getal, 1..n
1	Vaaktennis	Hoe vaak men per maand tennist	Aantal keren
2	Enkeldubbel	Enkel- en/of dubbelspel	1 = Enkel 2 = Dubbel 3 = Enkel & dubbel
3	Baanhuur	Uitgaven aan baanhuur (vorig jaar)	Aantal euro's
4	Kleding	Uitgaven aan tenniskleding (vorig jaar)	Aantal euro's
5	Newshoe	Nieuwe schoenen gekocht?	J = Ja N = Nee
6	Shoeprijs	Prijscategorie schoenen	1 = < 50 2 = 50 – 100 3 = > 100
7	Bezoektoernooien	Bezoek tennistoernooien	1 = Zelden 2 = Soms 3 = Vaak
8	Geslacht		M = Man V = Vrouw
9	Leeftijd		Aantal jaren
10	Inkomen	Netto maandinkomen	Aantal euro's