# 10 Frequentieverdelingen van één variabele: tabellen en grafieken

**Dit hoofdstuk gaat over tabellen en grafieken die gemaakt worden om de frequentieverdeling van één variabele weer te geven. Maar we gaan eerst wat verder in op een onderwerp dat in hoofdstuk 1 werd aangesneden, namelijk het meetniveau van een variabele. Daarna bespreken we de spelregels om een tabel er goed te laten uitzien. Soms is het nodig om tabellen die de computeroutput vormen te vereenvoudigen; we geven de redenen hiervoor aan, en bespreken hoe we dit het best kunnen doen. Maar frequentieverdelingen kunnen, in plaats van als tabel, ook grafisch worden voorgesteld, in de vorm van een 'plaatje'. Een grafische voorstelling is niet altijd duidelijker dan een tabel, en zeker niet zo precies. Maar soms kan een grafiek in een oogwenk duidelijk maken hoe de zaak in elkaar steekt, en in een publicatie worden ze meestal 'plezieriger' gevonden dan tabellen. We maken kennis met verschillende soorten plaatjes, en we leren wanneer we de ene, en wanneer we de andere voorstelling gebruiken.**

## Meetniveau

In hoofdstuk 1 zagen we dat een datamatrix het einde van de fase van dataverzameling voorstelt, en dat dit tegelijk het begin is van de analysefase. Datamatrices kunnen echter vele vormen aannemen. Sommige hebben alleen maar acroniemen of losse woorden als celvullingen, andere alleen numerieke waarden. In hoofdstuk 1 merkten we op dat meten bestaat uit het toekennen van een getal aan elke categorie of waarde van een variabele, en daarmee aan de eenheden die in die categorie vallen. De onderzoeker moet er wel voor zorgen dat *elke* eenheid gemeten wordt, dat wil zeggen wordt toegewezen aan een categorie (meten moet uitputtend zijn) en dat elke eenheid wordt toegewezen aan één en slechts één categorie (meten moet exclusief zijn). Kwalitatieve data worden soms niet-numeriek gelaten, maar vaak (vooral als het onderzoeksproject zowel kwantitatieve als kwalitatieve variabelen betreft) worden de verbale labels vervangen door getallen. Dus als je een matrix met alleen getallen tegenkomt, wil dit helemaal niet zeggen dat alle variabelen kwantitatief zijn.

Het begrip meetniveau heeft te maken met *hoe* getallen aan categorieën worden toegekend. En deze toekenning is gebaseerd op hoe de categorieën zich tot elkaar verhouden. De hier gemaakte onderscheidingen zijn belangrijk omdat de manier waarop we met de toegekende getallen omgaan (de analyse en de wijze van representatie van de data), er sterk van afhankelijk is. Daarom beginnen we dit hoofdstuk met een korte uitwerking van de verschillende meetniveaus. We onderscheiden er vier.

### *1. Het nominale niveau*

Sommige variabelen dienen alleen maar om eenheden te classificeren in verschillende categorieën. We kunnen die waarden of categorieën niet op een of andere manier in een rangorde zetten. De variabele 'sekse' heeft slechts twee waarden: man en vrouw. De variabele 'bloedgroep' heeft er vier: O, AB, A en B. Bij 'middel van vervoer' kunnen we diverse waarden onderscheiden: fiets, trein, bus, auto en nog een paar. Merk op dat de onderzoeker soms tamelijk vrij is in het onderscheiden van categorieën: voor bepaalde doeleinden zou een onderscheid tussen openbaar vervoer en eigen vervoer al voldoende zijn. Bij de variabele 'politieke voorkeur' zou elke categorie een politieke partij kunnen zijn, maar soms kunnen we volstaan met een onderscheid tussen 'rechts' en 'links', of tussen godsdienstig gefundeerde partijen en partijen die dat niet zijn.
Het doet er niet toe welke getallen aan de waarden worden toegekend. Of we nu een 1 aan mannen en een 2 aan vrouwen toekennen, of andersom (of het getal 107 aan mannen en het getal 1432 aan vrouwen), doet er niet toe, omdat deze getallen niet worden gebruikt om ermee te rekenen. Het ligt daarom voor de hand om de eenvoudigste getallen te gebruiken. Als we bijvoorbeeld mensen vragen wat hun meest gebruikte vervoermiddel is, kunnen we de getallen hierna aan de categorieën toekennen.

| | |
|---|---|
| te voet: | 1 |
| fiets: | 2 |
| bus: | 3 |
| trein: | 4 |
| eigen auto (passagier): | 5 |
| eigen auto (rijdt zelf): | 6 |
| andere: | 7 |

Elke andere verzameling van getallen is goed, als ten minste aan elke categorie één en slechts één getal wordt toegekend. Een dergelijke classificerende inde-

ling is de meest primitieve manier van meten; we noemen dit *meten op nominaal niveau* en we spreken van *nominale variabelen*. Het doel van het toekennen van getallen is alleen maar het vergemakkelijken van computerverwerking van de data. Het zou volledige onzin zijn om bijvoorbeeld het gemiddelde te berekenen van een frequentieverdeling. Er bestaat niet zoiets als een gemiddeld vervoermiddel (zo is er ook geen 'gemiddelde sekse') bij een feestje. We kunnen echter de frequentieverdeling in een rapport opnemen en daarmee iets zeggen over hoe de eenheden gespreid zijn over de categorieën; of bepaalde categorieën leeg zijn, en dergelijke. Als de frequentieverdeling klaar is voor het rapport worden de toegekende codecijfers weer weggelaten.
Een nominale variabele is altijd *discreet*, dat wil zeggen dat er geen waarden denkbaar zijn *tussen* de gegeven waarden. Meten op nominaal niveau wordt ook wel kwalitatief meten genoemd, in contrast met kwantitatief meten.

### *2. Het ordinale niveau*

We kennen daarnaast ook variabelen waarvan de categorieën of waarden een vaste rangorde hebben. Een voorbeeld is opleiding: als we aan mensen vragen wat hun laatst voltooide opleiding is, waarbij we werken met waarden zoals basisschool, vmbo, havo, vwo, hbo en wo. Een rangen- of een schalenstelsel bij een bedrijf of bij de overheid is een ander voorbeeld. Ook nu is het zinloos om zoiets als een gemiddelde van een frequentieverdeling te berekenen. Wel kunnen we weer aangeven welke categorie het meest voorkomt, of hoe hoog iemands score is in verhouding tot de anderen. Zo is de Cito-score een ordinale variabele: de toets meet wat een kind in vergelijking met andere kinderen in acht jaar basisonderwijs geleerd heeft.

Een veelvoorkomende soort van ordinale variabelen gaat over subjectieve oordelen van mensen, zoals:
- de mate van instemming (helemaal mee eens; mee eens; niet mee eens; helemaal niet mee eens) met een of andere bewering of opinie;
- de meningen van mensen of een bepaalde politicus links is, dicht bij het centrum staat of rechts is;
- het antwoord van de patiënt op de vraag of een therapie geholpen heeft, geen effect had of een tegengesteld effect had;
- het oordeel van de leraar over een essay: schiet helemaal tekort, moet flink bijgewerkt worden, redelijk goed, of heel goed;
- het oordeel van een Artsen zonder Grenzen-dokter over of een slachtoffer van een aardbeving 'onmiddellijke hulp nodig heeft', binnen een dag ge-

holpen moet worden, binnen een week geholpen moet worden, of helemaal geen hulp nodig heeft (wat ook kan betekenen: is al dood);
– het jaarlijks inkomen zoals surveyrespondenten dat aangeven: 'onder de € 15.000', '€ 15.000 tot € 40.000', of 'meer dan € 40.000'.

Welke getallen moeten nu worden toegekend? De rangorde van de toe te wijzen getallen moet overeenkomen met de empirische en/of de gepercipieerde rangorde van de categorieën, maar is verder vrij. Ook hier kiezen we natuurlijk de meest eenvoudige reeks van getallen, bijvoorbeeld 1, 2, 3, 4 en 5, of 0, 1, 2, 3 en 4.

We spreken van *meting op ordinaal niveau*, en van *ordinale variabelen*. Ordinale variabelen kennen geen vaste meeteenheid. Het verschil tussen twee militaire rangen kun je niet vergelijken met het verschil tussen twee andere rangen. Je kunt niet zeggen dat het ene verschil tweemaal zo groot is als het andere. Je kunt alleen maar zeggen dat kapitein hoger is dan luitenant, en luitenant hoger dan sergeant. Ook hier is het berekenen van een gemiddelde onzin. Op basis van de frequentieverdeling kunnen we wel zeggen wat de meest voorkomende categorie is, en kunnen we iets zeggen over de spreiding over de betrokken categorieën of aangeven welk percentage van de steekproef boven een bepaalde rang zit. Normaliter hebben verschillende eenheden in het bestudeerde domein eenzelfde rangnummer (we spreken wel over 'knopen' als dat zo is). Als er knopen zijn, spreken we van een 'zwakke rangorde'. Soms worden echter *alle* eenheden ten opzichte van elkaar in rangorde gezet. Een voorbeeld verkrijg je als je alle leerlingen van een kleine schoolklas in rangorde van hun (niet-afgeronde) gemiddelde cijfer plaatst, of wanneer je een serie foto's van huizen door iemand laat beoordelen op aantrekkelijkheid. Als alle objecten of eenheden ten opzichte van elkaar geordend zijn, spreken we van een sterke of strikte rangorde.
Ordinale variabelen zijn meestal discreet; denk aan een rangenstelsel of aan opleidingscategorieën. In principe zijn er ook wel continue variabelen denkbaar; continu wil zeggen dat er tussen twee gegeven waarden altijd nog tussenliggende waarden denkbaar zijn.

Tot zover hebben we alleen voorbeelden genoemd van variabelen die 'van nature' uitgedrukt zijn in 'kwalitatieve' waarden of categorieën, waarbij gemakshalve tijdens het onderzoek elk van de categorieën van een getal wordt voorzien. Met de hierna te noemen variabelen ligt de situatie anders: de onderzoeker hoeft geen getallen toe te kennen, die zijn al in het werkelijke leven gegeven of ze kunnen echt 'gemeten' worden. Wel kan de onderzoeker de 'natuur-

lijke' getallen eventueel vervangen door een wat meer gebruikersvriendelijke reeks, bijvoorbeeld van afgeronde getallen.

### *3. Het intervalniveau*

Er bestaan variabelen die zodanig gemeten worden dat niet alleen de waarden in een vaste rangorde staan, maar waarbij er bovendien een vaste meeteenheid bestaat, waardoor de afstanden of 'intervallen' tussen de waarden *in elkaar* uitgedrukt kunnen worden. Temperatuur gemeten in graden Celsius of Fahrenheit is een voorbeeld van een *variabele op intervalniveau*. Een verschil tussen ochtend- en avondtemperatuur van 9 graden is half zo groot als een verschil van 18 graden. Er is echter niet een vaste nulwaarde: nul graden Celsius is maar een willekeurige afspraak. Dat maakt dat we niet kunnen zeggen: het is vandaag tweemaal zo warm als gisteren. Behalve temperatuur, energie en kalenderdata zijn er niet veel variabelen die gemeten worden op intervalniveau. In de psychologie wordt door sommige wetenschappers beweerd dat de meting van het IQ, uitgedrukt in standaardscores, een voorbeeld is.

### *4. Het rationiveau*

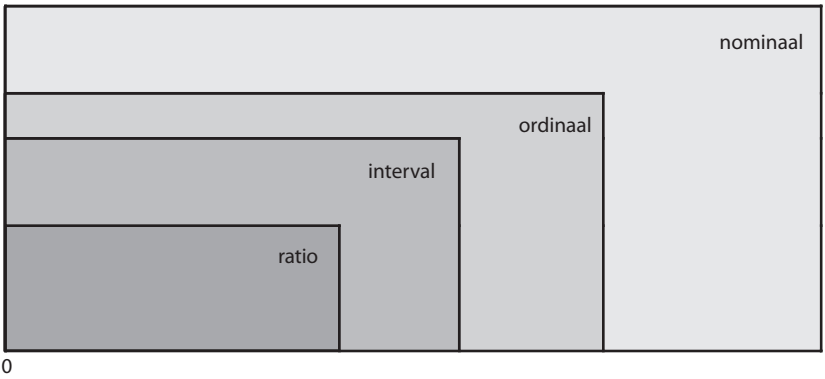
Op een nog hoger meetniveau zijn er variabelen waarvan de waarden niet alleen in rangorde staan en die een vaste meeteenheid kennen, maar die bovendien een natuurlijk nulpunt hebben. En omdat de waarde 'o' betekenis heeft, heeft de verhouding (Engels: *ratio*) tussen twee scores ook betekenis; daarom spreken we van het *rationiveau*. Bij een variabele zoals lengte is de bewering dat het ene potlood tweemaal zo lang is als het andere, zinvol. Andere voorbeelden vinden we heel veel in de natuurwetenschappen (lengte, gewicht, dichtheid, maar ook temperatuur gemeten in graden Kelvin). Een variabele zoals lengte wordt uitgedrukt in eenheden, bijvoorbeeld centimeters. Een onderzoeker die het oorspronkelijke getallenstelsel wil vervangen door een ander heeft niet zo veel keuze, omdat hij zowel rekening moet houden met de rangorde van de waarden, als met de vaste meeteenheid en met het vaste nulpunt. Om dezelfde verhoudingen te handhaven mag je de getallen alleen vermenigvuldigen met een constante: scores in centimeters kun je uiteraard zonder verlies aan informatie vervangen door scores in inches, en inkomens in euro's kun je evengoed in dollars uitdrukken. Andere veranderingen in de toegekende getallen – zoals een toevoeging van een constante – zijn niet mogelijk op rationiveau omdat ze de verhoudingen aantasten (bij meting op intervalniveau mag je daarentegen

zowel vermenigvuldigen als een constante bij alle scores optellen). Centraal staat de gedachte dat hoe hoger het meetniveau is, hoe 'exclusiever' de toekenning van bepaalde getallen aan de waarden en daarmee aan de eenheden.

In de sociale wetenschappen werken we met veel variabelen op dit meetniveau, zoals leeftijd, inkomen, rookfrequentie en aantal theaterbezoeken per jaar. En als we aan kenmerken van groepen, landen of organisaties denken, hebben we te maken met groepsomvang, aantal inwoners, ziekteverzuim, percentage van het budget aan onderwijs besteed en vele andere. De meeste van deze variabelen betreffen 'dimensieloze getallen', zoals aantallen, percentages en statistische waarschijnlijkheden. Zij zijn niet in een of andere eenheid uitgedrukt, maar ze zijn 'absoluut'. Je kunt deze getallen niet door andere getallen vervangen; geen enkele transformatie is toegestaan.

Variabelen op interval- en rationiveau zijn meestal continu (niet als de categorieën in absolute aantallen zijn uitgedrukt, zoals groepsgrootte of kindertal), dat wil zeggen dat tussen elke twee waarden altijd tussenliggende waarden denkbaar zijn, maar in de praktijk meten we op een of twee decimalen achter de komma en ronden we de resultaten af.
*Bij variabelen op interval- en rationiveau kunnen we met de toegekende getallen en frequenties echt rekenen, bijvoorbeeld door gemiddelde scores te berekenen* (in hoofdstuk 11 meer hierover).

Door sommige methodologen wordt meten op nominaal niveau aangeduid als 'categorisch meten', ook wel als 'kwalitatief meten'. Dit staat dan tegenover meten op interval- en op rationiveau, dat 'kwantitatief meten' wordt genoemd. Tot op zekere hoogte is deze eenvoudige tweedeling bruikbaar, maar een probleem vormt het meten op ordinaal niveau, dat er toch echt 'tussenin' ligt.
Zoals te zien is in figuur 10.1 zijn ordinale variabelen een deelverzameling van de nominale variabelen (ze 'maken niet alleen verschil', maar de categorieën staan ook in rangorde). De intervalniveauvariabelen zijn op zich weer een deelverzameling van de ordinale variabelen (de intervallen tussen categorieën zijn vergelijkbaar) en de rationiveauvariabelen zijn een deelverzameling van de intervalniveauvariabelen (ze kennen een natuurlijk nulpunt). Het plaatje laat zien dat elke variabele op zijn minst nominaal is ('een verschil maakt').




*Figuur 10.1 Relaties tussen meetniveaus*

Bedenk dat bepaalde manipulaties van de data kunnen resulteren in een lager meetniveau. Als we in een project ondernemingen indelen naar aantal werknemers, is deze variabele van nature op rationiveau gemeten, of nog liever op 'absoluut niveau'. Nu kan het zijn dat we niet geïnteresseerd zijn in een verfijnde indeling (de tabel zou in een onderzoeksrapport veel te groot worden), maar dat we tevreden zijn met een eenvoudige driedeling (een trichotomie, zoals dat in het jargon heet): 'minder dan tien werknemers', 'tien tot vijftig werknemers' en 'vijftig of meer werknemers'. Als we die vereenvoudiging toepassen, gaan we echter van het rationiveau naar het ordinale niveau. Merk op dat met een op deze manier vereenvoudigde frequentieverdeling we niet meer het precieze gemiddelde aantal werknemers over al die bedrijven kunnen berekenen. De redenering kan ook worden omgekeerd. Het is interessant om te zien dat in de loop van de eeuwen het meten van bepaalde variabelen op steeds hoger niveau kon plaatsvinden. 'When men recognised temperature only by sensation, when things were only "warmer" or "colder" than other things, temperature belonged only to the ordinal class of scales. It became an interval scale with the development of thermometry, and after thermodynamics had used the expansion ratio of gases to extrapolate to zero, it became a ratio scale.'[1]

Merk ook op dat het meetniveau van een variabele dus niet door de 'inhoud', het 'onderwerp' van die variabele wordt bepaald, maar door de wijze van meting. Een bekende ordinale variabele is 'het laatst gevolgde onderwijsniveau'

1 Stevens, S.S. (1951). Mathematics, measurement and psychophysics. In S.S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley. Ook te vinden in Churchman, C.W. & Ratoosh, P. (1959). *Measurement, definitions and theories*. New York: Wiley.

(lager onderwijs, voortgezet onderwijs of hbo/universiteit in de meeste landen). Maar er zijn technieken om opleiding te meten op rationiveau. Deze technieken zijn gebaseerd op het aantal jaren onderwijs dat met elke opleiding gemoeid is, en met de rangorde van de opleidingstypen. In de Verenigde Staten wordt opleiding eenvoudigweg gemeten via het 'aantal gevolgde schooljaren' (deze variabele wordt ook gebruikt in de GSS91-file;[2] zie Appendix B). Op deze manier wordt opleiding direct op rationiveau gemeten (zie ook de oefeningen 10.1 f, g en h). Maar er zitten wel wat haken en ogen aan: 'een jaar opleiding' heeft niet dezelfde inhoud en hetzelfde gewicht over het hele spectrum.

Je moet natuurlijk het meetniveau van een variabele vaststellen voordat je eventueel allerlei statistische bewerkingen op de frequentieverdeling loslaat, en voordat je in de rapportage tabellen en grafische voorstellingen maakt.

### Spelregels

Onderzoeksresultaten worden vaak gepresenteerd in de vorm van tabellen. We moeten dus een tabel kunnen *lezen*, en we moeten tabellen kunnen *maken* om gegevens van een (eigen) onderzoek weer te geven en die aspecten te laten zien welke van belang zijn voor de oplossing van de probleemstelling.
Een veel voorkomende presentatie van een frequentieverdeling is een staatje van alle waarden van een variabele met bij elke waarde de frequentie van voorkomen in de onderzochte verzameling van eenheden (zie tabel 10.1). Zo'n tabel laat dus zien hoeveel eenheden er in elke waarde van de variabele vallen. Links staan de waarden, rechts de frequenties. Onder alle frequenties staat een *streep*, en daaronder '*Totaal ...*'. In plaats van het woord 'Totaal' wordt vaak de letter N (van: *number*) gebruikt. Zo'n tabel wordt ook een univariate tabel of een univariate frequentieverdeling genoemd, dit in tegenstelling tot kruistabellen waarbij twee of meer variabelen betrokken zijn, die we dan ook bi- of multivariaat noemen (zie hiervoor hoofdstuk 12).

2 Dit is een van de voorbeelddatabestanden in SPSS. Het betreft de General Social Survey van 1991.

*Tabel 10.1* *Academische promoties in Nederland in het academisch jaar 2007/2008 naar disciplinegroepen (bron: CBS)*

| | (Absoluut) | (%) |
|---|---|---|
| Gezondheidswetenschappen | 1.029 | 32,0 |
| Sociale en gedragswetenschappen | 607 | 18,9 |
| Technische wetenschappen | 563 | 17,5 |
| Bètawetenschappen | 489 | 15,2 |
| Landbouwwetenschappen en diergeneeskunde | 289 | 9,0 |
| Alfawetenschappen | 37 | 7,4 |
| Totaal | 3.214 | 100,0 |

De *volgorde* van de waarden is zodanig dat bij *nominale* variabelen meestal met de waarde die de hoogste frequentie heeft, wordt begonnen, en zo verder aflopend tot de waarde met de laagste frequentie. Bij variabelen op *ordinaal of hoger meetniveau* ligt de volgorde vast. We beginnen dan meestal met de 'laagste' waarde: bij een rangenstelsel met de laagste rang, bij leeftijd met de jongste, bij lengte met de kortste, enzovoort. Psychologen hebben de gewoonte om het precies andersom te doen, dus te beginnen met de hoogste leeftijdsgroep, de intelligentste mensen, enzovoort.

In de *kop van de tabel* (en dit verhaal geldt precies zo voor elk *plaatje* dat we willen maken van de frequentieverdeling) wordt uiteraard de betrokken *variabele* vermeld. Gaat het om één, losstaande, tabel of om een bepaalde deelverzameling van eenheden die een andere is dan die in andere tabellen, dan vermelden we het *domein*, het *tijdstip van meting* en de *bron* in de kop. Maar als de tabel in een rapport staat tussen veel meer tabellen die alle betrekking hebben op eenzelfde domein, tijdstip en bron, dan worden deze zaken niet steeds weer bij elke tabel vermeld. Vermelding van nadere omschrijvingen, zoals 'Aantallen ...', 'Frequentieverdeling van ...', 'Vergelijking van ...', 'Overzicht van ...' enzovoort is geheel *overbodig*.

Wat is het nut van percentageberekening? Twee punten:
- Vergelijking tussen verschillende verzamelingen *die een verschillend absoluut totaal hebben*, wordt mogelijk; zo kan nu de verdeling van de academische promoties over vakgebieden in Nederland in 2009 worden vergeleken met die van 1990 en 2000.
- Eventuele zeer grote aantallen, die onoverzichtelijk zijn, worden vermeden.

Hoe worden percentages berekend? Is $N_i$ het aantal eenheden in waarde i, dan zijn de proporties of aandelen achtereenvolgens $N_1/N$, $N_2/N$ ... $N_k/N$. Uiteraard zijn de proporties samen gelijk aan 1. Om er percentages van te maken (breuken zijn ook weer onoverzichtelijk om te lezen) vermenigvuldigen we elke $N_i/N$ met 100. Frequentieverdelingen in proporties, percentages of promillages noemen we *relatieve frequentieverdelingen*.
Voordat we driftig met allerlei relatieve frequentieverdelingen gaan werken, past een grondige waarschuwing. Bij een N van pakweg minder dan 50 (dit is niet meer dan een vuistregel!) kunnen we aan percentages weinig betekenis toekennen. Immers, in verband met meetfouten en dergelijke hadden sommige eenheden best eens in een andere categorie kunnen vallen, en de verkregen absolute aantallen zijn daarom tot op zekere hoogte aan toeval onderhevig. Bij grote aantallen vallen meetfouten tegen elkaar weg; bij kleine aantallen is dit veel minder het geval. Als de gegevens als absolute aantallen worden gepresenteerd, ziet iedere lezer om welke kleine aantallen het gaat, en kan hij zich realiseren dat het feit dat 17 mensen waarde i hebben en 21 personen waarde j, de mogelijkheid openlaat dat het ook wel eens andersom had kunnen zijn. Zijn beide aantallen echter uitgedrukt in percentages (als de variabele maar twee waarden heeft, levert dit dus 45 en 55%), dan lijkt een verschil van 10% indrukwekkend, en dat is gezien de kans op toevalstreffers een onjuiste conclusie. Bij een kleine N kunnen we conclusies als volgt weergeven: '... van de N mensen antwoorden ...'; bijvoorbeeld: 17 van de 38 mensen stemmen op de PvdA. Conclusie: de computer rekent wel braaf allerlei percentages uit, maar het is lang niet altijd verstandig deze in de rapportage op te nemen.

Hoe vind je in een digitaal bestand via SPSS de tabel voor een bepaalde variabele?
- Klik eerst op ANALYZE in de menubalk.
- Klik vervolgens op DESCRIPTIVE STATISTICS.
- Klik ten slotte op FREQUENCIES en klik de gewenste variabele aan.

Een voorbeeld: we kunnen het GSS91-voorbeeldbestand van SPSS weer gebruiken. Voor de variabele REGION verschijnt dan tabel 10.2 op het scherm.

*Tabel 10.2 De GSS91-steekproef naar woongebied*

| Region of the United States | | | | | |
|---|---|---|---|---|---|
| | | Frequency | Percent | Valid percent | Cumulative percent |
| Valid | 1.00 North East | 679 | 44.8 | 44.8 | 44.8 |
| | 2.00 South East | 415 | 27.4 | 27.4 | 72.1 |
| | 3.00 West | 423 | 27.9 | 27.9 | 100.0 |
| | Total | 1517 | 100.0 | 100.0 | |

In SPSS-output bevat zo'n eenvoudige tabel niet minder dan zes kolommen. Het is een standaard-*format*, dat voor erg ingewikkelde frequentieverdelingen nuttig kan zijn, maar dat voor een onderzoeksrapportage veel te uitgebreid is. De kolom met de 'labels', de waarden of *categorieën* van de variabele, spreekt voor zichzelf, maar de codecijfers kun je weglaten. Merk op dat die hier zelfs in twee decimalen achter de komma genoteerd zijn! De frequentiekolom en de percentagekolom spreken voor zich. Wat betekent de kolom *valid percent*? In het geval dat er ontbrekende waarnemingen (*missing values*) zijn, moet *deze* kolom worden gebruikt. De ontbrekende waarnemingen zijn dan eerst van het totaal afgetrokken; pas daarna wordt gepercenteerd. Zijn er, zoals hier, geen ontbrekende waarnemingen, dan zijn de kolommen *percent* en *valid percent* identiek. Neem in de rapportage hoogstens één van beide op! En de kolom *cumulatief percentage*? Ook deze verschijnt automatisch in de output, maar we hebben hem vrijwel nooit nodig; zeker niet bij nominale variabelen zoals hiervoor. In zulke situaties: altijd weglaten!
Bij het rapporteren is het raadzaam tabellen te fatsoeneren door *alles wat niet strikt nodig is weg te laten.* En neem geen Engelstalige namen (*labels*) op in een rapport dat in de Nederlandse taal is geschreven, net zomin als je Nederlandse namen zou opnemen in een Engelstalig rapport. Ten slotte nog een opmerking over de lijnen in een tabel. In tabel 10.2 voldoen deze aan de 'APA-normen',[3] een optie binnen SPSS. Net zoals met de APA-normen voor bronvermeldingen (zie box 2.3) is het in het algemeen verstandig deze aan te houden, maar het is minder belangrijk. Soms kan het duidelijker zijn om ook verticale lijnen te gebruiken. In de meeste tabellen in dit boek zijn de lijnen iets anders dan volgens de APA-normen.

Willen we in SPSS meteen de frequentieverdelingen van meer dan één variabele, dan kunnen we die variabelen in één handeling aanklikken en laten verschijnen.

3 APA staat voor American Psychological Association.

### Het vereenvoudigen van een frequentieverdeling

De verscheidenheid aan empirische waarnemingen is vaak groter dan de verscheidenheid aan waarden waarin een onderzoeker werkelijk geïnteresseerd is. Gegevens zijn soms 'van nature' in een groot aantal verschillende waarden te onderscheiden. We voegen dan, nadat de frequentieverdelingen gemaakt zijn, ten behoeve van de rapportage verschillende waarden samen.

> *Denk erom dat de oorspronkelijke, niet-samengevoegde, variabele in de datamatrix blijft bestaan, dus dat je de 'gecomprimeerde' nieuwe variabele een andere, nieuwe, naam geeft. De oorspronkelijke scores worden door de computer gebruikt ten behoeve van de berekening van centrum- en spreidingsmaten en samenhangen met andere variabelen. Het zou immers verlies van informatie zijn als we voor die berekeningen al bij voorbaat de waarnemingen zouden gaan samenvoegen in klassen.*

Hierna staan mogelijke redenen voor samenvoeging.

- Het bereiken van overzichtelijkheid in de rapportage. Een tabel is alleen overzichtelijk als het aantal waarden de tien niet te boven gaat, en liefst beperkt blijft tot zo'n vijf of zes. De gebruiker heeft vaak alleen behoefte aan een eenvoudige twee- of driedeling. De wijze waarop wordt samengevoegd, wordt hierdoor bepaald. Een voorbeeld op ordinaal niveau: als we de Nederlandse bevolking indelen naar laatst gevolgde opleiding, ontstaat een groot aantal geordende waarden. We kunnen een dergelijke indeling vereenvoudigen tot een driedeling 'laag', 'midden' en 'hoog'. In de hiërarchie van schooltypen brengen we dan dus slechts twee snijpunten aan. Zo kunnen we, om te beginnen, 'alleen basisschool' combineren met 'basisschool plus vmbo'. Een voorbeeld op rationiveau: de individuele opgaven van genoten inkomen bij de Belastingdienst zijn tot op één euro nauwkeurig (wat iets anders is dan dat zij tot op één euro de juiste waarde weerspiegelen!). In de hierop gebaseerde *overheidsstatistieken* worden de inkomens ter wille van de overzichtelijkheid gepresenteerd in klassen van € 5.000 of € 10.000 'breed'. De *Belastingdienst* hanteert echter voor de aanslagen de tariefgrenzen van de belastingschijven.
- In een erg verfijnde waarde-indeling kunnen nogal wat meetfouten zitten doordat het meetinstrument 'fijn' was in verhouding tot de gegevens. Dit geldt vooral als het gemeten opinies, houdingen en gedragingen betreft. Bovendien is een verdeling met heel veel waarden vaak nogal onregelmatig (niet een mooie, gelijkmatige, verdeling, maar een met diverse pieken en dalen). Het heeft dan alleen maar voordelen om wat bredere klassen te

maken; fouten vallen tegen elkaar weg en de verdeling wordt daardoor regelmatiger.
– Indien sommige waarden heel weinig eenheden bevatten in verhouding tot de andere waarden, heeft het weinig zin dergelijke waarden apart te blijven onderscheiden. Als het een nominale variabele betreft, voegen we de kleintjes samen tot een categorie 'overige'; zorg daarbij dat de categorie 'overige' qua vulling kleiner blijft dan de kleinste van de apart genoemde waarden. Een voorbeeld is de categorie 'overige' bij een vraag naar motieven voor studiekeuze. Naast vier of vijf standaardmotieven zijn er veel motieven die maar door enkelen genoemd worden, en die komen dan samen in een categorie 'overige'. Op het ordinale meetniveau combineer je 'kleintjes' met *aangrenzende* grotere klassen. Bij variabelen op interval- en rationiveau zullen we, bij te kleine celvullingen, de klassen bijvoorbeeld twee aan twee samenvoegen.

Met SPSS wordt het samenvoegen van categorieën uitgevoerd door aanklikken van TRANSFORM in de menubalk, en dan RECODE. In het venster van RECODE geef je vervolgens de variabele(n) aan die gehercodeerd moet(en) worden. In Appendix B is hierover meer te lezen. Noteer ten behoeve van een rapport in de Nederlandse taal de namen van de variabelen en de categorieën meteen in het Nederlands, dat is mooi meegenomen bij het maken van de output.

### Gegroepeerde frequentieverdelingen; klassen; onder- en bovengrens

Als we de oorspronkelijke waarden van een *continue* variabele samenvoegen, spreken we van een *gegroepeerde frequentieverdeling*. Een tabel van een gegroepeerde frequentieverdeling (zie tabel 10.3) kunnen we gemakkelijk herkennen doordat er sprake is van *klassen*, elk met een *ondergrens* en een *bovengrens* (behalve soms de eerste en/of de laatste klasse). De klassen mogen niet overlappen, vandaar het woord *tot*, en dus *niet*: 'tot en met'. Iemand met een inkomen van precies €10.000 valt dus in de derde klasse in tabel 10.3. Het gebruik van het woordje *tot* verdient de voorkeur boven het vooral vroeger gebruikelijke *streepje*, omdat dit laatste óók gebruikt wordt in de betekenis van 'tot en met', en dus verwarring kan geven.

*Tabel 10.3* *Voltijdwerknemers in Nederland naar verdiend brutojaarloon (in €) in 2000 (in %) (bron: CBS StatLine)*

| | Jaarloonklasse | % |
|---|---|---|
| | Minder dan 5.000 | 0,4 |
| | 5.000 tot 10.000 | 1,5 |
| → | 10.000 tot 15.000 | 4,0 |
| | 15.000 tot 20.000 | 12,3 |
| | 20.000 tot 25.000 | 18,0 |
| | 25.000 tot 30.000 | 20,3 |
| | 30.000 tot 35.000 | 14,2 |
| | 35.000 tot 40.000 | 9,6 |
| | 40.000 tot 45.000 | 6,0 |
| | 45.000 tot 50.000 | 3,9 |
| | 50.000 en meer | 9,8 |
| | Totaal | 100,0 |

Hoewel het voordelen heeft om de klassenbreedte over de hele frequentieverdeling gelijk te houden, is dit helemaal niet verplicht. Het is best mogelijk een gegroepeerde frequentieverdeling te maken met ongelijke klassen. Als bijvoorbeeld de extreme klassen aan één kant van de verdeling slechts spaarzaam gevuld zijn, kunnen we de klassenbreedte daar tweemaal zo groot maken.

In het geval dat we achteraf – bijvoorbeeld op basis van een tabel die we in een rapport van een ander onderzoek aantreffen – 'met de hand' berekeningen willen uitvoeren, kunnen we dat in zo'n geval alleen benaderend doen. We geven dan aan *alle* eenheden in een klasse de waarden van het midden van die klasse (we nemen aan dat de eenheden wel gemiddeld op deze waarde zullen uitkomen). Het klassenmidden ligt precies halverwege de bovengrens en de ondergrens van de klasse. Bij open klassen (zoals de eerste en de laatste klasse in tabel 10.3) kunnen géén klassenbreedte en klassenmidden worden berekend; het gevolg is dat over een tabel waarin open klassen voorkomen, geen totaalgemiddelde berekend kan worden.

### *Drie redenen voor toch een streepje!*

Soms lijken bij een gegroepeerde frequentieverdeling de klassengrenzen niet goed op elkaar aan te sluiten, en lijkt er dus een gat tussen te zitten. Dat is niet

zo, want het 'streepje' in de klassenaanduiding moeten we dan lezen als '*tot en met*'. We geven drie voorbeelden.

a. Bij de veelgebruikte variabele *leeftijd* zien we vaak zoiets als:

| Jaar | Frequentie | Jaar | Frequentie |
|---|---|---|---|
| 0 - 14 | ... | 30 - 39 | ... |
| 15 - 19 | ... | 40 - 49 | ... |
| 20 - 24 | ... | 50 - 64 | ... |
| 25 - 29 | ... | 65 jaar en ouder | ... |

Even nadenken leert ons dat dit logisch is: ook iemand die 'de volgende dag vijftien jaar wordt', valt in de klasse 0-14 jaar. Het is dus duidelijk dat deze klasse doorloopt tot de volgende klasse, die begint met degenen die juist op de dag van het onderzoek vijftien jaar werden. Deze wijze van noteren is overigens niet meer dan een gewoonte. We zouden ook de klassen kunnen noteren als '0 tot 15 jaar', '15 tot 20 jaar', enzovoort. Maar om vergissingen te voorkomen ligt bij de variabele 'leeftijd' een notering zoals hiervoor wat meer voor de hand.

b. Soms moeten we werken met meetwaarden, scores die door andere onderzoekers of door beleidsambtenaren al zijn *afgerond* (meestal boven een halve meeteenheid naar boven, daaronder naar beneden). In dat geval noteren we de klassen zoals hierna:

| IQ | Frequentie | IQ | Frequentie |
|---|---|---|---|
| 82 - 87 | 2 | 106 - 111 | 20 |
| 88 - 93 | 23 | 112 - 117 | 10 |
| 94 - 99 | 22 | 118 - 123 | 0 |
| 100 - 105 | 65 | 124 - 129 | 8 |

De 'gaten' tussen de klassengrenzen zijn geen echte gaten, omdat bij het afronden als feitelijke klassengrenzen (we noemen dat de '*exacte klassengrenzen*') de getallen 87,5, 93,5 enzovoort zijn gebruikt. Persoon A met de ruwe score 87,4 (afgerond op 87) valt dus in de eerste klasse, persoon B met de ruwe score 87,7 (afgerond tot 88) in de tweede klasse. Daarbij zouden we ook nog moeten weten wat er gedaan is met een waarneming precies op de exacte klassengrens (bijvoorbeeld 87,5). De afspraak is meestal dat deze naar boven wordt afgerond.

Als we de klassen als volgt hadden genoteerd: '82 tot 88', '88 tot 93' enzovoort, dan was persoon B, als we alleen op diens afgeronde score waren afgegaan, ten onrechte in de tweede klasse terechtgekomen, die immers bij dit laatste notatiesysteem pas begint bij *precies* 88. *Veel maakt het allemaal niet uit*, natuurlijk, als we maar steeds consequent op dezelfde manier indelen. Maar we geven deze uitbreiding omdat, *als* we met afgeronde scores te maken krijgen, de notatie zoals hiervoor de beste is.

c. Een andere reden om de klassengrenzen niet aansluitend te maken vinden we bij discrete variabelen, waar het gaat om gehele getallen waar nu eenmaal niets tussen ligt. Als we gezinnen indelen naar *aantal kinderen*, als we gezinshoofden indelen naar *aantal verenigingen* waarvan zij lid zijn, en als we ambtenaren indelen naar *het aantal vergaderingen per week*, kunnen we als categorieën, als waarden van de variabele natuurlijk achtereenvolgens 0, 1, 2, 3, 4 enzovoort noteren. Maar als we wat beknopter willen rapporteren en dus waarden gaan samenvatten, krijgen we zoiets als:

| Aantal vergaderingen per week |
|---|
| 0 - 1 |
| 2 - 3 |
| 4 - 5 |
| 6 - 7 |
| 8 of meer |

Een betere notatie zou overigens zijn: '0 of 1', '2 of 3', enzovoort.

**Box 10.1**

### Een noot over de geschiedenis van klassengrenzen

**Vroeger waren er verschillende systemen in gebruik. Een paar voorbeelden, in de vorm van varianten op tabel 10.3:**

| *(a)* 0 tot 5.000 | *(b)* 0 – 4.900 | *(c)* 0 – 4.990 | *(d)* 0 – 4.999 |
|---|---|---|---|
| 5.000 tot 10.000 | 5.000 – 9.900 | 5.000 – 9.990 | 5.000 – 9.999 |
| 10.000 tot 15.000 | 10.000 – 14.900 | 10.000 – 14.990 | 10.000 – 14.999 |
| …… | …… | … | …… |

**De heldere en duidelijke grenzen van systeem a hebben de voorkeur, maar kunnen alleen gebruikt worden als de oorspronkelijke scores beschikbaar zijn. Systeem b kan worden gebruikt als de ruwe scores in een eerdere fase afgerond zijn op € 100. Systeem c bij eerdere afronding op € 10 en systeem d bij afronding op € 1. De indelingen b, c en d zijn dus alleen relevant als de scores eerder afgerond zijn. Maar als men, laten we zeggen, vijftig jaar geleden het gemiddelde van dertigduizend individuele scores moest berekenen, kostte dat erg veel tijd. Daarom werden de scores eerst samengenomen in klassen, en de klassengemiddelden werden gebruikt om het totaalgemiddelde te berekenen. Voor de afronding werden systemen als b, c en d gebruikt. Daarbij lijkt het alsof er 'gaten' zitten tussen elke twee op elkaar volgende klassen, maar dit is niet zo. In feite werden de punten halverwege de bovengrens van klasse a en de ondergrens van klasse a+1 gebruikt, de *exacte klassengrenzen* (zoals 4950 in b, 4995 in c en 4999,5 in d). Welke indeling werd gebruikt, hing dus af van de precisie van de oorspronkelijke meting (in hele getallen, één of meer decimalen, enzovoort).**

**Vroeger werden in leerboeken over statistiek heel wat pagina's besteed aan hoe je moest omgaan met zaken als gegroepeerde frequentieverdelingen en klassengrenzen. Dankzij de computer worden berekeningen nu altijd uitgevoerd op de oorspronkelijke data, en worden gegroepeerde frequentieverdelingen alleen maar gebruikt om de oorspronkelijke, uitgebreide, frequentieverdelingen samenvattend te presenteren.**

## Grafische voorstellingen

In plaats van, of naast, een tabel kunnen we gebruikmaken van een grafische voorstelling. Zo'n 'plaatje' is vaak instructiever dan een reeks getallen. De lezer krijgt meer vat op het geheel, en heeft minder kans zich in details te verliezen. Een bezwaar is dat degene die het plaatje maakt, ook meer gelegenheid heeft

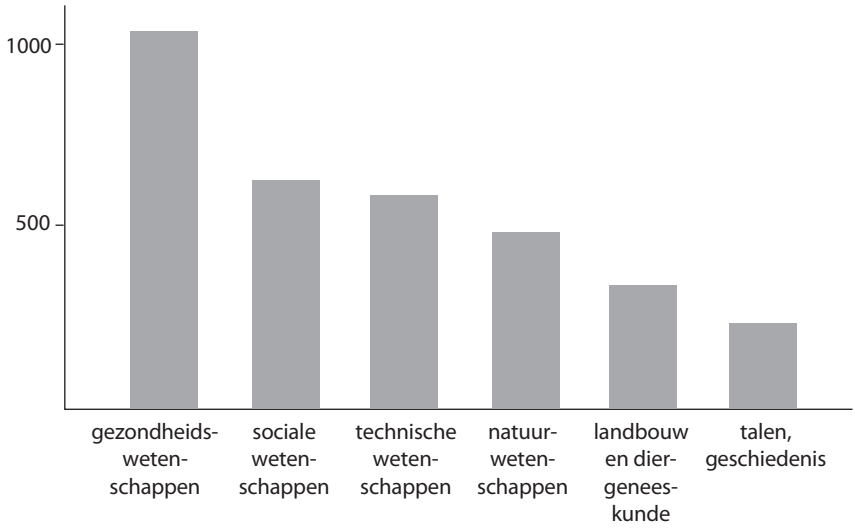
om de lezer of kijker te misleiden (denk aan het gebruik bij PowerPoint-presentaties). Manipulatie van de horizontale of verticale assen is daarvan een voorbeeld. Stel dat de scores op een variabele die in principe van 0 tot 100 loopt, variëren tussen 40 en 50. Geef je op de verticale as, zoals het hoort, het hele 0-100-bereik aan, dan blijken er dus maar kleine verschillen te zijn. Maar beperk je de scores op de verticale as tot bijvoorbeeld tussen de 30 en 60, dan lijken diezelfde verschillen, of schommelingen, heel indrukwekkend.

Er zijn allerlei manieren van grafische voorstelling. We onderscheiden vijf hoofdtypen:
- staafdiagrammen, met als bijzondere vorm het histogram;
- lijngrafieken;
- cirkeldiagrammen;
- stroomdiagrammen;
- symboolverzamelingen.

Op elk van deze hoofdtypen gaan we hierna in. Maar de mogelijkheden zijn hiermee niet uitgeput. Om er nog twee te noemen: het gebruik van verschillende arceringen (een PvdA-'kaart' van de Nederlandse gemeenten: hoe hoger het percentage stemmen op die partij, hoe 'dichter' de arcering) en het gebruik van kleuren. Het is een terrein waarop je je eigen inventiviteit naar hartenlust kunt uitleven.

### *Staafdiagrammen*

Figuur 10.2 is een voorbeeld van een staafdiagram (*bar chart*). Voor de kop van een staafdiagram – en dat geldt voor alle andere grafische voorstellingen – geldt precies hetzelfde als voor de kop van een tabel: geef duidelijk de variabele aan en afhankelijk van de context: domein, tijdstip en bron.




*Figuur 10.2 Academische promoties in Nederland in het academische jaar 2007/2008 naar disciplinegroepen (bron: CBS)*

De waarden van de variabele worden voorgesteld door staven *van gelijke breedte*. We zijn automatisch geneigd bij het bekijken van zo'n grafiek de lengte van de staven te vergelijken, en dat is prima zolang de staven maar even breed zijn; het gaat uiteindelijk om de oppervlakten.
Voor de volgorde van de waarden of staven in de figuur gelden dezelfde regels als voor de volgorde van de waarden in een tabel.

Bij nominale en bij discrete ordinale variabelen worden de staven voor de duidelijkheid van elkaar gescheiden door een overal gelijke tussenruimte. Verticaal worden de frequenties afgezet, in absolute aantallen of in percentages. De hoogte van de staven is evenredig met de frequenties. Dit betekent dat we op de verticale as altijd bij 0 moeten beginnen!

In SPSS klikken we op GRAPHS in de menubalk, daarna op BAR, en daarna op BAR CHARTS. Klik op het SIMPLE-plaatje en daarna op DEFINE. De variabele waarvan we de frequentieverdeling willen afbeelden, wordt nu gekozen. Deze komt op de CATEGORY AXIS; daarna hoef je alleen nog maar op OK te klikken.

We brengen nog meer informatie in het plaatje als we in figuur 10.2 achter elke staaf een 'schaduwstaaf' van een ander peilingsjaar, bijvoorbeeld 1997/1998, of

van een ander land zouden zetten. We kijken dan naar de verdeling over landen en tegelijkertijd naar de ontwikkeling in de tijd. Of we zouden in figuur 10.2 apart de staven voor de mannen en voor de vrouwen kunnen schetsen. Wanneer gegevens worden uitgesplitst op zo'n andere variabele wordt in SPSS-taal gesproken van een *clustered bar chart* (per jaar worden de staven van mannen en vrouwen 'geclusterd').

Staafdiagrammen zijn vooral zinvol bij variabelen op nominaal meetniveau, waarbij het aantal waarden tussen de vijf en vijftien ligt. Als het aantal waarden kleiner is, is een cirkeldiagram aantrekkelijker (zie verderop). Als het aantal waarden groter is, voegen we waarden meestal samen, bijvoorbeeld tot een categorie 'overige'.

Enkele voorbeelden voor toepassing van een staafdiagram:
- Nederlanders naar politieke partij of kerkelijke gezindte (y-as: # mensen);
- immigratie naar land van herkomst (y-as: # binnengekomen personen);
- aantal nieuw gebouwde woningen naar soort in een bepaald jaar (y-as: # woningen);
- dagbladartikelen, naar onderwerp (y-as: # artikelen);
- studenten naar opleiding (y-as: # studenten).

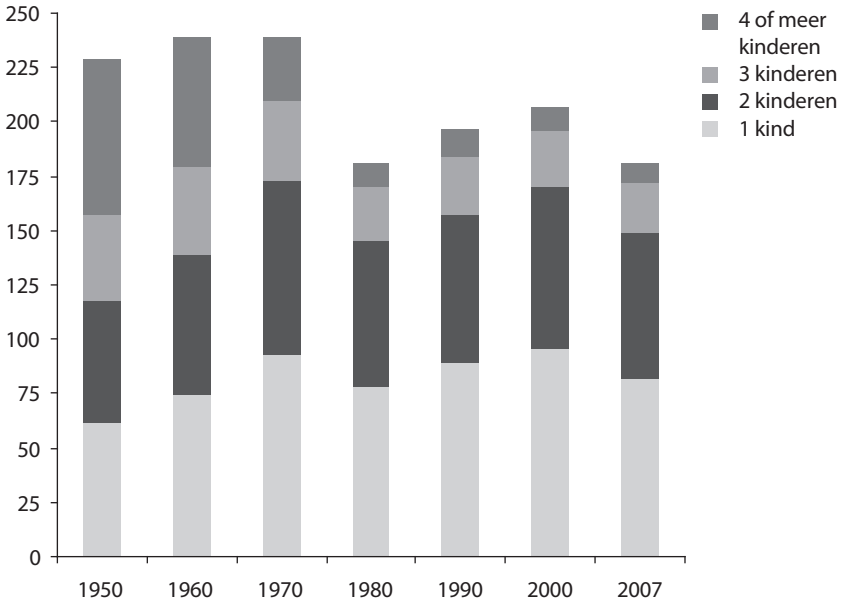
En op ordinaal niveau:
- Nederlanders naar opleidingsniveau (y-as: # personen);
- het personeel van een ambtelijke organisatie naar rang (y-as: # stafleden per rang).

Maar staafdiagrammen worden ook voor interval- en rationiveauvariabelen gebruikt:
- inkomensniveau (y-as: # mensen in elke klasse);
- leeftijd (denk aan de bekende 'leeftijdspiramide': dit is een 'dubbel' staafdiagram dat bovendien 'op zijn kant' staat.

Als je een staafdiagram gebruikt voor variabelen op het ordinale of hogere niveau is de volgorde natuurlijk die van de categorieën, die in het algemeen een andere is dan wanneer je de staven in volgorde van frequentie zou plaatsen.

Er zijn verschillende wat meer ingewikkelde uitvoeringen van het staafdiagram denkbaar.

Als elke staaf wordt onderverdeeld naar de waarden van een andere variabele (zie figuur 10.3) noemen we dat een 'gestapeld staafdiagram' (*stacked bar chart*, *segmented bar chart*).




*Figuur 10.3* *Aantal geboorten (Nederland 1950-2007) naar de rangpositie van het kind in het gezin (absolute aantallen x 1.000) (bron: CBS)*

Aan zo'n gestapeld staafdiagram kunnen we zien of de verhoudingen van jaar tot jaar ongeveer hetzelfde blijven, of dat er verschuivingen optreden. In feite hebben we hier met drie variabelen tegelijk te maken: aantal geboorten, jaartal en rangpositie. Duidelijker is het om in zo'n geval een *lijngrafiek* te tekenen (zie figuur 10.7), zeker als het over vijf of meer tijdstippen gaat (hier dus: per 'rangpositiegroep' een doorlopende lijn, in totaal een figuur met vier lijnen).

Bij continue variabelen op interval- en hoger niveau gemeten neemt een staafdiagram de vorm aan van een *histogram* (zie figuur 10.4, die gebaseerd is op tabel 10.4). Als je in SPSS een histogram maakt, denk er dan om de 'interactive mode' te gebruiken. Kijk goed naar het aantal categorieën en gebruik de 'edit'-mogelijkheden. Een histogram wijkt in enkele opzichten af van het staafdiagram:

– De waarden staan in een vaste volgorde, dus ongeacht de grootte van de frequenties.

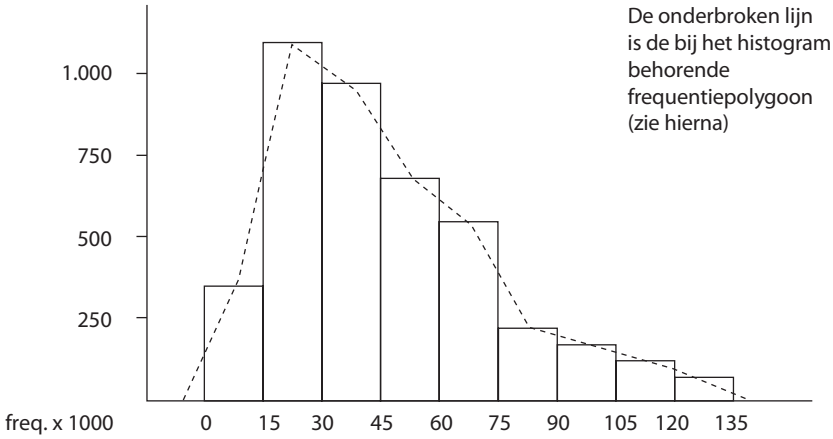
- De staven worden tegen elkaar geplaatst, zonder tussenruimte. In SPSS-output blijven de staven, ter wille van de herkenbaarheid, van elkaar gescheiden door een tussenruimte.
- Aan de voet van de staven worden de (exacte) klassengrenzen vermeld; soms zien we in plaats daarvan de klassenmiddens. Denk erom: als in de frequentietabel sprake is van 'afgeronde klassen' (dat wil zeggen dat de klassen elkaar niet raken), zetten we de *exacte klassengrenzen* af op de horizontale as. Immers, de staven horen onderling geen tussenruimte te hebben. In SPSS-output wordt niet aan deze voorwaarde voldaan; de getallen op de horizontale as zijn niet veel meer dan maatstreepjes.

Een histogram kan worden gebruikt voor bijvoorbeeld:
- leeftijdsklassen;
- testresultaten;
- schoolvorderingen;
- inkomen;
- aantal uren slaap;
- aantal bezoeken aan de huisarts per jaar.

Hierna geven we een meer uitgewerkt voorbeeld, waarbij ter controle eerst de frequentietabel wordt vermeld en daarna het histogram dat daarbij hoort.

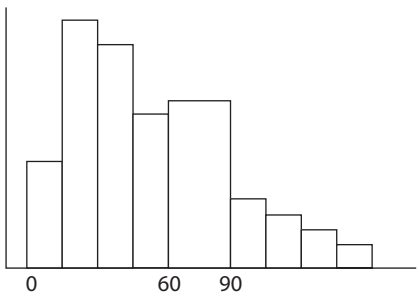
*Tabel 10.4* *Gemiddelde woon-werkreistijd van de Nederlandse werkende bevolking in 1980 (in veelvouden van een kwartier)*

| Reistijd | Frequentie x 1.000 | Reistijd | Frequentie x 1.000 |
|---|---|---|---|
| < 15 min. | 360 | 75 tot 90 min. | 220 |
| 15 tot 30 min. | 1.100 | 90 tot 105 min. | 200 |
| 30 tot 45 min. | 920 | 105 tot 120 min. | 90 |
| 45 tot 60 min. | 710 | 120 min | 40 |
| 60 tot 75 min. | 530 | | |
| | | Totaal | 4.170 |

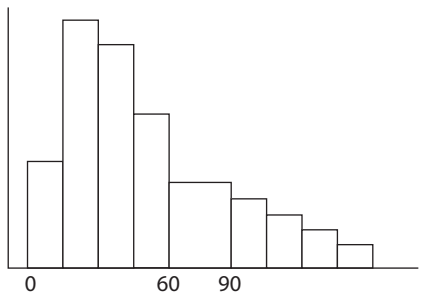



*Figuur 10.4 Gemiddelde woon-werkreistijd van de Nederlandse werkende bevolking in 1980 (in veelvouden van een kwartier)*

Meestal zullen de klassen langs de horizontale as even breed zijn. Als dat niet het geval is, moeten we de lengte van de staven overeenkomstig aanpassen. Met andere woorden: als een klasse tweemaal zo breed is als de andere klassen, moeten we de lengte van die staaf halveren. Immers, we willen in feite *oppervlakten* weergeven. Voorbeeld: als in tabel 10.4 om een of andere reden niet de aparte frequenties van de klassen 60 tot 75 min. en 75 tot 90 min. bekend zijn, zijn we wel gedwongen een combinatieklasse 60 tot 90 min. te maken. Figuur 10.5a is dan fout en figuur 10.5b geeft het juiste plaatje.




*Figuur 10.5a (fout)*




*Figuur 10.5b (goed)*

Als er vele klassen zijn, vervangen we het histogram liever door een *frequentiepolygoon*. Deze geeft een indruk van hoe het histogram eruit zou zien bij een

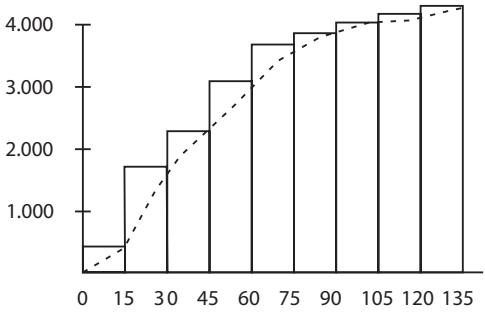
heel kleine klassenbreedte en een groot aantal waarnemingen. We kunnen een histogram simpel in een frequentiepolygoon omzetten door de *middens* van de bovenkanten van de staven met elkaar te verbinden (zie de onderbroken lijn in figuur 10.4). Als we dat doen, blijft de frequentiepolygoon echter links en rechts zweven boven de horizontale as. Daar is op zich geen bezwaar tegen. Maar we willen graag duidelijk maken dat de oppervlakte onder de curve van een polygoon gelijk is aan de som van de oppervlakten van de staven bij het histogram waarvan die polygoon is afgeleid. Daarom voegen we links en rechts een klasse toe (van gelijke breedte als de andere) met frequentie 0, en trekken we de polygoon tot de horizontale as door. Dat we daarmee aan die voorwaarde van gelijke oppervlakten voldoen, valt gemakkelijk aan te tonen.

Een klein probleempje ontstaat als er links en/of rechts een open klasse is. Met open klassen weten we in een histogram of een polygoon niet goed raad, omdat we in feite de breedte van het interval (en dus ook de hoogte, het gaat om de oppervlakte!) niet kunnen aangeven. Niettemin wordt een frequentieverdeling met open klassen wel door een histogram of polygoon voorgesteld, *mits er heel weinig eenheden in die open klasse zitten*; op het totale beeld van de grafiek kan een eventuele vertekening dan weinig invloed hebben. In ons voorbeeld betreft het mensen met een reistijd van twee uur of meer. In de figuur hebben we de gemiddelde reistijd van deze mensen op 127,5 minuten geschat; misschien is dit gemiddelde in werkelijkheid 145 minuten, maar veel maakt dit voor het totale beeld niet uit.

Hebben we als getalsbasis voor een histogram niet een absolute, maar een relatieve frequentieverdeling, dan maken we een *relatief histogram* en/of *een relatieve frequentiepolygoon*. Het enige verschil is alleen dat nu op de y-as de percentages van 0 tot 100 (of proporties, of promillages) zijn afgezet. Dit is vooral dan zinvol, wanneer we in één figuur vergelijkingen willen maken tussen groepen met heel verschillende totale aantallen. Zo kunnen we in een bedrijf de diensttijd van de mannen vergelijken met de diensttijd van de vrouwen.

Ten slotte kennen we nog het *cumulatieve histogram*. Het cumulatieve histogram in figuur 10.6 is gebaseerd op dezelfde data als het histogram van figuur 10.4. De lengte van een staaf wordt nu bepaald door het aantal eenheden in die klasse *plus* alle eenheden in alle lagere klassen. Let op: een cumulatief histogram ontstaat door de punten *rechtsboven* aan elke staaf met elkaar te verbinden; rechtsboven, omdat het, zoals bij elke cumulatieve tabel of grafiek, gaat om 'wat er onder de bovengrenzen zit'. Vooral bij een groot aantal klassen kan het cumulatieve histogram dienen als basis voor de constructie van een

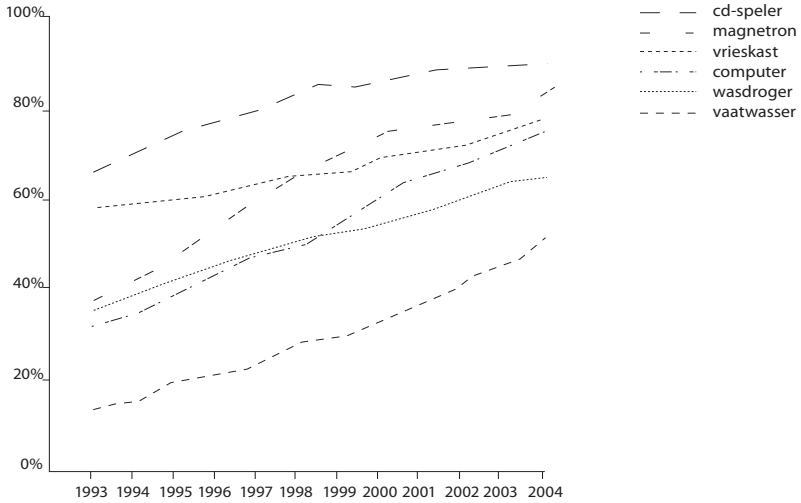
*cumulatieve frequentiepolygoon* (ook wel '*ogive*' genoemd, naar de vorm van een Gothisch gewelf), waarbij we een lijn trekken door de punten *rechtsboven* elke staaf. Op de horizontale as worden nu in elk geval de (exacte) klassengrenzen als meetpunten vermeld. Op de verticale as staan absolute aantallen, of, bij een relatieve cumulatieve polygoon, percentages of promillages. Een cumulatieve frequentiepolygoon stijgt altijd of blijft horizontaal (na een lege klasse) van links naar rechts gaande. Bij een cumulatief histogram zien we welk aantal, of welk percentage, van de eenheden beneden de bovengrens van een bepaalde klasse ligt, maar we kunnen minder snel iets zeggen over het percentage eenheden dat onder een bepaalde waarde ligt die niet toevallig met een klassengrens samenvalt.
Bij een cumulatieve frequentiepolygoon nu kan dit wél. Bij elke willekeurige waarde van de variabele kunnen we gemakkelijk de bijbehorende cumulatieve frequentie opzoeken, en andersom kunnen we bij elke willekeurige cumulatieve frequentie opzoeken welke waarde daarbij past. Meestal gaat het dan om de vraag onder welke waarde de laagste 25%, de laagste 50% of de laagste 75% ligt (respectievelijk het eerste, tweede en derde kwartiel).




*Figuur 10.6* *Een cumulatief histogram (de rechthoeken) en de bijbehorende cumulatieve frequentiepolygoon (de onderbroken lijn)*

### *Lijngrafieken*

Een lijngrafiek toont *het verloop in de tijd* van de absolute of relatieve score van één eenheid op één of enkele variabelen. Een lijngrafiek is dus 'het portret' van een *tijdreeks*. Figuur 10.7 geeft een voorbeeld.



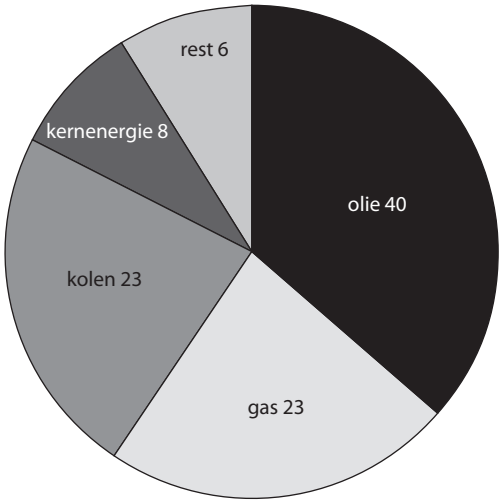

*Figuur 10.7 Bezit van apparatuur in Nederland, 1993-2004 (aantal apparaten per 100 huishoudens) (bron: CBS)*

Een lijngrafiek is vooral informatief als verschillende variabelen in dezelfde figuur worden gebruikt (zoals hier). Denk ook aan een schets van de aanhang van diverse politieke partijen in de loop van de tijd, aan het verloop van verschillende vormen van vrijetijdsbesteding, aan de ontwikkeling van het gemiddeld inkomen van verschillende groepen en aan de ontwikkeling van het percentage werklozen in het noorden, midden, westen en zuiden van het land. Bij voorkeur wordt op de verticale as met de frequentie of het percentage o begonnen; zo niet, dan moet zeer duidelijk op de verticale as een scheurlijn worden aangegeven (een stukje zigzag). Denk erom dat op de verticale as, net zoals bij alle grafieken, vermeld wordt waar het om gaat: absolute aantallen, percentages of wat dan ook.

### *Cirkeldiagrammen*

We kunnen een frequentieverdeling soms heel aanschouwelijk voorstellen door een oppervlakte van een bepaalde vorm in segmenten te verdelen die corresponderen met de relatieve frequenties. We zagen al het gebruik van 'opgedeelde staven' in gestapelde staafdiagrammen. Vooral variabelen van nominaal meetniveau lenen zich hier goed voor. Wil de voorstelling duidelijk zijn, dan mag het aantal waarden niet groter dan acht à tien zijn. Voorwaarde voor toepas-

sing is dat het niet zozeer gaat om absolute aantallen als wel om relatieve. Een aardige toepassing zien we als duidelijk gemaakt moet worden welk deel van de prijs van een liter benzine gaat naar het land van herkomst, naar de oliemaatschappij, naar transportkosten en naar de overheid: men kan dan de afbeelding van een jerrycan in stukjes verdelen. Evenzo bij een pak melk om te laten zien hoeveel cent de boer krijgt, hoeveel de melkfabriek, hoeveel de melkleverancier en hoeveel de overheid erop toelegt. Heel gebruikelijk, maar zeker niet noodzakelijk, zijn *cirkels*. Cirkeldiagrammen (ook wel taartdiagrammen genoemd) kunnen onder de naam PIE net zoals BAR CHARTS via SPSS worden aangemaakt.

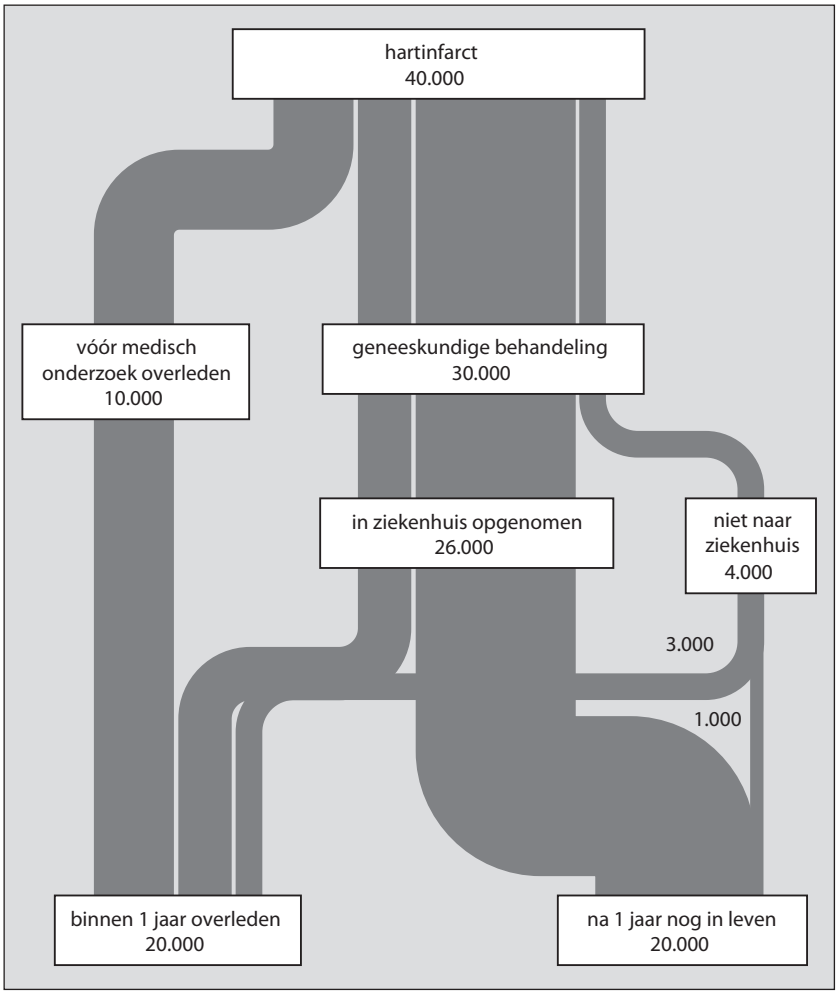



*Figuur 10.8* *Energieconsumptie naar bron in de Verenigde Staten in 2005 (bron: internet)*

### *Stroomdiagrammen*

Een stroomdiagram geeft door stromen van verschillende dikte aan hoe een bepaalde verzameling zich in de loop van de tijd opsplitst. Het is een soort tussenvorm van een lijngrafiek en een cirkeldiagram. Het essentiële ervan is dat niet een aantal momentopnamen los van elkaar wordt geportretteerd, maar dat één verzameling, één cohort, eenheden (leerlingen, werknemers) in de loop van de tijd wordt gevolgd. Stroomdiagrammen worden vaak gebruikt in de onderwijsstatistiek: hoe verdeelt een generatie achtstegroep-basisschoolleerlingen zich na elk jaar over de verschillende onderwijsvormen? Het is een voor-

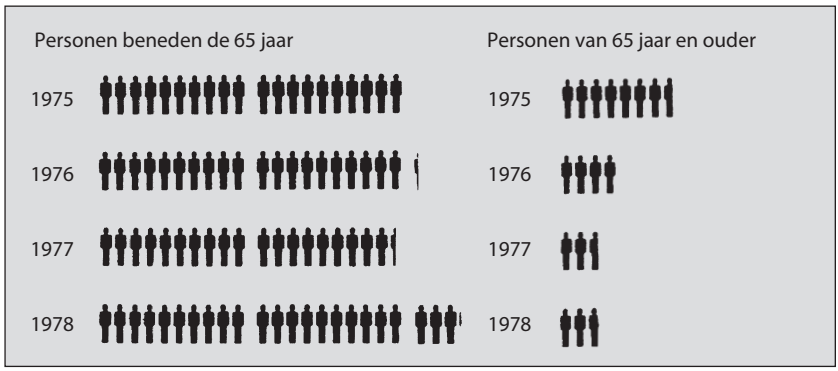
treffelijke wijze van visuele weergave van gegevens; helaas is de soort gegevens die men hiervoor nodig heeft, zelden beschikbaar.




*Figuur 10.9 Lotgevallen van slachtoffers van acuut hartinfarct (1973/1976) (bron: CBS, Statistisch Zakboek, 1981)*

### *Symboolverzamelingen*

Wat symboolverzamelingen[4] zijn, wordt onmiddellijk duidelijk met het voorbeeld hierna. Merk op dat de informatie van dit voorbeeld óók via twee lijnen in een lijngrafiek overgebracht kan worden. In andere voorbeelden zien we dat een 'symboolverzameling' een bijzondere vorm van een staafdiagram is, bijvoorbeeld als we het aantal marineschepen, tanks en vliegtuigen van twee landen willen uitbeelden.




*Figuur 10.10* *Periodieke bijstandsverlening aan thuiswonenden ingevolge de Algemene bijstandswet 1971-1980 (bron: Statistische Berichten, Gemeente Utrecht, 1980)*

### *De keuze van een grafische voorstelling*

Welke grafische voorstelling je nu het best kiest, hangt vooral af van het meetniveau van de variabele, maar we moeten ons ook afvragen of een plaatje werkelijk helpt om vat te krijgen op een frequentieverdeling, vergeleken met een tabel of met het presenteren van wat kenmerkende cijfers. Om een frequentieverdeling van een dichotome variabele (bijvoorbeeld sekse) voor te stellen heb-

4 Andere woorden zijn *pictogram sets*, 'visuele statistiek' of 'infographics'. Ze hebben een interessante geschiedenis. De basis ervoor werd gelegd door de Duitse ontwerper Gerd Arntz (1900-1988), die samenwerkte met de Weense statisticus Otto van Neurath, die het grootste deel van zijn leven in Nederland doorbracht. De techniek staat officieel bekend als ISOTYPE (International System of Typographic Picture Education). De belangrijkste toepassing van pictogrammen ligt tegenwoordig niet meer in de statistiek, maar in het gebruik op vliegvelden, stations en in andere openbare ruimten (Jansen, W. (2009). Neurath, Arntz and ISOTYPE: The legacy in art, design and statistics. *Journal of Design History,* 22, 227-242).

ben we aan één getal (47% is man) genoeg; het maken van een plaatje is overbodig.
Als een variabele meer dan twee categorieën heeft, kan een plaatje zinvol zijn. Bij een nominale variabele met drie tot zeven of acht categorieën kan een cirkeldiagram worden gebruikt. Zijn er meer categorieën, en ook bij ordinale variabelen, dan is een staafdiagram beter. En als een staafdiagram nuttig is, kun je ook geclusterde staafdiagrammen of gestapelde staafdiagrammen gebruiken.

Als de tijd een essentiële variabele is, moet deze op de horizontale as worden afgezet.
We hebben in dit hoofdstuk slechts enkele basisprincipes van grafische voorstellingen genoemd en toegelicht. Er zijn heel wat variaties en combinaties denkbaar. Bovendien kan de visuele aantrekkelijkheid worden verhoogd door de toepassing van driedimensionale plaatjes, kleuren en andere technieken. Een algemeen advies: maak het niet te mooi alleen maar om het 'mooi zijn', en zeker niet te ingewikkeld. Het doel van een plaatje is onmiddellijk duidelijk informatie overdragen; het moet geen zoekplaatje worden.

| *Nominaal niveau* | | |
|---|---|---|
| Twee categorieën | → | geen plaatje nodig |
| Drie tot zeven of acht categorieën | → | cirkeldiagram |
| Meer dan zeven of acht categorieën | → | staafdiagram |

| *Ordinaal niveau* | | |
|---|---|---|
| Twee categorieën | → | geen plaatje nodig |
| Meer dan twee categorieën | → | staafdiagram |

| *Interval- en rationiveau* | | |
|---|---|---|
| Twee categorieën | → | geen plaatje nodig |
| Discreet | → | staafdiagram |
| Continu, drie t/m tien categorieën | → | histogram/cumulatief histogram |
| Continu, meer dan tien categorieën | → | (cumulatieve) frequentiepolygoon |

## Oefeningen

**Oefening 10.1**

Welk meetniveau kenmerkt elk van de volgende variabelen?

a. opleiding naar het laatst behaalde diploma of certificaat;
b. opleiding naar aantal jaren;
c. politieke partij waarop je de laatste keer stemde;
d. rangorde van tijden op de 100 meter hardlopen;
e. politieke partijen naar hun positie op een links-rechtsschaal;
f. het gewicht van koffers op een weegschaal;
g. het gewicht van koffers zoals bepaald door een sterke man die je steeds het gewicht van twee koffers laat vergelijken;
h. het gewicht van koffers zoals bepaald door een sterke man, maar dan met de vraag of de twee koffers hetzelfde wegen, of verschillen;
i. de indeling van meetniveaus;
j. leeftijd (kinderen, jongvolwassenen, volwassenen);
k. leeftijd (geboortejaar);
l. de nummers van de 'boxen' in dit boek (hint: ze zijn er om gemakkelijk naar een box te kunnen verwijzen);
m. een aantal autotypen naar voorkeur, zoals genoemd door een surveyrespondent.

**Oefening 10.2**

Welke vorm heeft een datamatrix bij de oefeningen 10.1 g en h? Hoe ziet het patroon van de cellen bij opgave 10.1 g eruit, nadat je de data hebt verzameld en nadat je de rijen/kolommen in volgorde van gewicht hebt gezet?

**Oefening 10.3**

Temperatuur gemeten in Fahrenheit kan worden uitgedrukt in een Celsiusschaal. Wat is de formule die je daarbij gebruikt? Zoek het op of vind die zelf uit als je weet dat 60 °F = 15,6 °C, en 90 °F = 32.2 °C). Laat zien dat deze formule de afwezigheid van een nulpunt illustreert. Laat ook zien dat je zo'n formule in zijn algemene gedaante niet kunt gebruiken om de lengte van mensen gemeten in inches te transformeren naar centimeters.

**Oefening 10.4**

In veel vragenlijsten bij enquêtes worden beweringen gepresenteerd, en wordt aan de respondent gevraagd zijn of haar mening te geven door een van de op gelijke afstanden geplaatste hokjes aan te kruisen. Bijvoorbeeld:

'De AOW-plannen van het kabinet zijn oké.'
❑ helemaal mee eens
❑ mee eens

- ❑ niet mee eens, maar ook niet mee oneens
- ❑ niet mee eens
- ❑ helemaal niet mee eens

Om deze 'kruisjes' te scoren worden de getallen 1-5 aan de waarden toegekend. Wat is volgens jou het meetniveau van de variabele? Denk goed na over de argumenten voor en tegen 'intervalniveau'.

**Oefening 10.5** Leg uit waarom de volgende presentatie van riskante gewoonten van leerlingen van een school verkeerd is:

| | |
|---|---|
| Marihuanagebruik | 26,7 % |
| Alcoholgebruik | 50,3 % |
| Bovenmatig alcoholgebruik | 31,6 % |

**Oefening 10.6** Vereenvoudig op minstens twee manieren elk van de frequentieverdelingen hierna; geef de reden voor je keuze:

a. in Nederland wonende buitenlanders naar nationaliteit;
b. Nederlandse studenten naar gevolgde opleiding;
c. eindexamencijfers van een leerling;
d. doodsoorzaken (raadpleeg voor de oorspronkelijke cijfers de CBS-website!).

**Oefening 10.7** Bij een enquête naar sociale contacten wordt aan 81 respondenten gevraagd het aantal vrienden en bekenden aan te geven waaraan men minstens eenmaal per maand een bezoek brengt. De resultaten staan hierna:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3 | 5 | 2 | 3 | 3 | 4 | 1 | 8 | 4 |
| 2 | 4 | 2 | 5 | 3 | 3 | 3 | 0 | 3 |
| 5 | 6 | 4 | 3 | 2 | 2 | 6 | 3 | 5 |
| 4 | 14 | 3 | 5 | 6 | 3 | 4 | 2 | 4 |
| 9 | 4 | 1 | 4 | 2 | 4 | 3 | 5 | 0 |
| 4 | 3 | 5 | 7 | 3 | 5 | 6 | 2 | 2 |
| 5 | 4 | 2 | 3 | 6 | 1 | 3 | 16 | 5 |
| 3 | 11 | 4 | 5 | 19 | 4 | 5 | 2 | 2 |
| 4 | 3 | 14 | 5 | 2 | 1 | 4 | 3 | 4 |

a. Stel met SPSS een frequentieverdeling op; bereken het gemiddelde.
b. Hoe zou je deze frequentieverdeling in een rapport samenvatten? Geef argumenten voor deze samenvatting. Bereken opnieuw het gemiddelde (indien de laatste klasse 'open' is, kan dit eigenlijk niet; schat dan het gemiddelde in die klasse) en vergelijk met het eerder gevondene onder a.

**Oefening 10.8**

Maak een histogram bij de volgende IQ-verdeling. Dat kan met de hand, maar ook met Word!

| IQ | Frequentie |
|---|---|
| 82 – 87 | 2 |
| 88 – 93 | 23 |
| 94 – 99 | 22 |
| 100 – 105 | 65 |
| 106 – 111 | 20 |
| 112 – 117 | 10 |
| 118 – 123 | – |
| 124 – 129 | 8 |

**Oefening 10.9**

a. Waarom wordt bij het histogram als oorsprong altijd de nulwaarde gekozen voor de y-as, ook al ligt de laagste frequentie op bijvoorbeeld 30?
b. Hoe wordt het uiterlijk van een histogram veranderd als de verticale as wordt verlengd ten opzichte van de horizontale?
c. En andersom?
d. Als een histogram waarin een klasse met een afwijkende (dubbele) breedte voorkomt, moet worden omgezet in een cumulatief histogram, hoe zou je dan te werk gaan?
e. Beredeneer dat een 'leeftijdspiramide' (zie bijvoorbeeld *Statistisch Jaarboek*, 1990, p. 43) een 'dubbel' histogram is.
f. Een histogram kunnen we gemakkelijk in een frequentiepolygoon omzetten. Kan het omgekeerde ook?

**Oefening 10.10**

a. Waarom is het tekenen van een frequentiepolygoon alleen zinvol als de variabele op minstens intervalniveau gemeten is?
b. Kunnen we de vorm van een frequentiepolygoon afleiden uit de vorm van de cumulatieve frequentiepolygoon?
c. Als we twee frequentieverdelingen zouden willen vergelijken, waaraan geef je dan de voorkeur: aan een histogram of aan een polygoon?
d. Maak een relatieve frequentiepolygoon op basis van tabel 10.3, en in een volgende tekening de *ogive*.

**Oefening 10.11** Een bedrijf heeft 63 mannelijke en 182 vrouwelijke werknemers. De inkomensverdeling is als volgt:

| Maandsalaris in € | Frequentie | |
|---|---|---|
| | Mannen | Vrouwen |
| 1.000 tot 1.100 | 2 | 48 |
| 1.100 tot 1.200 | 4 | 62 |
| 1.200 tot 1.300 | 6 | 37 |
| 1.300 tot 1.400 | 24 | 16 |
| 1.400 tot 1.500 | 11 | 13 |
| 1.500 tot 1.600 | 5 | 4 |
| 1.600 tot 1.700 | 1 | 2 |

Maak in één tekening (kies een flink formaat) een cumulatieve frequentiepolygoon voor de mannen en idem voor de vrouwen.

**Oefening 10.12** Stel dat je een inkomensverdeling hebt van mannen en vrouwen. Hoe zou je deze grafisch weergeven? En een frequentieverdeling van het bezoek aan zes disco's in 2000, 2001 en 2002? En de mening van Nederlanders over het homohuwelijk, lopend van zeer afwijzend naar zeer positief?

**Oefening 10.13** De werkzame beroepsbevolking in Nederland naar leeftijd en geslacht in 2005 (in %):

| | Mannen | Vrouwen |
|---|---|---|
| 15-19 | 3,0 | 3,4 |
| 20-24 | 6,8 | 9,7 |
| 25-29 | 9,7 | 13,1 |
| 30-34 | 12,6 | 13,8 |
| 35-39 | 15,1 | 13,8 |
| 40-44 | 15,0 | 13,9 |
| 45-49 | 13,3 | 13,4 |
| 50-54 | 12,0 | 10,3 |
| 55-59 | 9,8 | 6,9 |
| 60-64 | 2,7 | 1,7 |
| Totaal | 100,0 | 100,0 |

a. Hoe zou je deze verdelingen grafisch willen voorstellen?
b. Als de beide N's bekend zijn; hoe zou je het dan willen doen?

## Literatuurtips

Algemene Nederlandstalige inleidingen:

Brink, W.P. van den & Koele, P. (2000). *Statistiek, Deel I: Datareductie* (8e druk). Amsterdam: Boom.

Buijs, A. (1993). *Statistiek om mee te werken*. Houten: Stenfert Kroese.

Ellis, J. (2004). *Statistiek voor de psychologie, Deel I*. Amsterdam: Boom.

Nijdam, B. & Buuren, H. van. (2002). *Statistiek voor de sociale wetenschappen – Deel I* (6e druk). Groningen: Wolters-Noordhoff.

Peet, A.A.J. van, Wittenboer, G.L.H. van & Hox, J.J. (1995). *Toegepaste statistiek. Deel 1: Beschrijvende technieken*. Groningen: Wolters-Noordhoff.

Slotboom, A. (1987). *Statistiek in woorden*. Groningen: Wolters-Noordhoff.

Specifiek gericht op het juist (en onjuist!) samenvatten en representeren van gegevens:

Ehrenberg, A.S.C. (1982). *A primer in data reduction: An introductory statistics textbook*. London: Wiley. Dit is een voortreffelijk boek, waarin, zonder dat van ingewikkelde statistiek sprake is, uitgelegd wordt hoe je met soms grote hoeveelheden numerieke data moet omgaan om deze overzichtelijk te maken en de kern ervan boven tafel te krijgen.

Hoaglin, D.C., Mosteller, F. & Tukey, J.W. (Eds.). (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

Holmes, N. (1984). *Designer's guide to creating charts and diagrams*. New York: Watson- Guftill. Een indrukwekkende verzameling grafische statistische voorstellingen uit de commercieel/artistieke hoek.

Huff, D. (1975). *How to lie with statistics*. London: Pelican Pocket. Een fameus boekje over misleiding door middel van grafische voorstellingen.

Maanen, H. van. (2009). *Goochelen met getallen. Cijfers en statistiek in krant en wetenschap.* Amsterdam: Boom.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Velleman, P.F. & Hoaglin, D.C. (1981). *Applications, basics and computing of exploratory data analysis*. Boston: Duxbury.

**Kernbegrippen**

spelregels voor tabelconstructie
cumulatief percentage
drie redenen voor vereenvoudiging
gegroepeerde frequentieverdeling
klassen, onder- en bovengrenzen
staafdiagram
lijngrafiek
cirkeldiagram
stroomdiagram
symboolvoorstelling
histogram
frequentiepolygoon
cumulatief histogram
cumulatieve frequentiepolygoon (ogive)