Maten voor het centrum, de spreiding en de vorm van een frequentieverdeling

11

In dit hoofdstuk bespreken we enkele kengetallen van een frequentieverdeling, zoals het gemiddelde, de mediaan en de standaardafwijking. We gebruiken zulke kengetallen om de frequentieverdeling van een variabele kort en bondig samen te vatten. We beginnen met het gemiddelde. Het gemiddelde is een maat voor het centrum van een verdeling, die vaak samen gerapporteerd wordt met een maat voor de spreiding rondom het gemiddelde, de standaardafwijking. Daarna spreken we over de vorm van een frequentieverdeling. We gebruiken liever de mediaan dan het gemiddelde als centrummaat als we te maken hebben met scheve, dus niet-symmetrische, verdelingen. De mediaan heeft ook een bijpassende maat voor spreiding, de zogeheten interkwartielafstand. Vervolgens wordt het basisprincipe van 'boxplots' uit de doeken gedaan. Een boxplot van een frequentieverdeling geeft ons een snelle indruk van een complete verdeling doordat diverse van de voornoemde maten in één plaatje te zien zijn. Het hoofdstuk wordt afgesloten met een leidraad om van geval tot geval de goede centrum- en spreidingsmaten te kiezen.

Inleiding

Tabellen en grafieken vormen op zich al vereenvoudigde voorstellingen van de werkelijkheid. Toch willen we vaak nog verder vereenvoudigen, vooral wanneer we met een groot aantal tabellen te maken hebben, of als het aantal waarden of klassen erg groot is. Als we een frequentieverdeling hebben met maar twee tot vijf waarden heeft het weinig zin om een centrummaat te gebruiken; zo'n frequentieverdeling is gemakkelijk te overzien, en aan reductie bestaat weinig behoefte. Maar als we bijvoorbeeld de gedetailleerde frequentieverdeling naar leeftijd van een groep mannen willen vergelijken met die van een groep vrouwen, is dit maar een moeizame bezigheid. We zijn meestal niet geïnteresseerd in alle specifieke verschilletjes, maar we zouden de hele verzameling mannen en de hele verzameling vrouwen elk willen typeren door een represen-

tatief getal, om vervolgens die twee getallen te vergelijken. Zo'n getal is dan een *model* (een datamodel dus) voor alle waarnemingen in die verzameling. Aan de andere kant moeten we bedenken dat elke vereenvoudiging of samenvatting van een frequentieverdeling verlies aan informatie met zich meebrengt. En soms is een frequentieverdeling te bijzonder of te ingewikkeld om in een of twee karakteristieke maten samen te vatten. Als dat het geval is, hebben we nog meer karakteristieke getallen nodig, of moeten we er zelfs wat uitleggende zinnetjes aan toevoegen.
Samenvattende maten onderscheiden we in:
- centrummaten;
- maten voor de spreiding;
- maten voor de vorm: bijvoorbeeld of de verdeling meer of minder scheef is.

Centrum- en spreidingsmaten bespreken we tegelijk omdat ze duidelijk bij elkaar horen. Daarna pas komen enkele aspecten van de vorm aan de orde.

Centrum en spreiding

Aan welke eisen moet een goede centrummaat, of maat voor centrale tendentie (want dat is de wat plechtiger naam van deze datamodellen) voldoen? Een centrummaat:
- moet een frequent voorkomende waarde of categorie zijn;
- moet ergens in het midden van de frequentieverdeling passen;
- moet zodanig worden gekozen dat de som van de afstanden tussen die maat en alle andere waarnemingen zo klein mogelijk is.

Deze formuleringen zijn niet erg precies, maar ze geven een idee van wat de bedoeling is. Maten voor het centrum van een verdeling zijn het gemiddelde, de mediaan en de modus. Hierna staan de definities:
- Het *gemiddelde* is de som van de scores gedeeld door hun aantal.
- De *mediaan* is de score van de middelste waarneming nadat alle waarnemingen in rangorde zijn geplaatst.
- De *modus* is de meest voorkomende score.

Naast centrummaten hebben we behoefte aan *maten voor spreiding*: zij drukken uit in hoeverre de waarnemingen op een kluitje rondom de centrummaat liggen, dan wel of ze verspreid liggen over de hele reikwijdte van de variabele. Variatie van waarnemingen is de basis voor alle statistiek. Als alle eenheden hetzelfde zouden zijn, is het overbodig om het gemiddelde of wat dan ook te

berekenen, omdat elke waarneming dezelfde score heeft. Een spreidingsmaat drukt dus uit in welke mate een centrummaat een goed model is voor de frequentieverdeling. Hoe kleiner de spreiding, des te geschikter de centrummaat is als model.

> Veronderstel dat zes jongens als volgt op een schoolvak scoren: 3, 5, 6, 6, 7 en 9, terwijl zes meisjes de volgende scores hebben: 5, 5, 6, 6, 7 en 7. Bij beide groepen is de gemiddelde score dan 6, maar de spreiding bij de meisjes is veel kleiner dan de spreiding bij de jongens. Voor de meisjes is de 6 een aardig representatief cijfer; voor de jongens is dat veel minder het geval.
> Ook kan het voor bepaalde onderzoeksdoeleinden zonder meer noodzakelijk zijn de spreiding te berekenen. Als we bijvoorbeeld geïnteresseerd zijn in het verschijnsel 'inkomensnivellering' tussen 1990 en 2000, zegt een centrummaat berekend op elk van deze twee tijdstippen niets, maar gaat het om de vraag of de spreiding van het ene tijdstip naar het andere kleiner of groter is geworden.

Een heel eenvoudige spreidingsmaat is uiteraard de *spreidingsbreedte* of variatiebreedte (*range*): het verschil tussen de hoogste en de laagste score (*maximum* en *minimum*). Omdat deze maat echter van slechts twee (misschien toevallig heel extreme) waarnemingen afhangt en niets zegt over de mate van clustering van alle andere, wordt hij weinig gebruikt. Andere spreidingsmaten zijn de *standaardafwijking* (die behoort bij het gemiddelde als centrummaat) en de *interkwartielafstand* (die behoort bij de mediaan als centrummaat). De definities volgen verderop.

In SPSS krijgen we – na aanklikken van ANALYZE in de menubalk, en DESCRIPTIVE STATISTICS – via DESCRIPTIVES een staatje te zien met daarin het totaal N, de hoogste en de laagste score, het gemiddelde en de standaardafwijking. Daar hebben we nog niet zo veel aan. Klikken we niet op DESCRIPTIVES, maar op FREQUENCIES, dan krijgen we veel meer te zien, als we tenminste het hulpmenu STATISTICS aanklikken. We zien nu een hele reeks maten, zoals kwartielen en (per)centielen, spreidingsmaten, zoals de standaardafwijking, de variantie, de *range* en de minimum- en maximumwaarde, centrummaten, zoals het gemiddelde, de mediaan en de modus, en bijzondere maten, zoals voor de spitsheid (*kurtosis*) en de scheefheid (*skewness*). Zouden we niet op DESCRIPTIVES of FREQUENCIES klikken, maar op EXPLORE, dan krijgen we een soortgelijk, maar iets ander, overzicht.

Welke centrum- en spreidingsmaat we gebruiken, hangt af van:
– het niveau waarop de variabele gemeten is;
– de vorm van de frequentieverdeling.

We beginnen met als centrummaat het gemiddelde, en als spreidingsmaat de standaardafwijking.

Het gemiddelde

Hoe we het rekenkundig gemiddelde van enkele getallen moeten berekenen, is bekend. Zo is de gemiddelde waarde van de drie waarnemingen 2, 6 en 7 gelijk aan (2 + 6 + 7) / 3 = 5; zie voor een voorbeeld tabel 11.1.

Tabel 11.1 Gewicht van 79 poststukken

Kg	Frequentie	
1	2	$\bar{X}$ = (2x1 + 6x2 + 17x3 + 29x4 + 16x5 + 8x6 + 1x7) / 79 = 4 kg.
		Let op streepje op de X!
2	6	
3	17	Toelichting op de berekening: uiteraard tellen we niet, als meerdere
4	29	eenheden dezelfde waarde (gewicht) hebben, al die gewichten
5	16	afzonderlijk, maar we vermenigvuldigen met de frequentie van
		voorkomen.
6	8	
7	1	
N	79	

In formuletaal:

$$\bar{X} = \sum_{i=1}^{N} X_i / N$$

Hierbij geldt:

$\bar{X}$ = het rekenkundig gemiddelde
X_i = de score van eenheid i op variabele X
N = het totaal aantal eenheden

$\sum_{i=1}^{N}$ = de som van de scores van alle N eenheden (omdat i alle waarden doorloopt van de eerste ($_i$ = 1) tot en met de laatste waarneming ($_i$ = N).

Als we achteraf 'met de hand' bij een gegroepeerde frequentieverdeling het gemiddelde willen berekenen, wordt het *klasmidden* aangemerkt als de score van *alle* eenheden in die klas.

Tabel 11.2 Lengte van een verzameling personen

Lengte	Klasmidden	Frequentie	
160 tot 170 cm	165	34	$\bar{X}$ = (34x165 + 93x175 +
170 tot 180 cm	175	93	70x185) / 197 =
180 tot 190 cm	185	70	176,82 is ongeveer 177 cm
N		197	

Merk op dat bij het berekenen van het gemiddelde op gegroepeerde gegevens een flinke vergroving optreedt. Zo wordt de score van alle 93 eenheden in de klas 170 tot 180 op 175 gesteld, terwijl van de oorspronkelijke scores best het grootste deel van de eenheden bijvoorbeeld score 173 of score 178 zou kunnen hebben. Het gemiddelde op basis van de gegroepeerde gegevens kan dus verschillen van het gemiddelde op basis van de oorspronkelijke gegevens. Bedenk dat het maken van een gegroepeerde frequentieverdeling achteraf gebeurt, en alleen ter wille van een overzichtelijke rapportage. De *berekening* van gemiddelden, en andere maten, door de computer vindt altijd plaats *op basis van de ruwe data*.

Variantie en standaardafwijking

Het belang van een centrummaat wordt mede bepaald door de mate van spreiding rondom die centrummaat. Naarmate die spreiding kleiner is, is de centrummaat meer *typerend* voor alle eenheden. Als we het gemiddelde als centrummaat gebruiken, is de mate van spreiding rondom dat gemiddelde de basis voor de berekening van een spreidingsmaat. We bepalen eerst voor elke waarneming X_i de afwijking van het gemiddelde X; dit wordt de *afwijkingsscore* genoemd. De variantie nu is het gemiddelde van de gekwadrateerde afwijkingen rondom het gemiddelde.

De variantie $S_x^2 = \frac{\Sigma (X_i - \bar{X})^2}{N}$

is dan en alleen dan gelijk aan 0 indien geen enkele waarneming van het gemiddelde afwijkt.[1]

De variantie wordt groter naarmate de waarnemingen meer van elkaar en dus ook van het gemiddelde afwijken.

> Waarom zo'n ingewikkelde formule? Waarom niet gewoon de afwijkingen bij elkaar opgeteld en gedeeld door N? Welnu, omdat de positieve afwijkingen tegen de negatieve wegvallen, komen we dan altijd op 0 uit. En als we nu eens de absolute afwijkingen zouden optellen en delen door N? We zijn dan inderdaad van dat probleem van het tegen elkaar wegvallen af, maar we geven toch de voorkeur aan de wat ingewikkelder formule, omdat we de extreme scores zwaarder willen laten wegen en omdat kwadratering voorts bepaalde algebraïsche voordelen oplevert.

Een bezwaar van S_x^2 is dat hij in vierkante eenheden is gemeten. Als een lengteverdeling is uitgedrukt in centimeters, is de variantie uitgedrukt in vierkante centimeters. We kunnen het ook zo begrijpen: als we de 'afstanden' tussen de scores verdubbelen (bijvoorbeeld bij de scores 2, 4 en 8 in plaats van 1, 2 en 4) wordt de variantie niet tweemaal zo groot, maar viermaal (reken maar na!). Dit is de reden dat als beschrijvende spreidingsmaat doorgaans niet de variantie zelf, maar de wortel daaruit wordt gebruikt, die we de *standaardafwijking* noemen (*standard deviation*, in SPSS: *stddev*):

$$S_x = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{N}}$$

De standaardafwijking van de frequentieverdeling van tabel 11.1 berekenen we als volgt:

1 Gaan we ervan uit dat onze data een steekproef uit een populatie betreffen, dan geeft deze formule als we een kleine steekproef (n <20) hebben, een geringe onderschatting van de variantie in de populatie. We corrigeren dit door in de noemer N-1 in plaats van N te gebruiken. Aangezien dit voor grotere steekproeven toch nauwelijks verschil maakt, wordt in handboeken en in software (en bij handrekenmachientjes!) veelal N-1 in plaats van N gebruikt.

$\bar{X} = \Sigma X_i / N = 4$ kg (reeds berekend)

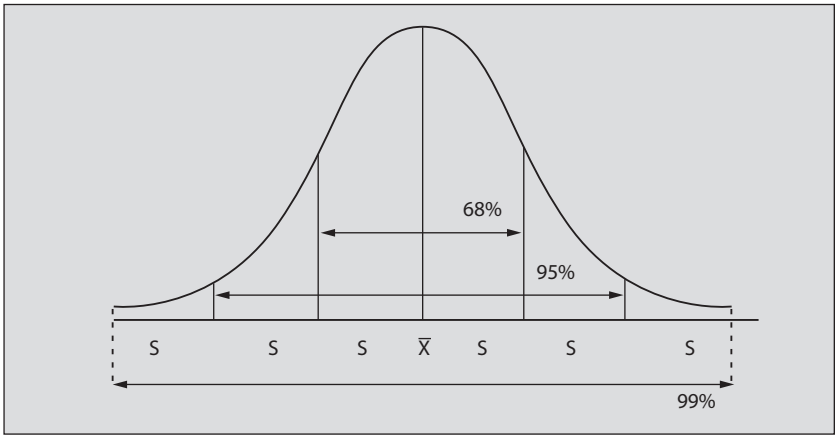
$S_x^2 = ((1-4)^2 \times 2 + (2-4)^2 \times 6 + (3-4)^2 \times 17 + (5-4)^2 \times 16 + (6-4)^2 \times 8 + (7-4)^2 \times 1) / 79$

$= (9x2 + 4x6 + 1x17 + 1x16 + 4x8 + 9x1) / 79$

$= 116 / 79 = 1{,}468$

$S_x = \sqrt{1{,}468} = 1{,}21$

De standaardafwijking heeft een belangrijke interpretatie in termen van de kansrekening en de normale verdeling. In een zogeheten normale verdeling (een Gauss- of klokvormige verdeling) ligt 68% van de waarnemingen binnen één standaardafwijking links en rechts van het gemiddelde, 95% van de eenheden binnen twee standaardafwijkingen links en rechts van het gemiddelde, en 99% binnen drie standaardafwijkingen links en rechts van het gemiddelde. Ook wanneer een verdeling niet helemaal normaal is, is deze interpretatie een bruikbare vuistregel. Onderzoekers hebben deze getallen in het hoofd als ze bekijken of een standaardafwijking van een verdeling groot of klein is. Bij oefening 11.2 weet je uit de gegevens dat dus zo'n 68% van de studenten tussen de 18 en 22 jaar zal zijn, en zo'n 95% tussen de 16 en 24 jaar, vooropgesteld dat het een grote groep is. De normale verdeling komt in de werkelijkheid veel voor, bijvoorbeeld als we spreken over lengte of gewicht van mensen, over afwijkingen van het gemiddelde bij industriële producten, over de opbrengst van landbouwgronden bij een bepaalde bemesting, of als je driehonderd keer een dobbelsteen opgooit en het aantal keren '4' zou noteren, enzovoort.



Figuur 11.1 De normale verdeling

Standaardscores

Het gemiddelde en de standaardafwijking zijn belangrijke karakteristieken van een verdeling. Met behulp van deze twee maten kunnen we uitstekend de *relatieve plaats* van een eenheid op een variabele bepalen, dat wil zeggen dat we kunnen bepalen of een eenheid veel of weinig boven of beneden het gemiddelde van een verzameling eenheden ligt. Denk aan de score van een leerling ten opzichte van alle anderen in de klas.

Tabel 11.3 Gewicht en lengte van 200 gekeurde personen

Gewicht in kg	Frequentie	Lengte in cm	Frequentie
50 - 54	5	161-164	1
55 - 59	8	165-168	4
60 - 64	14	169-172	10
65 - 69	30	173-176	21
70 - 74	42	177-180	35
75 - 79	46	181-184	48
80 - 84	28	185-188	40
85 - 89	14	189-192	28
90 - 94	8	193-196	12
95 - 99	5	197-200	1

Het gemiddelde gewicht blijkt na berekening te zijn 74,475 kg; de standaardafwijking is 9,53 kg. De gemiddelde lengte is 182,68 cm, met als standaardafwijking: 6,77 cm. Veronderstel nu dat een van de gekeurde personen – hij weegt 62 kg en zijn lengte is 170 cm – wil weten of zijn 'scores' erg afwijkend zijn. Nu is zijn afwijking van het gemiddelde op de beide variabelen snel berekend, maar als de standaardafwijking van een verdeling groter is, weegt als het ware zo'n afwijking minder sterk. Daarom delen we het verschil met het gemiddelde door de standaardafwijking van de frequentieverdeling, en komen dan tot het begrip *standaardscore*.

$$z_x = (X_i - \bar{X}) / S_x$$

Hier: de standaardscore van deze persoon voor gewicht is (62-74,5) / 9,53 = -1,31; zijn standaardscore voor lengte is (170-182,68) / 6,77 = -1,87. Dat wil dus zeggen dat hij voor lengte veel verder onder het groepsgemiddelde zit dan voor gewicht. Standaardscores van één persoon op verschillende variabelen zijn onderling vergelijkbaar, omdat alle in standaardscores omgezette frequentieverdelingen eenzelfde gemiddelde (= 0) en standaardafwijking (= 1) hebben.

Box 11.1

Het gemiddelde is in de eerste plaats bedoeld voor continue variabelen. Bij een continue variabele kunnen er altijd waarden worden gevoegd tussen de al gegeven waarden (intervallen kunnen verder worden verdeeld, zoals centimeters in millimeters). De meeste variabelen op interval- en rationiveau zijn continu.

Maar het gemiddelde kan ook een goede beschrijvende centrale maat zijn bij discrete variabelen, zoals 'aantal kinderen per gezin' of 'aantal verenigingen waarvan iemand lid is'. Het ziet er wat vreemd uit als je zegt dat het gemiddeld aantal kinderen 2,4 is, of dat mensen van gemiddeld 4,17 verenigingen lid zijn. Maar we zijn langzamerhand gewend geraakt aan zulke gebroken getallen. Ze worden natuurlijk gebruikt omdat het gebruik van alleen hele getallen veel te grof is. Wat we hier eigenlijk doen, is het behandelen van een continue variabele alsof hij discreet is. En waarom zouden we dat niet doen, als het de precisie ten goede komt?

De vorm van een frequentieverdeling

We weten nu wat het gemiddelde en wat de standaardafwijking is, en hoe ze berekend worden. Maar zijn het altijd geschikte maten?

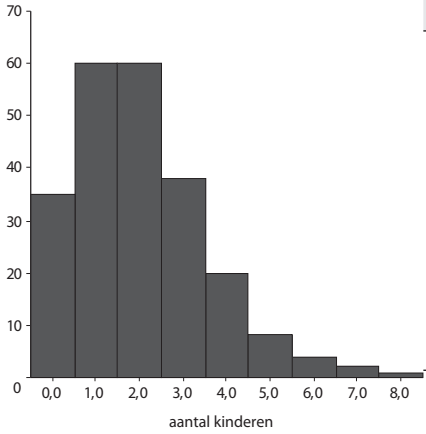
Bij een mooie, symmetrische, verdeling zoals die in tabel 11.1, hebben we er weinig moeite mee de frequentieverdeling samen te vatten in het gemiddelde. Het gemiddelde is immers bij *deze* frequentieverdeling niet zo gek: er zijn ongeveer even veel, en even grote, afwijkingen naar boven en naar beneden; het gemiddelde is bovendien een waarde die veel meer voorkomt dan alle andere waarden, en zo zijn er nog een paar eigenschappen. Maar helaas, vaak missen verdelingen een dergelijke fraaie vorm. We noemen enkele voorbeelden.

a. Een verdeling kan meer of minder scheef zijn.
b. Een verdeling kan heel vlak zijn; in alle categorieën zitten ongeveer evenveel waarnemingen. Het gemiddelde is dan niet een bijzondere, eruit springende maat.[2]
c. Een verdeling heeft niet één duidelijke top, maar twee of meer pieken.
d. Soms zijn er enkele waarnemingen met een heel hoge of juist een heel lage score, de zogenoemde uitschieters of uitbijters.

2 'Spitsheid' of, het omgekeerde, 'platheid' van een frequentieverdeling duiden we aan met de term *kurtosis*. Een positieve kurtosis wil zeggen dat de verdeling een nogal hoge spits heeft (zo'n verdeling noemen we leptokurtosisch; *lepto* betekent dun) in vergelijking met een normale verdeling, die een kurtosis van 0 heeft (mesokurtosisch; *meso* betekent midden). Een negatieve kurtosis wil zeggen dat de verdeling vlakker is dan een normale verdeling; we noemen zo'n verdeling platykurtosisch (*platy* = vlak).

En dan zijn er natuurlijk nog de vele variabelen die op het nominale of ordinale niveau zijn gemeten. Bij deze variabelen hebben het gemiddelde en de standaardafwijking geen enkele zin. Wat we met deze variabelen doen, wordt verderop besproken. In al deze gevallen worden gemiddelde en standaardafwijking dus niet gebruikt.

Tabel 11.4 en figuur 11.2 laten ons een scheve verdeling zien. We kunnen het gemiddelde berekenen, maar er liggen veel meer waarnemingen *onder* dan *boven* het gemiddelde. Het gemiddelde (2,0 kinderen; reken na) is gevoelig voor de 'lange staart' van de verdeling. Qua aantallen gezinnen is deze staart niet belangrijk, maar door de vermenigvuldiging met de relatief grote getallen 6, 7 en 8 wordt het gemiddelde sterk door de staart bepaald.

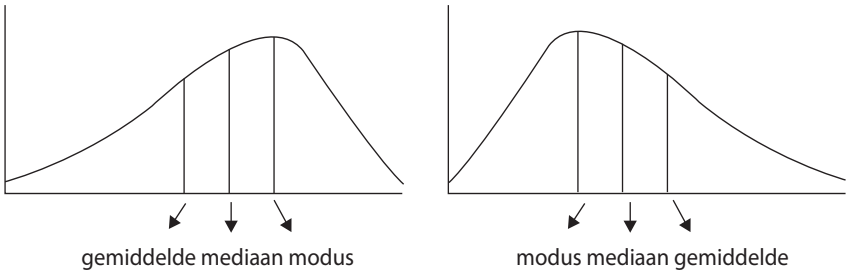


Aantal kinderen	Frequentie
0	35
1	60
2	60
3	38
4	20
5	8
6	4
7	2
8	1

Figuur 11.2 en tabel 11.4 Een scheve verdeling: gezinnen naar aantal kinderen

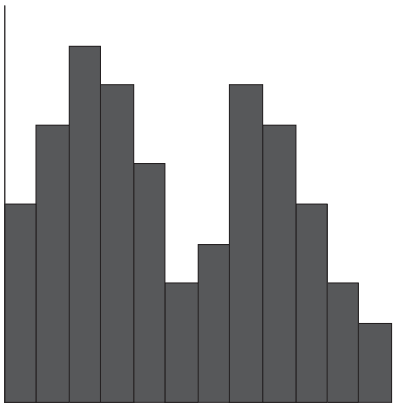
Typisch scheve verdelingen vind je in het GSS91-bestand bij de variabelen 'aantal broers en zusters' en 'inkomen'. *Laat jezelf wat statistieken van deze variabelen zien, en maak ook van elk een histogram.* Inkomensverdelingen zijn in ontwikkelingslanden heel scheef, maar zijn ook scheef in alle Europese landen. Andere scheve verdelingen betreffen bijvoorbeeld het aantal aidspatiënten dat iemand kent, en het aantal sekspartners. De scheefheid (*skewness*) wordt voor je berekend in SPSS. De scheefheid hoort te liggen tussen -1 en +1, wil je het gemiddelde mogen berekenen. Als een variabele schever is verdeeld, is een van de mogelijkheden om deze te transformeren naar een meer normale verdeling door de logaritmen te nemen, en het gemiddelde van de logscores te gebruiken. Maar voor veel mensen is dit een stap te ver. Voor eenvoudige beschrijvende doeleinden doet men er beter aan om zowel de mediaan als het gemiddelde te

rapporteren. Als deze centrummaten flink van elkaar verschillen is dat een gevolg van een scheve verdeling.[3]
Een verdeling kan links of rechts scheef zijn (zie figuur 11.3).



Figuur 11.3 Een links scheve en een rechts scheve verdeling

Evenmin geslaagd als centrummaat is het gemiddelde bij een verdeling zoals in figuur 11.4. Hier is het gemiddelde een relatief weinig voorkomende waarneming. Het merkwaardige van deze verdeling is de tweetoppigheid, en het is daarom goed dit aspect, als we denken aan een beknopte samenvatting van de verdeling in een 'model', vast te houden, bijvoorbeeld door de twee toppen (twee *modi*) apart te noemen. Meertoppigheid wijst er vaak op dat de steekproef uit twee verschillende populaties is getrokken, elk met hun eigen top. Zoek in ieder geval naar een interpretatie van zo'n merkwaardige verdeling.



Figuur 11.4 Een meertoppige verdeling

3 De grootte van het verschil tussen gemiddelde en mediaan is een indicator voor de scheefheid van de verdeling. Een formele maat voor scheefheid is 3 x (gemiddelde minus de mediaan) / stdafw. In een links scheve verdeling is het gemiddelde kleiner dan de mediaan, en is deze waarde negatief; in een rechts scheve verdeling ligt het gemiddelde rechts van de mediaan, en is de scheefheid positief. In een symmetrische verdeling is hij gelijk aan 0, omdat mediaan en gemiddelde samenvallen (tenzij er uitschieters zijn, zoals in tabel 11.5).

Een soortgelijk verschijnsel treedt op als er bij een overigens redelijke verdeling een uitschieter is (of enkele uitschieters zijn); zie tabel 11.5. Ook in zo'n geval kiezen we de mediaan als centrummaat.

Tabel 11.5 Een verzameling werknemers naar inkomen

Inkomen	Frequentie
€ 3.000	1
€ 5.000	4
€ 7.000	1
€ 8.000	1
€ 9.000	1
€ 10.000	1
€ 15.000	1
€ 38.000	1
N	11

Wat is hier een goede centrummaat? Een vakbondsman zou wellicht zeggen: € 5.000 (de modus); een werkgever zou antwoorden: € 10.000 (gemiddelde); een bemiddelaar houdt, zoals te verwachten, het juiste midden en geeft de mediaan (€ 7.000). Maar het gemiddelde is sterk beïnvloed door die ene uitschieter van € 38.000 en is hier per se niet goed. De mediaan is niet gevoelig voor uitschieters; ook als de hoogste score € 380.000 zou zijn, blijft de mediaan € 7.000. Een beter, maar iets minder overzichtelijk 'model' van deze verdeling zou dan ook de constatering zijn dat het gemiddelde inkomen € 6.200 bedraagt, met de vermelding dat één persoon met een inkomen van € 38.000 buiten beschouwing is gelaten. Het 'data'-model is daarmee wat ingewikkelder dan we zouden wensen (niet één kenmerkende maat, maar twee gegevens!), maar het is veel beter dan blindelings het gemiddelde berekenen.

Naast al deze bijzondere eigenschappen van verdelingen is er nog een ander aspect dat hier genoemd moet worden, namelijk het al dan niet aanwezig zijn van open klassen.
Bij open klassen (zie bijvoorbeeld tabel 11.5) *kan* het gemiddelde niet worden berekend, omdat het klassenmidden niet bepaald kan worden. Overigens hangt dit een beetje van de frequentieverdeling af. Stel dat we een leeftijdsverdeling van mensen hebben met een klasbreedte van vijf jaar en een hoogste klas van '95 jaar en ouder', die uiteraard slechts een gering percentage van de waarnemingen omvat, dan is er weinig bezwaar tegen om deze klas een fictief

klasmidden van 97,5 te geven en vervolgens het gemiddelde te berekenen. Maar is de hoogste klasse '65 jaar en ouder', en omvat deze een behoorlijk percentage van de waarnemingen, dan zou een gegokt klasmidden van 67,5 veel te laag zijn. In dat geval kunnen we beter de mediaan gebruiken als centrummaat.

Andere centrummaten: modus en mediaan

Hiervoor zagen we dat onder bepaalde omstandigheden het gemiddelde niet geschikt is als centrummaat. We bespraken deze omstandigheden uitvoerig, maar gingen slechts kort in op het meetniveau. We komen daar nu op terug.

Op het nominale niveau is het gemiddelde zinloos, omdat de getallen die gebruikt worden om de waarden van de variabele aan te duiden, niet meer zijn dan kengetallen, zoals telefoonnummers. Het heeft duidelijk weinig zin om het gemiddelde van drie telefoonnummers uit te rekenen. De mediaan is ook zinloos omdat er geen vaste rangorde van waarden is. Daarom blijft alleen de *modus* over. Met het noemen van de modus worden vooral twee doeleinden bereikt:

- Het 'meest karakteristieke' van een verdeling wordt aangegeven, bijvoorbeeld de haardracht of het menu van een bevolkingsgroep, de 'meest typerende' boot op de Loosdrechtse Plassen, het doorsneetenttype op een camping, de populairste politieke partij in een gemeente, het meest verkochte automerk in een bepaald jaar.
- Het is een maat die in de werkelijkheid ook voorkomt: een gemiddelde of mediane autobandenmaat kunnen we wel berekenen, maar we hebben meer aan het 'meest verkochte type'.

Is het altijd zinvol om de modus (bij een gegroepeerde verdeling spreken we van een *modale klasse*) te vermelden? We moeten daarvoor goed naar de verdeling kijken. De modus is helaas een nogal instabiele maat. Bij een gegroepeerde frequentieverdeling is hij erg gevoelig voor de ligging van de klassengrenzen. Als de modus niet heel duidelijk een 'piek' vertoont ten opzichte van alle andere waarden heeft het weinig zin om de modus te vermelden. Als alle waarden ongeveer evenveel voorkomen, spreken we van een vlakke verdeling (denk aan een staafdiagram met even lange staven). Als er twee, of misschien wel drie waarden zijn die veel sterker gevuld zijn dan de andere waarden, terwijl ze onderling niet veel verschillen in frequentie, vermelden we niet de ene modus, omdat die misschien net iets meer waarnemingen telt dan de andere, maar is het verstandiger ze alle twee of drie apart te noemen.

Als we nadenken over een *spreidingsmaat* op het nominale niveau, letten we op frequentieverschillen tussen de verschillende waarden. Intuïtief is duidelijk dat de spreiding het geringst is als alle waarnemingen in één categorie zitten, en dat de spreiding het grootst is als de waarnemingen gelijkelijk over alle categorieën verdeeld zijn (een perfect vlakke verdeling). Op dit principe zijn wel wat maten gebaseerd, maar ze zijn te ingewikkeld om hier uit te leggen en het loont nauwelijks om deze te gebruiken. Willen we toch iets over spreiding zeggen, dan is het natuurlijk nuttig om niet alleen *de modus* te noemen, maar ook *het percentage van de waarnemingen dat in de modus valt*. Het maakt nogal wat uit of 65% in de modus of modale klasse valt of 30%. Vaak ziet men ook een uitdrukking als: '80% van de waarnemingen valt in de categorieën a, b en c', waarmee aangegeven wordt dat er – op een totaal van misschien wel twaalf categorieën – slechts drie categorieën zijn die samen toch het leeuwendeel van de waarnemingen trekken. Zo'n categorieënstelsel kan bijvoorbeeld horen bij een verzameling motieven voor een bepaald gedrag, of bij een verzameling vakantielanden. Zo'n uitdrukking is dan een mooie reductie van een onoverzichtelijke tabel. Van deze redenering leren we ook dat het reduceren van een frequentieverdeling tot een paar 'kengetallen' niet iets mechanisch is, waarbij je simpel een paar regeltjes volgt. Soms moeten we ook constateren dat een frequentieverdeling te ingewikkeld of te bijzonder is om in één of enkele kengetallen samen te vatten. In zo'n geval doen we dat dan ook niet!

In dezelfde situaties als hiervoor genoemd, gebruiken we ook bij ordinale variabelen de modus. Heeft het zin om – daarnaast – de mediaan te gebruiken? Immers, als een variabele op ordinaal niveau gemeten is, betekent dit dat de waarnemingen in rangorde kunnen worden geplaatst, en we zouden dus kunnen nagaan welke de waarde is die evenveel waarnemingen boven als onder zich heeft; die waarde zouden we met recht als 'centrummaat' kunnen zien.

Een heel bekend type ordinale variabelen zijn enquêtevragen met een vaste reeks van geordende antwoordcategorieën, zoals al vermeld in hoofdstuk 6, bijvoorbeeld:

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

De bedoeling is dat de respondent een van de vijf hokjes, die keurig op gelijke afstanden van elkaar staan, aankruist. Onderzoek heeft aangetoond dat de

afstanden tussen de antwoordcategorieën als nagenoeg gelijk worden gezien door de respondenten. *Daarom worden dergelijke variabelen meestal als gemeten op intervalniveau beschouwd*. De waarde 'helemaal mee eens' krijgt dan een 1, de volgende een 2, en de laatste (helemaal niet mee eens) een 5. De mediaan wordt bij zulke enquêtevragen weinig gebruikt; vrijwel altijd wordt het gemiddelde berekend. Een twijfelgeval zijn schoolcijfers op een schaal van 1-10. Het is natuurlijk de bedoeling dat de verschillen tussen 2 en 3 en die tussen 6 en 7 aan elkaar gelijk zijn; met andere woorden: de bedoeling is een intervalschaal. Aangetoond is echter dat er veel meer zessen en veel minder vijven door leraren worden uitgedeeld dan 'volgens statistische verwachting' het geval zou zijn, wat erop wijst dat het interval tussen 5 en 6 door de leraren als groter wordt beoordeeld dan bijvoorbeeld dat tussen 7 en 8 (geen wonder, want tussen 5 en 6 ligt de scheiding tussen onvoldoende en voldoende!). Als we van twee klassen de prestaties vergelijken, nemen we dan ook het best de medianen. Maar ... heel vaak wordt dit verschijnsel verwaarloosd en berekent men toch gewoon gemiddelden.

Zijn er nog andere ordinale variabelen in de praktijk van het sociaalwetenschappelijk onderzoek? De belangrijkste zijn opleiding naar schooltype, en rangenstelsels. Deze variabelen zijn discreet en kennen meestal een gering aantal categorieën. Bij de belangrijke variabele 'laatst gevolgde opleiding' krijgen we categorieën als 'vmbo', 'havo', 'gymnasium', 'hbo' enzovoort; een typisch ordinale variabele. Als we de frequentieverdeling naar opleiding van een verzameling mensen zouden willen samenvatten in een centrummaat, is alleen de modus soms goed bruikbaar (bijvoorbeeld als meer dan 45% 'havo' noemt). Een mediaan is nietszeggend, dus we volstaan met de modus, of we zouden van de steekproef kunnen zeggen dat bijvoorbeeld 70% van de mensen havo of minder heeft.

Conclusie: de meest voorkomende ordinale variabelen, namelijk enquêtevragen voorzien van een vaste reeks geordende antwoordcategorieën, worden meestal behandeld als variabelen op intervalniveau; als centrummaat berekenen we gewoonlijk het gemiddelde. Bij de overige ordinale variabelen is de situatie bijna hetzelfde als bij nominaal gemeten variabelen, dat wil zeggen dat er weinig keus is: alleen de modus komt in aanmerking. En wat voor centrummaten op ordinaal niveau geldt, geldt ook voor spreidingsmaten: er is weinig specifieks op ordinaal niveau.

Bij variabelen op interval- en rationiveau kunnen in principe de modus, de mediaan en het gemiddelde worden gebruikt, elk onder de voornoemde restric-

ties. Eerder zagen we al dat de mediaan heel geschikt is voor scheef verdeelde variabelen, zoals het inkomen. Er zijn nu eenmaal relatief weinig mensen met een hoog inkomen, maar er zijn veel mensen in de lagere inkomensklassen. Het inkomensniveau waarboven 50% en waaronder 50% van de inkomens ligt, lijkt daarom een heel geschikte maat. Ook is de mediaan ongevoelig voor uitschieters. Om een scheve verdeling samen te vatten in een enkel getal is de mediaan ongetwijfeld de beste centrummaat. Maar kan de mediaan altijd worden berekend?

Voor gegroepeerde frequentieverdelingen is er een probleem (zie tabel 11.6). De mediaan ligt bijna altijd ergens in een klas. Maar als je zou zeggen: de mediaan van de inkomensverdeling in tabel 11.6 ligt ergens in de klas € 3.000 tot € 4.000 (immers, zowel de 110e als de 111e waarneming ligt in deze klas), dan is die conclusie niet erg informatief. Bovendien, als de mediaan dicht bij een grens tussen klassen ligt, kan het om een of andere reden weglaten van één waarneming de mediaan net over de rand van een klas duwen, zodat hij in een andere klas belandt. Als we een gegroepeerde frequentieverdeling met klassen van € 1.000 zouden hebben, zou het al wat preciezer kunnen, en bij een in de oorspronkelijke scores gegeven verdeling is er helemaal geen probleem. Eenzelfde moeilijkheid duikt op als we de mediaan bij een frequentieverdeling zoals in tabel 11.4 zouden willen bepalen. De mediaan ligt hier bij twee kinderen, maar erg bevredigend is dat niet, omdat hij niet precies 50% boven en 50% onder de waarde 2 ligt.

We zien dat ook bij een gering aantal waarden de mediaan niet erg informatief is. Dat is de reden dat bij een gegroepeerde frequentieverdeling, of bij een gering aantal waarden van de variabele, soms een meer precieze formule wordt gebruikt om de mediaan te berekenen, een formule die berust op de aanname dat we met een continue variabele te maken hebben, en dat alle waarnemingen in elke klas keurig met gelijke intervalletjes over die klas verdeeld zijn. Maar die formule wordt zelden gebruikt, en is ook in SPSS afwezig.[4] Voor verdelingen zoals in tabel 11.6 is de mediaan niet erg aantrekkelijk, tenzij ook nog ergens de oorspronkelijke data beschikbaar zijn. Het rapporteren van verschillende centrummaten tegelijk met een grafiek (zoals in figuur 11.3) of een boxplot (zie verderop) is dan waarschijnlijk de beste oplossing.

4 De afwezigheid van deze berekening, die vroeger vaak werd gebruikt, kan deels worden verklaard uit het feit dat de berekening van de mediaan over een zeer groot aantal niet-afgeronde scores voor een computer een fluitje van een cent is, terwijl vroeger eerst een gegroepeerde verdeling werd gemaakt voordat allerlei maten werden berekend, en pas bij zo'n gegroepeerde verdeling krijgen we met het vinden van een 'meer precieze' mediaan te maken.

Tabel 11.6 Inkomen van een verzameling huishoudens in €

Inkomen	Frequentie	Cumulatieve frequentie	Percentage	Cumulatief percentage
Minder dan 1.000	16	16	8,0	8,0
1.000 tot 2.000	37	53	18,5	26,5
2.000 tot 3.000	45	98	22,5	49,0
3.000 tot 4.000	36	134	18,0	67,0
4.000 tot 5.000	26	160	13,0	80,0
5.000 tot 6.000	21	181	10,5	90,5
6.000 tot 7.000	7	188	3,5	94,0
7.000 tot 8.000	3	191	1,5	94,5
8.000 tot 9.000	2	193	1,0	95,5
9.000 tot 10.000	1	194	0,5	97,0
10.000 tot 11.000	2	196	1,0	98,0
11.000 tot 12.000	1	197	0,5	98,5
12.000 tot 13.000	1	198	0,5	99,0
13.000 tot 14.000	1	199	0,5	99,0
14.000 tot 15.000	1	200	0,5	100,0

En vervolgens: zijn er bijpassende spreidingsmaten? Om weer de kern van een spreidingsmaat aan te duiden vragen we ons af of de waarnemingen voor een groot deel dicht bij de mediaan liggen, of dat ze verspreid zijn over het hele bereik van de variabele. Je kunt de vraag als volgt specificeren: als we nu eens kijken naar die 50% van de waarnemingen die het dichtst bij de mediaan liggen, hoe ver liggen die dan uit elkaar? Wat zijn de begrenzingen? Met andere woorden: we vragen ons af wat de waarde van de 25e waarneming is, en wat de waarde van de 75e waarneming, en bepalen dan de afstand tussen die twee (denk eraan: de score van de 50e waarneming is de mediaan!). Die afstand tussen de 25e en de 75e noemen we de interkwartielafstand (IQR). Maar ook hier geldt dat deze op een gegroepeerde frequentieverdeling als in tabel 11.6 alleen heel ruw bepaald kunnen worden, en we dus eigenlijk de oorspronkelijke niet-afgeronde scores nodig hebben.

In bepaalde omstandigheden zou het misschien nuttig kunnen zijn om niet met de 25e en de 75e waarneming te werken, maar bijvoorbeeld met de 10e en de 90e: het gaat dan om de vraag hoe ver de middelste 80% van de waarnemingen van elkaar liggen. In alle gevallen is de opzet om extreme waarnemingen buiten beschouwing te laten.

Bij het zoeken naar grenzen hebben we te maken met decielen en percentielen. Bijvoorbeeld: het 6e deciel is het inkomen dat past (zie tabel 11.6) bij de 60/100 x 200 = de 120e waarneming. Percentielen splitsen je steekproef in honderd gelijke aantallen, decielen in tien gelijke aantallen waarnemingen.

Als we te maken hebben met een omvangrijke populatie van leerlingen die aan zo'n test deelnemen, krijgen we een normale verdeling van de scores. We begrijpen nu misschien ook dat het verschil tussen de uitkomsten 94 en 95 enerzijds, en dat tussen 51 en 52 anderzijds, niet gelijk is. In het midden van de verdeling is het verschil erg klein, aan de randen veel groter. In beide gevallen is de afstand tussen twee grenzen bepaald door 1% van de waarnemingen, maar in het midden zijn er veel waarnemingen (bij een histogram kunnen we aan hoge staven denken) en is de afstand langs de horizontale as klein, en aan de randen zijn de staven laag en is de afstand dientengevolge groot. Het xe percentiel wordt aangeduid als P_x. Dus voor het eerste kwartiel, Q1, schrijven we P_{25}, voor het tweede kwartiel, de mediaan, schrijven we P_{50}, en voor het derde, Q3, P_{75}. De interkwartielafstand IQR is $P_{75} - P_{25}$.

In SPSS kun je de percentielen, het minimum en het maximum, lezen door STATISTICS te klikken in het FREQUENCIES-venster, en dan de desbetreffende maten te selecteren.

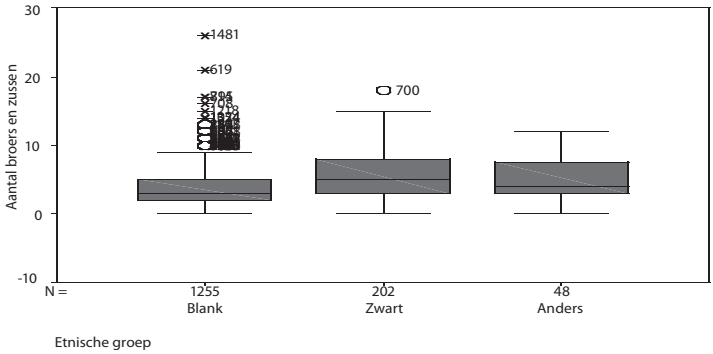
Quantielen vergeleken met standaardscores

Met het gebruiken van quartielen, decielen en percentielen (tezamen quantielen genoemd) heb je een techniek die, net zoals bij standaardscores, gebruikt kan worden om de positie van een eenheid ten opzichte van een verzameling eenheden aan te geven. Beide technieken beantwoorden vragen als 'Als Jantjes score 112 is, is dat dan een hoge of een lage score?' Bij het berekenen van quantielen gebruiken we echter alleen de rangorde van de scores. Quantielen zijn daarom onafhankelijk van de eenheid waarin gemeten is (centimeters of inches, om een voorbeeld te noemen). Standaardscores zijn daarentegen gebaseerd zowel op de rangorde als op de grootte van de intervallen.

Boxplots

Om snel een visueel overzicht van de frequentieverdeling van een variabele te hebben – en vooral als je dat wilt voor verschillende categorieën van een an-

dere variabele, bijvoorbeeld mannen en vrouwen apart of inkomensklassen apart – wordt tegenwoordig vaak een zogeheten boxplot gebruikt. Dit is een plaatje waarin zowel de mediaan, het eerste en het derde kwartiel, als de uitschieters staan aangegeven. Figuur 11.5 geeft de boxplots van 'aantal broers en zussen' naar etnische herkomst van de respondent weer (uit het 'General Social Survey 1991'-bestand). Langs de horizontale as staan de totalen voor drie herkomstgroepen naast elkaar vermeld. Langs de verticale as geeft de onderste horizontale lijn van de rechthoek het eerste kwartiel (Q1) aan, de middelste horizontale lijn de mediaan (Q2) en de bovenste het derde kwartiel (Q3). In de box zit de middelste 50% van de waarnemingen. De dikte van de rechthoek geeft dus de interkwartielafstand aan. De breedte van de box doet er niet toe. Dan zijn er nog twee horizontale lijnen, boven en onder de box, die de scores omsluiten die liggen op minder dan 1,5 x de interkwartielafstand van respectievelijk het derde en het eerste kwartiel. Vrijwel alle waarnemingen vallen binnen deze twee lijnen, en hun functie is dan ook alleen om uitschieters (*outliers*) te definiëren: alles wat buiten die twee lijnen valt. Deze extreme waarnemingen worden per stuk met hun volgnummer apart aangegeven. Uit dit voorbeeld leren we in één oogopslag dat de blanke respondenten het laagste mediane aantal broers en zussen hebben, en dat ook de spreiding in die aantallen kleiner is dan bij de andere etnische groepen. Wel zijn er enkele forse uitschieters; zo is er iemand (nummer 1481) met 26 broers en zussen (althans dat wordt door die respondent beweerd!).
In SPSS kom je via GRAPHS bij BOX PLOTS.



Figuur 11.5 Aantal broers en zussen naar etnische herkomst

De keuze van een centrum- en een spreidingsmaat

Voor de keuze tussen de maten geven we hierna een algemeen advies. In de praktijk van een onderzoek kan het wel eens moeilijk zijn om de aard van een bepaalde frequentieverdeling aan te geven in termen van de hierna genoemde criteria. Maar in de grote meerderheid van de gevallen is het hiernavolgende goed bruikbaar.

Nominaal niveau

Vlakke verdeling:	**geen centrummaat noemen**
Meertoppige verdeling:	**noem alle toppen + % in de toppen**
Eentoppige verdeling:	**noem de modus + % in de modus**

Ordinaal niveau

Beslis of je de ordinale informatie wilt negeren en de variabele als nominaal behandelen, of dat je gelijkheid van de intervallen *veronderstelt* en de variabele behandelt als op intervalniveau. Ordinale variabelen zoals het niveau van de laatst gevolgde opleiding of een rangenstelsel worden gewoonlijk als nominaal behandeld. Dus wordt hoogstens de modus gebruikt, met dezelfde restricties als hiervoor onder 'nominaal'. Ordinale variabelen zoals allerlei beoordelingen worden gewoonlijk als van intervalniveau beschouwd. Als centrummaat wordt het gemiddelde gebruikt met de standaardafwijking als spreidingsmaat, vooropgesteld dat er geen open klassen, extreme waarden of scheve verdelingen zijn.

Interval- en rationiveau

Beslis of er uitschieters zijn. Zo ja, laat deze weg uit de berekeningen, maar noem deze in het rapport afzonderlijk; ga door naar 'scheefheid'. Beslis of de verdeling scheef is. Zo ja, gebruik de mediaan en IQR voor beschrijvende doeleinden. Als je ook het gemiddelde en de standaardafwijking berekent, kun je het verschil tussen de mediaan en het gemiddelde beschouwen als een maat voor scheefheid. Maar met het oog op verdere multivariate analyse kun je beter een logaritmische transformatie toepassen, en het gemiddelde en de standaardafwijking van de logscores berekenen. Indien niet scheef, bereken dan het gemiddelde en de standaardafwijking. Bij een gegroepeerde verdeling: beslis of er open klassen zijn. Zo ja, bepaal het percentage in de open klasse(n). Indien minder dan 5%, bepaal dan een fictief midden van de open klasse(n), en ga door met het berekenen van het gemiddelde en misschien de mediaan. Indien meer dan 5%, kies dan de mediaan en IQR.

Oefeningen

Oefening 11.1

Toon met een (fictieve) frequentieverdeling van een nominale variabele aan dat het zinloos is om een mediaan of een gemiddelde te berekenen doordat deze kunnen veranderen na verwisseling van twee waarden.

Oefening 11.2

De gemiddelde leeftijd van een groep studenten is twintig jaar, en de standaardafwijking is twee jaar. Hoe groot zijn het gemiddelde en de standaardafwijking tien jaar later? Geef een argumentatie op basis van de formule!

Oefening 11.3

Gegeven de volgende frequentieverdelingen op een variabele van rationiveau:

	A	B	C	D
Waarden	**Frequentie**	**Frequentie**	**Frequentie**	**Frequentie**
1	10	20	-	-
2	10	-	20	10
3	10	10	10	30
4	10	-	20	10
5	10	20	-	-
N	50	50	50	50

a. Voorspel zonder berekening welke variantie groter is dan welke andere. Bereken daarna de variantie en de standaardafwijking bij elke verdeling.
b. De vijftig eenheden worden nu over een variabele met de waarden 1 tot en met 10 geheel vlak verdeeld, waarbij er dus in elke waarde vijf eenheden komen. Wat denk je: is de variantie van de nieuwe variabele kleiner dan, groter dan of gelijk aan de variantie van A? Reken variantie en standaardafwijking vervolgens uit.

Oefening 11.4

Schat de percentielscores van de genoemde persoon met betrekking tot lengte en gewicht (zie de tekst aan het einde van box 11.1). Neem aan dat per klasse de eenheden op gelijke afstandjes van elkaar liggen. Vergelijk de conclusies die gebaseerd zijn op percentielen, met de scores gebaseerd op standaardscores. Beredeneer dat dit kon worden verwacht.

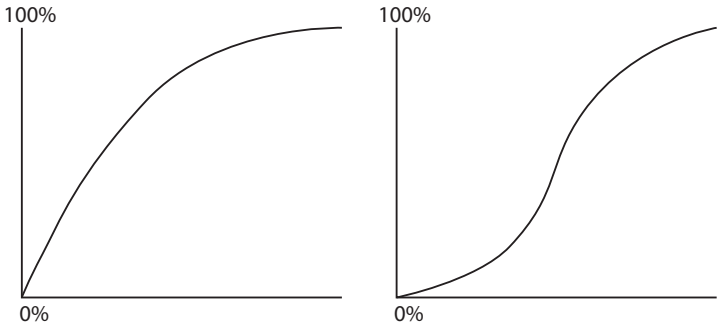
Oefening 11.5

Bereken de mediaan van de frequentieverdeling van tabel 11.6 als:
a. tien personen uit de klasse 3.000-4.000 verschuiven naar de klasse 2.000-3.000;
b. tien personen uit de klasse 3.000-4.000 verschuiven naar de klasse 4.000-5.000.

Voorspel eerst, voordat je gaat rekenen, of de mediaan hoger of lager zal worden!

Oefening 11.6 In een onderzoeksrapport worden de inkomensverdelingen van twee groepen A en B met elkaar vergeleken aan de hand van de relatieve cumulatieve frequentieverdelingen.

a. Geef je conclusie in termen van mediaan en interkwartielafstand.
b. Wat kun je zeggen over de vorm van de (niet-cumulatieve) frequentieverdelingen?



Oefening 11.7 Een steekproef van elf eenheden levert waarnemingen op de intervalniveauvariabele X die lopen van 17 tot 38. Hoe veranderen het gemiddelde, de mediaan, de modus en de standaardafwijking als je:

a. de scores van alle waarnemingen met 10 verhoogt?
b. de scores van alle waarnemingen met 2 vermenigvuldigt?
c. de score van de op twee na hoogste waarneming met 11 verhoogt?

Oefening 11.8 Na een door de regering afgekondigde verhoging van het collegegeld heeft men bij een steekproef uit de bevolking de opinie hierover gemeten met behulp van een schaal, bestaande uit zeven 'zespuntsitems', die elk gescoord werden van 1 (zeer mee eens) tot 6 (helemaal niet mee eens). Opgeteld kon elk dus minimaal 7 en maximaal 42 punten halen. Een gegroepeerde frequentieverdeling ziet er als volgt uit:

Score	Frequentie
7 - 9	50
10 - 12	130
13 - 15	150
16 - 18	50
19 - 21	50
22 - 24	40
25 - 27	30
28 - 30	30
31 - 33	20
34 - 36	10
37 - 39	10
40 - 42	10
N	580

a. Teken een histogram en een cumulatieve frequentiepolygoon.
b. Bereken alle centrum- en spreidingsmaten en geef aan welke de voorkeur genieten.

Oefening 11.9

Voor de fans van puzzels van het sudoku-type: gegeven twee niet-identieke frequentieverdelingen met hetzelfde gemiddelde en dezelfde standaardafwijking; probeer twee zulke frequentieverdelingen op te stellen. Hint: gebruik zeven categorieën.

Literatuurtips

Blalock, H.M. (1979). *Social statistics* (2nd ed.) (International Student Edition). Tokyo: McGraw-Hill Kogakusha.

DeVeaux, R.D. & Velleman, P.F. (2003). *IntroStats*. Boston: Addison Wesley.

Hays, W.L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart & Winston.

Tufte, E.R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tufte, E. (2001). *The visual display of quantitative information*. Cheshire, CT: Graphics Press.

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

Kernbegrippen

centrummaten
gemiddelde
mediaan
modus
spreidingsmaten
variantie
standaardafwijking
interkwartielafstand
standaardscoreberekening
scheefheid
boxplots