Centrummaten 2

In hoofdstuk 1 is het meetniveau van variabelen besproken (nominaal, ordinaal, interval of ratio). Verder zijn de begrippen continu en discreet toegelicht. Zoals gezegd is dit van belang voor de verschillende analyses die mogelijk zijn met die variabelen. In dit hoofdstuk staan de meest elementaire analyses centraal: centrummaten.
Een centrummaat is een getal dat aangeeft rond welke (centrale) waarde de uitkomsten van een serie waarnemingen liggen. Je onderscheidt drie centrummaten: modus, mediaan en gemiddelde.

2.1 Modus

De modus is de waarde die het meest voorkomt, de waarde met de hoogste frequentie. De modus kun je bij alle meetniveaus gebruiken, maar geeft niet altijd zinnige informatie. Op nominaal niveau is de modus de meest geschikte en de enig mogelijke centrummaat.
Laten we nog eens kijken naar het aantal drankjes dat geturfd is op het terrasje.

Tabel 2.1 Centrummaat op nominaal niveau: de modus

Drankje	Aantal (geturfd)	Absolute frequentie
1: bier	𝍸 III	⑧
2: rosé	𝍸	5
3: cola light	I	1
4: cappuccino	III	3
Totaal		17

Uit tabel 2.1 blijkt dat de meeste mensen bier willen (acht mensen), de modus is dus 1, omdat dat de waarde is die we aan bier hebben toegekend. Nogmaals, de waarden hebben op nominaal niveau geen getalsmatige betekenis.
Je kunt de modus bepalen aan de hand van een frequentieverdeling. We hebben bijvoorbeeld in een databestand een variabele die aangeeft wat de favoriete televisieserie is van de respondenten. Favoriete televisieserie is een nominale variabele omdat er slechts sprake is van classificatie (naamgeving), maar niet van een rangordening. In de output van SPSS ziet de frequentieverdeling eruit als in tabel 2.2.

Tabel 2.2 Frequentieverdeling van de variabele televisieserie (SPSS-output)

Serie favoriete tvserie

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 True Detective	22	33,8	33,8	33,8
	2 Game of Thrones	20	30,8	30,8	64,6
	3 Dr. Who	23	35,4	35,4	100,0
	Total	65	100,0	100,0	

Zowel aan de absolute frequentie als aan het percentage is te zien dat de meeste mensen in dit onderzoek *Dr. Who* als favoriete serie hebben. *Dr. Who* heeft in dit onderzoek de waarde 3. De modus is dus 3.

Een modus mag je ook uitrekenen voor variabelen die op ordinaal, interval- of rationiveau zijn gemeten. Zo zou je kunnen kijken welke opleiding (lager, middelbaar of hoger onderwijs – ordinaal) het meest voorkomt of hoe oud (leeftijd in jaren – ratio) de meeste van je respondenten zijn.
Een nadeel van de modus is dat dit kengetal geen informatie geeft over de overige waarden van een variabele. Daardoor geeft de modus soms geen informatie waar je wat aan hebt. Als we bijvoorbeeld kijken naar de leeftijdsverdeling van vijftig mensen die variëren van 18 tot 81 jaar, dan zou de modus 18 kunnen zijn. Het is mogelijk dat maar vijf personen die leeftijd hebben, en dat van alle andere leeftijden er steeds vier of minder zijn. De modus is ook 18 als veertig van de vijftig personen 18 jaar zijn, maar dan is er sprake van een geheel andere leeftijdsverdeling binnen die groep. Een modus van 18 geeft in dit geval beperkte en niet erg nuttige informatie.

2.2 Mediaan

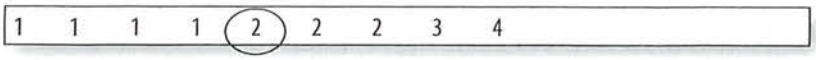
De mediaan is de middelste waarneming na rangordening van de data van laag naar hoog. Het is de waarneming waar 50% van de onderzoekseenheden onder ligt en 50% boven. Uit deze definitie blijkt al dat je de mediaan niet op nominaal niveau kunt gebruiken, want er moet een rangorde in de waarden zitten. Je gebruikt de mediaan op ordinaal of hoger niveau. De mediaan is in de regel de meest geschikte centrummaat voor ordinale variabelen.
Stel, negen personen hebben aangegeven hoeveel televisie ze per dag kijken. Televisiekijken is in dit geval op een ordinale schaal gemeten, namelijk:
1: één uur of minder;
2: meer dan één, maar minder dan anderhalf uur;
3: anderhalf tot twee uur;
4: meer dan twee, maar minder dan tweeënhalf uur;
5: tweeënhalf tot drie uur;
6: meer dan drie uur.

NB: De categorieën moeten elkaar uitsluiten!

Dit leverde de volgende gegevens op voor de meting van de televisiekijktijd:

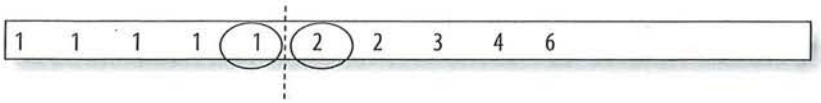
1 2 1 4 2 1 2 1 3

De eerste stap is het rangschikken van de gegevens. De middelste waarneming is de mediaan (zie figuur 2.1).



Figuur 2.1 Mediaan bij oneven aantal waarnemingen

De mediaan is hier 2. Dat wil zeggen dat 50% van de personen een score had van 2 of lager en dat 50% van de personen een score had van 2 of hoger. Met andere woorden, 50% van de personen kijkt minder dan anderhalf uur per dag televisie, en 50% van de personen kijkt meer dan één uur per dag televisie. De waarde 2 is immers 'meer dan één, maar minder dan anderhalf uur'.
Wanneer er een even aantal waarnemingen is, is er geen middelste waarneming. Toch kun je wel een mediaan berekenen. Eerst sorteer je de waarnemingen weer op grootte. Vervolgens tel je de twee middelste waarden bij elkaar op en deel je dat getal door twee (zie figuur 2.2).



Figuur 2.2 Mediaan bij even aantal waarnemingen

De mediaan is hier (1 + 2) / 2 = 1,5. Dat wil zeggen dat de mediaan tussen de 1 en de 2 ligt. Dit betekent dat 50% van de onderzoekseenheden voor die variabele een waarde heeft van 2 of meer en 50% een waarde van 1 of minder. Als het weer gaat om de eerder gebruikte ordinale schaal voor kijktijd, betekent dit dat 50% een uur of minder en 50% langer dan een uur per dag televisiekijkt. Het gebruik van de mediaan bij interval- en ratiovariabelen kan zinnig zijn omdat deze centrummaat ongevoelig is voor uitschieters, terwijl het rekenkundig gemiddelde daar wel gevoelig voor is (zie paragraaf 2.3).
Ook de mediaan is op basis van een frequentietabel in de SPSS-output eenvoudig te bepalen. SPSS geeft bijvoorbeeld de in tabel 2.3 weergegeven frequentietabel van de variabele opleiding.

Aan het cumulatieve percentage is af te lezen dat bij de waarde 5 (vwo) de 50%-grens wordt gepasseerd. De mediaan is derhalve 5. Dat wil zeggen dat 50% van de respondenten een opleidingsniveau heeft van vwo of lager en dat 50% een opleidingsniveau heeft van vwo of hoger.
De mediaan kan dus nooit berekend worden bij variabelen op nominaal niveau, je kunt immers niet zeggen: '50% van de respondenten leest het liefst detectives of minder, en 50% van de respondenten leest het liefst detectives of meer'. In

de variabele 'favoriete genre boek' zit immers geen rangordening en heeft de mediaan geen betekenis. De mediaan kan wel op een hoger niveau dan ordinaal worden berekend, dus bij numerieke variabelen (interval en ratio). Wanneer je 'aantal uur per week een boek lezen' hebt gemeten (een ratiovariabele), is het geoorloofd om te zeggen: '50% van de jongeren leest 5 uur of minder per week een boek en 50% van de jongeren leest 5 uur of meer per week een boek'.

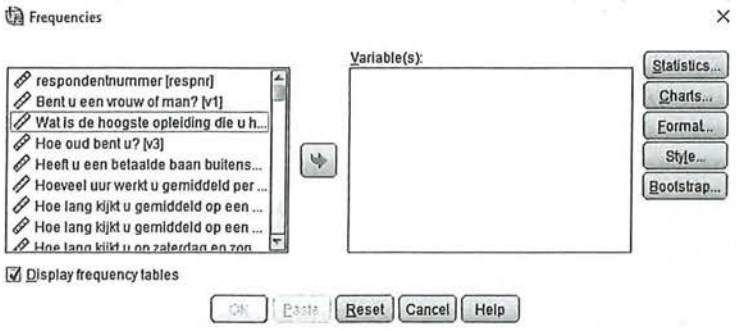
Tabel 2.3 Frequentieverdeling van de variabele opleiding (SPSS-output)

v2 opleiding

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 lager onderwijs	26	3,6	3,6	3,6
	2 mavo	45	6,2	6,2	9,8
	3 mbo	83	11,4	11,5	21,3
	4 havo	100	13,8	13,8	35,1
	5 vwo	209	28,8	28,9	64,0
	6 hbo	158	21,8	21,8	85,8
	7 universiteit	103	14,2	14,2	100,0
	Total	724	99,9	100,0	
Missing	System	1	,1		
Total		725	100,0		

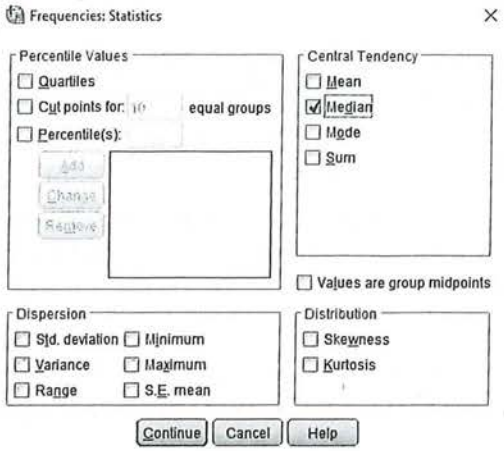
SPSS Centrummaten

Voor het berekenen van centrummaten in SPSS ga je eerst weer naar *Frequencies (Analyze → Descriptive Statistics → Frequencies)*. Hier selecteer je de variabele waarvan je een centrummaat wilt laten berekenen, bijvoorbeeld opleiding (zie figuur A).



Figuur A Frequencies-venster

Via *Statistics* kun je door SPSS een centrummaat laten berekenen (zie figuur B).



Figuur B Statistics-venster

In het *Statistics*-venster kun je aanvinken welke centrummaat (onder *Central Tendency*) je wilt laten berekenen. Afhankelijk van het meetniveau en je eigen wensen kun je kiezen voor het gemiddelde (*Mean*), de mediaan (*Median*) of de modus (*Mode*). In dit voorbeeld gaat het om de variabele opleiding. Opleiding is een ordinale variabele, de meest geschikte centrummaat is dan de mediaan. Overigens raden wij aan om het gemiddelde op een andere manier door SPSS te laten berekenen, zie daarvoor kader 3.1 in hoofdstuk 3.

Om de analyse uit te voeren kun je klikken op OK of op PASTE. In hoofdstuk 4 (Bewerken van je data) leggen we uit waarom je beter op PASTE kunt klikken.

Kader 2.1

2.3 (Rekenkundig) gemiddelde

De laatste centrummaat is het gemiddelde. Het gemiddelde gebruik je alleen op interval- en rationiveau. Bij nominale en ordinale variabelen is het uitrekenen van het gemiddelde niet geoorloofd. Je kunt niet zeggen dat het gemiddelde opleidingsniveau 3,4 is, aangezien de voor de afzonderlijke opleidingen gekozen waarden en de intervallen tussen die waarden geen betekenis hebben. Bij het interval- en rationiveau hebben de getallen van de waarden wel een betekenis. Met die waarden mogen we dan ook rekenen.
Het rekenkundig gemiddelde bereken je door alle waarnemingen bij elkaar op te tellen en te delen door het totaal aantal waarnemingen (n). Het symbool voor het gemiddelde is $\overline{x}$ (x streep). In wetenschappelijke artikelen wordt ook vaak de letter M, van het Engelse woord *Mean*, gebruikt om het gemiddelde aan te geven.
In formulevorm ziet de berekening er als volgt uit:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Formule voor gemiddelde

De Griekse hoofdletter sigma (Σ) wordt gebruikt als sommatieteken en betekent: neem de som van. Boven de sigma staat een n, en onder de sigma i = 1. Dit betekent: neem van elke i-de onderzoekseenheid, vanaf de eerste (i =1) tot en met de n-de (i = n), de waarde van x. De sigma betekent dat je al die waarden bij elkaar moet optellen.[1] De n staat voor het totaal aantal waarnemingen. De gehele formule zegt dus: om $\bar{x}$ te berekenen neem je de som van alle waarden van x (van waarneming 1 tot en met n) en deel je deze door n (het totaal aantal waarnemingen).

Laten we eens kijken naar tien personen die een statistiektest hebben afgelegd met daarin twintig vragen. De reeks hierna toont het aantal fouten dat gemaakt is.

0 0 1 1 1 4 4 4 6 6

Voor de berekening van het gemiddelde tellen we alle *x*'en (alle individuele scores) bij elkaar op en delen deze door 10.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0+0+1+1+1+4+4+4+6+6}{10} = \frac{27}{10} = 2{,}7$$

Het gemiddeld aantal fouten dat de personen hebben gemaakt in de statistiektest is dus 2,7.
Ook vanuit een frequentietabel is het gemiddelde te berekenen. De formule is dan iets anders, maar het principe werkt hetzelfde.

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j f_j}{n}$$

Formule gemiddelde berekend op basis van groepsfrequenties

In plaats van een *i* staat er nu een *j* onder de sigma en een *k* erboven. Achter de *x* is een *f* toegevoegd. De *j* betekent 'de groep' en er zijn in totaal *k* groepen. De f_j is de frequentie waarmee een waarde van *x* in groep *j* voorkomt. Om het gemiddelde te berekenen vermenigvuldig je voor elke groep (*j* = 1 tot en met *k*) de waarde *x* met de frequentie (f_j = het aantal onderzoekseenheden in die groep); de producten van alle groepen tel je vervolgens bij elkaar op en deel je door het totale aantal onderzoekseenheden (*n*).
Een voorbeeld zal dit duidelijker maken. De frequenties van het 'aantal fouten in de test' staan in tabel 2.4.

Tabel 2.4 Frequentietabel 'fouten in test' in absolute aantallen (n = 10)

Fouten (x)	Frequentie (f)
0	2
1	3
4	3
6	2
	10

Tabel 2.4 laat zien dat er twee keer nul fouten werden gemaakt, drie keer één fout enzovoorts. Om het gemiddelde te berekenen zou je volgens de eerste formule (waarbij achter de x een i stond) alle x'en (de fouten van alle individuen) uit moeten schrijven, en bij elkaar optellen. Je krijgt dan de eerder gegeven rij cijfers: 0 0 1 1 1 4 4 4 6 6. Het is eenvoudiger om de x te vermenigvuldigen met de frequentie f en daarna de uitkomsten daarvan bij elkaar op te tellen, zoals in tabel 2.5 is gedaan.

Tabel 2.5 Berekenen van gemiddeld 'aantal fouten' (n=10)

Fouten (x)	Frequentie (f)	$x_j * f_j$
0	2	0 * 2 = 0
1	3	1 * 3 = 3
4	3	4 * 3 = 12
6	2	6 * 2 = 12
		Σ 27

De formule is nu eenvoudig in te vullen:

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j f_j}{n} = \frac{27}{10} = 2{,}7$$

Bij tien onderzoekseenheden is het nog mogelijk om de waarden van alle eenheden uit te schrijven en bij elkaar op te tellen. In het volgende voorbeeld gaat het om een groter aantal onderzoekseenheden. We willen nu de gemiddelde leeftijd weten van 73 personen. Gegeven is de frequentietabel (tabel 2.6).

Tabel 2.6 Frequentietabel van leeftijd

Leeftijd (*x*)	Frequentie (*f*)
21	1
22	6
23	22
24	19
25	20
26	3
28	1
31	1
	73

Het is te veel werk om alle scores apart uit te gaan schrijven. Je zou dan een lange rij met getallen krijgen: 21 22 22 22 22 22 22 23 23 23 23 23 enzovoort. We kiezen er daarom voor om het gemiddelde te berekenen op basis van de groepen (tabel 2.7).

Tabel 2.7 Berekenen van gemiddelde leeftijd ($n = 73$)

Leeftijd (*x*)	Frequentie (*f*)	$x_j * f_j$		
21	1	21 * 1	=	21
22	6	22 * 6	=	132
23	22	23 * 22	=	506
24	19	24 * 19	=	456
25	20	25 * 20	=	500
26	3	26 * 3	=	78
28	1	28 * 1	=	28
31	1	31 * 1	=	31
	73	Σ	=	1752

$$\bar{x} = \frac{\sum_{j=1}^{k} x_j f_j}{n} = \frac{1752}{73} = 24$$

De gemiddelde leeftijd van deze 73 respondenten is 24 jaar.

2.4 Keuze tussen centrummaten

Als een variabele op nominaal niveau is gemeten, is de enige centrummaat die je kunt gebruiken de modus. Een andere keuze is er niet. Bij variabelen op ordinaal niveau geven modus en mediaan informatie die voor een onderzoeker nuttig kan zijn. Bij ratio- en intervalvariabelen is het gebruik van alle drie de centrummaten mogelijk, maar die informatie is niet altijd zinnig.
Stel, je wilt door middel van een kengetal iets zeggen over de leeftijden van de leden van een huishouden. Het huishouden bestaat uit een man, vrouw, vier kinderen (waarvan een tweeling) en een inwonende grootmoeder. Hun leeftijden zijn:

15 15 16 17 45 50 80

In dit geval is de modus 15, de mediaan 17 en de gemiddelde leeftijd 34. Welke centrummaat geeft in dit geval de meest zinnige informatie?
Als niet de jongste maar de oudste toevallig een tweeling was geweest, was de modus 17 geweest. De modus als kengetal zegt hier niet zoveel over de leeftijdsverdeling. Aan de mediaan hebben we meer. Dat 50% 17 jaar of jonger en 50% 17 jaar of ouder is, geeft een aardige indruk van de leeftijdsverdeling. De gemiddelde leeftijd van 34 zegt weer minder, want dit gemiddelde wordt sterk beïnvloed door de leeftijd van de grootmoeder. Zonder deze uitschieter zou de gemiddelde leeftijd 26,3 jaar zijn.

15 15 16 17 45 50

Zonder de leeftijd van de grootmoeder verandert de modus niet en ligt de mediaan tussen de 16 en 17 (16,5). De keuze voor een centrummaat is dus afhankelijk van de informatie die je nodig hebt, en de uitschieters die eventueel in een verdeling voorkomen.

2.5 Samenvatting

De eerste stap voor het kiezen van een centrummaat is het vaststellen van het meetniveau van de variabelen. Het meetniveau beperkt de keuzemogelijkheden. Als we de kenmerken van het meetniveau maximaal willen benutten, is de meest geschikte centrummaat voor interval- en ratiovariabelen het rekenkundig gemiddelde, voor ordinale variabelen is het de mediaan en voor nominale variabelen de modus.

Tabel 2.8 Meetniveaus en centrummaten

Nominaal	Ordinaal	Interval	Ratio
modus	*modus*	*modus*	*modus*
	mediaan	*mediaan*	*mediaan*
		gemiddelde	***gemiddelde***

In tabel 2.8 staat cursief voor 'geoorloofd'. Cursief en vet staat voor 'meest geschikt', dat wil zeggen dat er maximaal gebruik wordt gemaakt van de kenmerken van het meetniveau.

Ga naar de website om de opdrachten bij dit hoofdstuk te maken.

Noot

1 We zullen bij het uitwerken van de formules niet altijd deze informatie rondom de sigma vermelden; over het algemeen wordt ook uit de formule zelf al duidelijk wat er gedaan moet worden.