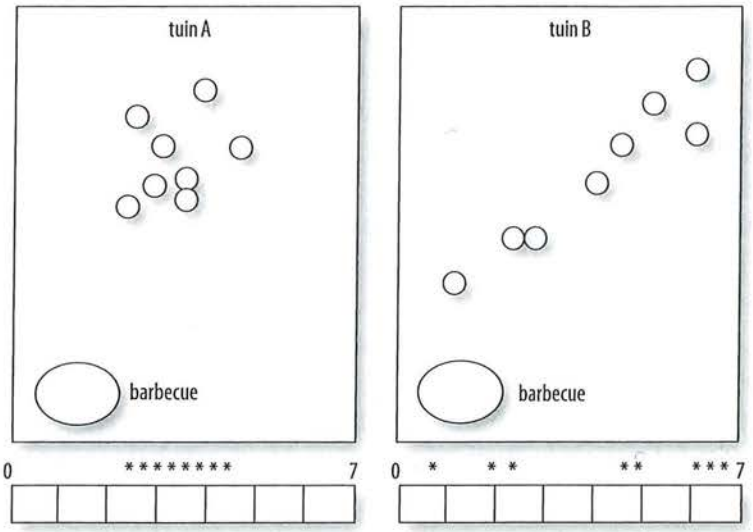
Spreiding 3

In dit hoofdstuk staan de spreidingsmaten centraal. In het vorige hoofdstuk zagen we dat centrummaten aangeven rond welke waarde op de meetschaal de waarnemingen zich centreren, oftewel rond welke centrale waarde op de meetschaal de waarnemingen verspreid zijn. Spreidingsmaten geven aan hoe sterk de waarden zich concentreren: liggen deze dicht bij elkaar of zijn ze juist erg verdeeld?
Wat is nu eigenlijk spreiding? Spreiding is in feite niets anders dan de afstand tussen de verschillende waarnemingen. Stel, je bent op een tuinfeestje met in een hoek van de tuin een barbecue (zie figuur 3.1).



Figuur 3.1 Voorbeeld van tuin A, waarin er weinig spreiding is van de personen over de tuin, en tuin B, waarin de personen veel meer verspreid zijn, terwijl de gemiddelde afstand tot de barbecue ongeveer gelijk is

Wanneer iedereen dicht bij elkaar staat, al of niet bij de barbecue, is er weinig spreiding (tuin A). Wanneer iedereen is verdeeld over de hele tuin, is er sprake van veel spreiding (tuin B). Omdat de afstand tussen de waarden centraal staat in veel spreidingsmaten, zijn deze maten vooral van belang bij metingen op interval- of rationiveau. Op nominaal niveau kun je niet spreken over afstanden tussen de waarden. Bij nominale variabelen is het aantal mogelijke waarden een manier om een indicatie van spreiding te geven. Er werden bijvoorbeeld vier verschillende soorten drankjes besteld. Als er een volgende keer meer verschillende soorten drankjes besteld worden, is de spreiding gemeten in het *aantal categorieën* (het

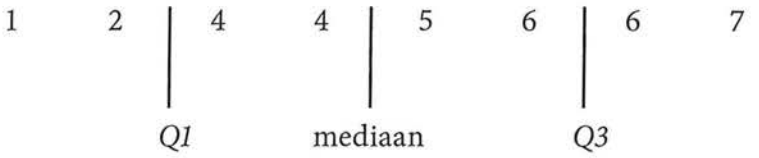
aantal verschillende drankjes) groter. Een andere manier om bij nominale variabelen een indicatie van de spreiding te geven is de *variatieratio*, waarmee je het aandeel van de onderzoekseenheden aangeeft dat niet in de modale categorie valt. De meest simpele indicatie van spreiding bij variabelen vanaf ordinaal niveau is de *range*. Dit is het verschil tussen de hoogste en laagste waarde van de variabele.

3.1 Kwartielen

Bij variabelen van minimaal ordinaal niveau zijn er meer mogelijkheden om iets over spreiding te zeggen, omdat de waarden die je gebruikt een bepaalde rangordening hebben. Plaats je de uitkomsten van een serie waarnemingen in oplopende volgorde, dan ontstaat een geordende getallenreeks. In het vorige hoofdstuk hebben we al gezien dat de 50%-grens van deze getallenreeks de mediaan is. We kunnen de getallenreeks ook in vier stukken verdelen, die elk 25% van de waarnemingen bevatten. Deze stukken noem je de *kwartielen*. Informatie over deze kwartielen geeft een indruk van de spreiding bij ordinale variabelen.
Het eerste kwartiel (aangeduid als *Q1*) is de waarde waarbij 25% van de onderzoekseenheden een kleinere of gelijke waarde heeft, en 75% een gelijke of grotere waarde. Het tweede kwartiel is de mediaan: 50% is gelijk of kleiner dan die waarde en 50% is gelijk of groter dan die waarde. Het derde kwartiel (*Q3*) is de waarde waarbij 25% van de onderzoekseenheden grotere of gelijke waarden heeft (en daarmee heeft 75% een gelijke of kleinere waarde).
Het verschil tussen het eerste en derde kwartiel noem je de *interkwartielafstand* (*Q3 – Q1*). Er moet sprake zijn van waarden die groter of kleiner zijn, en omdat voor het berekenen van de interkwartielafstand het verschil tussen twee waarden (*Q1* en *Q3*) wordt bepaald, moet dat verschil ook betekenis hebben. Deze spreidingsmaat is dus alleen geschikt als het meetniveau minimaal interval is.
Een voorbeeld: op het tuinfeestje zijn de mensen verspreid over de hele tuin (die 7 meter lang is) en is de afstand tussen acht mensen en de barbecue berekend in meters. Deze gegevens zijn vervolgens gerangschikt van laag naar hoog.

1 2 4 4 5 6 6 7 (afstand in meters ten opzichte van de barbecue)

Er is één persoon die één meter van de barbecue staat, er is ook één persoon op een afstand van twee meter, er zijn twee mensen die vier meter van de barbecue staan enzovoort. De mediaan is hier 4,5 (= (4 + 5) ÷ 2). Om de mediaan te vinden heb je de rij in twee gelijke stukken opgedeeld. Het eerste kwartiel ligt op de helft van het eerste stuk, het derde kwartiel ligt op de helft van het tweede stuk.



Het eerste kwartiel (*Q1*) is hier 3 (= (2 + 4) ÷ 2) en het derde kwartiel 6 (= (6 + 6) ÷ 2).
De interkwartielafstand is dan: $Q3 - Q1 = 6 - 3 = 3$.

De waarde van de interkwartielafstand is beter te interpreteren wanneer je deze vergelijkt met een andere interkwartielafstand. Stel dat je na een paar uur nog een meting doet, en dan de volgende data krijgt:

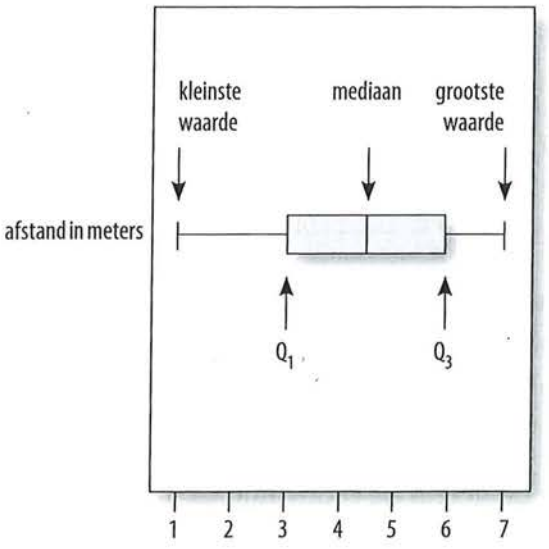
1 4 | 4 4 | 5 5 | 6 7

Q1 mediaan *Q3*

Het eerste kwartiel (*Q1*) is nu 4 (= (4 + 4) ÷ 2) en het derde kwartiel 5,5 (= (5 + 6) ÷ 2).
De interkwartielafstand is dan: $Q3 - Q1 = 5{,}5 - 4 = 1{,}5$ meter.
De interkwartielafstand is kleiner geworden (gedaald van 3 naar 1,5). Je kunt nu een vergelijking maken tussen de interkwartielafstanden. De spreiding is dus eerder op de avond groter dan later op de avond.

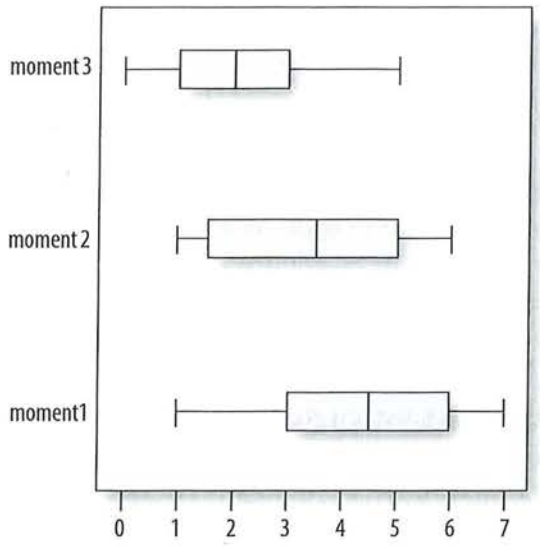
3.1.1 Boxplot

Een *boxplot* is een grafische weergave van de kwartielen. Het is een handig hulpmiddel om zowel de centrale tendentie (centrummaat) als de spreiding in één oogopslag te zien. In een boxplot worden de laagste waarde, het eerste kwartiel, de mediaan, het derde kwartiel, en de hoogste waarde weergegeven. Figuur 3.2 geeft een boxplot van de positie van de gasten op de barbecue 'eerder op de avond'.



Figuur 3.2 Boxplot van de afstand tot de barbecue 'eerder op de avond'

Gedurende de avond lopen deze acht mensen door de tuin heen, waardoor de spreiding kan variëren. Wanneer je op drie momenten de posities van deze mensen bijhoudt, kun je door middel van een boxplot snel een overzicht krijgen van het verloop van de avond.



Figuur 3.3 Boxplot van drie momenten

Naarmate de avond vordert, wordt de spreiding kleiner (op moment 3) en neemt langzamerhand de afstand tot de barbecue af.

3.2 Variantie

Door een centrummaat te berekenen geef je een beschrijving van een groep onderzoekseenheden. Een spreidingsmaat voegt hier belangrijke informatie aan toe.
We zagen al eerder aan het voorbeeld van de barbecue op het tuinfeest dat spreiding iets zegt over de afstand van de onderzoekseenheden ten opzichte van een bepaald centrum. De meest gebruikte manier om iets over spreiding te zeggen is de standaarddeviatie. Deze wordt berekend aan de hand van de variantie en de variatie. We zullen daarom eerst in deze paragraaf de *variatie* en *variantie* bespreken.

We kijken naar een voorbeeld waarin aan acht jongeren is gevraagd hoeveel uur zij per week online het nieuws lezen. Dit zijn de gemeten waarden:

0 1 2 2 4 6 6 8

Op interval- en rationiveau is de meest geschikte centrummaat het rekenkundig gemiddelde. Het gemiddeld aantal uur dat jongeren online het nieuws lezen is 3,625.[1]

$$\bar{x} = \frac{\sum x_i}{n} = \frac{0+1+2+2+4+6+6+8}{8} = 3{,}625$$

We kunnen nu per persoon (per onderzoekseenheid) kijken in hoeverre hij of zij in afstand verschilt (of afwijkt) van het gemiddelde. Persoon 1 leest geen nieuws online, hij of zij heeft de waarde 0. Deze persoon leest dus 3,625 minder uur online het nieuws dan het gemiddelde van deze acht personen. Oftewel: 0 – 3,625 = –3,625.
Persoon 2 leest 1 uur per week online het nieuws. Het verschil met het gemiddelde is 1 – 3,625 = –2,625. Dit verschil kan voor elke onderzoekseenheid berekend worden. De notatie hiervoor is $(x_i - \bar{x})$. Letterlijk staat hier: van elke *i*ndividuele x (en x is aantal uur online nieuws kijken voor alle afzonderlijke personen) wordt het gemiddelde van x afgetrokken.
Wanneer we dat voor elke onderzoekseenheid doen, hebben we acht verschillende afstanden ten opzichte van het gemiddelde, de 'individuele verschillen met de gemiddelde afstand' (tabel 3.1). Dit zegt nog steeds niet zoveel. Over die verschillen met de gemiddelde afstand kun je ook een gemiddelde berekenen: het gemiddelde verschil met de gemiddelde afstand. Om dit te berekenen, zou je alle individuele verschillen ten opzichte van het gemiddelde bij elkaar op willen tellen $\sum(x_i - \bar{x})$ en delen door n (het totaal). Het probleem daarbij is dat deze som (Σ) *altijd* uitkomt op nul.[2]

Tabel 3.1 Individuele verschillen met de gemiddelde afstand

x_i	$(x_i - \bar{x})$
0	(0 – 3,625) = –3,625
1	(1 – 3,625) = –2,625
2	(2 – 3,625) = –1,625
2	(2 – 3,625) = –1,625
4	(4 – 3,625) = 0,375
6	(6 – 3,625) = 2,375
6	(6 – 3,625) = 2,375
8	(8 – 3,625) = 4,375
Σ	0

Om van de nul af te komen kunnen we elk verschil kwadrateren. Door te kwadrateren raken we het minteken kwijt, zodat we de negatieve waarden bij de eerste vier onderzoekseenheden kwijt zijn. Daarna kunnen we alle kwadraten bij elkaar optellen, sommeren. In formulevorm ziet dat er als volgt uit:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Formule voor variatie

Deze kwadratensom noem je de *variatie*. Maar dit getal is als spreidingsmaat moeilijk te interpreteren, omdat het getal sterk afhankelijk is van het aantal onderzoekseenheden. Hoe meer onderzoekseenheden er zijn, hoe hoger de waarde van de variatie wordt. Dat maakt interpretatie van het getal moeilijk. Je lost dit probleem op door de *variantie* te berekenen. De variantie is een soort gemiddelde kwadratische afwijking ten opzichte van het gemiddelde. Het symbool voor variantie is s^2. De formule is:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Formule voor variantie

Letterlijk staat er: elke individuele x min het gemiddelde van x wordt gekwadrateerd, de kwadraten van alle onderzoekseenheden worden bij elkaar opgeteld, en de som wordt gedeeld door het totaal aantal waarnemingen (n) min 1.[3]
Om de variantie uit te rekenen berekenen we dus eerst de variatie (de teller in de formule voor variantie).

Tabel 3.2 Berekenen van de variatie ($n = 8$)

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
0	(0 – 3,625) = –3,625	13,141
1	(1 – 3,625) = –2,625	6,891
2	(2 – 3,625) = –1,625	2,641
2	(2 – 3,625) = –1,625	2,641
4	(4 – 3,625) = 0,375	0,141
6	(6 – 3,625) = 2,375	5,641
6	(6 – 3,625) = 2,375	5,641
8	(8 – 3,625) = 4,375	19,141
Σ	0	55,878

De variatie is hier 55,878.
Vervolgens vullen we de formule in om de variantie te berekenen:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} = \frac{55{,}878}{8-1} = 7{,}983$$

De variantie is hier 7,983. Maar ook dit getal is moeilijk te interpreteren. We hebben immers een kwadratensom gebruikt. Daardoor verduidelijkt de uitkomst van de variantie niet direct hoe het aantal uur dat online nieuws wordt gelezen is verspreid over de acht jongeren.

Wat we wel kunnen doen, is de variantie van deze groep jongeren vergelijken met de variantie van een groep ouderen. We vragen aan acht ouderen hoe vaak zij per week online het nieuws lezen, en berekenen eerst weer het gemiddelde.

$$\bar{x} = \frac{0+0+2+2+2+6+6+8}{8} = 3{,}25$$

Vervolgens berekenen we de variatie. Deze is 63,504 (zie tabel 3.3).

Tabel 3.3 Berekenen van de variatie (voor ouderen)

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
0	(0 – 3,25) = –3,25	10,563
0	(0 – 3,25) = –3,25	10,563
2	(2 – 3,25) = –1,25	1,563
2	(2 – 3,25) = –1,25	1,563
2	(2 – 3,25) = –1,25	1,563
6	(6 – 3,25) = 2,75	7,563
6	(6 – 3,25) = 2,75	7,563
8	(8 – 3,25) = 4,75	22,563
Σ	0	63,504

De variantie is hier:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} = \frac{63{,}504}{8-1} = 9{,}072$$

De variantie is voor de groep ouderen dus groter dan voor de groep jongeren. Bij ouderen is er dus meer spreiding in het aantal uur dat online nieuws wordt gelezen dan bij jongeren. Het gemiddeld aantal uur dat online nieuws gelezen wordt, verschilt echter niet veel van elkaar (3,63 voor de jongeren en 3,25 voor de ouderen).

Laten we nog een voorbeeld bekijken. Je zoekt een kamer in Amsterdam en bekijkt de prijzen van zes kamers.
kamer 1: € 175
kamer 2: € 180
kamer 3: € 190
kamer 4: € 240
kamer 5: € 350
kamer 6: € 550

Wat is nu de variantie van deze kamerprijzen? De eerste stap is het berekenen van het gemiddelde.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{175+180+190+240+350+550}{6} = 280{,}833$$

De gemiddelde kamerprijs van deze zes kamers in Amsterdam is dus € 280,833. Vervolgens neem je de kwadratensom van de verschillen van elke individuele kamerprijs ten opzichte van het gemiddelde (de variatie).

Tabel 3.4 Berekenen van de variatie (kamerprijzen Amsterdam)

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
175	(175 – 280,833) = –105,833	11200,624
180	(180 – 280,833) = –100,833	10167,294
190	(190 – 280,833) = –90,833	8250,634
240	(240 – 280,833) = –40,833	1667,334
350	(350 – 280,833) = 69,167	4784,074
550	(550 – 280,833) = 269,167	72450,874
Σ	0	108520,834

Nu kun je de formule voor variantie invullen.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1} = \frac{108520{,}834}{6-1} = 21704{,}167$$

De variantie in kamerprijzen is in dit geval € 21.704,167. Dit is natuurlijk een raar getal: hoe moet je dit interpreteren in combinatie met de gemiddelde prijs van € 280,833?
De standaarddeviatie is een kengetal dat je wel gemakkelijk kunt interpreteren. Deze wordt uitgelegd in de volgende paragraaf.

3.3 Standaarddeviatie

Door de afstand tussen de individuele score en het gemiddelde te kwadrateren, krijgen we getallen die moeilijk te interpreteren zijn. Het is daarom nuttig om het kwadraat in de variantie op te heffen. Dit doe je door de wortel te trekken uit de waarde die we voor de variantie hebben berekend. Op die manier berekenen we de *standaarddeviatie*, ook wel standaardafwijking genoemd.
De formule voor de standaarddeviatie (aangeduid met de letter *s*) is:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Formule voor de standaarddeviatie

In het eerste voorbeeld bij variantie (het aantal uren dat door jongeren online nieuws wordt gelezen) was de variantie 7,983. Om daar de standaarddeviatie van te berekenen, nemen we de wortel van dat getal: $\sqrt{7,983} = 2,825$. De standaardafwijking is dus 2,83. Dit getal is beter te interpreteren in combinatie met het gemiddelde dan de variantie. Jongeren lezen gemiddeld 3,63 uur per week online het nieuws, en wijken daar gemiddeld 2,83 uur van af.
Dit wordt helemaal duidelijk in het voorbeeld van de kamerprijzen. De variantie van de kamerprijzen was € 21704,167, een getal dat in geen enkele verhouding staat tot het gemiddelde van € 280,833.
De standaarddeviatie is veel beter te interpreteren: $\sqrt{21704,167} = 147,323$. De standaarddeviatie is de gemiddelde afwijking ten opzichte van het gemiddelde en die is hier € 147,32. De kamerprijzen in Amsterdam zijn gemiddeld € 280,833 en daar wordt gemiddeld € 147,32 van afgeweken.

Hoe lager de standaarddeviatie, hoe dichter de individuele scores zich rondom het gemiddelde concentreren. En andersom, hoe hoger de standaarddeviatie, hoe verder (hoe meer verspreid) de individuele scores van het gemiddelde af liggen.
Laten we eens twee steden met elkaar vergelijken qua kamerprijzen. Behalve naar kamers in Amsterdam kijk je ook naar kamers in Utrecht. In Utrecht treffen we de volgende kamerprijzen aan:
kamer 1: € 75
kamer 2: € 130
kamer 3: € 200
kamer 4: € 230
kamer 5: € 400
kamer 6: € 650

Het gemiddelde van deze kamerprijzen is hetzelfde als in Amsterdam, namelijk € 280,83. Enkel op basis van het gemiddelde zou je denken dat er geen verschil is tussen de prijzen van kamers in Amsterdam en Utrecht. Maar de standaarddeviatie van de kamerprijzen in Utrecht verschilt wel van de standaarddeviatie van de kamerprijzen in Amsterdam, deze is in Utrecht namelijk € 212,00. Hoewel het gemiddelde in beide steden dus gelijk is, is de spreiding in Utrecht groter dan de spreiding in Amsterdam. Dit is ook goed te zien in de figuren 3.4 en 3.5.

Figuur 3.4 Spreiding kamerprijzen Amsterdam

Figuur 3.5 Spreiding kamerprijzen Utrecht

De statistische resultaten van een univariate analyse vermeld je doorgaans in de tekst. Je rapporteert altijd het totale aantal onderzoekseenheden, de (meest geschikte) centrummaat en, zo mogelijk, de spreidingsmaat. Bijvoorbeeld: 'Aan het onderzoek deden meer meisjes (56%) dan jongens mee (44%). De 180 jongeren keken gemiddeld ongeveer 2,5 keer per week naar soaps (M = 2,58, SD = 2,26).'

Wil je meerdere centrummaten en spreidingsmaten tegelijk rapporteren, dan is het handig om een overzichtstabel te presenteren. Onderstaande tabel (tabel 3.5) is een voorbeeld van hoe zo'n overzichtstabel er in een wetenschappelijk artikel uit kan zien. In dit artikel is een experiment uitgevoerd waarin werd gekeken naar de voorkeur voor omslagen van kinderboeken, waar steeds een realistisch omslag werd vergeleken met een niet realistisch omslag en een niet complex omslag met een complex omslag. De waarden die in de tabel staan, geven de gemiddelde voorkeur weer voor het boekomslag. Uit de tabel is onder andere af te lezen dat boekomslagen met een realistische foto hoger worden gewaardeerd dan boekomslagen met een onrealistische foto.[4]

Tabel 3.5 Gemiddelde scores voor realisme en complexiteit per set omslagen, in artikel van Hartman et al. (2014)[5]

Set	Realisme				Complexiteit			
	Onrealisme		Realistisch		Niet complex		Complex	
	M	SD	M	SD	M	SD	M	SD
Met voeten	1.87*	1.05	4.23*	.95	2.45*	1.19	3.56*	1.12
Met detectives	2.32*	1.13	3.65*	1.10	2.25*	1.00	3.42*	1.10
Met springende persoon	2.14*	1.12	3.92	1.21	2.04*	.99	3.61	1.05

3.4 Centrum- en spreidingsmaten in SPSS

Net als de modus en de mediaan kan het gemiddelde in SPSS berekend worden door een frequentietabel uit te laten draaien. Hierbij kan er voor gekozen worden om de juiste spreidingsmaten aan te vinken. Omdat interval- en ratiovariabelen over het algemeen erg veel waarden hebben, is het echter niet altijd overzichtelijk om een frequentieverdeling te maken. Het is ook mogelijk om via het commando *Descriptives* de informatie op te vragen, of via het commando *Explore*. Wij raden deze laatste manier aan bij het beschrijven van numerieke variabelen (zie ook kader 3.1).

SPSS Centrum- en spreidingsmaten

Bij het beschrijven van numerieke variabelen (interval- of ratio meetniveau) wordt gebruikgemaakt van *Analyze → Descriptive Statistics → Explore*. In het venster dat verschijnt kun je de variabele(n) die je wilt beschrijven invoeren in de *Dependent List*. Eventueel kan onder *Plots* gekozen worden om een histogram te laten maken.

Figuur A: Explore-venster

Kader 3.1

Je hebt in een enquête gevraagd hoe vaak respondenten per week naar de televisie kijken, en je hebt dat gemeten in het aantal minuten dat ze dat doen (zie ook paragraaf 4.3 voor het samenstellen van variabelen). Wanneer je door middel van *Explore* deze variabele gaat beschrijven, krijg je eerst de volgende twee tabellen:

Tabel 3.6 Beschrijving van de variabele minuten tv-kijken via Explore (SPSS-output)

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
minutentv	1468	100,0%	0	0,0%	1468	100,0%

Descriptives

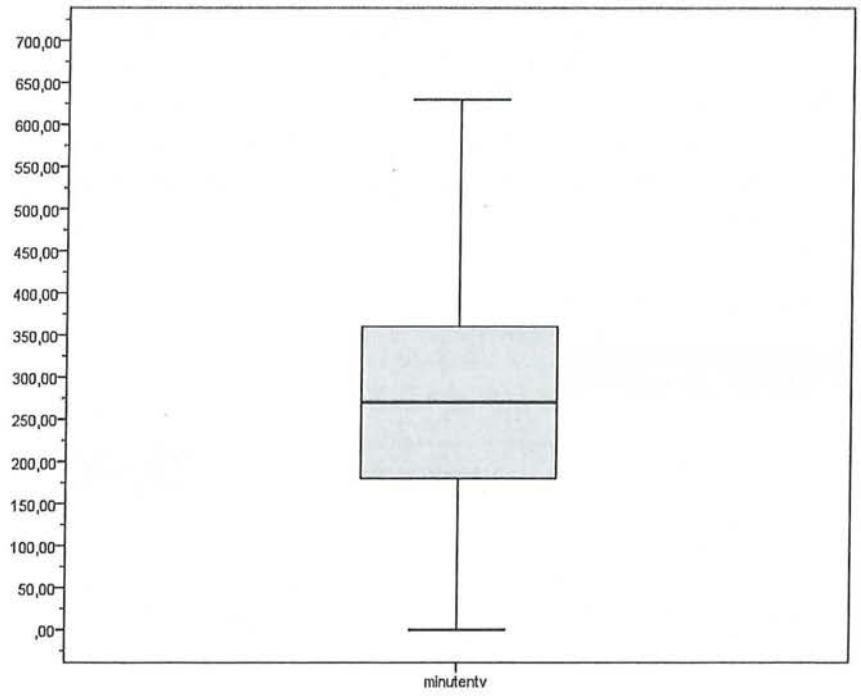
			Statistic	Std. Error
minutentv	Mean		284,1335	3,62596
	95% Confidence Interval for Mean	Lower Bound	277,0209	
		Upper Bound	291,2461	
	5% Trimmed Mean		279,6783	
	Median		270,0000	
	Variance		19300,652	
	Std. Deviation		138,92678	
	Minimum		,00	
	Maximum		630,00	
	Range		630,00	
	Interquartile Range		180,00	
	Skewness		,464	,064
	Kurtosis		-,434	,128

Je ziet hierin bijna alle informatie die je nodig hebt. In de tabel *Case Processing Summary* zien we dat 1468 respondenten deze vraag hebben beantwoord en dat er geen *missing values* zijn. In de bovenste rij van de tabel *Descriptives* staat het gemiddelde: gemiddeld kijken de onderzoekseenheden 284,13 minuten per week naar de televisie. Iets verder in de tabel zien we de standaarddeviatie: de gemiddelde afstand ten opzichte van het gemiddelde is 138,93. Dat wil dus zeggen dat er een grote mate van spreiding is, er zijn mensen die veel lager en mensen die veel hoger dan het gemiddelde scoren. De standaarddeviatie werd berekend door de wortel te trekken uit de variantie (die staat daarboven vermeld: 19300,65). De minimumwaarde die op deze variabele is gescoord is nul (er zijn mensen die niet televisiekijken) en de maximumwaarde is 630 (de maximale tijd dat per week televisie wordt gekeken is 10,5 uur). De mediaan is 270: 50% van de respondenten kijkt 270 minuten of minder per week naar de televisie, 50% van de respondenten kijkt 270 minuten of meer per week naar de televisie. Tot slot kun je de range uit de tabel aflezen (het verschil tussen de hoogste en laagste waarde), en de interkwartielafstand, die is hier 180.

Van de gepresenteerde spreidingsmaten geeft de standaarddeviatie in dit geval de meeste informatie, vooral als deze gekoppeld wordt aan het gemiddelde. De interkwartielafstand geeft pas nuttige informatie wanneer je bijvoorbeeld de

interkwartielafstand van vrouwen zou vergelijken met de interkwartielafstand van mannen, of een andere vergelijking zou maken.

Behalve een overzichtelijke beschrijvende tabel, krijg je bij het uitvoeren van deze analyse een boxplot in je output (zie figuur 3.6).



Figuur 3.6 Boxplot van aantal minuten tv-kijken (via Explore)

3.5 Standaardiseren (z-scores)

Z-scores zijn gestandaardiseerde scores van een variabele die we voor elke onderzoekseenheid apart kunnen uitrekenen. Door standaardisatie zijn de waarden van variabelen die een verschillende meeteenheid hebben met elkaar te vergelijken. De z-scores zijn gebaseerd op de standaarddeviatie en het gemiddelde van een variabele. Aangezien de z-scores gebaseerd zijn op het rekenkundig gemiddelde, kunnen ze alleen maar uitgerekend worden voor variabelen die op interval- of rationiveau zijn gemeten. De z-score geeft aan hoeveel maal de standaarddeviatie de waarde van de betreffende onderzoekseenheid afwijkt van het gemiddelde van een variabele. Een negatieve z-score betekent dat de waarde van de onderzoekseenheid voor die variabele kleiner is dan het gemiddelde van de groep, een positieve z-score betekent dat deze waarde groter is dan het gemiddelde van de groep. Zo betekent $z = 1$ dat de waarde van de onderzoekseenheid op de variabele één standaarddeviatie groter is dan het gemiddelde, en $z = -2$ betekent dat de waarde twee standaarddeviaties kleiner is dan het gemiddelde.

De formule voor z is:

$$z = \frac{x - \bar{x}}{s}$$

Formule voor de z-score

Om z uit te rekenen moet dus eerst het gemiddelde ($\bar{x}$) worden berekend, en de standaarddeviatie (s). Stel, je wilt twee gegevens van een aantal personen met elkaar vergelijken: hun intelligentie (IQ) en hun inkomen. Deze variabelen hebben een verschillende meeteenheid. Het inkomen is gemeten in euro's per week, en de intelligentie met de scores van een IQ-test. De waarden van de variabelen zijn in eerste instantie moeilijk met elkaar te vergelijken. Dit wordt eenvoudiger als je per onderzoekseenheid z-scores berekent.

Tabel 3.7 Datamatrix inkomen en IQ

	IQ	Inkomen
A	115	460
B	85	340
C	100	400
$\bar{x}$	100	400
s	15	60

De twee variabelen hebben een interval/ratio meetniveau, we kunnen dus het gemiddelde en de standaarddeviatie berekenen. Voor het IQ is het gemiddelde 100 en de standaarddeviatie 15, voor inkomen is het gemiddelde 400 en de standaarddeviatie 60 (zie tabel 3.8). Met deze informatie kun je per onderzoekseenheid de z-score uitrekenen. Voor persoon A geldt bijvoorbeeld voor IQ een z-score van

$$z = \frac{x - \bar{x}}{s} = \frac{115 - 100}{15} = 1$$

en voor inkomen een z-score van

$$z = \frac{x - \bar{x}}{s} = \frac{460 - 400}{60} = 1$$

Op deze manier kun je voor de drie onderzoekseenheden per variabele een z-score berekenen zoals weergegeven is in tabel 3.8.

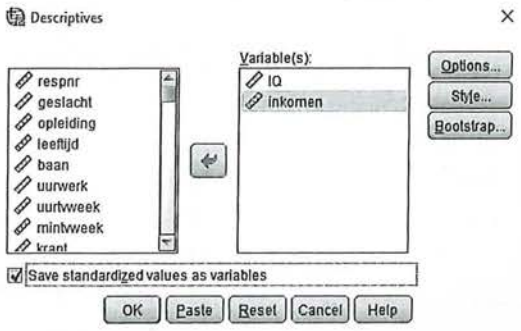
Tabel 3.8 Z-scores voor IQ en inkomen

	IQ	Inkomen	Z-IQ	Z-Inkomen
A	115	460	1	1
B	85	340	–1	–1
C	100	400	0	0
$\bar{x}$	100	400		
s	15	60		

Uit tabel 3.8 blijkt dat onderzoekseenheid A één standaarddeviatie boven het gemiddelde scoort voor zowel IQ als inkomen, en dat voor onderzoekseenheid C de waarden voor beide variabelen gelijk zijn aan de gemiddeldes. Door naar de z-scores te kijken wordt meteen duidelijk dat er een sterke samenhang is tussen IQ en inkomen. Dit is bij de werkelijke waarden van de variabelen minder direct te zien.

SPSS Berekenen van z-scores

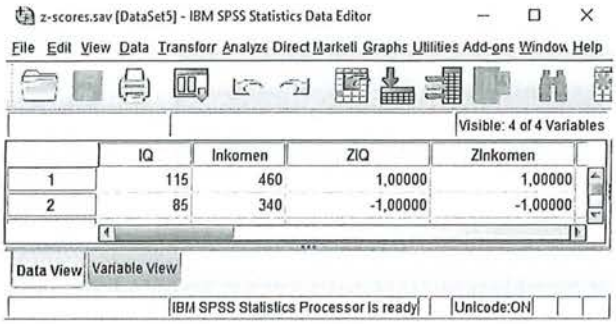
Wanneer je in SPSS z-scores berekent, worden deze aan de datamatrix toegevoegd. Om de z-scores te laten berekenen volg je de volgende procedure: *Analyze → Descriptive Statistics → Descriptives*. Hier voer je de variabelen in waar je z-scores van wilt hebben en vink je vervolgens het vakje *Save standardized values as variables* aan.



Figuur A Z-score via Descriptives

Wanneer je nu op PASTE klikt en de syntax runt (zie paragraaf 4.1), zijn de z-scores aan de datamatrix toegevoegd (zie figuur B).

SPSS geeft zelf een naam aan de nieuwe variabelen, in dit geval ZIQ voor de z-scores van IQ en ZInkomen voor de z-scores van inkomen.

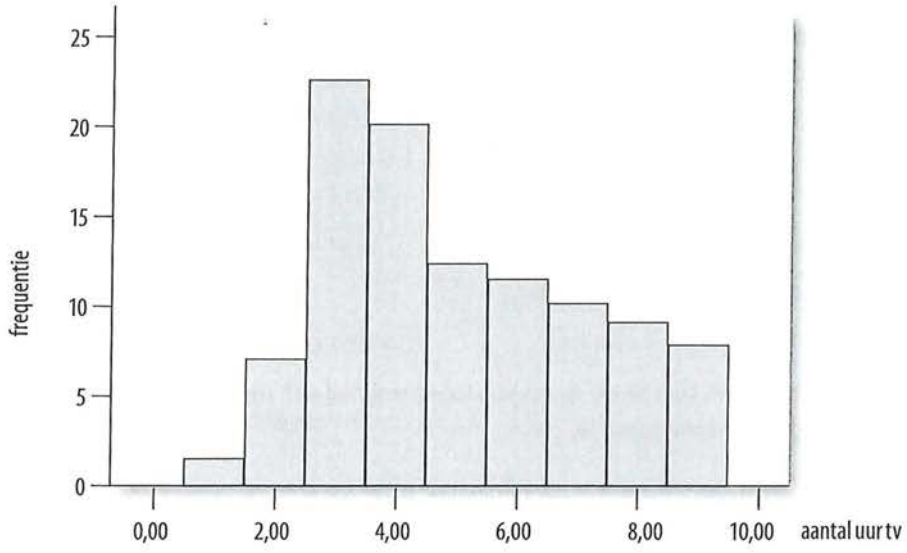


Figuur B Data View met z-scores

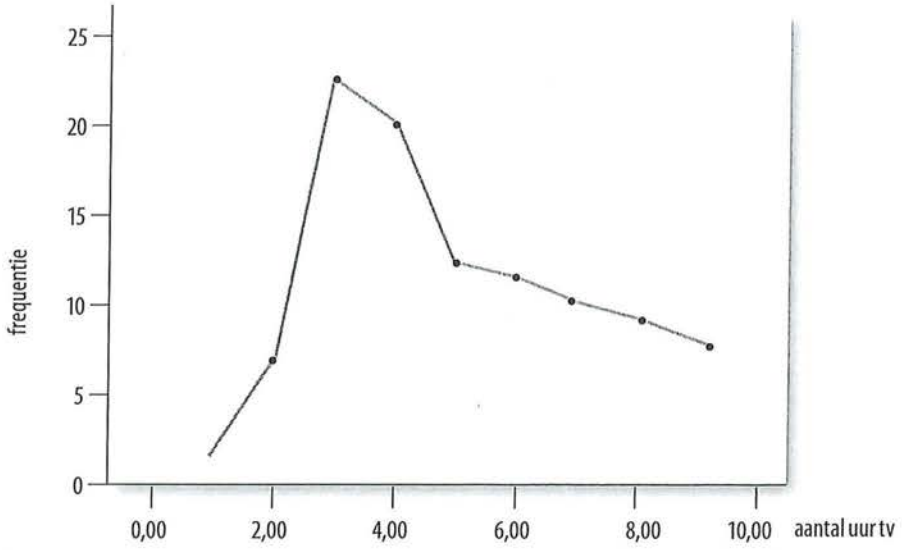
Kader 3.2

3.6 Normale en scheve verdelingen

In hoofdstuk 1 hebben we gezien dat je een frequentieverdeling grafisch kunt weergeven in een taartdiagram of in een staafdiagram (paragraaf 1.2.3). Deze figuren gebruik je beide voor nominale variabelen en/of wanneer het aantal waarden van een variabele beperkt is. Wanneer het meetniveau minimaal ordinaal is, is een histogram of frequentiepolygoon mogelijk. Een histogram is een grafische weergave van de frequentieverdeling van (in klassen) geordende data. Hieruit blijkt al dat de meetschaal dan minimaal op ordinaal niveau moet zijn. Een histogram toont in kolommen hoe vaak een waarde voorkomt (zie figuur 3.7).



Figuur 3.7 Histogram van aantal uur televisiekijken

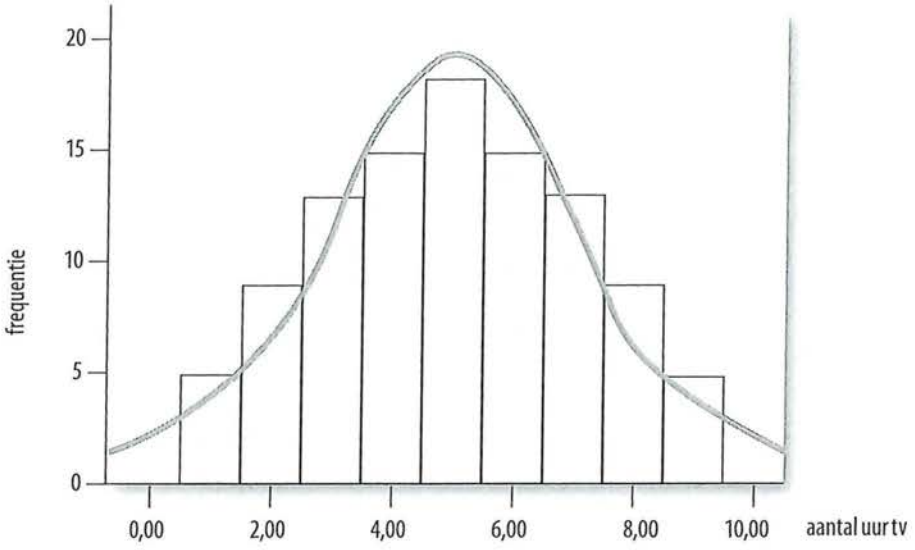


Figuur 3.8 Frequentiepolygoon van het aantal uur televisiekijken (behorende bij histogram figuur 3.7)

Zoals te zien is in figuur 3.7, wordt in een histogram de waarde van de variabele op de *x*-as in het midden van de kolombreedte (klassenmidden) aangegeven. De middens van de klassen boven in de kolommen zou je door middel van een lijn met elkaar kunnen verbinden. Op die manier ontstaat een frequentiepolygoon (zie figuur 3.8). In een histogram en een frequentiepolygoon is te zien hoe de frequentieverdeling eruitziet.

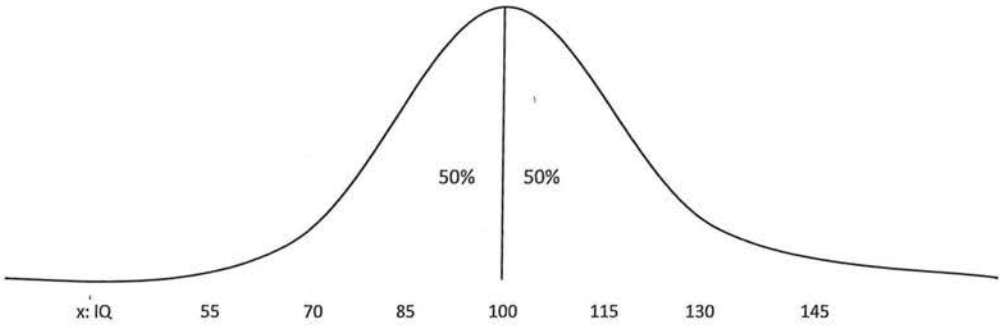
3.6.1 Normale verdeling

Behalve de centrummaten en de spreiding van een frequentieverdeling is ook de symmetrie dan wel scheefheid een kenmerk van verdelingen. Wanneer de verdeling symmetrisch is, spreken we van een *normale verdeling*. Wanneer we een vloeiende lijn tekenen, krijgen we een klokvormige figuur. In een normale verdeling is er één top waar het gemiddelde, de mediaan en de modus samenvallen. In figuur 3.9, waar een verdeling wordt gegeven van aantal uur televisiekijken, is dat goed te zien. Het gemiddelde, de mediaan en de modus zijn hier 5 (uur). De mensen die 'extreem' scoren, zitten in de staartjes van de verdeling.



Figuur 3.9 Normale verdeling van uur televisiekijken

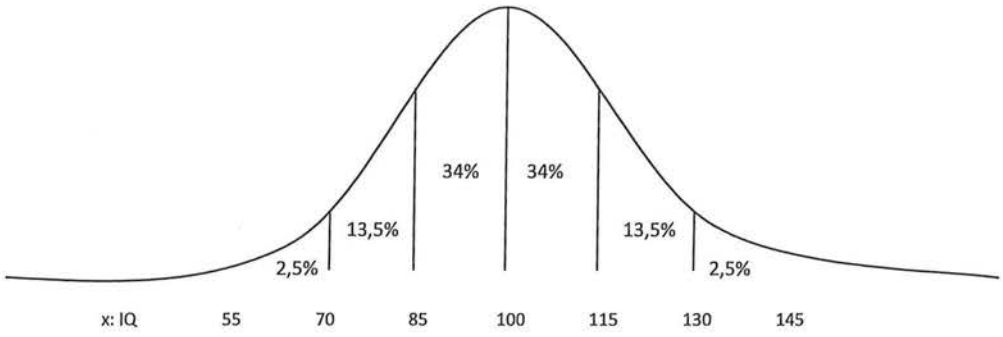
Stel dat je de frequentieverdeling van de IQ-scores van alle metingen onder Nederlandse volwassenen in een grafiek zou weergeven, dan zou deze de normale verdeling benaderen (zie figuur 3.10). Dat wil dus zeggen dat er weinig mensen zijn met een erg lage IQ-score, er weinig mensen zijn met een erg hoge IQ-score, en dat de meeste mensen in de verdeling niet erg ver boven of onder het gemiddelde van 100 zouden zitten. De meeste mensen (modus) hebben een IQ van 100, en omdat de verdeling symmetrisch is, is ook de mediaan 100 en kunnen we zeggen dat 50% van de volwassenen een IQ heeft van 100 of lager en 50% een IQ van 100 of hoger.



Figuur 3.10 Normale verdeling van IQ-scores

We kunnen aan de hand van een normale verdeling iets zeggen over de kans dat een bepaalde waarde voorkomt. Wanneer een variabele normaal verdeeld is, kun je de kans berekenen dat bepaalde waarden voorkomen.

Bij een normale verdeling kunnen we stellen dat 50% hoger en 50% lager scoort dan het gemiddelde. De kans dat een willekeurige volwassene een IQ heeft van 100 of hoger (dit is zowel de mediaan als het gemiddelde) is 50% (zie figuur 3.10). Het is ook mogelijk om de afzonderlijke helften weer verder op te delen in kansen (figuur 3.11). Bij een normale verdeling ligt ongeveer 34% van de scores altijd tussen het gemiddelde en één standaarddeviatie lager of hoger verwijderd van het gemiddelde. Metingen die een tot twee standaarddeviaties verwijderd zijn van het gemiddelde komen in ongeveer 13,5% van de gevallen voor. Ongeveer 2,5% van de scores is minimaal drie standaarddeviaties hoger of lager dan het gemiddelde. Dit wordt de *empirische regel* genoemd (figuur 3.11).

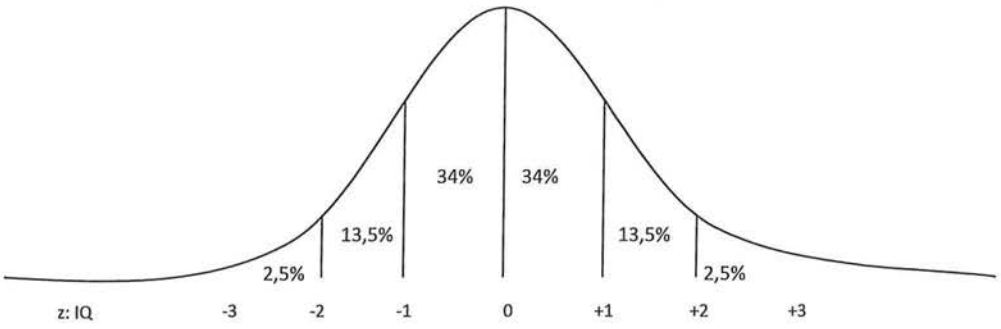


Figuur 3.11 De empirische regel

In figuur 3.10 en 3.11 is te zien dat IQ verdeeld is in stapjes van 15. Dat is omdat in dit voorbeeld het gemiddelde IQ 100 was, met een standaarddeviatie van 15. Op basis van de empirische regel (figuur 3.11) zouden we kunnen stellen dat ongeveer 34% van de volwassen mensen in deze steekproef een IQ heeft tussen de 100 en 115, en ongeveer 34% een IQ heeft tussen de 85 en 100. We kunnen ook zeggen dat 2,5% een IQ heeft van 130 of hoger. Oftewel: de kans dat iemand een IQ heeft van 130 of hoger is 2,5%. Dit is de *overschrijdingskans.*

Voor elke waarde van IQ kunnen we de overschrijdingskans bepalen. Hiervoor maken we gebruik van de z-scores. We standaardiseren de waarden van het IQ in z-scores die aangeven hoeveel maal de standaarddeviatie van die waarde afwijkt van het gemiddelde. We zetten de normale verdeling van de waarden van het IQ om in een verdeling van z-scores. Dit is de *standaardnormale verdeling.*

In een standaardnormale verdeling (figuur 3.12) zijn de waarden van de oorspronkelijke variabele gestandaardiseerd door middel van z-scores. Het gemiddelde van deze z-scores is altijd nul, en de standaarddeviatie is altijd 1. Dit geldt voor elke variabele die we standaardiseren. In het voorbeeld van IQ was de standaarddeviatie 15. Zoals we konden zeggen dat ongeveer 68% van de volwassenen een IQ heeft tussen de 85 en 115, kunnen we ook zeggen: 68% van de volwassenen wijkt maximaal één standaarddeviatie, één z-score af van het gemiddelde. Voor elke waarde van het IQ kunnen we nu bepalen hoe groot de kans is dat een willekeurige volwassene een IQ heeft dat hoger is dan die waarde. Daarvoor kun je een tabel gebruiken met de overschrijdingskansen voor mogelijke z-scores (zie tabel 3.9).



Figuur 3.12 Standaardnormale verdeling

Hieronder vind je een gedeelte van deze tabel. De gehele tabel is te vinden na het formuleblad in de bijlage.

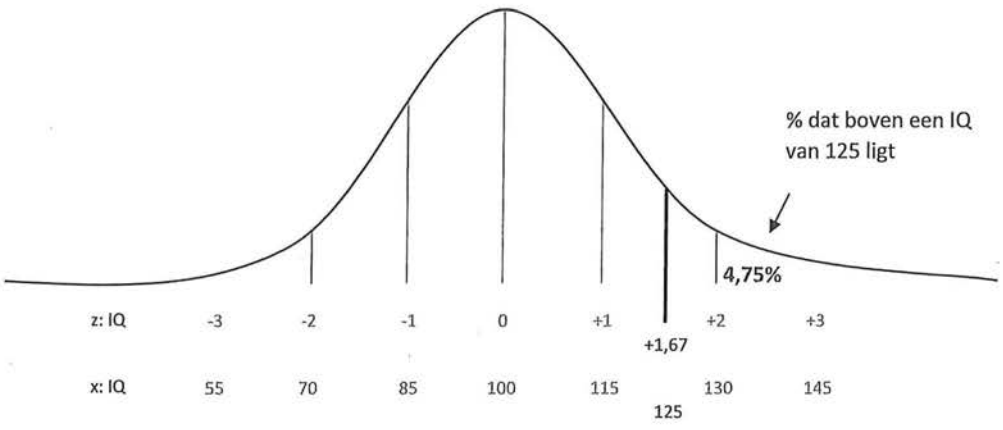
Tabel 3.9 Tabel met rechter overschrijdingskansen in de standaardnormale verdeling

z	$P_R(z)$	z	$P_R(z)$	z	$P_R(z)$	z	$P_R(z)$
0,00	0,5000	0,38	0,3520	0,76	0,2236	1,14	0,1271
0,01	0,4960	0,39	0,3483	0,77	0,2206	1,15	0,1251
0,02	0,4920	0,40	0,3446	0,78	0,2177	1,16	0,1230
0,03	0,4880	0,41	0,3409	0,79	0,2148	1,17	0,1210
0,04	0,4840	0,42	0,3372	0,80	0,2119	1,18	0,1190
0,05	0,4801	0,43	0,3336	0,81	0,2090	1,19	0,1170
0,06	0,4761	0,44	0,3300	0,82	0,2061	1,20	0,1151
0,07	0,4721	0,45	0,3264	0,83	0,2033	1,21	0,1131
0,08	0,4681	0,46	0,3228	0,84	0,2005	1,22	0,1112
0,09	0,4641	0,47	0,3192	0,85	0,1977	1,23	0,1093
0,10	0,4602	0,48	0,3156	0,86	0,1949	1,24	0,1075
0,11	0,4562	0,49	0,3121	0,87	0,1922	1,25	0,1056
0,12	0,4522	0,50	0,3085	0,88	0,1894	1,26	0,1038
0,13	0,4483	0,51	0,3050	0,89	0,1867	1,27	0,1020
0,14	0,4443	0,52	0,3015	0,90	0,1841	1,28	0,1003
0,15	0,4404	0,53	0,2981	0,91	0,1814	1,29	0,0985
0,16	0,4364	0,54	0,2946	0,92	0,1788	1,30	0,0968
0,17	0,4325	0,55	0,2912	0,93	0,1762	1,31	0,0951
0,18	0,4286	0,56	0,2877	0,94	0,1736	1,32	0,0934
0,19	0,4247	0,57	0,2843	0,95	0,1711	1,33	0,0918
0,20	0,4207	0,58	0,2810	0,96	0,1685	1,34	0,0901
0,21	0,4168	0,59	0,2776	0,97	0,1660	1,35	0,0885
0,22	0,4129	0,60	0,2743	0,98	0,1635	1,36	0,0869
0,23	0,4090	0,61	0,2709	0,99	0,1611	1,37	0,0853
0,24	0,4052	0,62	0,2676	1,00	0,1587	1,38	0,0838
0,25	0,4013	0,63	0,2643	1,01	0,1562	1,39	0,0823
0,26	0,3974	0,64	0,2611	1,02	0,1539	1,40	0,0808
0,27	0,3936	0,65	0,2578	1,03	0,1515	1,41	0,0793
0,28	0,3897	0,66	0,2546	1,04	0,1492	1,42	0,0778
0,29	0,3859	0,67	0,2514	1,05	0,1469	1,43	0,0764
0,30	0,3821	0,68	0,2483	1,06	0,1446	1,44	0,0749
0,31	0,3783	0,69	0,2451	1,07	0,1423	1,45	0,0735
0,32	0,3745	0,70	0,2420	1,08	0,1401	1,46	0,0721
0,33	0,3707	0,71	0,2389	1,09	0,1379	1,47	0,0708
0,34	0,3669	0,72	0,2358	1,10	0,1357	1,48	0,0694
0,35	0,3632	0,73	0,2327	1,11	0,1335	1,49	0,0681
0,36	0,3594	0,74	0,2296	1,12	0,1314	1,50	0,0668
0,37	0,3557	0,75	0,2266	1,13	0,1292	1,51	0,0655

Boven de tabel zie je staan: 'Rechter overschrijdingskansen in de standaardnormale verdeling'. Dat betekent dat je uit deze tabel alleen maar de rechterkant van de tabel kunt berekenen. Omdat de standaardnormale verdeling geheel symmetrisch is, geldt de rechterkant van de verdeling echter ook voor de linkerkant van de verdeling, maar dan zijn de z-scores negatief. Kijk je bijvoorbeeld bij een z-score van 0,00, dan is de kans dat iemand hoger of lager dan die waarde scoort, 50%. Kijk je bij een z-score van 1, dan zie je dat de rechter overschrijdingskans 0,1587 is. Met andere woorden: de kans dat iemand minimaal één standaarddeviatie hoger of lager scoort dan het gemiddelde, is 15,87%. De kans dat iemand minimaal twee standaarddeviaties hoger of lager scoort dan het gemiddelde, is maar 2,28%.
Stel, je hebt je eigen IQ laten berekenen, dat is 125, en je wilt weten hoeveel procent van de volwassenen een hoger IQ heeft dan jij. Je moet dan eerst je IQ-score omzetten in een z-score

$$z = \frac{(x - \bar{x})}{s} = \frac{125 - 100}{15} = \frac{25}{15} = 1{,}667$$

Jij scoort dus 1,67 standaarddeviaties boven het gemiddelde. Deze waarde opzoeken in de tabel leert ons dat 4,75% van de volwassenen deze waarde of hoger scoort. Dat betekent dus ook dat 100 – 4,75 = 95,25% een lager IQ dan jij heeft.



Figuur 3.13 Berekenen van kansen via de standaardnormale verdeling

Hoe meer standaarddeviaties een onderzoekseenheid afwijkt van het gemiddelde, hoe minder groot de kans is dat die waarde vaak voorkomt. Het is waarschijnlijker dat je iemand treft met een IQ van 115 of hoger (namelijk 15,87%) dan dat je iemand treft met een IQ van 130 of hoger (namelijk 2,28%). We spreken van een *extreme waarde* wanneer een onderzoekseenheid vijf standaarddeviaties ($z < -5$ of $z > 5$) onder of boven het gemiddelde scoort, en van een *uitbijter* (*outlier*) wanneer een onderzoekseenheid drie standaarddeviaties onder of boven het gemiddelde scoort ($z < -3$ of $z > 3$). Een extreme waarde of uitbijter kan ervoor zorgen dat de verdeling scheef wordt.

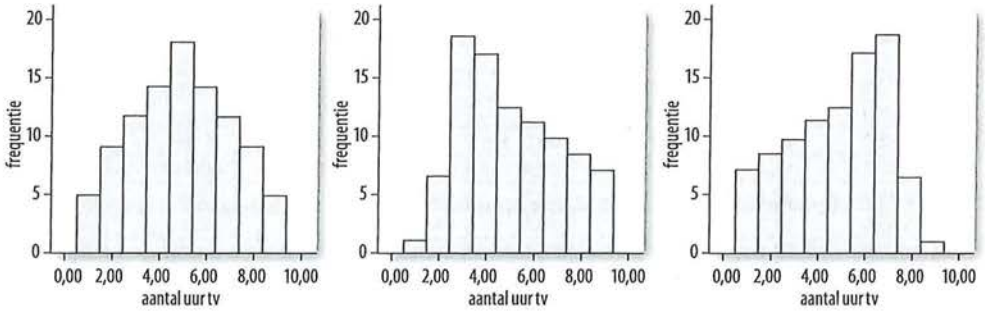
3.6.2 Scheve verdelingen

Een *scheve verdeling* is een verdeling die niet symmetrisch is. Scheve verdelingen kunnen ofwel scheef naar rechts, ofwel scheef naar links zijn. Scheefheid (in het Engels *skewness*) ontstaat wanneer ten opzichte van de modus aan één kant van de verdeling meer afwijkende waarden voorkomen dan aan de andere kant. Er zijn dan aan één kant extreme waarden, ofwel waarden die ver afliggen van de modus. Liggen de extreme waarden aan de linkerkant, dan is de verdeling scheef naar links, en liggen de extreme waarden rechts, dan is de verdeling scheef naar rechts (zie figuur 3.14).

Uur tv	Frequentie
1	5
2	9
3	12
4	15
5	18
6	15
7	15
8	9
9	5
N	100
Modus	5
mediaan	5
gemiddelde	5
skewness	0

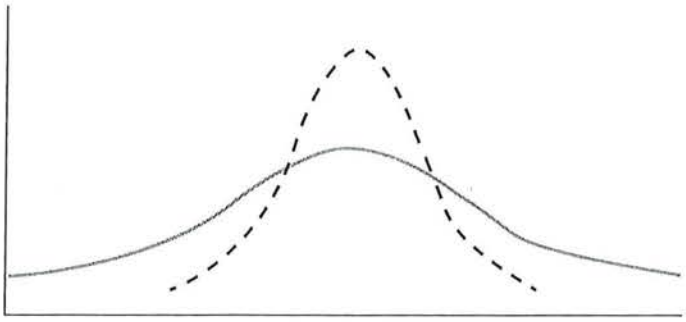
Uur tv	Frequentie
1	1
2	7
3	22
4	20
5	12
6	11
7	10
8	9
9	8
N	100
Modus	3
mediaan	4,5
gemiddelde	5
skewness	0,412

Uur tv	Frequentie
1	8
2	9
3	10
4	11
5	12
6	20
7	22
8	7
9	1
N	100
Modus	7
mediaan	5,5
gemiddelde	5
skewness	−0,412



Figuur 3.14 Histogrammen met bijbehorende frequentieverdeling van een normaal verdeelde variabele, een verdeling die scheef is naar rechts en een verdeling die scheef is naar links

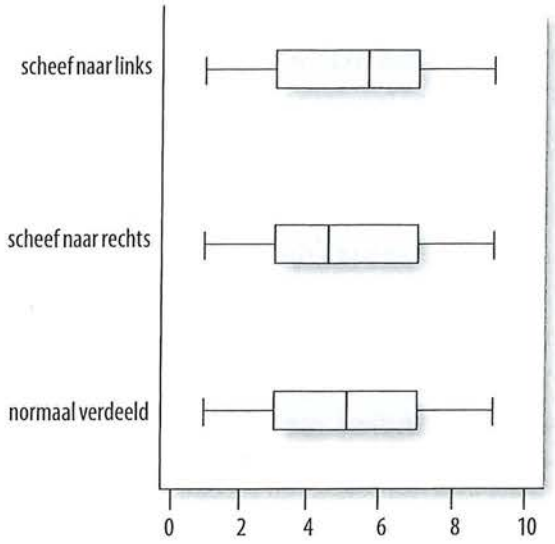
Behalve over de scheefheid van de verdeling kun je ook iets zeggen over de gewelfdheid van de verdeling, dat wil zeggen hoe plat of spits de verdeling is. Deze gewelfdheid noem je *kurtosis*. Een hoge kurtosis wijst op een verdeling met een sterke piek, een lage kurtosis wijst op een platte verdeling (zie figuur 3.15).



Figuur 3.15 Een platte en een spitse verdeling

Hoe spitser de verdeling (in figuur 3.16 de gestippelde lijn), hoe minder spreiding er is, en hoe minder extreme waarden zich in de verdeling bevinden.
De mate van scheefheid en kurtosis kunnen we door SPSS laten berekenen. In ons voorbeeld van aantal uur televisiekijken is sprake van een normale verdeling. Modus, mediaan en gemiddelde zijn 5. De standaarddeviatie is 2,17. De *skewness* is nul, dat wil zeggen dat de verdeling geheel symmetrisch is. Als de *skewness* een positieve waarde heeft, is de verdeling scheef naar rechts en als de *skewness* een negatieve waarde heeft, is de verdeling scheef naar links (figuur 3.14[6]). Over het algemeen hanteren we een marge van 1 (zowel naar links als naar rechts) om te bepalen of een verdeling te scheef is of niet. Wanneer een variabele te scheef is verdeeld, gaat de empirische regel niet meer op en is het niet meer zomaar mogelijk om kansen te berekenen aan de hand van een standaardnormaalverdeling.

In een histogram en een boxplot is het verschil tussen normale en scheve verdelingen snel te herkennen. In de onderste boxplot in figuur 3.16 zien we een normaal verdeelde variabele voor aantal uur televisiekijken, in de twee boxplots daarboven zien we scheve verdelingen voor de variabele aantal uur televisiekijken. De boxplot bovenaan is iets scheef naar links, de boxplot daaronder is iets scheef naar rechts. In alle drie de gevallen zijn honderd respondenten ondervraagd over hun kijkgedrag (uren televisiekijken). De laagste en hoogste waarden zijn in alle drie de boxplots gelijk. In de scheve verdeling naar links zitten meer extreme waarden in de linkerkant van de verdeling, en zit de mediaan meer rechts van het midden. Andersom zitten in de scheve verdeling naar rechts meer extreme waarden in de rechterkant van de verdeling, en zit de mediaan meer links van het midden. Bij de normale verdeling is er een symmetrische verdeling van waarden rond de mediaan.



Figuur 3.16 Boxplots van normale en scheve verdelingen

3.7 Samenvatting

Spreiding is de mate waarin de waarden van een variabele variëren. Bij de variatie, de variantie en de standaarddeviatie bereken je dit op basis van de afstand ten opzichte van het gemiddelde. Daarom gebruik je deze spreidingsmaten enkel op interval- en rationiveau. Ook voor de berekening van de interkwartielafstand is minimaal een meting op intervalniveau nodig.
De variatie is de kwadratensom van de afstanden van alle onderzoekseenheden tot het gemiddelde. Bij de variantie deel je deze kwadratensom door het aantal onderzoekseenheden minus 1. Omdat een kwadratensom moeilijk te interpreteren is (de waarden die verkregen worden, staan niet in verhouding tot de oorspronkelijke waarden), wordt voor de berekening van de standaarddeviatie de wortel getrokken uit de berekende variantie: $s = \sqrt{s^2}$

Om variabelen met verschillende meeteenheden met elkaar te kunnen vergelijken, worden de waarnemingen bij de onderzoekseenheden op deze variabelen gestandaardiseerd. Dit gebeurt door middel van een z-score. Deze score geeft aan hoeveel standaarddeviaties de waarneming van het gemiddelde afligt. Met een z-score kan in de tabel voor standaardnormaalverdelingen een overschrijdingskans worden opgezocht die aangeeft hoeveel kans er is dat een waarde boven (of onder) die z-score wordt gevonden. Een voorwaarde voor deze manier van kansberekening is dat de variabele normaal verdeeld is, en niet te scheef. Wanneer er te veel extreme waarden zijn, gaat de empirische regel niet meer op.

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

Noten

1 Ter herinnering: in dit boek zullen wij altijd rekenen met drie decimalen achter de komma. Bij het interpreteren van de waarden (meestal in de conclusie) ronden we (pas) af naar twee decimalen.
2 Door afrondingsverschillen is dit soms iets meer of iets minder dan exact nul.
3 Er wordt door $n - 1$ gedeeld en niet door n omdat je meestal wilt dat de variantie in een steekproef een schatting geeft van de variantie in de populatie. De $n - 1$ geeft een correctie die maakt dat de variantie als schatter kan dienen voor de variantie in de populatie. Als je de variantie niet als schatter voor een populatiewaarde gebruikt, dan kun je door n delen (σ^2). Omdat wij onze berekeningen met SPSS willen controleren en in SPSS bij de berekeningen door $n - 1$ wordt gedeeld, gebruiken we s^2.
4 Hartman, L., Okken, V. & Rompay, T. van (2014). Evaluating books by their covers; de invloed van realisme en complexiteit in fotografiegebruik op de waardering van tweens. *Tijdschrift voor Communicatiewetenschap, 42*(2), pp. 221-243.
5 In dit boek zullen we niet stilstaan bij de betekenis van de asterisken bij de cijfers in de tabel. Voor de beschrijvende statistiek volstaat het overzichtelijk weergeven van de gemiddelden en standaarddeviaties.
6 Overigens valt het met de scheefheid van de verdelingen in de voorbeelden in figuur 3.15 en figuur 3.17 nog wel mee. Frequentieverdelingen zijn soms veel schever verdeeld.