Hoofdstuk 3
De univariate beschrijvende statistiek

1. Inleiding

Een datamatrix voor statistische analyse bestaat uit één of meerdere metingen, scores of waarden voor alle verschillende onderzoekseenheden (individuen, objecten, gebieden, criminele gebeurtenissen,…). De methodologie voor het organiseren en samenvatten of beschrijven van de gegevens voor een steekproef of de gehele populatie, wordt beschrijvende statistiek genoemd. Welke statistische technieken men hierbij gebruikt, is afhankelijk van **twee zaken**: (1) de **onderzoeksvraag** en (2) het **meetniveau** van de kenmerken van de eenheden waarover we uitspraken willen doen. Eenheden verschillen van elkaar in termen van bepaalde kenmerken. Studenten verschillen bijvoorbeeld in scores op hun examens, in leeftijd, geslacht,… Men kan deze verschillende waarden bestuderen en dit doet men door ***frequentieverdelingen*** en ***grafieken*** te maken die deze gegevens samenvatten en visualiseren. Door gebruik te maken van de frequentieverdelingen worden twee zaken duidelijk: de grote ***centrale tendensen*** en ***bijzondere observaties***, i.e. observaties die weinig voorkomen. We leren dus nogal wat uit wat ogenschijnlijk maar een eenvoudige frequentieverdeling is.

Verder kiest men voor parameters van ***centraliteit***, om de centrale tendensen weer te geven en parameters van **spreiding** om de waargenomen verschillen samen te vatten. Bij elk van de mogelijkheden moet telkens eerst de vraag gesteld worden wat het meetniveau is van de variabele. Immers, het meetniveau bepaalt welke mogelijkheden er bestaan. Vergelijk het geheel van statistische analysetechnieken met een gereedschapskist waaruit je diverse soorten van gereedschap kan kiezen. Je kiest wat je nodig hebt, maar daarin word je beperkt door de mogelijkheden van de variabelen. Hierbij geldt de algemene regel dat de mogelijkheden die we hebben bij het laagste meetniveau ook voor de hogere meetniveaus mogen gebruikt worden. Het omgekeerde is echter niet mogelijk.

2. Over absolute en relatieve frequenties en hun grafische voorstelling

Kenmerken van onderzoekseenheden kunnen verschillende waarden hebben. Bij sommige kenmerken zijn de waarden al een getal, bij andere kenmerken zou je voor de voorkomende categorieën een getal kunnen verzinnen. De waarden van bijvoorbeeld het kenmerk leeftijd zijn getallen die direct gerelateerd zijn aan de werkelijkheid. Als een persoon 21 jaar oud is, is het logisch dat deze persoon de waarde 21 krijgt voor de variabele ‘leeftijd in jaren’. Geslacht heeft

geen vaststaande numerieke waarde. Om in de statistiek toch op een geordende wijze iets te kunnen zeggen over de onderzoekseenheden, krijgen de categorieën ‘man’ en ‘vrouw’, waarin de variabele ‘geslacht’ kan worden onderverdeeld, wel een numerieke waarde om de dataverwerking te vergemakkelijken. Je zou kunnen besluiten vrouwen de waarde 1 te geven en mannen de waarde 2. Op die manier kun je alle onderzoekseenheden voorzien van een numerieke waarde voor het kenmerk ‘geslacht’.

Een variabele (kenmerk) met de daarbij behorende waarden kun je op een overzichtelijke manier presenteren in een **frequentietabel**. Stel, in de zomervakantie zit je met wat vrienden op een terrasje en jij bent aangewezen om de drankjes te halen. Je vraagt je vrienden wat ze willen drinken. Je kunt proberen alles te onthouden, maar op een bierviltje de drankjes turven is gemakkelijker. Door te turven maak je een overzicht van het aantal keer dat een waarde voorkomt. Het tellen van de streepjes brengt je op de **absolute frequentie** van het kenmerk “te consumeren drankje” bij de onderzoekseenheden “vrienden”. Jan en Piet drinken een pint, Johan, Katrien en Mohammed drinken een koffie en Jana drinkt een thee. Die variabele die we hier beschreven hebben is “drankje”. Dit is een variabele van het nominale niveau. Hoezo nominaal niveau? Je kan toch ordenen, bijvoorbeeld op basis van het alcoholpercentage? Dat is juist, maar dan zou de variabele “alcoholpercentage” zijn en de eenheid het drankje. Hier zijn de eenheden jouw vrienden en het drankje is gewoon wat ze besteld hebben op het terras. Dit kan allemaal heel banaal klinken, we zouden het hier niet neerschrijven mochten studenten hier niet zo frequent fouten tegen maken. Je bent sneller verstrooid dan je denkt. Je leest een vraag maar half en je denkt het antwoord al te weten. Zo werkt het niet. Statistiek heeft te maken met nauwkeurigheid en dat begint vanaf het begin, van de meest banale analyse tot de meest complexe. Het moet juist zijn.

In kwantitatief criminologisch onderzoek is de situatie analoog als de situatie die we daarnet hebben beschreven. Je stelt in een criminologisch onderzoek geen gewone alledaagse vraag, maar een onderzoeksvraag die gerelateerd is aan het fenomeen dat je wil bestuderen. Je bent criminoloog en je wilt weten hoeveel verschillende delicten jongeren binnen de tijdspanne van een jaar plegen. Je stelt dus een **beschrijvende onderzoeksvraag**: hoeveel verschillende delicten heeft u het afgelopen jaar gepleegd?
Hoeveel jongeren rapporteren als misdrijf omschreven feiten, of meer algemeen ‘hoeveel elementen van de steekproef hebben een bepaalde waarde op de variabele criminaliteit?’ is een elementaire onderzoeksvraag in jeugdcriminologisch onderzoek van beschrijvende aard. Het

aantal elementen met een bepaalde waarde van een variabele noemt men **de absolute frequentie van die waarde**. Men kan deze bepalen door een frequentietabel op te stellen. Dit kan voor elk meetniveau.

We komen terug op het voorbeeld van het delinquent gedrag dat jongeren plegen. Het is een voorbeeld uit een masterproef in de criminologische wetenschappen. Een oud-student deed onderzoek naar jeugdcriminaliteit en vroeg in verschillende scholen hoe vaak de jongeren bepaalde delicten gepleegd hadden in een bepaalde periode (twaalf maanden). De vraag uit de vragenlijst luidde: "hoe vaak heb je het afgelopen jaar iets gestolen": de antwoordcategorieën waren 0 = nul keer, 1= één keer,… , 6 = zes tot tien keer en 7 = meer dan tien keer. De ruwe antwoorden van de jongeren waren:
0 0 1 3 0 0 4 1 "geen antwoord" …

Om een overzicht te krijgen over al deze waarnemingen, plaatsen we ze in een tabel waarin in de eerste kolom de verschillende waarden worden genoteerd die voorkomen. In een tweede kolom tellen we hoeveel keren elke categorie voorkomt. In de oude dagen dienden onderzoekers deze antwoorden te turven, nu gebeurt dit aan de hand van software. Hieronder zie je een tabel zoals deze gemaakt wordt met het statistische verwerkingspakket SPSS. Het voorbeeld is gebaseerd op de variabele diefstal. Let op de verschillende waarden die je in de rijen ziet staan.

Variabele: Hoe vaak heb je het afgelopen jaar iets gestolen?

Waarde		Frequentie Absolute aantallen	Percentages Relatieve aantallen	Geldige Percentages	Cumulatieve geldige percentages
Valid	nul keer	2848	92,4	94,4	94,4
	één keer	60	1,9	2,0	96,4
	twee keer	33	1,1	1,1	97,5
	drie keer	21	,7	,7	98,2
	vier keer	10	,3	,3	98,5
	vijf keer	10	,3	,3	98,9
	zes tot tien keer	11	,4	,4	99,2
	meer dan tien keer	23	,7	,8	100,0
	Totaal	3016	97,9	100,0	
Missing	System	66	2,1		
n (steekproefgrootte)		3082	100,0		

In de eerste kolom zie je de waarden die de vraag “hoe vaak heb je iets gestolen” in ons voorbeeld heeft. SPSS maakt een onderscheid tussen geldige waarden (valide) en ontbrekende waarden (missing). In de tweede kolom (Frequentie) staan de absolute frequenties, het aantal keer dat een bepaalde waarde voorkomt. De waarde “nul keer” komt klaarblijkelijk het meest voor. 2848 respondenten hebben een score nul keer. Dat wil zeggen dat 2848 respondenten het afgelopen jaar niets gestolen hebben.

Daarnaast staan de **percentages**. Percentages gaan van nul procent tot honderd procent. Het percentage dat een bepaalde waarde voorkomt is de verhouding tussen het absolute aantal keer dat een waarde voorkomt, gedeeld door het *steekproefeffectief (n),* vermenigvuldigd met honderd. Als je aandachtig kijkt naar deze tabel, dan zie je dat 2,1 procent deze vraag niet heeft ingevuld of niet beantwoord heeft. 2,1 procent is het percentage ontbrekende informatie. Hier valt dit nog mee. De vraag is niet zo bedreigend dat een meerderheid er niet op geantwoord heeft. In ander onderzoek gebeurt het dat 20% de vraag niet beantwoordt. Je kan je dan afvragen of dat geen invloed heeft op de resultaten. Voorzichtigheid is dus zeker geboden. In ons voorbeeld hebben 66 respondenten de vraag niet ingevuld. In het valide percentage worden deze niet meegeteld.

De som van alle **absolute frequenties** is gelijk aan het totaal aantal elementen in de steekproef en wordt voorgesteld door de kleine letter “n”. Soms wordt de hoofdletter N gebruikt. De hoofdletter wordt gebruikt als het gaat om een populatie, de kleine letter wordt gebruikt als het gaat om een steekproef. De steekproefgrootte bedraagt 3082 respondenten. De statistische notatie is de volgende:

$$n = f_1 + f_2 + .. + f_m = \sum_{i=1}^{m} f_i = n$$

Hier staat: 2848+60+33+…+66 = 3082.

We kunnen ook de **relatieve frequentie** berekenen door **percentages** (kolom 3) te presenteren. Deelt men elke absolute frequentie door het totaal aantal waarnemingen (hier 3082) dan bekomt men **proporties**. Deze zijn hier niet weergegeven.

$$f_i' = \frac{f_i}{n}$$

Vermenigvuldigt men deze proporties met de waarde honderd dan spreken we over **percentages**. We geven dit als volgt weer:

$$f_i' = \frac{f_i}{n} x100$$

In de kolom **geldige percentages** zien we dat de respondenten die niet op de vraag hebben geantwoord, niet meetellen. Voor de rest is de berekening dezelfde. Het steekproefeffectief (n) is dus iets kleiner. Vaak houdt men bij het maken van een tabel ook rekening met de mensen die 'geen antwoord' gaven, de zogeheten *weigeraars*. Men herberekent de percentages enkel voor diegenen die een antwoord hebben gegeven. Bij deze berekeningen wordt het aantal mensen die niet antwoordden of niet moesten antwoorden, afgetrokken van het totaal aantal waarnemingen. Dit nieuwe totaal (de geldige percentages) is dan het getal waardoor elke absolute frequentie wordt gedeeld. Het aantal **geldige percentages** vinden we in de derde kolom terug. Dit zijn de percentages berekend op diegenen die een *geldig antwoord* hebben op de vraag.

In de kolom daarnaast zien we de ***cumulatieve percentages****:* hier worden de percentages van elke volgende waarde bij de voorgaande opgeteld. Een criminoloog die statistieken analyseert laat zich niet misleiden door wat in één kolom staat, maar brengt in rekening wat in elke kolom staat.

Het voordeel van het gebruik van relatieve frequenties en percentages is dat de frequentieverdelingen voor verschillende, niet even grote groepen personen beter vergelijkbaar worden. Wanneer **minstens op een ordinaal niveau** werd gemeten, heeft het verder ook zin om de **cumulatieve percentages** weer te geven. *Anders gezegd: het heeft absoluut geen zin om voor nominale kenmerken cumulatieve percentages te berekenen.* Aan de hand van deze cumulatieve verdeling kunnen we zien hoeveel waarnemingen kleiner dan of gelijk aan een bepaalde waarde zijn. We stellen dit als volgt voor:

$$K(x_i) = \sum_{x_j \leq x_i} F_j$$

Zo kunnen we onmiddellijk aflezen dat meer dan 90% van de bevraagde jongeren het afgelopen jaar geen enkele diefstal heeft gepleegd. We onderscheiden *absolute cumulatieve frequenties* (voorgesteld door de hoofdletter **K**) en *relatieve cumulatieve frequenties* of cumulatieve percentages (voorgesteld door de kleine letter **k**).

Opgelet: **percentages zijn niet altijd geschikt om gegevens voor te stellen**. **Bevat de totale steekproef minder dan dertig eenheden, dan werkt men beter met absolute aantallen.** De reden hiervoor is eenvoudig: je kan misleidende resultaten presenteren. Als één respondent op drie antwoordt dat deze al cocaïne gesnoven heeft, betekent dat 33.33% van de steekproef. Dat percentage is misleidend en totaal betekenisloos aangezien je maar drie respondenten hebt. Deze fout wordt heel vaak gemaakt door criminologen die kwalitatief onderzoek doen, en die denken dat ze hun onderzoek wat wetenschappelijker kunnen maken door een statistiekje te maken op basis van hun beperkte steekproeven.

Wanneer gepercenteerde tabellen worden weergegeven in een rapport, moet absoluut het steekproefeffectief (n) worden vermeld. Enkel dan kan men zelf de gegevens nog (beperkt) herordenen. In criminologisch onderzoek heeft men het vaak over **incidentie** en **prevalentie** van criminaliteit. Incidentie is het aantal nieuwe gevallen van een bepaalde conditie dat voorkomt in een populatie gedurende een bepaalde periode. Bijvoorbeeld het totaal aantal nieuwe gedetineerden in de Gentse gevangenis gedurende een jaar. Prevalentie is het totaal aantal personen in een bepaalde conditie in een populatie op een bepaald moment. Bijvoorbeeld het aantal gedetineerden in de Gentse gevangenis op 1 januari 2015.

Grafische voorstellingen

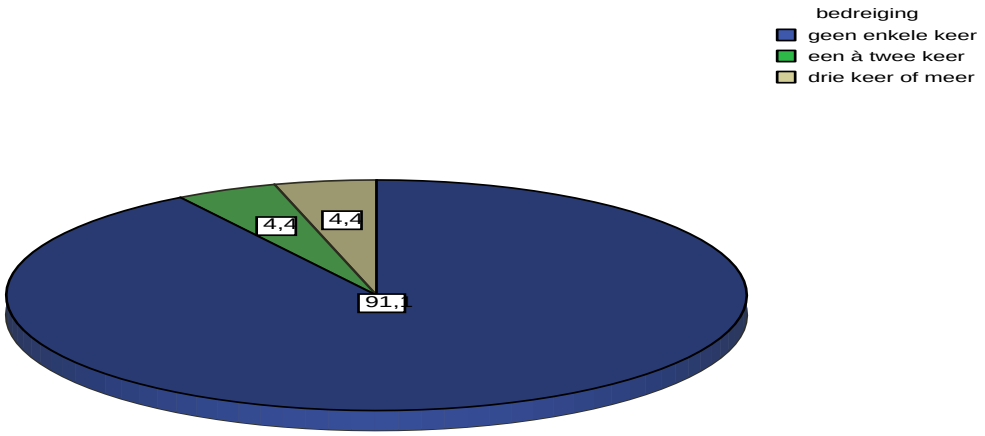
Een grafiek zegt soms meer dan duizend woorden. Dat is een heel terechte uitspraak. In een oogopslag zie je wat opvalt: de trends in criminaliteit, de stijging of daling. De grillige vormen of stabiele patronen inzake misdadigheid en de kenmerken waarmee misdadigheid samenhangt worden in een oogopslag zichtbaar gemaakt. Nog anders gezegd: met grafische voorstellingen weet je precies welk soort vlees je in de kuip hebt, of anders gezegd wat voor steekproef je hebt, in termen van de belangrijke variabelen die je bestudeert.

Een grafische voorstelling streeft altijd naar het overbrengen van de informatie van complexe gegevens via beelden (tekeningen). Hierbij is het belangrijk dat het geconstrueerde beeld in overeenstemming is met de reële informatie in de gegevens. Verder mag het informatiegehalte van een grafiek niet te klein zijn maar ook niet te groot. Bij het tekenen van grafieken is het belangrijk om ook rekening te houden met het **meetniveau** van de variabele. Grafische voorstellingen worden aan de hand van statistische verwerkingspakketten gemaakt.

Taartdiagram of cirkelgrafiek (pie chart)

Bij een taartdiagram wordt op basis van de verschillende frequenties of percentages een cirkelschijf verdeeld in sectoren.

Figuur: pie chart van de variabele "hoe vaak heb je het afgelopen jaar iemand bedreigd"



Deze grafiek is gemakkelijk te interpreteren en visualiseert de informatie uit de tabel. **Taartdiagrammen** zijn populair bij kenmerken gemeten op het **nominale** en **ordinale** niveau. We gebruiken deze grafiek eigenlijk wanneer we met een beperkt aantal categorieën werken. Dit is belangrijk met betrekking tot de duidelijkheid. Bij meer dan 5 categorieën wordt het moeilijk om de verschillende cirkelsectoren van elkaar te onderscheiden. De volgorde van de categorieën hangt af van het meetniveau. Bij nominale variabelen is de volgorde in principe willekeurig. Toch is het interessant om categorieën die inhoudelijk bij elkaar aansluiten naast elkaar te plaatsen. Voor ordinale en metrische variabelen volgt men het ordeningscriterium bij het weergeven van de categorieën.

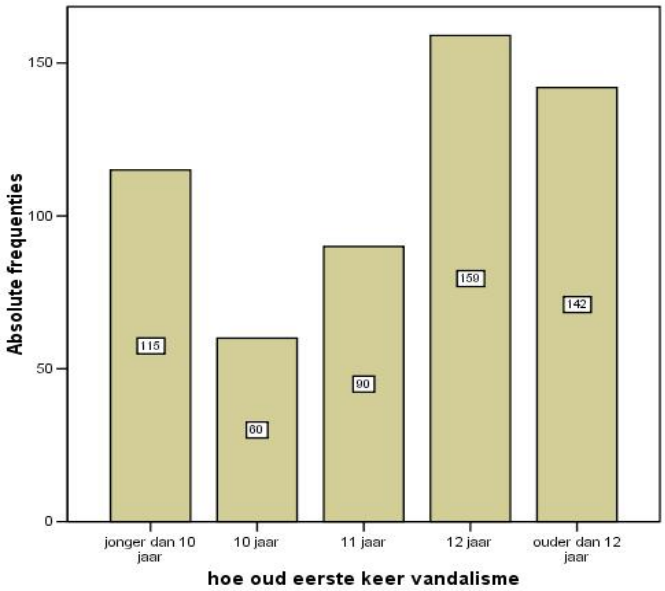
Staafdiagram (bar chart)

Bij een staafdiagram worden de gegevens op ***twee assen*** voorgesteld. Op de ***horizontale as*** *of de X-as,* worden **de verschillende categorieën van de variabele X weergegeven**. De ***verticale as****, of de Y-as* geeft de aantallen weer. Dit kan onder de vorm van absolute frequenties of percentages. Op elke waarde van X tekent men een staafje. De oppervlakte van deze verschillende staven drukken de absolute aantallen of percentages uit. De staafjes kunnen zowel horizontaal als verticaal worden voorgesteld.
Als de staafjes driedimensionaal worden voorgesteld, spreken we van een *blokgrafiek*.

Er zijn een aantal regels die in acht moeten worden genomen.

- Staafdiagrammen worden gebruikt bij de grafische voorstelling van kenmerken gemeten op het **nominale** en **ordinale** niveau.
- We plaatsen de staafjes los van elkaar omdat de variabele nominaal of ordinaal is.
- Indien het om een metrische variabele gaat, mogen de staafjes bij elkaar aansluiten en wordt hierdoor visueel weergegeven dat de categorieën zich op ***een continuüm*** bevinden (we spreken dan van een *histogram* – zie verder).
- In een staafdiagram kunnen meer categorieën worden voorgesteld.

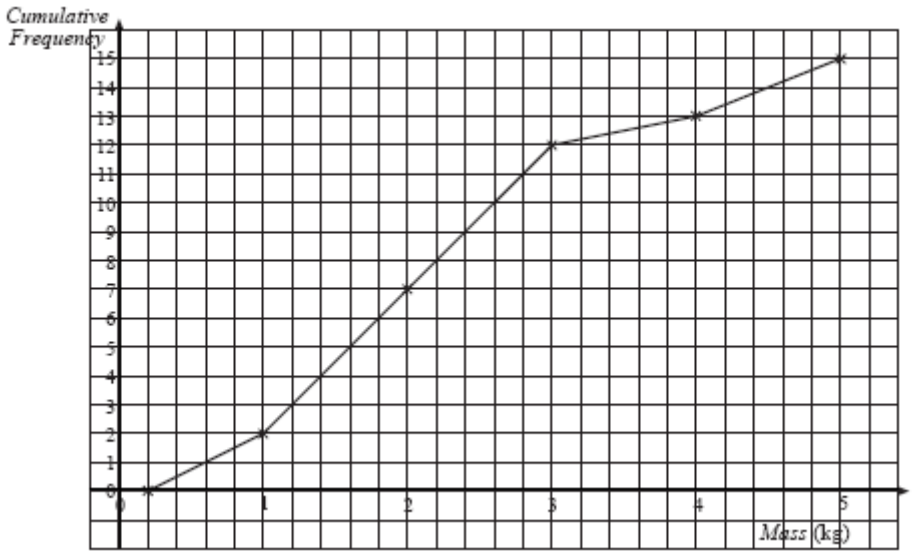
Figuur: staafdiagram van het kenmerk "leeftijd bij het eerste delict"



Cumulatief frequentiediagram

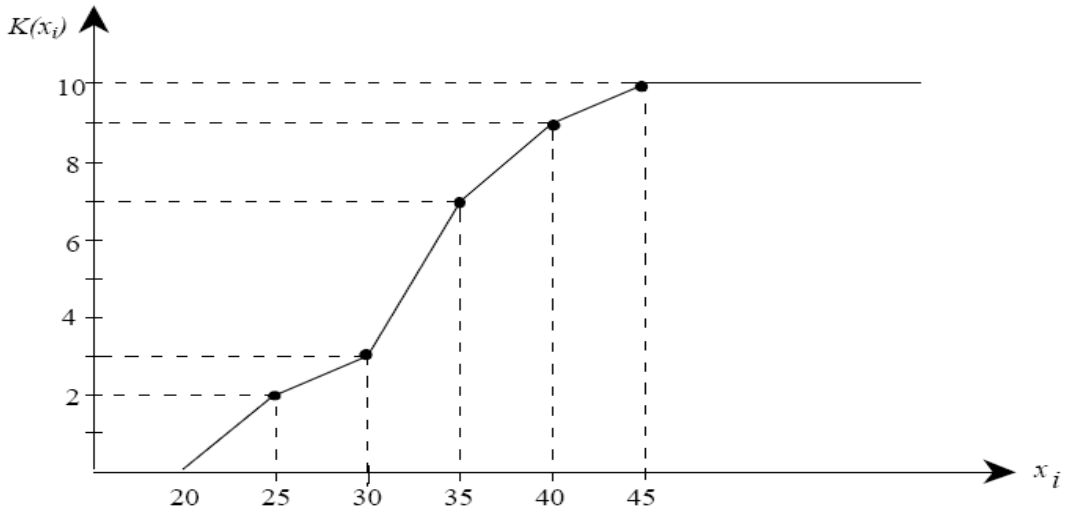
De absolute of relatieve cumulatieve frequenties kunnen worden voorgesteld in een cumulatief frequentiediagram en dit vanaf het **ordinale** niveau. Voor gegevens die niet in klassen werden ingedeeld, tonen we hieronder een voorbeeld.

Figuur: cumulatief frequentiediagram



We zien op dit cumulatieve diagram dat op de verticale as de cumulatieve frequenties staan en dat op de horizontale as de verschillende waarden staan. De variabele is hier massa (uitgedrukt in kg). Wanneer de waarden in klassen zijn ingedeeld, zal het cumulatieve frequentiediagram er enigszins anders uitzien. Bij indeling in klassen gaan we immers uit van de hypothese dat de waarnemingen gelijkmatig verdeeld zijn over de verschillende klassen. Het gevolg hiervan is dat het diagram niet langer een trapfunctie zal zijn, maar een gebroken lijn.

Figuur: cumulatief frequentiediagram

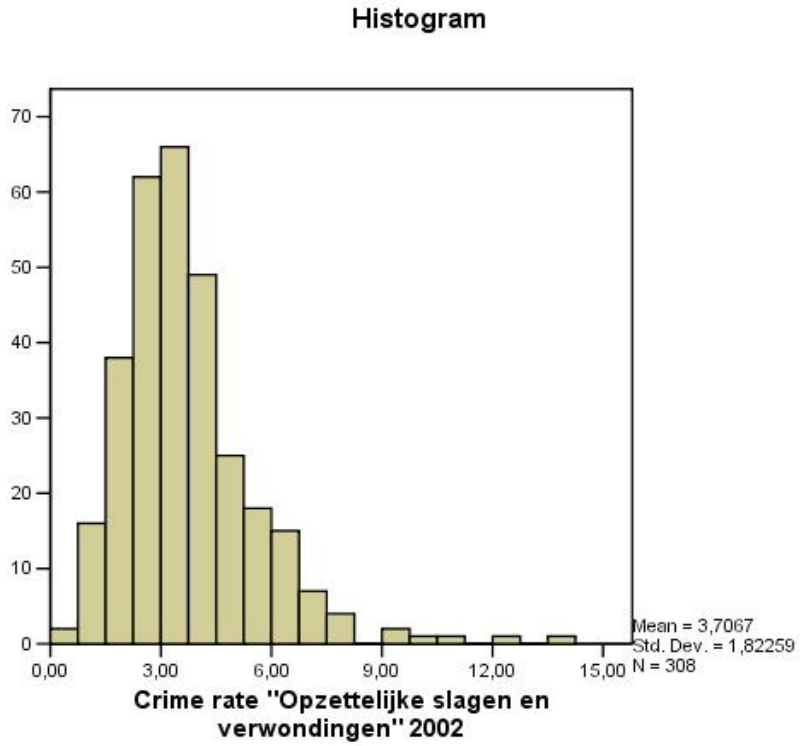


Histogram

Het histogram is een veelgebruikte visuele voorstelling voor **metrische** kenmerken. **Metrische** gegevens in een klassentabel worden vaak voorgesteld d.m.v. een **histogram**: de breedte van elk balkje komt overeen met de klassenlengte; de hoogte van elk balkje met het aantal gegevens dat binnen die klasse valt. Bemerk dat de balkjes elkaar moeten raken! Mochten er spaties tussen de balkjes zitten, zouden mensen verkeerdelijk kunnen denken dat de bijbehorende lengtes niet voorkomen of niet mogelijk zijn.

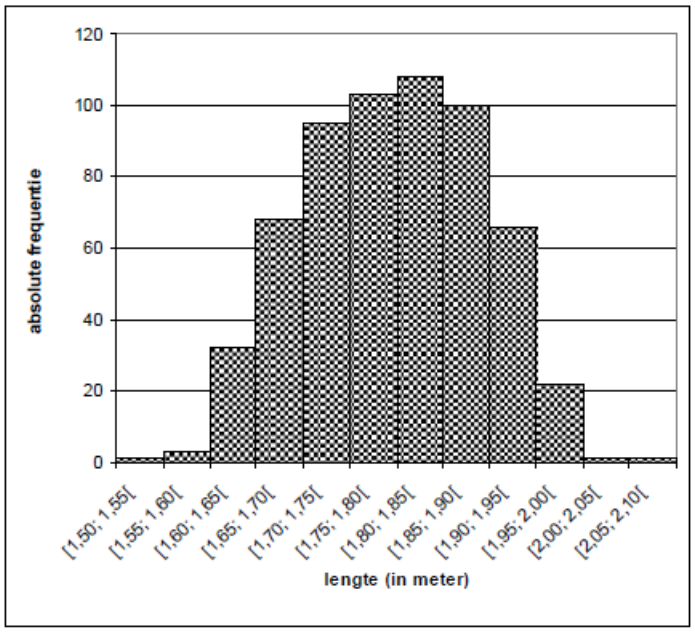
Dit is een belangrijk verschil met het staafdiagram. De blokjes worden aan elkaar getekend omdat de waarden van X elkaar opvolgen. Door deze voorstelling wordt duidelijk gemaakt dat de categorieën op een ***continuüm*** liggen. De oppervlakte van de kolom is steeds gelijk aan de frequentie van de waarde die de kolom voorstelt. De totale oppervlakte van het histogram is dan gelijk aan het totaal van het aantal elementen. De totale oppervlakte bevat honderd procent van de waarnemingen. Bij gegevens in klassen ingedeeld, worden de klassengrenzen weergegeven. Het voorbeeld dat we presenteren is het histogram van de crime rate voor opzettelijke slagen en verwondingen in Vlaamse gemeenten.

Figuur: histogram voor de criminaliteitsgraad voor geweld



Figuur: histogram voor lengte (in klassen gegroepeerd)

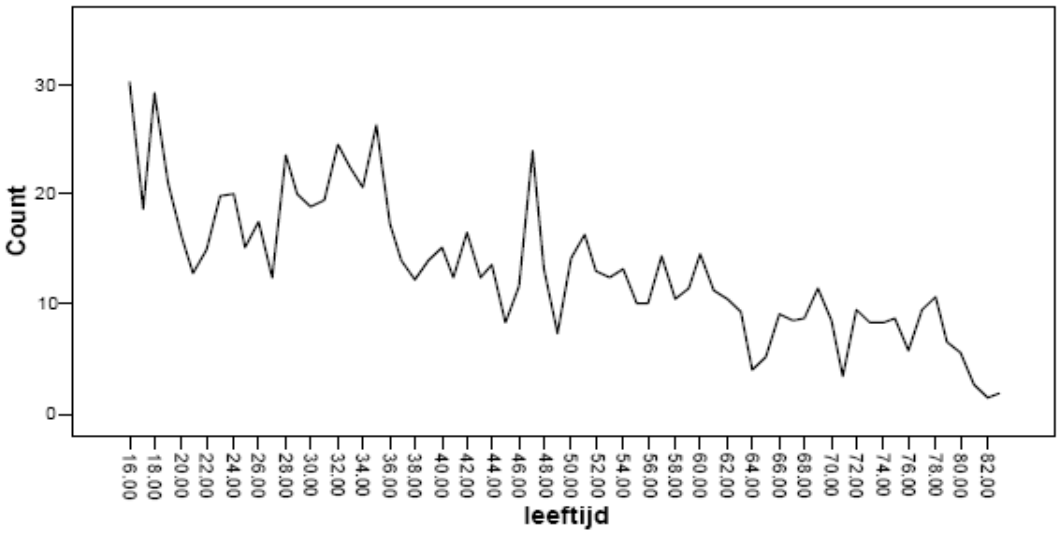
Lengte (in m)	aantal
[1,50; 1,55[	1
[1,55; 1,60[	3
[1,60; 1,65[	32
[1,65; 1,70[	68
[1,70; 1,75[	95
[1,75; 1,80[	103
[1,80; 1,85[	108
[1,85; 1,90[	100
[1,90; 1,95[	66
[1,95; 2,00[	22
[2,00; 2,05[	1
[2,05; 2,10[	1
totaal	*600*



Lijndiagram

In een lijndiagram worden de niet in klassen gegroepeerde gegevens visueel voorgesteld door hun frequentie op een verticale as aan te duiden en de waarden op de x-as te stellen. De punten worden vervolgens door middel van een lijn met elkaar verbonden.

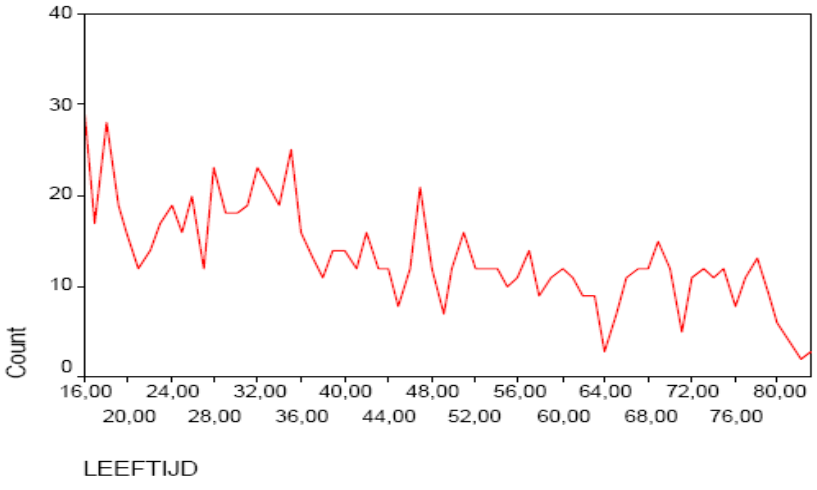
Figuur: lijndiagram voor de variabele leeftijd



Frequentiepolygoon

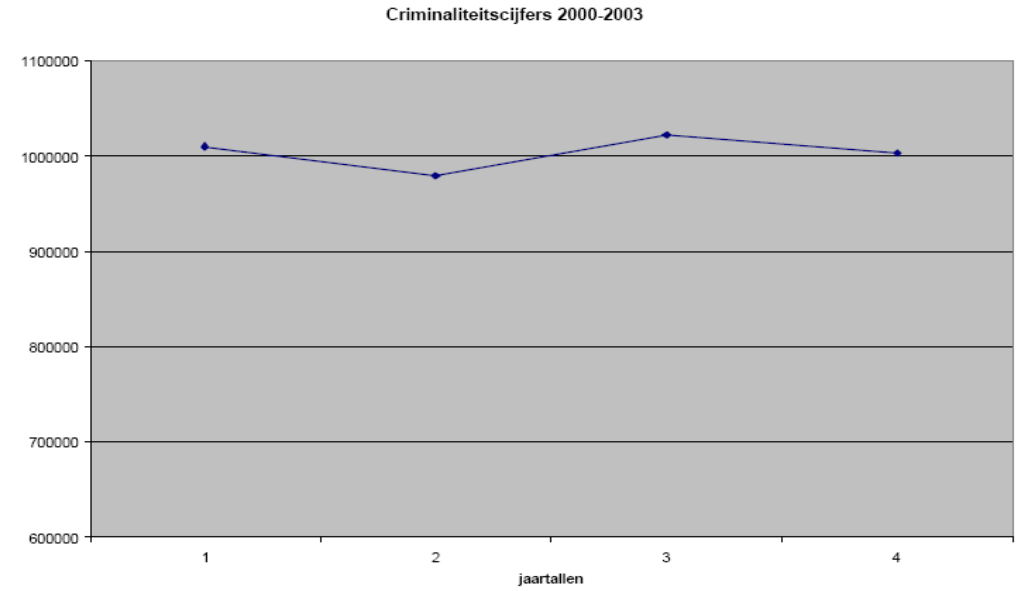
De frequentiepolygoon is nauw verwant aan het lijndiagram. Het is in feite een lijndiagram voor in klassen gegroepeerde gegevens. In een frequentiepolygoon worden de categorieën, voorgesteld door hun klassenmiddens en punten gevormd door het aanduiden van de hoogte van de frequentie, met elkaar verbonden.

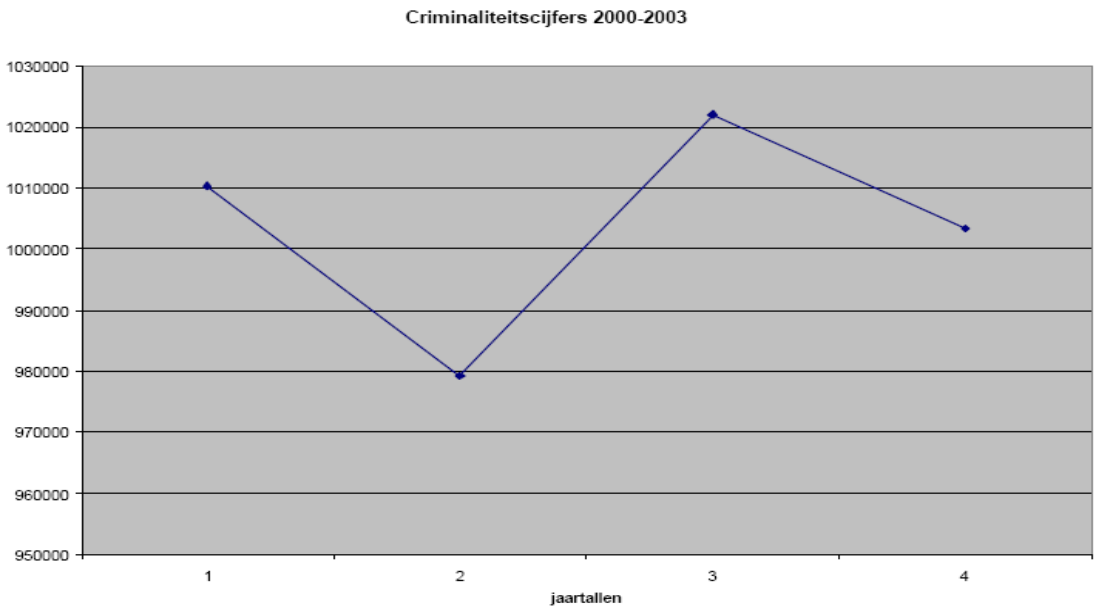
Figuur: frequentiepolygoon voor de variabele leeftijd



Opgelet met grafische voorstellingen

Grafieken kunnen heel verhelderend werken, doch kunnen ook heel bedrieglijk zijn. Vergeet daarom nooit ook de cijfers zelf van dichterbij te bekijken. De wijze waarop men de x- en y-as ijkt, is immers bepalend voor de mate waarin men detaillering ziet in het verloop van de cijfers en dus ook voor de wijze waarop het beeld zal overkomen bij de lezer. Onderstaand voorbeeld illustreert dit. In de eerste grafiek lijken de criminaliteitscijfers niet echt sterk te schommelen, er lijkt sprake te zijn van een vrij stabiele trend, daar waar de tweede grafiek veel grotere schommelingen lijkt te vertonen. Beide zijn nochtans opgesteld met exact dezelfde cijfers. Alles heeft uiteraard te maken de manier waarop de **Y -as** geijkt is.





Bron: Federale politie, criminaliteitscijfers 2000-2003

3. Parameters van centraliteit

Parameters van centraliteit geven een antwoord op **beschrijvende onderzoeksvragen**. Deze onderzoeksvragen zijn er op gericht **de centrale tendensen** te ontdekken. Centrale tendensen vinden we door te kijken naar centrale waarden. We spreken ook van ***centrummaten***. Een centrummaat hanteren we wanneer we de frequentieverdeling willen kenmerken aan de hand van een *centraal gelegen waarde*. Deze waarde wordt dan als een representatieve maat beschouwd die de volledige verdeling van de waarnemingen zo goed mogelijk karakteriseert. Het gebruik van een bepaalde maat hangt af van het meetniveau van de variabele.

De modus

De modus is een centrummaat en betreft de categorie van de variabele met de ***frequentie die het vaakst voorkomt.*** Aangezien de modus enkel verwijst naar de waarde met de hoogste frequentie is er geen enkel probleem om de modus te bepalen bij nominale variabelen, ordinale en metrische variabelen. De modus heeft echter een beperkte betekenis en verwijst enkel naar de meest voorkomende waarde. Het geeft de meest in het oog springende categorie aan. De modus is meestal heel stabiel en dit is haar sterkte. Er mogen al heel wat verschuivingen optreden in de frequenties van de verschillende categorieën voordat de modus wijzigt. Het is mogelijk dat er twee of meer modi voorkomen wanneer er twee of meer categorieën voorkomen met een even hoge frequentie.

Laten we een voorbeeld geven. In een jongerenbevraging werd gepeild naar spijbelgedrag van vrienden van de respondenten. De frequentietabel wordt hieronder weergegeven.

Hoeveel van je vrienden hebben ooit gespijbeld?

		Frequentie	Percentage	Geldig Percent	Cumulatief percentage
Valid	**geen enkele**	**1953**	**63,4**	**63,5**	63,5
	één vriend	924	30,0	30,0	93,5
	twee vrienden	144	4,7	4,7	98,1
	drie of meer vrienden	57	1,8	1,9	100,0
	Total	3078	99,9	100,0	
Missing	**System**	4	,1		
Total		3082	100,0		

Hieruit blijkt dat de **meest voorkomende waarde "geen enkele"** is. 1953 jongeren of 63.5 procent geeft dit antwoord. De meeste jonge adolescenten geven te kennen dat ze geen vrienden

hebben die spijbelen. De categorie “geen enkele” is hiermee de modus. De modus is niet 1953, maar de categorie die met dit aantal overeenkomt!

De mediaan

De mediaan van een statistische verdeling is het *midden van die verdeling*. De mediaan is een centrummaat die het punt in de frequentieverdeling aangeeft waaronder 50% van de gevallen en waarboven de andere 50% van de gevallen liggen. De frequentieverdeling wordt als het ware in twee gelijke stukken gedeeld. De mediaan vormt met andere woorden het *middelpunt van de verdeling*. De mediaan is de middelste van de (oneven aantal) waarden in de rangschikking naar grootte. Bij een even aantal waarden is de mediaan het gemiddelde van de beide middelste waarden. Om de mediaan te kunnen bepalen van een kenmerk moeten de categorieën in oplopende volgorde gerangschikt zijn. Dit betekent dat men minstens een **ordinaal meetniveau** moet hebben om de mediaan te mogen gebruiken. Strikt genomen is de definitie enkel van toepassing als het aantal elementen (n) in de frequentieverdeling oneven is. In die situatie is er één waarde die in het midden ligt. Bij een even aantal waarnemingen zijn er in feite twee middelste waarden. Gaat het om dezelfde waarden dan is dit de mediaan. Gaat het om twee verschillende waarden dan kan men één van de waarden per toeval selecteren of indien het meetniveau het toelaat het gemiddelde van de twee berekenen.
Een voorbeeld wordt hieronder uitgewerkt voor ruwe gegevens die niet in tabellen verwerkt zijn. We vragen aan 13 respondenten (n = 13) hoeveel keer ze slachtoffer werden van vandalisme aan hun fiets. We bekomen de volgende antwoorden:

5; 1; 3; 3; 4; 5; 1; 2; 3; 4; 4 ;5; 2;

Alle waarnemingen worden eerst van laag naar hoog geordend:

Waarden	1	1	2	2	3	3	3	4	4	4	5	5	5
Respondenten	1	2	3	4	5	6	**7**	8	9	10	11	12	13

n = 13 of oneven waardoor de mediaan gelijk wordt aan de waarde van de zevende respondent (n+1)/2. In dit voorbeeld is dit de waarde 3. Indien het hier om 12 eenheden zou gaan, dan zijn de 6de en 7de waarneming de middelste. Beide zijn 3 dus ook hier is de mediaan gelijk aan 3.

In de hierboven gepresenteerde tabel die spijbelgedrag van vrienden meet, is de mediaan gelijk aan de waarde *“geen enkele”*. **Hoe kan ik dat afleiden uit een frequentietabel?** Het is

eenvoudig. 3078 respondenten vulden een geldig antwoord in. De waarde die overeenkomt met de waarneming die zich situeert tussen de 1539ste en 1540ste waarneming (3078+1)/ 2), is eveneens **"geen enkele**". De helft van het aantal respondenten heeft geen enkele vriend die ooit gespijbeld heeft, de andere helft heeft geen enkele tot en met drie of meer vrienden die ooit gespijbeld hebben.

De kwantielen

De mediaan is een speciaal geval van de maten die we **kwantielen** noemen. Men kan een geordende rij elementen niet alleen in twee gelijke groepen indelen (zoals we doen wanneer we de mediaan gebruiken) maar in principe in om het even welk aantal groepen met een gelijk aantal elementen. Net zoals bij de mediaan verwachten we dat de variabele gemeten is op het ordinale meetniveau. Zo kan men geordende gegevens in drie gelijke groepen indelen: het laagste derde, het middelste derde en het hoogste derde. De twee waarden van de variabele die gebruikt worden om de drie groepen af te bakenen, zijn T1 en T2 (eerste en tweede terciel) Het is meer gebruikelijk om **kwartielen** te gebruiken. Hiervoor zijn drie waarden nodig: het eerste kwartiel (Q1 of de waarde waaronder zich 25% van alle eenheden bevindt), het tweede kwartiel (Q2 of de waarde waaronder 50% van alle eenheden valt of ook de mediaan) en het derde kwartiel (Q3; de waarde waaronder zich 75% bevindt). Bij voldoende eenheden kan men ook *decielen* (1ste deciel= 1ste 10 %, 2de deciel= 1ste 20 % etc.) of *percentielen* bepalen.

Kwartielen van het aantal verschillende zelfgerapporteerde feiten

n	**Valid (geldig)**	3015
	Missing (ontbrekende data)	67
Kwartielen	**Q1 (25%) of eerste kwartiel**	0
	Q2 (50%) of tweede kwartiel	1
	Q3 (75%) of derde kwartiel	3

De mediaan is gelijk aan 1. De helft van de respondenten heeft dus 0 tot 1 delict op jaarbasis gepleegd. Het eerste kwartiel is gelijk aan 0, het derde is gelijk aan 3. Dit betekent dat een kwart van de respondenten geen delict heeft gepleegd. Driekwart van de respondenten heeft geen enkel tot en met drie verschillende delicten (uit een reeks van negen) gepleegd.

Het rekenkundig gemiddelde

Het rekenkundig gemiddelde is een centrummaat dat gebruikt wordt bij variabelen gemeten op het metrische niveau, dus bij interval en ratio-variabelen. Het rekenkundig gemiddelde van een kenmerk wordt verkregen door alle voorkomende waarden bij elkaar op te tellen en vervolgens het totaal te delen door het aantal respondenten. Als er n respondenten zijn, wordt hun rekenkundig gemiddelde gegeven door de formule:

Voor *individuele waarnemingen*:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Of in het geval van *absolute frequenties*:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{m} f_i \times x_i$$

Of in het geval van *relatieve frequenties:*

$$\bar{x} = \sum_{i=1}^{m} p_i \times x_i$$

Aantal zelfgerapporteerde delicten uit een reeks van 9 delictvragen

Aantal verschillende misdrijven (uit een reeks van 9 vormen)		Frequentie	Procent	Geldig procent	Cumulatief procent
Geldig	0	1021	33,1	33,9	33,9
	1	708	23,0	23,5	57,3
	2	469	15,2	15,6	72,9
	3	284	9,2	9,4	82,3
	4	196	6,4	6,5	88,8
	5	146	4,7	4,8	93,7
	6	72	2,3	2,4	96,1
	7	49	1,6	1,6	97,7
	8	36	1,2	1,2	98,9
	9	34	1,1	1,1	100,0
	Totaal geldig	**3015**	**97,8**	**100,0**	
Missing	System	67	2,2		
Totaal		3082	100,0		

In een zelfrapportagestudie werd gevraagd aan jonge adolescenten of ze een reeks van negen delicten gepleegd hadden. Het gemiddeld aantal gerapporteerde delicten bedraagt 1,78. Dit kan ook als volgt uit de frequentietabel worden afgeleid door de som te nemen van de producten van elke waarde met diens absolute frequentie:
[(1021*0) + (1*708) + (2*469) + (3*284) + (4*196) + (5*146) +(6*72) + (7*49) + (8*36) + (9*34)]/3015 = 1.78

Het **rekenkundig gemiddelde** heeft enkele belangrijke kenmerken:
(1) Het rekenkundig gemiddelde wordt enkel voor metrische variabelen gebruikt;
(2) het rekenkundig gemiddelde is het evenwichtspunt (of zwaartepunt) van de verdeling;
(3) het rekenkundig gemiddelde is geen resistente (robuuste) maat: dit betekent dat de waarde gevoelig is voor uitschieters of extreme waarden;[2]
(4) de som van alle afwijkingen tegenover het rekenkundig gemiddelde is nul.
Dit laatste vergt enige uitleg. De afwijkingen tegenover het gemiddelde noemen we **deviatiescores**.

Hieronder presenteren we deze eigenschap aan de hand van de formule en maken een kleine rekensom om dit te bewijzen.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

	x_i	$x_i - \bar{x}$
	12	1,17
	14	3,17
	7	-3,83
	10	-0,83
	9	-1,83
	13	2,17
Som	65	0,00
$\bar{x}$	10,83	

[2] Extreme waarden noemen we outliers en komen heel weinig voor. Door hun extreem karakter beïnvloeden ze wel mee de uitkomst.

Verantwoord kiezen tussen centrummaten

Welke centrummaat zullen we kiezen bij de rapportage van gegevens? Zowel de modus, de mediaan als het rekenkundige gemiddelde zijn centrummaten. De vraag stelt zich welke het best gehanteerd kan worden. Stel dat we deze 3 centrummaten berekenen voor een zelfde verdeling en we bekomen 3 verschillende resultaten. Hoe moeten we met dergelijke bevindingen omgaan? Dit is bijvoorbeeld het geval voor het aantal zelfgerapporteerde delicten uit het hierboven vermelde onderzoek.

Parameters van centraliteit: Aantal zelfgerapporteerde feiten

N	Valid	3015
	Missing	67
Gemiddelde		1,7847
Mediaan		1,0000
Modus		,00

Belangrijk is te weten dat bij de berekening van het rekenkundige gemiddelde alle waarnemingen worden betrokken. Dit is voordelig, want het informatiegehalte is daarom zeer groot. Maar er is een nadeel: en dat nadeel is precies dat *extreme waarden* de resultaten van het gemiddelde beïnvloeden. De mediaan is minder gevoelig aan extreme waarden, aangezien deze maat afhankelijk is van waar de meeste frequenties zich bevinden. Het is belangrijk om ook eens de verdeling van een kenmerk waarin je geïnteresseerd bent vormelijk te gaan bestuderen. Bij *niet-symmetrische verdelingen* (en dat komt vaak voor in de criminologie: zie verder, wanneer we de parameters van vorm presenteren) wordt het gemiddelde sterk beïnvloed door extreme waarden: liggen deze rechts, dan schuift het gemiddelde mee op naar rechts, m.a.w: dan zal het rekenkundig gemiddelde groter zijn dan de mediaan. Bij niet-symmetrische verdelingen is het rekenkundige gemiddelde dus niet altijd een betrouwbare centrummaat. Ze is te groot wanneer de 'staart' van de verdeling rechts ligt en te klein als deze links ligt.

De mediaan is vaak een meer betrouwbare centrummaat bij niet-symmetrische verdelingen.

Het is van het grootste belang dit te beseffen, aangezien veel kenmerken die door criminologen bestudeerd worden een niet-symmetrisch karakter hebben. Zo zijn slachtofferschap en daderschap gemeten binnen een tijdspanne zeer ongelijk verdeeld. Een voor de criminologie typerend voorbeeld is de vaststelling dat een kleine groep van delictplegers heel actief is. Berekenen we het gemiddelde dan beïnvloeden deze extreme waarden, die we *"outliers" of "uitschieters*" noemen, de resultaten.

4. Parameters van spreiding: vive la difference!

Wat betekent spreiding? Vrijwel alle kenmerken in onze samenleving kennen spreiding. Een bepaald kenmerk heeft dan meerdere categorieën en na een bevraging van respondenten (of andere eenheden) blijken de respondenten niet in één en dezelfde categorie terecht te komen, maar (heel vaak ongelijk) verdeeld te zijn over de verschillende categorieën. M.a.w. verschillende mensen vallen in de verschillende categorieën van het kenmerk dat we bestuderen. Hoe meer de waarnemingen verspreid zijn over alle categorieën, hoe groter de spreiding. Parameters van spreiding bieden een antwoord op beschrijvende onderzoeksvragen die de ongelijke spreiding van criminologisch relevante fenomenen willen bestuderen. Hoe ongelijk is slachtofferschap verdeeld in de samenleving? Hoe sterk verschillen jonge adolescenten in hun betrokkenheid bij delinquent gedrag? Zo zijn er nog meer voorbeelden te bedenken.

In de criminologische theorie speelt de ongelijke verdeling van respondenten over de categorieën van een kenmerk een grote rol. Men stelt zich de vraag waaraan deze variatie tussen de onderzochte eenheden kan worden toegeschreven. Wanneer we de criminele carrières van de Belgen bestuderen, bemerken we heel grote verschillen. Een reeks van mensen komt in de georganiseerde misdaad terecht, de meesten echter niet. Hoe komt dat? We zullen ons afvragen hoe we deze grote verschillen kunnen verklaren. Opleiding, hard werken, een hoog IQ, bindingen met de samenleving, iets te verliezen hebben,… het zijn allemaal mogelijke verklaringen voor de variatie in de betrokkenheid bij crimineel gedrag die we trachten te meten. In de criminologie stellen we ook vast dat veel mensen zelden of nooit slachtoffer worden van bepaalde delicten, en een kleine groep herhaaldelijk slachtoffer wordt. We stellen vast dat in veel gemeenten de geregistreerde criminaliteit aan de lage kant is, en in een paar gemeenten eerder hoge concentraties kent. Deze variabiliteit interesseert de criminoloog. Waarom is het zo dat criminaliteit zulke ongelijke verdelingen kent? Al jaren proberen criminologen, met de hulp van theorieën hiervoor verklaringen te bieden en verklaringen te gaan toetsen.

Hieruit vloeit voort dat een variabele voldoende spreiding moet hebben (variatie) om ze te onderzoeken. Als iedereen hetzelfde gedrag vertoont (bv. iedereen pleegt met een zelfde frequentie delicten) dan kan men geen statistisch onderzoek opzetten met de vraag hoe de frequentie van delictplegen te verklaren valt. Men beschouwt een variabele als verklaard wanneer de *spreiding* of *variatie* in die variabele kan toegeschreven worden aan een identificeerbare bron. Deze bron is vanuit statistisch oogpunt een andere variabele (bivariate

beschrijvende statistiek) of een set van andere variabelen (multivariate beschrijvende statistiek). Dit komt in latere hoofdstukken van deze cursus aan bod. **Afhankelijk van het meetniveau van een variabele** wordt een spreidingsmaat gekozen.

De variatieratio (VR)

Op het **nominale meetniveau** treffen we **de spreidingsmaat 'de variatieratio'** aan. Alle spreidingsmaten voor nominale schalen voldoen aan het principe dat de waarde groter wordt naarmate de heterogeniteit groter wordt. *De variatieratio is de proportie waarnemingen die niet tot de modale categorie behoort*. De variatieratio neemt de waarde van nul aan indien alle waarnemingen tot de modale categorie behoren. Een minder aangename eigenschap is het feit dat er geen vaste bovengrens is. Er is geen maximumwaarde. Dus is het moeilijk de spreiding te gaan interpreteren. De maximale waarde van de variatieratio benadert de waarde één wanneer iedere waarneming een verschillende waarde heeft. De variatieratio is dus een eenvoudige maat om te berekenen, maar is niet genormeerd en houdt enkel rekening met de proportie van waarnemingen die tot de modale klasse behoort.[3] In het voorbeeld werken we een situatie uit.

Slachtofferschap afgelopen vijf jaar	**Percentage**
Geen slachtoffer	75
Minstens één maal slachtoffer de afgelopen vijf jaar	25

We zien dat de modale categorie de categorie "niet" is. Daartoe behoort 75% . De variatieratio is dus 100% -75% oftewel 25%. Of in proporties uitgedrukt: 1 - 0.75 = 0.25. De spreiding is aan de lage kant want is ver van 1 verwijderd. Wat betekent dat nu inhoudelijk: mensen worden niet zo vaak slachtoffer, de waargenomen variatie is niet zo groot. Een minderheid wordt slachtoffer.

De index van diversiteit (ID)

Op het **nominale meetniveau** treffen we ook de **index van diversiteit** aan. De index van diversiteit is een spreidingsmaat die net zoals de variatieratio is gebaseerd op de relatieve frequenties van de categorieën, maar het enige verschil is nu dat rekening wordt gehouden met de proportie van waarnemingen binnen iedere categorie of equivalentieklasse. De waarde van

[3] Normering betekent in deze situatie bij nominale kenmerken dat een vaste waarde wordt bereikt bij maximale spreiding en dat het aantal categorieën niet van invloed is op de spreidingsmaat.

deze parameter geeft een idee van de *mate van concentratie van de waarnemingen over de categorieën* van de variabele. Meer concreet betekent dit dat als dit cijfer nul bedraagt, alle waarnemingen dezelfde waarde hebben en er dus ook geen sprake kan zijn van spreiding van waarnemingen.

De ID wordt als volgt berekend: ID = 1 – (f'1*f'1 + f'2*f'2 +f'3*f'3 +…f'n*f'n).
Oftewel: de waarde één minus de som van de gekwadrateerde proporties in elke categorie. Door te kwadrateren wordt aan de categorieën met een hogere frequentie een hoger gewicht toegekend. We geven een voorbeeld.

Regio van herkomst van vreemdelingen in het Vlaamse gewest	**Freq. (Absolute frequentie)**	**Prop. (Proportie)**	**Gekwadrateerde proportie (= prop * prop)**
Europa	155 098	0.55	0.30
Afrika	57 065	0.20	0.04
Azië	55 794	0.20	0.04
Oceanië	369	0.00	0.00
Amerika	8 020	0.03	0.00
Onbekend	4 741	0.02	0.00
Som	281 087	1	0.38

Wanneer we nu de spreidingsmaat uitrekenen, moeten we voor elke categorie eerst de proportie berekenen en dan het kwadraat van deze proportie.
Vervolgens rekenen we uit: ID = 1 – (0.30+0.04+0.04+0.00+0.00+0.00)= 0.62. De index van diversiteit is interessant omdat deze in termen van kansen kan geïnterpreteerd worden. De ID betekent: **“de kans dat twee willekeurig gekozen vreemdelingen in het Vlaamse Gewest uit een verschillende regio afkomstig zijn, bedraagt 0.62”**

De variatiebreedte

Wanneer het meetniveau van een variabele ordinaal is, worden **ordinale spreidingsmaten** gebruikt. Een gekend voorbeeld is de **variatiebreedte**. De variatiebreedte is het verschil tussen de grootste en de kleinste waargenomen waarde.

$$V = \max_i x_i - \min_i x_i$$

Laat ons een voorbeeld geven. We hernemen de vraag uit ons onderzoek naar spijbelgedrag.

Hoeveel van je vrienden hebben ooit gespijbeld?

		Frequentie	Percentage	Geldig Percent	Cumulatief percentage
Valid	**geen enkele**	**1953**	**63,4**	**63,5**	63,5
	één vriend	924	30,0	30,0	93,5
	twee vrienden	144	4,7	4,7	98,1
	drie of meer vrienden	57	1,8	1,9	100,0
	Total	3078	99,9	100,0	
Missing	**System**	4	,1		
Total		3082	100,0		

De variatiebreedte kan berekend worden want de antwoordcategorieën zijn gemeten op het ordinale niveau. *De variatiebreedte gaat van "geen enkele" tot "drie of meer vrienden".*

De variatiebreedte is een zeer rudimentaire parameter van spreiding. Ze geeft immers enkel aan over welke afstand de waarnemingen verspreid zijn. Het zou best kunnen dat de variatiebreedte groot is, en het gevolg is van het feit dat er een aantal respondenten in een hoge rang zitten. Wanneer de gegevens in klassen gegroepeerd zijn, wordt de variatiebreedte gedefinieerd als het verschil tussen de bovengrens van de hoogste klasse en de ondergrens van de laagste klasse. Sommige auteurs hanteren een andere definitie, namelijk als het verschil tussen de klassenmiddens van beide uiterste klassen.

De interkwartiel-afstand (K3-K1)

De **interkwartiel-afstand** kan ook gehanteerd worden vanaf het ordinale meetniveau. De interkwartielafstand is het verschil tussen het derde en eerste kwartiel. Aangezien het eerste kwartiel het punt aangeeft waaronder 25% van de verdeling ligt, en het derde kwartiel het punt waaronder 75% van de verdeling ligt, bevat de interkwartielafstand de helft van het totale aantal waarnemingen. Extreme waarden hebben geen invloed op de interkwartiel-afstand. De **interdeciel-afstand** is het verschil tussen het negende en eerste deciel.

Spreidingsmaten op metrisch niveau

Als we te maken hebben met kenmerken die tenminste op het intervalniveau worden gemeten, dan kunnen we de afwijkingen tegenover het rekenkundig gemiddelde berekenen. De **metrische spreidingsmaten** zijn dus allemaal op eenzelfde principe gebaseerd. Men bepaalt

eerst het rekenkundig gemiddelde op basis van alle observaties. Vervolgens willen we weten elke waarneming verschilt van het rekenkundig gemiddelde. De logica hierachter is heel eenvoudig. Observaties die heel ver verwijderd liggen van het rekenkundig gemiddelde zijn heel bijzonder of afwijkend (deviant). Het is wel handig om deze verschillen te kunnen uitdrukken in een getal. Op die manier weten we meteen hoe homogeen of heterogeen onze steekproef is. Het motto hier is dus: **vive la difference**! Leve de variabiliteit. Variatie is de bron van alle theorie. Er zijn verschillen tussen individuen in termen van criminaliteit die men zelf begaat, maar ook in termen van criminaliteit waarvan men zelf het slachtoffer wordt. Er zijn naast verschillen tussen individuen ook verschillen binnen individuen. Een individu is geen statisch organisme. Individuen ontwikkelen zich doorheen de levensloop en dus plegen zij in sommige perioden van hun leven meer regelovertredend gedrag dan in andere perioden. Ook dat is spreiding. We willen in de criminologie verklaringen bieden voor variatie. In feite is dat zoeken naar variatie eigen aan de wetenschap. Ook in andere wetenschappen, zoals de biologie, wil men variatie beschrijven. Darwin beschreef de variatie en ontwikkeling van soorten, waaronder de mens, en het is op basis van variaties dat Darwin zich vragen begon te stellen die hebben geleid tot zijn evolutietheorie. Variatie doet er dus wel degelijk toe. De spreidingsmaten die we gebruiken op het metrische niveau zijn allemaal gebaseerd op de afwijkingen tegenover het rekenkundig gemiddelde. Dat betekent dat zij ook beïnvloed worden door de eigenschappen van dat rekenkundig gemiddelde. Wat geldt voor het rekenkundig gemiddelde, geldt ook voor de spreidingsmaten die er op gebaseerd zijn: net zoals het gemiddelde zijn de metrische spreidingsmaten niet zo robuust. Ze zijn zeer afhankelijk van extreme waarden. Een paar extreme waarnemingen (statistische "**outliers"**), die voorkomen wanneer een kenmerk heel scheef verdeeld is in een steekproef, kunnen de spreiding sterk beïnvloeden. Laat criminaliteit nu zo een variabele zijn die heel scheef verdeeld is. Eén van de belangrijkste variabelen uit de criminologische theorievorming is enorm scheef verdeeld. Dat heeft consequenties, maar daar komen we ten gepasten tijde nog op terug. We bespreken achtereenvolgens de *gemiddelde absolute afwijking, de variatie, de (steekproef)variantie, de (steekproef)standaardafwijking* en *de variatiecoëfficiënt*.[4]

[4] Het is goed het onderscheid te maken tussen de steekproefstandaardafwijking en de standaardafwijking gebaseerd op gegevens uit een volledige onderzoekspopulatie. Details horen eigenlijk niet in dit hoofdstuk thuis, maar we zullen er in de les wel op terugkomen.

De gemiddelde absolute afwijking

Een klassieker, die echter niet veel meer gebruikt wordt, is de **gemiddelde absolute afwijking**. Dit is de som van de *absolute waarden* van de afwijkingen van elke waarde ten aanzien van het rekenkundig gemiddelde, gedeeld door het aantal waarnemingen. Men gebruikt de absolute waarden, omdat de som van alle (positieve en negatieve) afwijkingen tegenover een gemiddelde altijd nul wordt.

De variatie

De **variatie** of nog de **"Sum of Squares" (afgekort: SS of soms ook var)** genoemd, is de som van de gekwadrateerde afwijking van elke waarde tegenover het gemiddelde.

$$\boxed{SS = \sum (X - \overline{X})^2} \text{ of } SS = \sum_{i=1}^{N} (x_i - \overline{x})^2$$

Men **kwadrateert** de verschillen tussen elke geobserveerde waarde van een metrische variabele X en het gemiddelde omdat de som van de afwijkingen van elk individu ten opzichte van het gemiddelde altijd nul is. Kwadrateren lost dit probleem op. De optelsom van de kwadraatafwijkingen geeft een indicatie van spreiding: hoe groter de maat, hoe groter de verschillen tussen de statistische eenheden. Bij een constante zal er nooit spreiding zijn, en dus zullen alle spreidingsmaten nul zijn. De waarde nul betekent dus: iedereen heeft een gelijke score.

De (steekproef)variantie

De **variantie** (s^2) is de variatie gedeeld door het aantal onderzoekseenheden wanneer we over populatiegegevens beschikken en gedeeld door het aantal steekproefeenheden minus één.[5] In dit handboek wordt steeds gewerkt met de formule van de *steekproefvariantie*. De formule voor de berekening van de variantie uit een steekproef ziet er als volgt uit:

[5] Dit noemen we het aantal vrijheidsgraden (degrees of freedom) wanneer we ons baseren op gegevens verkregen uit een steekproef. Het aantal vrijheidsgraden van een statistiek kan algemeen gesproken het aantal ongebonden variabelen worden genoemd. Of nog: het aantal variabelen die verschillende waarden kunnen aannemen of variëren. Voor het berekenen van de spreiding van een populatie is het aantal vrijheidsgraden gelijk aan het aantal van deze populatie. In het geval van een steekproef verliezen we één vrijheidsgraad. Statistische verwerkingspakketten berekenen trouwens ook de variantie en standaardafwijkingen van kenmerken op basis van de formule die we toepassen op steekproeven.

$$\boxed{s^2 = \frac{\sum (X - \overline{X})^2}{n-1}}$$

De (steekproef)standaardafwijking

De **steekproefstandaardafwijking** is de *vierkantswortel van de steekproefvariantie*. De *populatiestandaardafwijking* is de vierkantswortel van de *populatievariantie*. In deze syllabus hanteren we steeds de formule voor de steekproefstandaardafwijking en gebruiken we hiervoor soms de afkorting "**std**" van het Engelse begrip "**standard deviation**". [6]

$$\boxed{s = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}} \text{ of } s = \sqrt{s^2}$$

Uit de wiskundige notatie kunnen we één en ander afleiden: bereken voor elke meetwaarde de afstand tot het gemiddelde. Neem het kwadraat hiervan voor elke waarde. Tel deze bij elkaar op en deel ze door het aantal observaties (of onderzoekseenheden) min één. De vierkantswortel (square root) is nodig om ook de negatieve afstanden correct mee te tellen.

5. Zelf uitrekenen van gemiddelde, variantie en standaardafwijking

Hoewel het zelf uitrekenen van statistische coëfficiënten in het beroepsleven gebeurt aan de hand van statistische verwerkingspakketten, is het toch cruciaal dat het gemiddelde, de variantie en standaardafwijking nog steeds kunnen berekend worden. Deze vormen immers de belangrijkste parameters die de basis vormen voor vele meer geavanceerde vormen van statistische analyses. Als u de berekening zelf uitvoert, ziet u wat er gebeurt achter de schermen van een statistische analyse. U begrijpt de uitkomst beter omdat u weet hoe u aan de uitkomst bent geraakt. We tonen hieronder een tabel die informatie verschaft over vijftien studenten. Van elke student is hun score op T1 (toets 1) genoteerd. Deze score is niet voor iedereen gelijk. De test staat op 50 punten. An heeft dus 30 op 50, Arno heeft 45 op 50, enz.. We stellen vast dat sommige studenten het heel goed hebben gedaan, maar er zijn toch studenten die het wat minder hebben gedaan. Het rekenkundig gemiddelde is 30. Hoe varieert nu elke student tegenover dat

[6] We wijzen hier op het belang van een perfect begrip van het statistische concept 'standaardafwijking'. Een goed begrip ervan is de basis voor het standaardiseren van waarnemingen (i.e. het omzetten van scores naar z-scores) in het volgende hoofdstuk 4 (de standaardnormale verdeling en diens eigenschappen). Een z-score is immers niets anders dan het *aantal* standaardafwijkingen dat een bepaalde score boven of onder het gemiddelde van de reeks scores ligt.

gemiddelde? Laten we eens kijken. An wijkt niet af van het gemiddelde. Haar deviatiescore is nul. Arno wijkt wel af van het gemiddelde. Hij presteert beter. Zijn deviatiescore is 15. Hiermee zit hij fel boven het gemiddelde. Als we nu voor iedereen de deviatiescore hebben berekend, dan kunnen we de som nemen van die deviatiescores. Die is nul. Dat mag geen verassing zijn, want het is precies een eigenschap van het rekenkundig gemiddelde, met name dat de som van diens afwijkingen steeds nul bedraagt.
Precies omdat de som van alle afwijkingen tegenover het rekenkundig gemiddelde nul bedraagt, kunnen we de deviatiescores niet gebruiken om de spreiding of afwijking rond het rekenkundig gemiddelde uit te drukken. Daar is wat op gevonden. In de kolom ernaast ziet u dat we het kwadraat hebben berekend van elke afwijking tegenover het rekenkundig gemiddelde. Deze oplossing is bijzonder eenvoudig en toch geniaal: de waarden verschillen van nul en men kan de som nemen van al deze waarnemingen. Die som is nu niet meer gelijk aan nul.

Hoewel dit zeer handig is, voelt u wellicht intuïtief aan dat er iets vervelends gebeurt: grote afwijkingen worden uitvergroot door te kwadrateren. Dat is juist. Daarom heeft men hier ook iets op gevonden: we nemen gewoon die som en delen deze door het aantal waarnemingen, of nog beter, we nemen van deze laatste maat gewoon de vierkantswortel. Dit zorgt voor de correctie op de uitvergroting van verschillen. Reken even mee:

Tabel: studentscores, deviatiescores en het kwadraat daarvan

Student	ScoreT1	$x_1 - \bar{x}$	$(x_1 - \bar{x})^*$ $(x_1 - \bar{x})$
An	30,00	0	0
Arno	45,00	15	225
Bart	35,00	5	25
Björn	20,00	-10	100
Delphine	40,00	10	100
Hanne	35,00	5	25
Henk	30,00	0	0
Ines	30,00	0	0
Jeroen	25,00	-5	25
Jurgen	20,00	-10	100
Kim	40,00	10	100
Robert	25,00	-5	25
Nele	20,00	-10	100
Sara	25,00	-5	25
Sofie	30,00	0	0
N= 15 $\bar{x}$ = 30			Sum of squares 850

Var X= Sum of squares X/(N-1)= 850/14= 60.71

Std X= SQRT (60.71)= 7.79

Werkwijze :	
Stap 1:	Bereken het rekenkundig gemiddelde van de variabele
Stap 2:	Trek het rekenkundig gemiddelde af van iedere waarde (kolom 2)
Stap 3:	Kwadrateer de verschillen: dit gebeurt omdat anders positieve verschillen tegen negatieve zouden wegvallen. (het totaal van de afwijkingen zou dan nul zijn!)
Stap 4:	Tel de gekwadrateerde verschillen op.
Stap 5:	Bereken de variantie: deel het totaal van de gekwadrateerde verschillen door het aantal waarnemingen min één.
Stap 6:	Bereken de standaardafwijking: trek de wortel uit de variantie. De standaardafwijking wordt hierdoor weer in vergelijkbare hoeveelheden gegeven als het gemiddelde.

De variatiecoëfficiënt

Het nadeel van hiervoor besproken spreidingsmaten is hun afhankelijkheid van de meeteenheid. Als we nu twee totaal verschillende variabelen beschouwen: criminaliteit en IQ. Beide worden gemeten in een andere meeteenheid. Je kan van elke variabele de standaardafwijking berekenen. Je zal ongetwijfeld vaststellen dat het ene kenmerk een grote standaardafwijking heeft dan het andere kenmerk. Dat is juist, maar dat kan heel misleidend zijn. Want: precies omdat de beide kenmerken in een eigen meeteenheid zijn uitgedrukt, kunnen we die toch niet gaan vergelijken met elkaar? We riskeren een kanjer van een fout te begaan als we dat zouden doen. Voor dat probleem hebben statistici een handige oplossing bedacht. Aangezien beide kenmerken verschillen in termen van meeteenheid, en dus een eigen gemiddelde hebben, delen we gewoon de standaardafwijkingen door de respectieve gemiddeldes. Het resultaat is de variatiecoëfficiënt v. De **variatiecoëfficiënt (v)** is een **gestandaardiseerde spreidingsmaat**. Wanneer we de spreiding van de prijs van een aantal producten willen weten en hiervoor de standaardafwijking berekenen van de prijs uitgedrukt in Belgische frank en daarna uitgedrukt in Euro, dan zal de bekomen maat verschillend zijn, terwijl ze in feite op dezelfde gegevens berust. Om die reden werd de variatiecoëfficiënt ingevoerd. Men zegt wel eens van de variatiecoëfficiënt dat deze ***dimensieloos*** (niet afhankelijk van de meeteenheid) is en dit laat toe de spreiding van

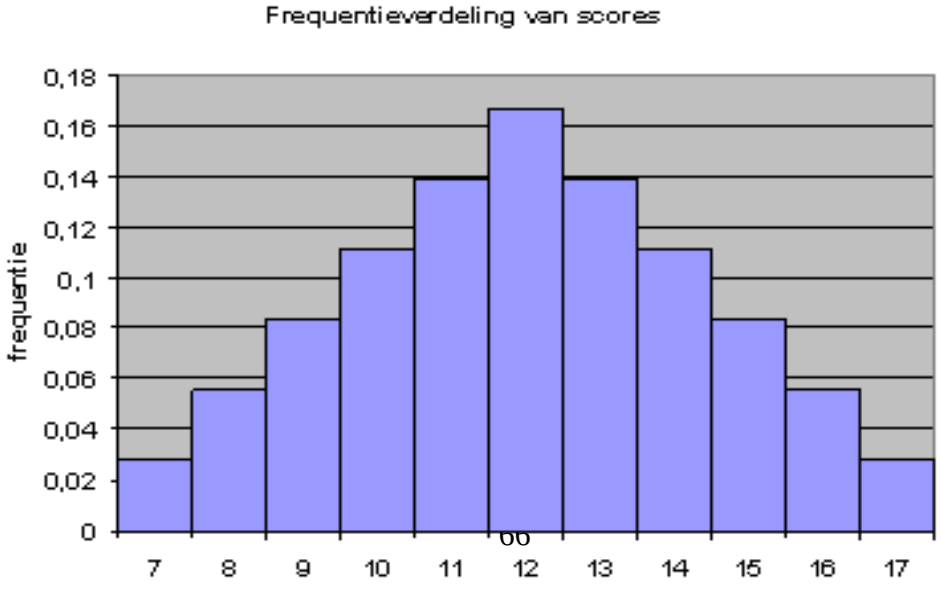
verdelingen die worden uitgedrukt in *verschillende meeteenheden te vergelijken*. De variatiecoëfficiënt wordt berekend door de standaardafwijking te delen door het rekenkundige gemiddelde (beide worden immers in dezelfde meeteenheid uitgedrukt). Onthoud deze belangrijke eigenschap van de variatiecoëfficiënt.

$$v = \frac{s}{\overline{X}}$$

6. Parameters van vorm

Naast centraliteit en spreiding kunnen we ook de ***vorm*** van de *verdeling van kenmerken* samenvatten aan de hand van enkele parameters. Vergelijken we verdelingen inzake vorm, dan kunnen we vaststellen dat verdelingen verschillen in de mate waarin zij afwijken van een *symmetrische verdeling*. Bij het bestuderen van de symmetrie van een verdeling bekijken we in feite hoe de gegevens verdeeld zijn ten opzichte van het rekenkundig gemiddelde. Uit dit laatste kunnen we reeds afleiden dat symmetrie enkel kan bestudeerd worden voor **metrische meetschalen**. Een verdeling kan symmetrisch, links asymmetrisch of rechts asymmetrisch zijn. Laten we dit illustreren aan de hand van enkele voorbeelden. Een verdeling is **symmetrisch** als het rekenkundig gemiddelde en de mediaan aan elkaar gelijk zijn. Een symmetrisch verdeelde variabele is de normaal verdeelde variabele. Dit is een variabele die de gekende Gauss-curve volgt. De Gauss-curve werd door de statisticus Quetelet ontdekt voor sociale kenmerken.

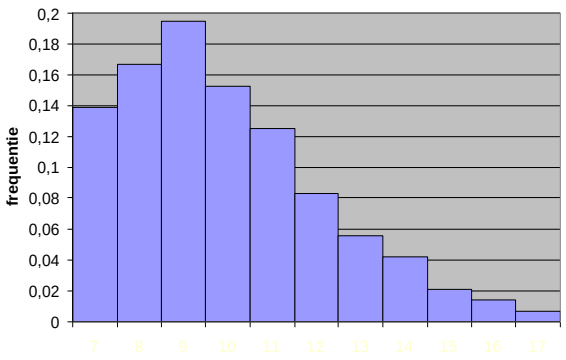
Een symmetrische verdeling ziet er als volgt uit:

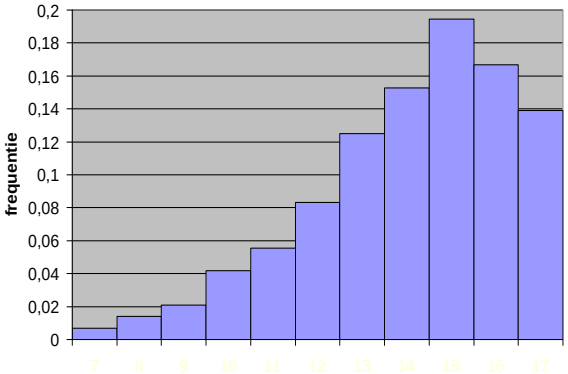


Deze symmetrische verdeling noemen we symmetrisch omwille van het feit dat de linkerhelft en de rechterhelft perfect op elkaar lijken. Ze zij elkaars spiegelbeeld.

Een verdeling is **positief asymmetrisch** als de verdeling een langere staart naar rechts heeft. In dit geval is het rekenkundig gemiddelde groter dan de mediaan. Positieve asymmetrie betekent dat de hogere waarden minder vaak voorkomen dan de lagere waarden. Dit is heel vaak het geval bij criminaliteitsmaten. Vele jongeren hebben geen enkel delict gepleegd in een periode van 12 maanden, een niet onaanzienlijk deel van de jeugd pleegt wel eens een delict, maar slechts heel weinig jongeren plegen heel veel delicten. Een verdeling is **negatief asymmetrisch** als de verdeling links een langere staart heeft. In dat laatste geval komen de lage waarden minder voor dan de hoge waarden.

Positieve asymmetrie **Negatieve asymmetrie**





De vraag is nu: kunnen we de symmetrie die een variabele kenmerkt ook uitdrukken op basis van een eenvoudig getal, zodat we in een oogopslag, ook zonder naar de figuur te kijken kunnen aflezen hoe symmetrisch een variabele is. Het antwoord hierop is ja. We spreken dan van **parameters van vorm**. Er bestaan verschillende statistische parameters van vorm. Dit zijn de parameters die we nodig hebben om de symmetrie te berekenen. We beperken ons hieronder tot de ***empirische coëfficiënt van Pearson***.

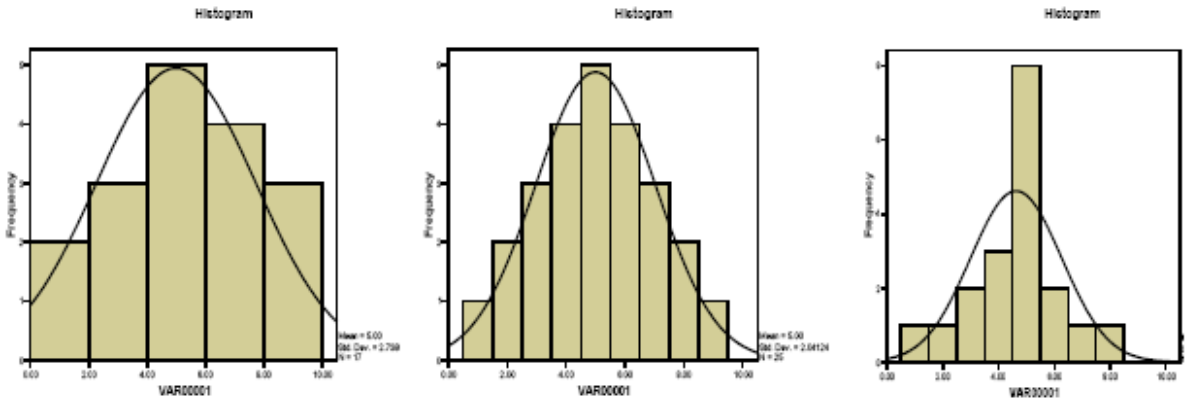
De **empirische coëfficiënt van Pearson** wordt als volgt berekend: (1) bereken het verschil tussen het gemiddelde en de mediaan en (2) deel deze waarde door de standaardafwijking.

$$S = \frac{\overline{\overline{X}} - \tilde{x}}{s}$$

Een verdeling is positief asymmetrisch als de coëfficiënt een positieve waarde heeft en negatief asymmetrisch als de coëfficiënt een negatieve waarde heeft.

Naast de symmetrie kan men ook de mate van **afplatting** van een verdeling bestuderen. De afplatting of **kurtosis** is de mate van afplatting van de gegevens rondom het rekenkundige gemiddelde. De kurtosis wordt steeds met de **standaardnormale** of **Gauss-verdeling** als standaard vergeleken. We besteden verderop in deze syllabus meer aandacht aan deze verdeling. We onderscheiden, **mesokurtische** verdelingen (gemiddelde afplatting), **leptokurtische** verdelingen (scherper) en **platykurtische** verdelingen (platter). Een platykurtische vedeling is een verdeling die lijkt op een normaal verdeeld kenmerk waar iemand spreekwoordelijk met een hamer op heeft geslagen. Daardoor ziet deze er platter uit. Een leptokurtische verdeling is een verdeling waar de hoogst voorkomende waarde zo veel vaker voorkomt, dat deze er echt uitschiet. Om de vergelijking met de normale verdeling mogelijk te maken, werd deze ook getekend op de histogrammen in kwestie.

Figuur: de kurtosis



Platykurtische verdeling | Mesokurtische verdeling | Leptokurtische verdeling

7. De box plot

Een overzichtelijke manier om gegevens **vanaf ordinaal meetniveau** visueel voor te stellen, is de **box plot**. Een standaard box plot laat toe om een snelle visuele evaluatie te maken van de informatie die vervat zit in een frequentieverdeling.

De box plot is een grafiek van de vijf-getallensamenvatting. De vijf-getallensamenvatting van een verdeling bestaat uit de mediaan M, de kwartielen Q1 en Q3 en de minimale en maximale waarnemingen, genoteerd als:

(Q3-Q1)*1.5 of minimale niet-uitschietende waarde - Q1 - Mediaan - Q3 - (Q3-Q1)*1.5 of maximale niet-uitschietende waarde[7].

Opgelet: minimum en maximum betekenen hier niet de allerkleinste of allergrootste waargenomen waarde. Ze betekenen respectievelijk de hoogste en laagste NIET-UITSCHIETENDE waarde. De combinatie van deze vijf getallen is een snelle manier om een samenvatting te krijgen van zowel het centrum als de spreiding van een variabele. De mediaan beschrijft het centrum van de verdeling, de kwartielen tonen de spreiding van de middelste helft van de gegevens (de centrale 50%), de waarde die overeenstemt met anderhalve keer de interkwartielafstand het minimum en maximum tonen de volledige spreiding van de gegevens. Deze vijf-getallensamenvatting leidt tot een grafische verdeling: de box plot.

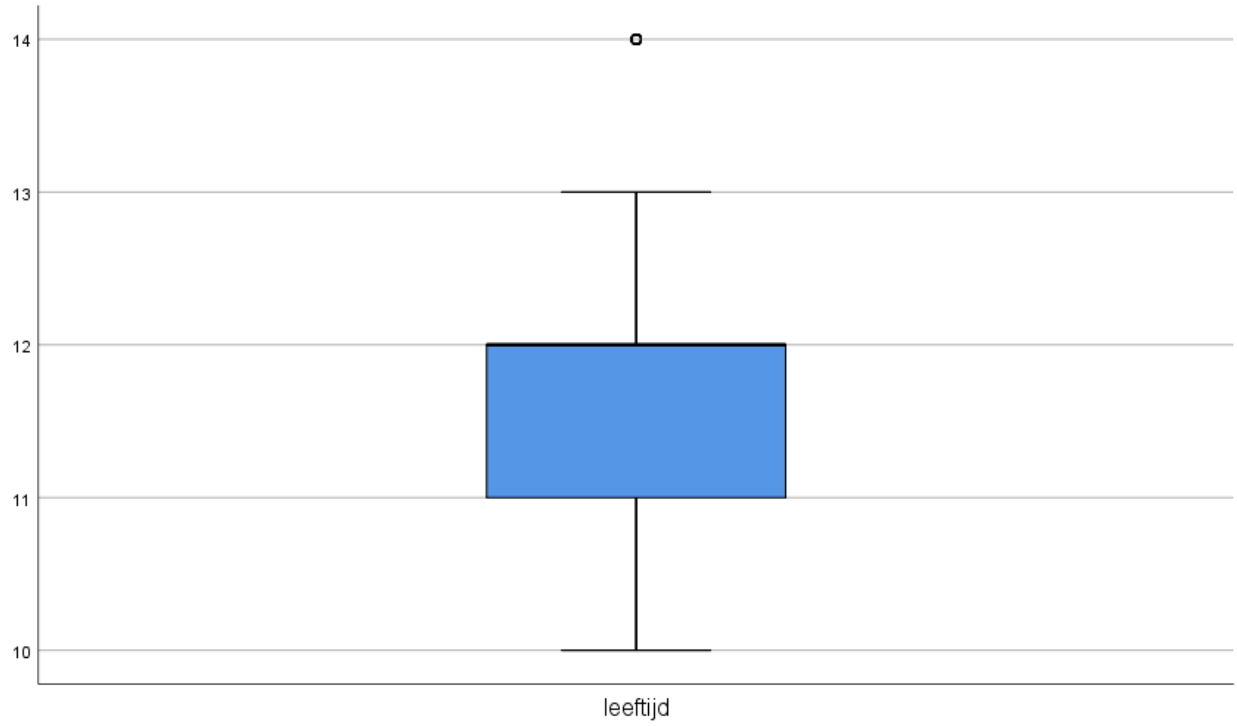
De box plot bestaat uit:

- De centrale box (rechthoek) die zich uitstrekt van het eerste kwartiel Q1 tot het derde kwartiel Q3
- De mediaan M die in de rechthoek gemarkeerd wordt door een lijn
- Onder en boven de box zie je twee lijnen (whiskers). Deze tonen de waarden die overeenkomen met anderhalve keer de interkwartielafstand. De waarde die overeenkomt met de onderste en bovenste lijn komt niet noodzakelijk overeen met de allerlaagste en allerhoogste waarneming. Deze komt enkel overeen met de laagste en hoogste "niet-uitschietende" waarneming die zich op de whisker bevindt. Waarnemingen buiten de whiskers zijn dus steeds mogelijk! **Let dus goed op want als je gevraagd wordt de variatiebreedte af te lezen, mag je je niet exclusief door de whiskers laten leiden.** De allerlaagste en allerhoogste waarneming kunnen

[7] Eerder mathematische handboeken gebruiken soms de term minimum en maximum maar bedoelen daarmee de minimale en maximale niet-uitschietende waarde. Om verwarring met het echte minimum in de betekenis van de allerlaagste waarde en echte maximum (maximumwaarde) te vermijden spreken we in dit handboek van minimale niet-uitschietende waarde als waarde die overeenkomt met de onderste whisker en maximale niet-uitschietende waarde als we het hebben over de maximale niet-uitschietende waarde.

immers buiten de whiskers vallen. We spreken dan van outliers of uitbijters. Statistisch gezien komen deze waarden zeer zelden voor, omwille van hun grote afstand tot anderhalve keer de interkwartielafstand (i.e. Q3-Q1) *1.5.

Figuur 1: Box plot van een metrische variabele leeftijd



De vijf-getallensamenvatting voor de variabele 'leeftijd' in *figuur 1* bedraagt:

10 11 12 12 13

Hoe lezen we best een box plot ?

Eerst kijken we naar de mediaan. In *figuur 1* zien we dat de mediaan 12jaar bedraagt: het is het punt dat precies in het midden valt van de verdeling. Daarna kijk je naar de spreiding. De kwartielen tonen de spreiding van de middelste helft van de gegevens. In bovenstaand voorbeeld is de waarde van Q1 = 11 (=waarde waaronder zich 25% van alle eenheden bevinden) en de waarde van Q3= 12 (=waarde waaronder zich 75% van alle eenheden bevinden). Bemerk dat in dit voorbeeld de mediaan samenvalt met het derde kwartiel Q3. We herhalen dat de box zelf de centrale 50% van de eenheden vormt. In dit voorbeeld heeft 50% van de respondenten een leeftijd tussen 11 en 12jaar.

De uiterste waarden tenslotte (de kleinste en de grootste waarnemingen) tonen de spreiding van de hele gegevensverzameling. De allerlaagste waarneming valt hier samen met het minimum en bedraagt 10jaar en het maximum bedraagt 13jaar. De allerhoogste

waarneming is 14 jaar. De variatiebreedte komt overeen met het verschil tussen de allerhoogste en allerlaagste waarneming (14-10= 4).

Box plots worden in de praktijk gebruik voor het opsporen van uitzonderlijk lage of hoge scores. Er is hier duidelijk sprake van een uitschieter. Een aantal respondenten hebben een leeftijd van 14jaar.

De variatiebreedte is echter gevoelig voor uitschieters. De afstand tussen de kwartielen (=de spreidingsbreedte van de middelste helft van de gegevens) is een meer *resistente* spreidingsmaat[8]. We noemen deze afstand de interkwartielafstand (*IKA*= de afstand tussen het eerste en het derde kwartiel. *IKA* = Q3 – Q1). In ons voorbeeld hierboven geldt dat *IKA* = 12 – 11 = 1.

De interkwartielafstand wordt vooral gebruikt als vuistregel voor het opsporen van verdachte uitschieters.

Een waarneming is een verdachte uitschieter als deze ten minste 1,5 x *IKA* boven het derde kwartiel of onder het eerste kwartiel ligt. We noemen dit het *1,5 x IKA-criterium* voor uitschieters.

In ons voorbeeld in figuur 1 geldt:

$$1{,}5 \text{ x } IKA = 1{,}5 \text{ x } 1 = 1{,}5$$

De waarden onder 11 – 1,5 = 9,5 en boven 12 + 1,5 = 13,5 zijn herkenbaar als mogelijke uitschieters. In ons voorbeeld zijn er geen uitschieters naar beneden maar wel naar boven. Er zijn respondenten met een leeftijd van 14jaar.

Statistische software zoals SPSS maken gebruik van de 1,5 x *IKA*. Box plots die door software zijn getekend zijn vaak **gemodificeerde box plots** die verdachte uitschieters afzonderlijk weergeven. In SPSS worden uitzonderlijk lage/hoge scores aangegeven met ° en met een *.

Waarnemingen aangeduid als ° bevinden zich tussen Q1 – 1,5 *IKA* en Q1 – 3 *IKA* enerzijds en tussen Q3 + 1,5 *IKA* en Q3 + 3 *IKA* anderzijds. Dit zijn ‘zwakke’ uitschieters.

Waarnemingen gemarkeerd met een * bevinden zich buiten Q1 – 3 *IKA* en Q3 + 3 *IKA*.

Deze waarnemingen zijn mogelijks uitschieters die de uitkomsten te sterk beïnvloeden. Het zijn ‘extreme’ uitschieters.

De lijnen die uit de centrale box komen, hebben dan alleen betrekking op kleinste en grootste waarnemingen die niet voldoen aan de 1.5 x *IKA*-regel. In ons voorbeeld zijn de waarnemingen

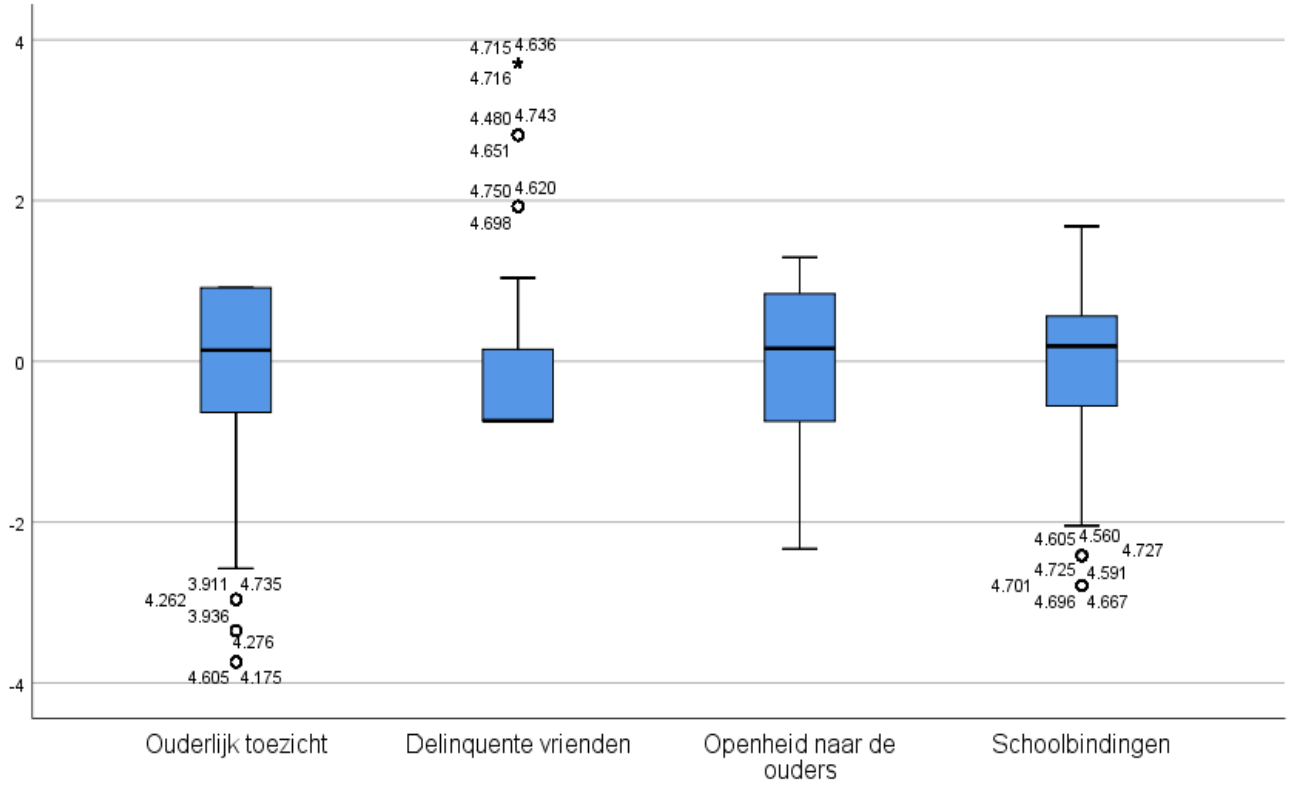
[8] Een resistente maat van een verdeling wordt nauwelijks beïnvloed door veranderingen in de numerieke waarden van een klein gedeelte van het totale aantal waarnemingen. Bijvoorbeeld: de mediaan en kwartielen zijn resistente maten, maar het gemiddelde en de standaardafwijking zijn dat niet. We komen hier verder in de cursus op terug.

die hieraan niet voldoen respectievelijk 10 en 13 en zijn er uitschieters naar boven met een waarde 14.

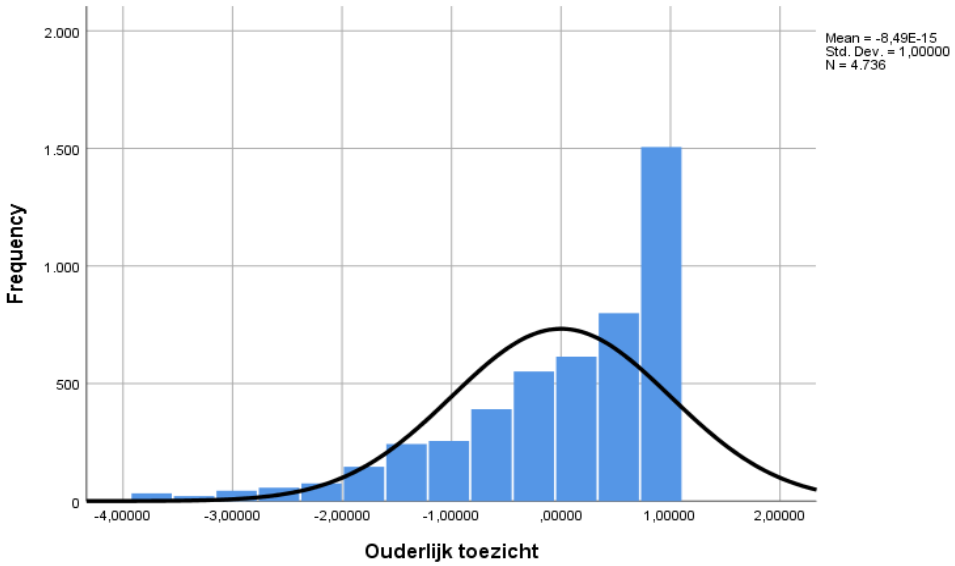
Box plots detecteren niet enkel uitschieters. Omdat box plots minder details laten zien dan bijvoorbeeld een histogram, worden ze bij voorkeur gebruikt om verschillende verdelingen met elkaar te vergelijken zoals afgebeeld in *figuur 2*. Figuur 2 toont de afzonderlijke box plots voor 4 variabelen 'ouderlijk toezicht', 'delinquente vrienden', 'openheid naar de ouders' en 'schoolbindingen'. De variabelen werden gemeten door aan respondenten verschillende vragen te stellen en hun scores op elke vraag samen te tellen. De variabelen werden gestandaardiseerd om de interpretatie eenduidig te maken.

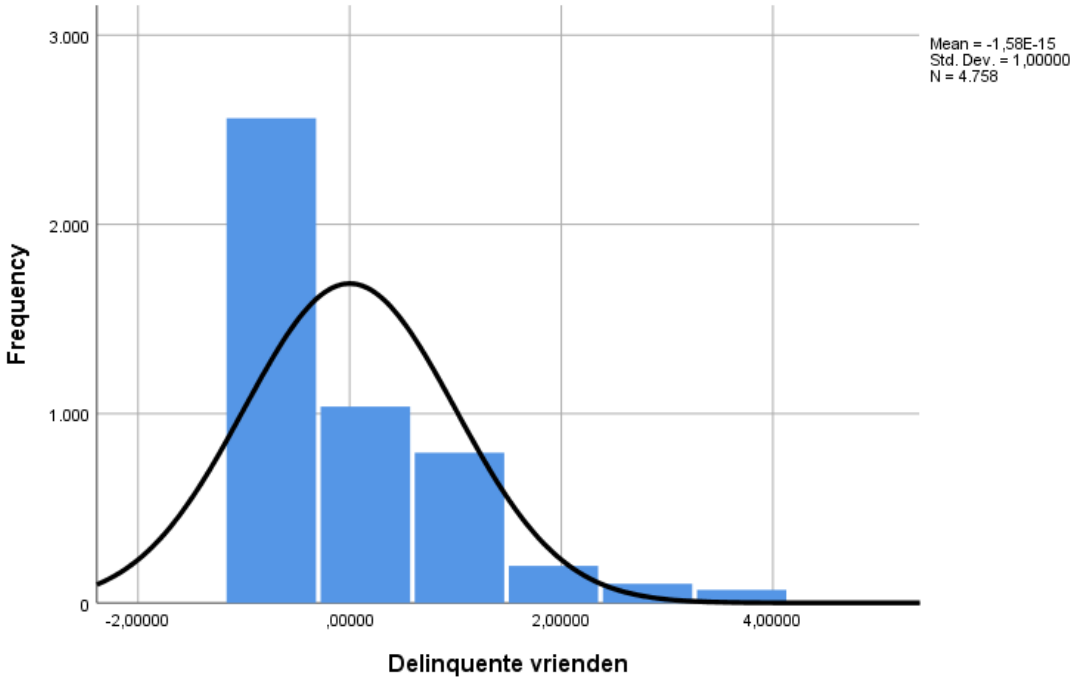
Kijken we eerst naar de mediaan, dan zien we dat deze het laagst is voor de variabele 'Delinquente vrienden'. Kijken we naar de kwartielen die ons de spreiding tonen van de middelste helft van de verdeling, dan zien we dat deze het kleinst is voor de variabele 'Delinquente vrienden'. Kijken we naar de uiterste waarden of de kleinste en de grootste waarnemingen dan zien we voor de variabele 'openheid naar de ouders' geen uitschieters. Er zijn geen respondenten die op dit kenmerk hoger of lager scoorden dan 1,5 x *IKA* en 3 x *IKA*. Er zijn wel uitschieters voor de variabelen 'ouderlijk toezicht' en 'schoolbindingen', uitschieters naar beneden, en voor de variabele 'delinquente vrienden' zijn er uitschieters naar boven.

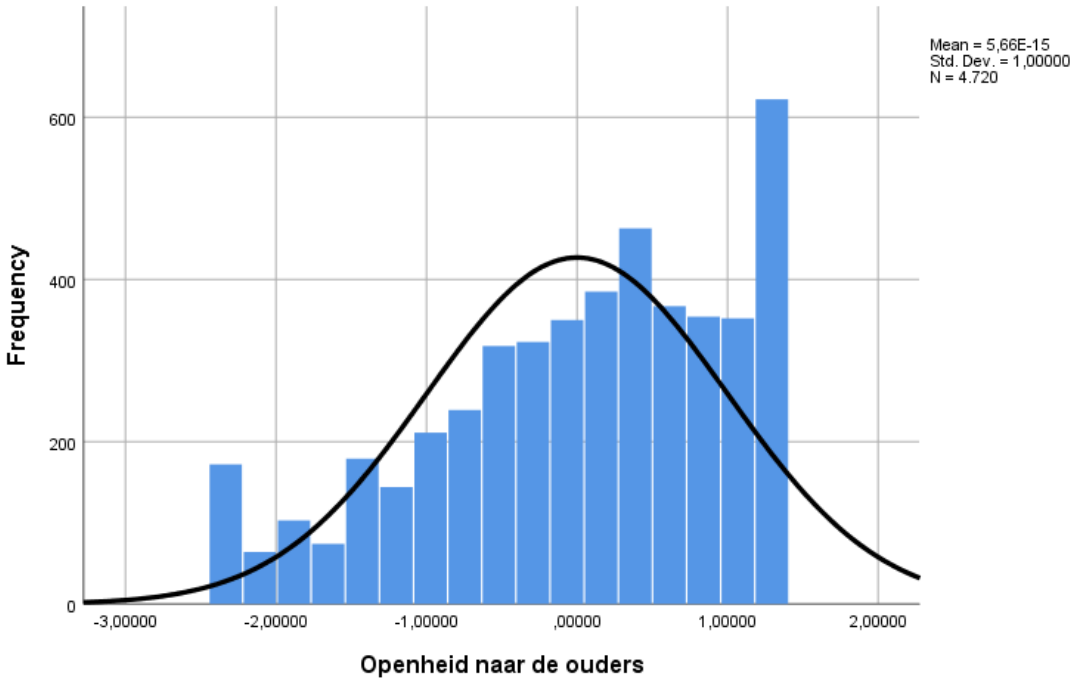
Figuur 2: Box plots voor ouderlijk toezicht, delinquente vrienden, openheid naar de ouders en schoolbindingen.

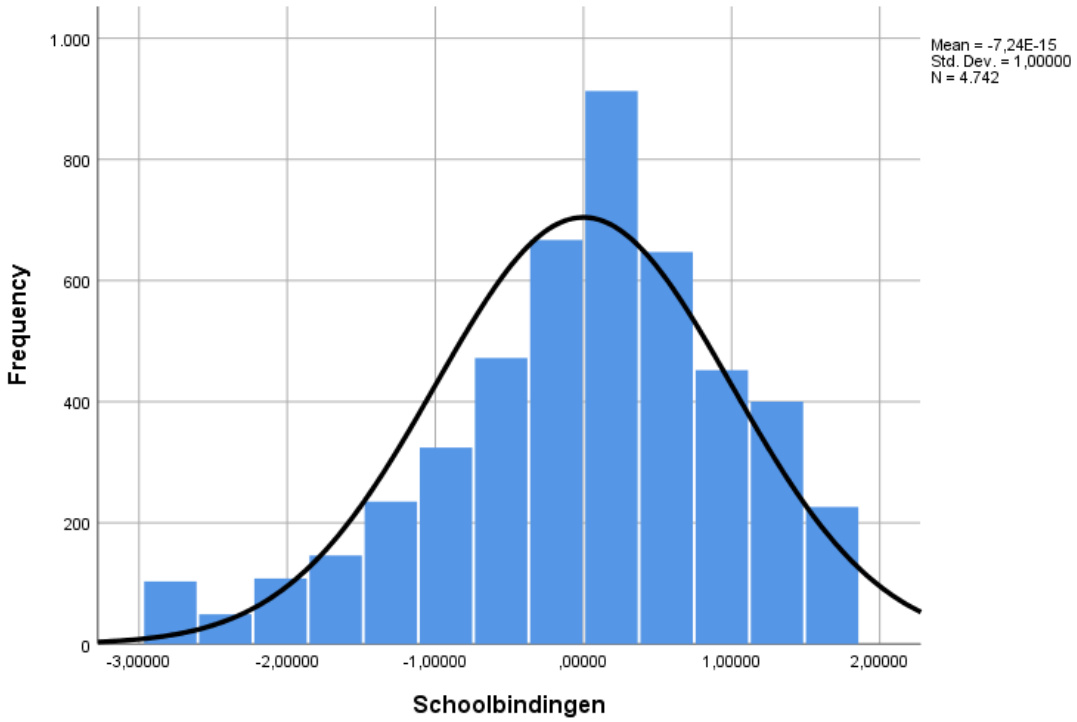


Deze box plots zeggen ons ook iets over de vorm van de verdelingen. De variabelen ‘ouderlijk toezicht’ en ‘schoolbindingen’ hebben nogal wat uitschieters naar beneden en zijn links scheef verdeeld. De variabele ‘openheid naar de ouders’ is links scheef verdeeld. De variabele ‘delinquente vrienden’ is rechts scheef verdeeld. Ter controle en vergelijking presenteren we hierna tenslotte het histogram met aanduiding van de normaalcurve van elke variabele.




Mean = -1,58E-15
Std. Dev. = 1,00000
N = 4.758
Frequency
3.000
2.000
1.000
0
-2,00000
,00000
2,00000
4,00000
Delinquente vrienden


Mean = 5,66E-15
Std. Dev. = 1,00000
N = 4.720
Frequency
600
400
200
0
-3,00000
-2,00000
-1,00000
,00000
1,00000
2,00000
Openheid naar de ouders



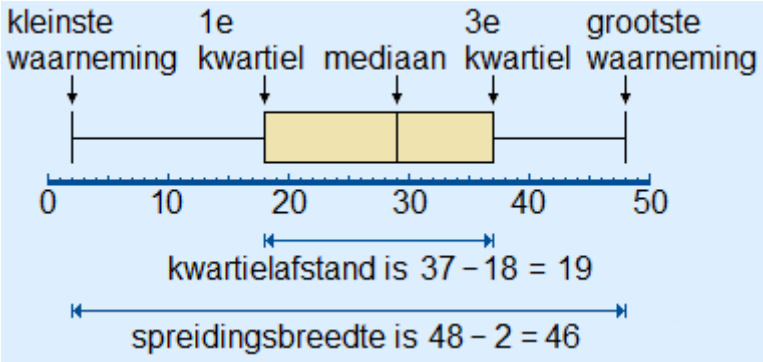
Samenvatting

De **vijf-getallensamenvatting,** bestaande uit de mediaan, de kwartielen en minimale en maximale niet-uitschietende waarde, verschaft een snelle en globale beschrijving van een verdeling. De mediaan beschrijft het centrum en de kwartielen en de whiskers worden gebruikt om aan te geven wat statistisch gezien abnormaal is: waarden die daarbuiten vallen, komen uiterst zelden voor.

De **interkwartielafstand** (*IKA*) is het verschil tussen de kwartielen. Het is de spreiding van de middelste helft van de gegevens. Het 1,5 x *IKA-criterium* merkt waarnemingen die ten minste 1,5 x *IKA* voorbij de kwartielen vallen als mogelijke uitschieters.

Box plots die berusten op de vijf-getallensamenvatting zijn nuttig om verschillende verdelingen met elkaar te vergelijken. De centrale rechthoek strekt zich uit van het eerste tot het derde kwartiel en geeft de spreiding weer van de middelste helft van de verdeling. De mediaan wordt in de centrale rechthoek gemarkeerd. Lijnen strekken zich uit tot aan de kleinste en de grootste waarneming en geven de volledige spreiding van de gegevens weer, behalve de punten gemerkt door het 1,5 x *IKA- criterium*, deze worden vaak afzonderlijk weergeven als uitschieters.

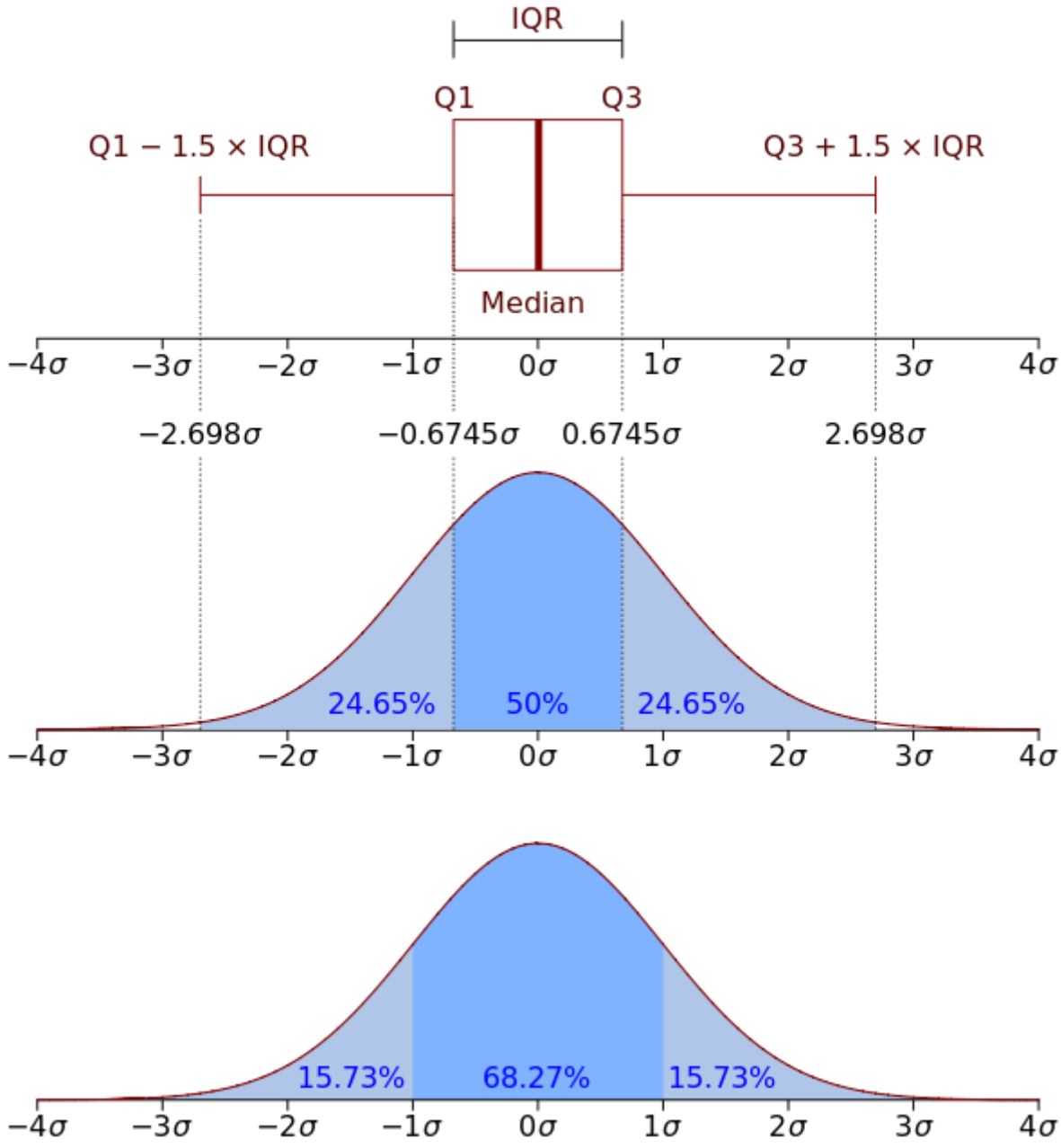
Samenvattende figuur



We herhalen dat in deze cursus onder 'kleinste waarneming' wordt verstaan : minimale niet-uitschietende waarde of (Q3 – Q1) * 1,5 en onder 'grootste waarneming': maximale niet-uitschietende waarde of (Q3 – Q1) * 1,5.

De vijf-getallensamenvatting is niet de meest gangbare numerieke beschrijving van een verdeling. Een geschikter middel hiervoor is een combinatie van het gemiddelde om het centrum te meten en de standaardafwijking om de spreiding te meten. Ter herhaling: de standaardafwijking meet de spreiding door te kijken hoe ver de waarnemingen van hun gemiddelde zijn verwijderd. De standaardafwijking s is de wortel van de variantie s^2. Het gemiddelde en de standaardafwijking zijn goede beschrijvingen voor symmetrische verdelingen zonder uitschieters. Ze zijn bijzonder nuttig voor de normale en standaardnormale verdelingen die in het volgende hoofdstuk aan bod komen. In *figuur 3* presenteren we een vergelijking van een box plot met een standaardnormale verdeling en leggen hiermee de link als overgang naar het volgende hoofdstuk.

Figuur 3: vergelijking van een box plot met een standaardnormale verdeling



8. Testvragen

Hieronder vind je enkele uitspraken over de univariate statistiek. Deze vragen kan je gebruiken om je parate kennis te toetsen over de basiskennis die tot nog toe werd meegegeven. **In gewijzigde vorm kunnen dergelijke theorievragen ook op het examen voorkomen.** Deze vragen zijn afkomstig uit vroegere examens. De correcte antwoorden zijn rechtstreeks uit de cursus afleidbaar. De antwoorden vind je achteraan dit theorieboek. Dit is niet meer dan een test om te zien of je mee bent met de leerstof. Faal je op deze testvragen, dan is het hoogdringend tijd om in actie te schieten.

1. Deviatiescores zijn

- De som van de gekwadrateerde afwijkingen tegenover het rekenkundig gemiddelde
- De som van de afwijkingen tegenover het rekenkundig gemiddelde

2. De mediaan is een robuuste parameter van centraliteit

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

3. De Index van Diversiteit kan geïnterpreteerd worden als de kans dat twee willekeurig gekozen onderzoekseenheden tot een verschillende categorie behoren

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

4. De variatieratio neemt de waarde van nul aan indien alle waarnemingen tot de modale categorie behoren

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

5. Een variabele van het metrisch niveau kan bestudeerd worden op ordinaal niveau

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

6. Een variabele van het ordinale niveau kan bestudeerd worden op het metrische niveau

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

7. De variatiebreedte is het verschil tussen de maximale waarde en de minimale waarde

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

8. De variatiebreedte is de som van de maximale en minimale waarde

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

9. Uitspraken extrapoleren van de steekproef naar de populatie doe je via

- de beschrijvende univariate statistiek
- de inferentiële statistiek

10. De steekproefstandaardafwijking wordt berekend op basis van de formule van de populatiestandaardafwijking maar in de noemer staat n+1

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

11. Een bar chart wordt gebruikt

- Voor variabelen op het metrische niveau
- Voor variabelen op het categorische niveau

12. Variabelen die op een histogram worden gepresenteerd zijn steeds ratio niveau

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

13. Een lijndiagram wordt gebruikt voor de

- Visuele voorstelling van metrische gegevens die niet in klassen zijn gegroepeerd (ruwe scores)
- Visuele voorstelling van metrische gegevens die wel in klassen zijn gegroepeerd

14. Een frequentiepolygoon wordt gebruikt voor de

- Visuele voorstelling van metrische gegevens die niet in klassen zijn gegroepeerd (ruwe scores)
- Visuele voorstelling van metrische gegevens die wel in klassen zijn gegroepeerd

15. Een platykurtische variabele is

- Platter dan een normaal verdeelde variabele
- Scherper dan een normaal verdeelde variabele

16. Een box-plot kan worden gebruikt voor variabelen vanaf

- Het nominaal niveau
- Ordinaal niveau
- Interval niveau
- Ratio niveau

17. De mediaan komt overeen met

- Het vijftigste percentiel
- Het eenenvijftigste percentiel

18. Als we een kenmerk dat perfect normaal verdeeld is voorstellen via een box-plot, dan is de afstand tussen de mediaan en de hoogste waarde even groot als de afstand tussen de mediaan en de laagste waarde

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

19. Een variabele die rechtsscheef verdeeld is

- Heeft een langere staart naar rechts
- Heeft een langere staart naar links
- Heeft geen staart

20. De interkwartielafstand is een maat van

- Centraliteit
- Spreiding
- Vorm

21. Een frequentieverdeling kunnen we opvatten als een kansverdeling

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

22. Een onderzoekseenheid heeft een z-score van -2.20 voor het metrisch kenmerk "studieresultaat" .

- De onderzoekseenheid valt buiten de centrale 95% van de waarnemingen.
- De onderzoekseenheid doet het beduidend beter dan de gemiddelde onderzoekseenheid

23. Centreren wil zeggen dat men een kenmerk uitdrukt als een afwijking tegenover het rekenkundig gemiddelde

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

24. De frequentieverdeling (histogram) verandert vormelijk niet wanneer men standaardiseert

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

25. Het rekenkundig gemiddelde is een spreidingsmaat die gevoelig is voor uitschieters

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

26. Variabelen van het categorische niveau bevatten categorieën. Een onderzoekseenheid mag tegelijkertijd in twee categorieën van dezelfde variabele zitten

- Dit mag niet als we de regels van de statistiek volgen
- Dit mag wel, er zijn hier geen regels voor

27. Operationalisering betekent

- Dat we een kenmerk meetbaar maken
- Dat we een kenmerk van een conceptuele definitie voorzien

28. De populatievariantie is de som van de gekwadrateerde afwijkingen tegenover het rekenkundig gemiddelde, gedeeld door het steekproefeffectief

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

29. De variatieratio is een spreidingsmaat die we moeten gebruiken wanneer twee metrische kenmerken in een verschillende meeteenheid werden gemeten en men wil de spreiding vergelijken

- Deze uitspraak is juist
- Deze uitspraak is verkeerd

30. De keuze voor een univariate parameter wordt ingegeven door

- De onderzoeksvraag
- Het meetniveau
- Beide

31. Stel: je bent activist die ijvert voor meer inkomensgelijkheid. Om de inkomensongelijkheid te demonstreren, kan je kiezen tussen een aantal parameters. Inkomen is zeer scheef verdeeld.

- Je presenteert vanuit je standpunt de mediaan
- Je presenteert vanuit je standpunt het rekenkundig gemiddelde

9. Leerdoelen

Het hoofdstuk over de univariate beschrijvende statistiek is een belangrijk hoofdstuk. De kennisopbouw in dit hoofdstuk is cumulatief. Je zal gemerkt hebben dat je de elementaire kennis en basisbegrippen uit het vorige hoofdstuk dient te begrijpen vooraleer je de essentie van de univariate statistiek kent en kan toepassen. Univariate statistiek wordt gebruikt om onderzoeksvragen te beantwoorden. Het gaat om univariate beschrijvende onderzoeksvragen die peilen naar de centrale tendensen, naar spreiding en naar vorm.

Het is belangrijk dat je volgende begrippen eigen maakt: absolute frequenties, proporties, relatieve frequenties en cumulatieve absolute en relatieve frequenties. Het is belangrijk dat je inziet dat het meetniveau van een variabele heel belangrijk is met betrekking tot de keuze voor een beschrijvende analysetechniek. Als onderzoeker heb je de vrijheid. Gebruik die en baseer je daarbij op wat je weet. Het indelen in klassen is theoretisch belangrijk, maar wordt in de criminologische praktijk vaak gedaan aan de hand van informatieverwerkingspakketten en op basis van kwartielen en meer inhoudelijke gronden.

We hebben een aantal manieren voorgesteld om statistische gegevens grafisch voor te stellen. De belangrijkste zijn het taartdiagram en het staafdiagram voor nominale en ordinale kenmerken. Bedenk dat het aantal categorieën best beperkt is bij taartdiagrammen. In een staafdiagram kunnen meer categorieën worden voorgesteld op een duidelijke manier dan in een taartdiagram. Voor de cumulatieve voorstelling van gegevens kunnen we het cumulatief frequentiediagram gebruiken. Metrische gegevens worden aan de hand van een histogram voorgesteld. Andere manieren om metrische gegevens voor te stellen zijn lijndiagrammen en frequentiepolygonen. We komen ze echter minder vaak tegen dan het histogram. Wees steeds alert bij grafieken. De manier waarop gegevens in assen worden voorgesteld kan al even

misleidend als verhelderend zijn. Uitspraken zoals “met statistiek kan je alles bewijzen” zijn partieel gebaseerd op misleidende voorstellingen.

Het meetniveau van een variabele is ook bepalend voor de keuze van een beschrijvende analysetechniek. Je moet heel goed het onderscheid kennen tussen de parameters van centraliteit, spreiding en vorm. Je moet heel goed beseffen welke parameters kunnen gehanteerd worden en daarbij moet je steeds de link leggen met het meetniveau van het bestudeerde kenmerk.

De parameters van centraliteit die belangrijk zijn in criminologisch onderzoek zijn de modus, de mediaan, de kwantielen en het rekenkundig gemiddelde. Elk geven zij op hun manier de centrale tendensen weer. Je moet weten hoe je deze dient te berekenen en je moet deze kunnen interpreteren. Je moet de eigenschappen van het rekenkundig gemiddelde zeer goed kennen.

De parameters van spreiding die we gezien hebben bij de studie van spreiding op het nominale niveau zijn de variatieratio, de index van diversiteit. Deze moet je zelf kunnen berekenen. Vanaf het ordinale niveau hebben we gezien dat de variatiebreedte en de interkwartielafstand interessante spreidingsmaten zijn. Deze moet je uit tabellen kunnen afleiden. De parameters van spreiding die gehanteerd worden op het metrische niveau zijn zeer belangrijk en zullen steevast terugkeren in deze syllabus. Deze moet je dus zeer goed begrijpen. Je moet het onderscheid kennen tussen de gemiddelde absolute afwijking, de variatie, de variantie, de standaardafwijking en de variatiecoëfficiënt. Je dient deze zelf te kunnen berekenen en je dient de uitkomsten van univariate analyses zelf inhoudelijk te kunnen interpreteren.

Tot slot hebben we de parameters van vorm gezien. We hebben aandacht besteed aan symmetrie en kurtosis. Je dient te weten wat symmetrie betekent en hoe afwijkingen ten opzichte van symmetrie kunnen worden vastgesteld. Je dient natuurlijk wel te weten wat kurtosis is, maar we hechten in deze starterscursus meer belang aan asymmetrie. We hebben ook gezien dat we een kenmerk kunnen voorstellen aan de hand van een box plot. Je dient te weten wat je kan afleiden uit een box plot.

In concreto betekent dit dat je niet enkel kennis dient te hebben over de centrale begrippen uit de univariate beschrijvende statistiek, maar dat je ook inzichten dient te hebben bekomen. Je dient correct enkele tabellen en beschrijvende coëfficiënten te kunnen interpreteren en steeds

de link te kunnen leggen met de onderzoeksvraag die achter het gebruik van een beschrijvende statistische parameter gaat.

Hoe bereid je dit hoofdstuk best voor met betrekking tot het examen? Het examen bestaat uit meerkeuzevragen. Voor concrete voorbeelden verwijzen we naar deel II van dit handboek, met name de praktische oefeningen en voorbeelden. Je zal situaties voorgelegd krijgen, zoals tabellen met onvolledige informatie. Je kan een uitspraak krijgen over een situatie, bijvoorbeeld over de verdeling van een kenmerk, of over de relatie tussen meetniveau en de keuze voor een parameter van centraliteit of spreiding. Je kan gevraagd worden een variantie zelf te berekenen uit een beperkt aantal gegevens. Je bent best voorbereid als je klaar en duidelijk weet welk antwoord een statistische beschrijvende parameter kan geven. Ten tweede stel je best automatisch de vraag of de techniek je wel toelaat om je probleemstelling te beantwoorden. Is het meetniveau van de variabele die je wil bestuderen wel in overeenstemming met de keuze die je zou maken?
Leer univariate grafische voorstellingen bekijken en herinterpreteer de tabellen ook in het licht van de essentiële informatie waaraan tabellen dienen te voldoen. Dit hebben we in het inleidend hoofdstuk reeds besproken en dit wordt dus niet meer herhaald. In de lessen zal je ook geleerd worden om aan de hand van SPSS de belangrijkste univariate beschrijvende parameters te berekenen en te interpreteren.