STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Exploratieve statistiek

Infoboekje

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Inleiding

Dit infoboekje bevat achtergrondinformatie bij de tekst “Exploratieve statistiek voor het secundair onderwijs”. Statistische begrippen worden hier met duidelijke voorbeelden uitvoerig geïllustreerd.

Terwijl je in dit boekje leest, kan je misschien een markeerstift bij de hand houden om belangrijke begrippen aan te duiden. Zo heb je meteen een goed overzicht.

Superhandig is dat je dit boekje ook gewoon kan bijhouden om later nog wat dingen weer op te zoeken.

Inhoud

1 De structuur van een dataset

1.1 Gegevensverzameling of dataset

De oorspronkelijke opmetingen, samen met informatie over de manier waarop je die hebt verzameld, zijn het basismateriaal voor elke studie.
Als voorbeeld kijk je even naar een studie die in de jaren zeventig plaats vond in Californië. Daar werd, in het kader van de volksgezondheid, een uitgebreide gegevensverzameling aangelegd bij de geboorte van een kind. Heel wat kenmerken van dat kind werden opgeschreven (geslacht, bloedgroep, gewicht, lengte, tijdstip van geboorte, …), samen met kenmerken van de vader en de moeder (leeftijd, gewicht, lengte,...). Tien jaar later werd elk gezin opnieuw onderzocht. Een heel klein stukje uit die gegevensverzameling **(**ook **gegevensbank**, **databank** of **dataset** genoemd) ziet er als volgt uit.

	Veranderlijken					
	ID	*SEX*	*BLGK*	*LGTK1*	*GEWK1*	*GEWM2*
Elementen = gezinnen	1	J	B	53.3	3.810	60.3
	2	J	AB	55.9	3.720	73.9
	3	M	O	50.8	3.180	66.2
	4	M	O	50.8	2.990	59.0
	5	J	A	50.8	2.900	47.2
	6	M	A	55.9	4.350	78.5
	7	M	AB	49.5	2.770	53.1
	8	M	A	53.3	3.670	57.6

De tabel die je hier ziet is typisch voor elk statistisch onderzoek. Bij gegevens moet je ook altijd zeggen in welke context ze zijn opgemeten. Die context is, samen met de getallen, belangrijk voor het verdere onderzoek. Daarom moet je bij elke dataset minstens kunnen antwoorden op de vragen: “Welke elementen zijn er hier onderzocht?”, “Welke veranderlijken zijn er bij die elementen opgemeten?” en “Hoe zijn die gegevens verzameld?”.

1.2 Elementen

"Elementen" is de verzamelnaam voor de objecten die in een statistische studie worden onderzocht. Dit kunnen personen zijn (kinderen, Vlamingen,…) of dieren (paarden, muizen,…) of planten (irissen, eiken,…) of zaken (gemeenten, auto's,…), enz. De elementen schrijf je op de **rijen** van een rechthoekig schema (matrix). Bij elke rij hoort juist één element.

In ons voorbeeld bestaan de elementen uit Californische gezinnen die in 1971 een baby kregen. Elke rij stelt dus zo'n gezin voor. Deze gezinnen hebben in de gegevensbank geen naam gekregen maar enkel een identificatienummer (afgekort door ID). Het is niet ongewoon dat elementen enkel met een code worden geïdentificeerd wanneer de gegevens te maken hebben met "privacy" of met "medisch geheim". Afhankelijk van het type onderzoek kom je voor het woord "element" ook meer specifieke namen tegen zoals "respondent" (bij een enquête), "patiënt" (bij een klinische studie) of "individu", "deelnemer", "geval", enz.

1.3 Veranderlijken

Per element meet je bepaalde eigenschappen op en de resultaten hiervan schrijf je in de kolommen van de matrix. **Elke kolom draagt een naam om aan te geven over welke eigenschap het juist gaat. Elke eigenschap die je zo opmeet wordt een veranderlijke genoemd**.
Let op: de **naam** van de veranderlijke en de **waarden** van de veranderlijke zijn twee verschillende dingen! "Bloedgroep" is een voorbeeld van een naam van een veranderlijke, terwijl "AB" een voorbeeld is van een mogelijke waarde van deze veranderlijke.
De naam is dikwijls afgekort en dan weet je nog niet juist waarover het gaat. Daarom voeg je een precieze beschrijving van de veranderlijken toe aan je gegevensbank. Als je bijvoorbeeld ziet staan dat het gewicht van iemand gelijk is aan 100, dan moet je wel weten of dit gewicht opgemeten is in kilogram of in Engelse ponden.

In het voorbeeld van de Californische databank kan je de veranderlijken als volgt omschrijven:

- ID identificatienummer van het gezin
- SEX geslacht van het kind (M=meisje, J=jongen)
- BLGK bloedgroep van het kind (O, A, B of AB)
- LGTK1 lengte (in cm) van het kind bij de geboorte
- GEWK1 gewicht (in kg) van het kind bij de geboorte
- GEWM2 gewicht (in kg) van de moeder tien jaar later

2 Discreet numerieke gegevens: gemiddelde, mediaan en staafdiagram

Numerieke gegevens kan je opschrijven met een getal en je kan er bovendien zinvolle wiskundige bewerkingen (zoals som of product) mee maken. Je noemt zo'n gegevens discreet als niet alle tussenliggende getallen kunnen voorkomen. De mogelijke uitkomsten maken telkens een sprong. Tussen 2 en 3 ligt bijvoorbeeld ook het getal 2.25 maar dit kan je niet uitkomen als je wil weten hoeveel kinderen er in een gezin zijn.

2.1 Gemiddelde en mediaan

Gemiddelde en mediaan zijn twee kengetallen die gebruikt worden om het "centrum" van een verzameling getallen aan te duiden. Soms geven zij bruikbare informatie, soms ook niet. Om dat te weten te komen, moet je een figuur tekenen. In combinatie met een figuur krijgen het gemiddelde en de mediaan pas echt betekenis.

2.1.1 Het gemiddelde

Bij het gemiddelde van *n* getallen $\{x_1, x_2, ..., x_n\}$ zijn drie grootheden met elkaar verbonden:

- het **aantal** getallen, wat je algemeen noteert door n
- de **som** van die n getallen, namelijk $x_1 + x_2 + ... + x_n$, wat je op een korte manier opschrijft als $\sum_{i=1}^{n} x_i$
- het **gemiddelde** van die n getallen, wat gelijk is aan $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

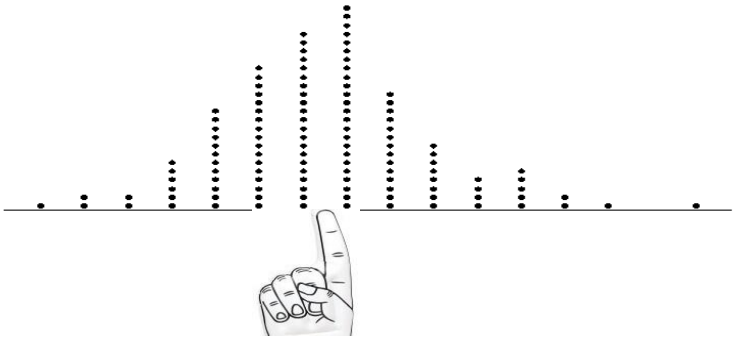
In woorden zeg je: *"het gemiddelde van een verzameling getallen is gelijk aan de som van die getallen gedeeld door het aantal getallen"*.

Voorbeeld
Het gemiddelde $\bar{x}$ van de volgende 9 getallen {8, 23, 6, 9, 13, 10, 12, 7, 11} is gelijk aan 11, want 8+23+6+9+13+10+12+7+11 = 99 en 99/9 = 11.

Waar afronden nodig is, schrijf je het gemiddelde op met één decimale plaats meer dan de oorspronkelijke getallen. Dat is gewoon een afspraak. Als je je aan die afspraak houdt, dan is het niet moeilijk om zo'n gemiddelde op te schrijven.

Maar het is wel moeilijk om er de juiste interpretatie aan te geven. Je moet immers goed in het oog houden dat het gemiddelde, als nieuw getal, ontstaan is door een wiskundige bewerking uit te voeren (de som delen door het aantal). En wiskundige bewerkingen storen zich niet aan de context. Als je dat weet, dan weet je ook hoe je een gemiddelde moet interpreteren. Wat dacht je anders van de uitspraak: “in ons land heeft een gezin gemiddeld 2.1 kinderen”? Waarschijnlijk denk jij daar anders over dan een kind uit de lagere school in Australië dat zei: “dat betekent dat elk gezin twee kinderen heeft en dat de mama terug in verwachting is”!

Je zou aan het gemiddelde ook een fysische betekenis kunnen geven: het is het zwaartepunt op de x-as van een staafdiagram of histogram (zie verder). Als je de figuur zou uitknippen en je zou aftasten tot zij in evenwicht op je vinger rust, dan heb je daar de plaats van het gemiddelde ontdekt.



Figuur 1

2.1.2 De mediaan

Heel dikwijls zie je de foutieve uitspraak die zegt dat de mediaan het middelste getal is. Als je naar {8, 23, 6, 9, 13, 10, 12, 7, 11} kijkt, dan zou de mediaan 13 moeten zijn want dat getal staat mooi in het midden, met 4 getallen ervoor en 4 getallen erna. Maar dit is fout.

Hoe kom je dan wel te weten wat de mediaan is? Wel, je moet beginnen met eerst al je getallen te rangschikken, van klein naar groot. In dit voorbeeld krijg je dan {6, 7, 8, 9, 10, 11, 12, 13, 23}. En pas nu mag je het middelste getal nemen. Dat is hier 10 en dat is dan ook de mediaan van die 9 getallen. De mediaan is dus het middelste getal van een verzameling getallen die geordend zijn van klein naar groot. We noteren: $Me = 10$.

Als je een even aantal getallen hebt, wat moet je dan doen? Ook hier moet je ze eerst ordenen van klein naar groot. En dan zal je bemerken dat er eigenlijk twee getallen in het midden staan. Zo heb je voor {6, 7, 8, 9, 10, 11, 12, 13} dat er

drie getallen vóór 9 staan en dat er ook juist drie getallen na 10 staan. Negen en tien staan dus beide in het midden en voor de mediaan neem je dan het gemiddelde van die "twee middelste", namelijk $Me = \frac{9+10}{2} = 9.5$.
Een leuk opdrachtje dat je met je klas kan doen: ga eens op een rij van klein naar groot staan. Wie heeft de mediaanlengte?

2.2 Staafdiagrammen

Staafdiagrammen kan je tekenen voor categorische gegevens. Dat zijn gegevens waarbij je de waarnemingen in verschillende categorieën klasseert. Soms hebben deze categorieën geen logische volgorde (zoals kleuren of bloedgroepen), maar soms hebben zij dat wel (zoals "goed, beter, best" of getalwaarden zoals "1, 2, 3"). In beide gevallen worden de balkjes in het staafdiagram los van elkaar getekend.
Als de categorieën geen logische volgorde hebben, dan laat je een andere karakteristiek (zoals hun frequentie) de volgorde bepalen. Vaak zie je dat men kiest voor een alfabetische volgorde maar dat is meestal geen goed idee.

In dit hoofdstuk bekijk je discreet numerieke opmetingen en die hebben een logische volgorde. Die volgorde gebruik je bij het tekenen van je staafdiagram.

2.2.1 Symmetrisch rond één top

Op een donderdag werd in een school aan 120 leerlingen gevraagd hoeveel boeken (handboeken, schriften en ringmappen) zij die dag hadden meegebracht. Het antwoord was als volgt.

12	13	12	13	13	12	17	20	13	12	12	12
11	14	10	11	9	10	9	13	7	9	11	12
12	12	12	11	10	9	7	9	11	11	13	11
12	9	10	16	16	14	8	11	10	11	13	8
13	12	15	12	13	10	17	16	12	12	16	14
11	15	9	11	12	12	6	8	12	12	11	10
16	5	9	8	10	15	10	8	9	12	13	12
11	11	11	18	11	11	12	12	12	13	9	8
10	13	12	14	14	11	10	11	10	9	10	15
9	10	6	11	10	13	14	14	14	10	10	13

Het gaat hier over het "aantal" boeken, wat een discreet numerieke veranderlijke is met een beperkt aantal verschillende uitkomsten (het minimum is 5 en het maximum is 20 in dit voorbeeld).
Als je al die getallen samentelt en het resultaat deelt door 120 krijg je het gemiddelde $\bar{x}$ en dat is hier gelijk aan 11.5.

Om de mediaan te vinden, moet je eerst de getallen ordenen van klein naar groot. Dat ziet er als volgt uit.

5	6	6	7	7	8	8	8	8	8	8	9
9	9	9	9	9	9	9	9	9	9	9	10
10	10	10	10	10	10	10	10	10	10	10	10
10	10	10	10	11	11	11	11	11	11	11	11
11	11	11	11	11	11	11	11	11	11	11	**11**
12	12	12	12	12	12	12	12	12	12	12	12
12	12	12	12	12	12	12	12	12	12	12	12
12	13	13	13	13	13	13	13	13	13	13	13
13	13	13	14	14	14	14	14	14	14	14	15
15	15	15	16	16	16	16	16	17	17	18	20

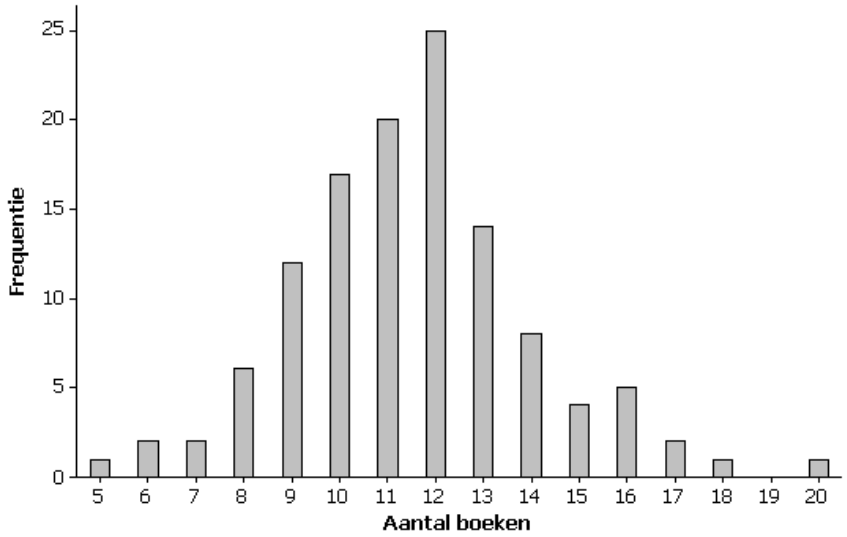
In het midden staan twee getallen, op plaats 60 (want daar staan 59 getallen vóór) en op plaats 61 (want daar staan 59 getallen achter). Maak het gemiddelde van de getallen op plaats 60 (dat is hier het getal 11) en op plaats 61 (dat is hier het getal 12). Zo krijg je de mediaan $Me = \frac{11+12}{2} = 11.5$.

Wanneer je beschikt over een geordende lijst getallen, dan is het heel eenvoudig om een frequentietabel op te stellen. Let erop dat je begint bij de kleinste observatie (dat is hier 5) en stapsgewijs verder gaat tot aan je grootste observatie (dat is hier 20), **zonder een stap over te slaan**. Dat wil hier bijvoorbeeld zeggen dat je in je frequentietabel ook het getal 19 moet zetten, hoewel dat getal niet in je dataset voorkomt. Wat niet voorkomt geef je een frequentie gelijk aan nul. Controleer dat de som van alle frequenties gelijk is aan het totale aantal observaties, dat is hier 120.

Aantal boeken	Frequentie = hoeveel leerlingen met dit aantal boeken
5	1
6	2
7	2
8	6
9	12
10	17
11	21
12	24

Aantal boeken	Frequentie = hoeveel leerlingen met dit aantal boeken
13	14
14	8
15	4
16	5
17	2
18	1
19	0
20	1
	SOM = 120

Een staafdiagram is een grafische voorstelling van de informatie in je frequentietabel. De verschillende waarden van de veranderlijke staan geordend op de x-as. De frequentie of relatieve frequentie vind je op de y-as.



Figuur 2

Als je naar de "globale" vorm van dit staafdiagram kijkt, dan bemerk je dat er geen opvallende "pieken of gaten" in zitten. Er is natuurlijk wat schommeling en perfect symmetrisch is de figuur ook niet. Maar dat is niet erg, want echt uitgesproken scheef ziet hij er zeker niet uit. Het "globale" patroon zegt dat er heel wat waarnemingsgetallen in het centrum liggen (ergens tussen 10 en 13) en dat er minder en minder waarnemingsgetallen voorkomen naarmate je verder van dit centrum weggaat (zowel naar links als naar rechts).

Bij figuren die er globaal uitzien zoals dit staafdiagram is het gemiddelde "typisch" voor je opmetingen. Dat gemiddelde is hier $\bar{x} = 11.5$ en je kan zeggen dat een "typische" leerling zo'n 11 à 12 boeken bij heeft. Daarbij verwacht je ook heel wat leerlingen met 10 of 13 boeken maar veel minder leerlingen met een aantal boeken dat ver van het gemiddelde ligt, zoals 6 of 18.

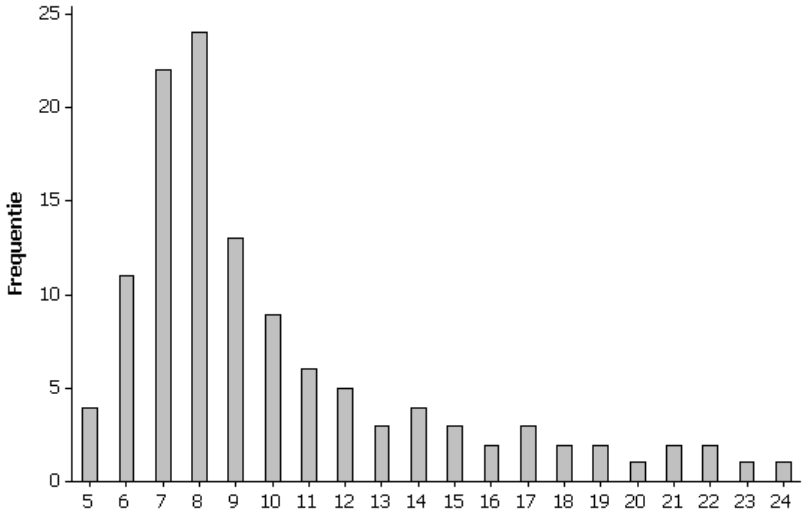
Ook de mediaan is hier een goede maat voor het centrum, in de betekenis van "typisch" voor deze opmetingen. De mediaan is gelijk aan 11.5, wat toevallig exact samenvalt met het gemiddelde.

Bij staafdiagrammen die een globale vorm hebben zoals in figuur 2 (met één uitgesproken top en bovendien redelijk symmetrisch dalend naar links en naar rechts) vind je meestal dat het gemiddelde en de mediaan niet veel van elkaar verschillen. Zij zijn beide een goede maat voor het "centrum". Zij geven je al een eerste indruk voor wat "typisch" is bij je onderzoek.

2.2.2 Eén top en scheef

Scheef naar rechts

Een ander onderzoek zou een volgend beeld kunnen geven.



Figuur 3

Op deze figuur zie je dat veel waarnemingsgetallen de waarde 7, 8, of 9 hebben en hier ligt duidelijk de top als je naar de globale vorm van de grafiek kijkt. Links en rechts van die top worden de staafjes kleiner maar dat gebeurt helemaal niet op een symmetrische manier. Naar rechts is de figuur veel verder uitgespreid dan naar links. Zo'n vorm heet "scheef naar rechts".

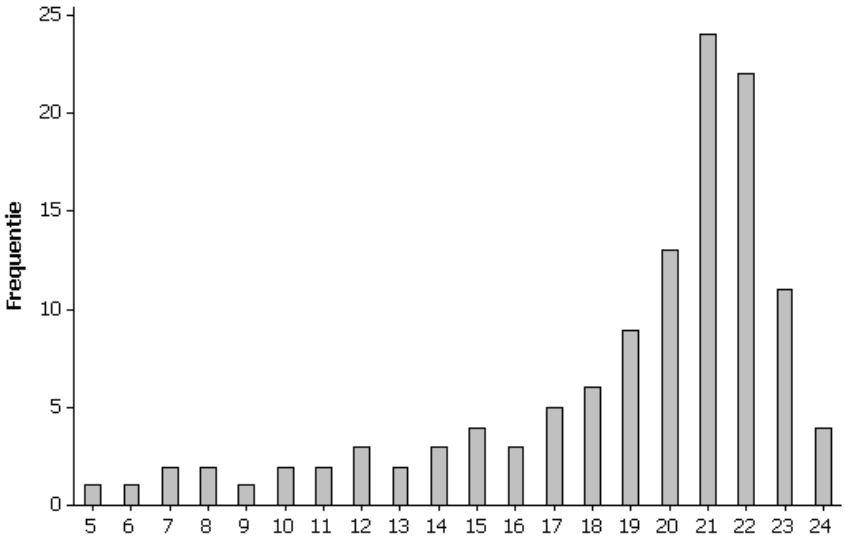
Waar liggen gemiddelde en mediaan bij een grafiek die "scheef naar rechts" is?

Voor de 120 observaties van figuur 3 is de mediaan *Me* gelijk aan 8 terwijl het gemiddelde $\bar{x}$ gelijk is aan 10.1. Het gemiddelde is hier groter dan de mediaan.

Wat je op dit voorbeeld ontdekt hebt, is ook algemeen waar. Staafdiagrammen die "scheef naar rechts" zijn, stellen een dataset voor waarbij het gemiddelde groter is dan de mediaan. Je kan dit als volgt begrijpen. Het gemiddelde houdt rekening met de "waarde" van alle getallen. Als er dus veel getallen naar rechts verschuiven, dan verschuift het gemiddelde ook naar rechts. De mediaan houdt geen rekening met de "waarde" van die "naar rechts verschoven" getallen, alleen met het "aantal" (er moeten er evenveel links als rechts van de mediaan liggen).

Scheef naar links

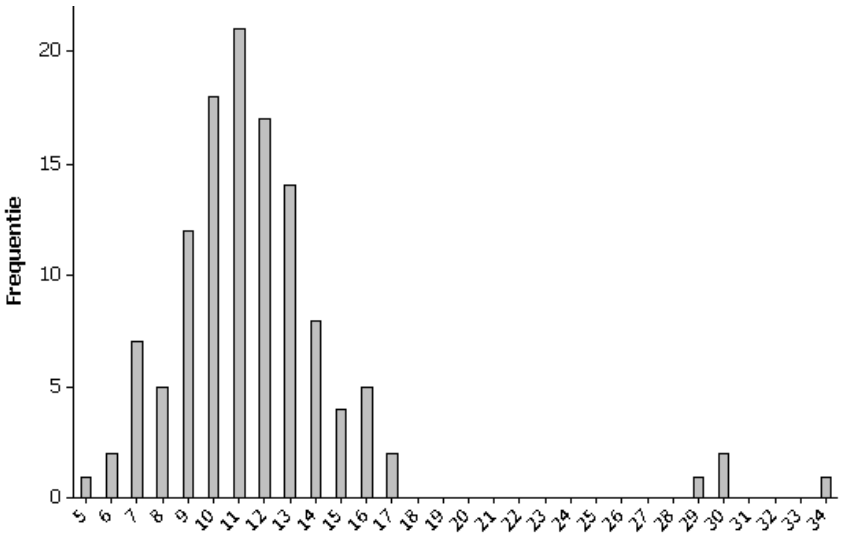
Het staafdiagram in figuur 4 stelt 120 getallen voor met mediaan $Me = 21$ en gemiddelde $\bar{x} = 18.9$. Hier is het gemiddelde dus kleiner dan de mediaan. Dat had je natuurlijk al verwacht, na de vorige paragraaf.



Figuur 4

2.2.3 Uitschieters

Uitschieters zijn getallen die ver weg liggen van de overgrote meerderheid van je opmetingen en die meestal een tussenruimte (een "gat" in de figuur) creëren. Als je een grafiek hebt met uitschieters, dan moet je goed opletten. Meestal trekken zij zeer sterk je aandacht zodat je wel eens een verkeerde indruk kan krijgen van het "globaal" gedrag van de meerderheid van je waarnemingen.



Figuur 5

Als je naar figuur 5 kijkt, zou je wel eens aan “scheef naar rechts” kunnen denken. Maar eigenlijk is dat niet zo en je mag zeker niet besluiten dat “scheef naar rechts” de globale eigenschap is van je opmetingen. Van die 120 getallen zijn er maar 4 die uitzonderlijk ver naar rechts liggen. De andere 116 getallen liggen mooi tussen 5 en 17 en hun staafdiagram zou je kunnen omschrijven als “ongeveer symmetrisch en met één top”.

Uitschieters moet je altijd speciale aandacht geven. Ga terug naar je oorspronkelijk onderzoek en probeer te achterhalen hoe die uitschieters tot stand zijn gekomen. Het zou niet de eerste keer zijn dat bij het intikken van een dataset er hier en daar een tikfout wordt gemaakt. Een getal dat oorspronkelijk is opgeschreven als 14 is misschien ingebracht als 144.

Ook als je geen tikfout ontdekt, dan nog moet je in je rapport de uitschieters afzonderlijk vermelden. Probeer, als je kan, er een zinvolle uitleg aan te geven.

Het gemiddelde is gevoelig voor uitschieters, zelfs als er slechts 4 zijn op een totaal van 120. Zonder die uitschieters is, voor de resterende 116 getallen, de mediaan $Me = 11$ en het gemiddelde $\bar{x} = 10.8$. Er is bijna geen verschil en dat verwacht je ook bij het staafdiagram van die 116 getallen. Maar voor alle 120 getallen, dus met de uitschieters erbij, is het gemiddelde gestegen tot $\bar{x} = 11.8$, terwijl de mediaan nog steeds $Me = 11$ is.

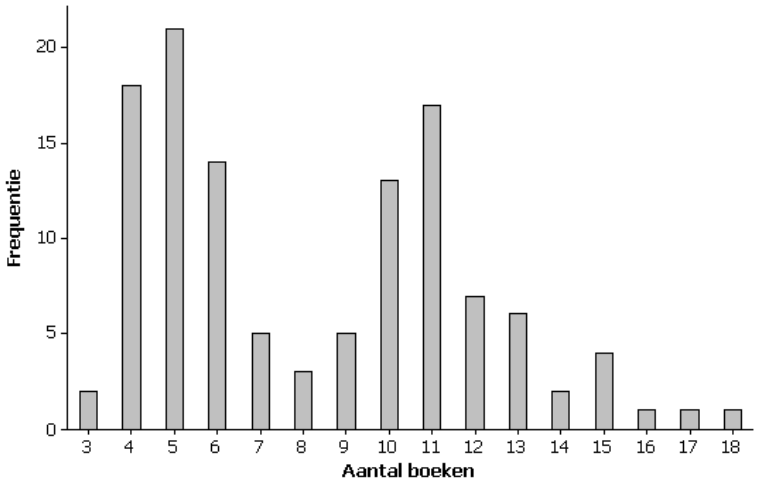
Op figuur 5 geeft de mediaan iets beter weer “waar je getallen globaal liggen”. In andere situaties verandert het gemiddelde nog veel meer door uitschieters. Dan is $\bar{x}$ helemaal niet goed om te zeggen “waar je getallen globaal liggen”.

2.2.4 Clusters

Aan 120 leerlingen werd gevraagd hoeveel boeken zij bij hadden. Twee leerkrachten (die niet in dezelfde klassen kwamen) ondervroegen elk 60 leerlingen en legden hun getallen samen. Hier zijn de 120 antwoorden.

8	4	5	10	11	12	16	5	5	11	6	14
4	5	11	17	6	5	10	14	6	4	13	12
6	4	7	10	11	6	11	12	5	3	10	9
4	11	6	11	8	13	11	8	5	10	10	11
4	18	7	10	10	5	10	10	6	11	9	11
10	7	11	5	5	5	5	15	4	6	7	4
5	9	13	9	5	5	6	7	4	5	15	5
15	4	13	4	4	6	3	4	4	11	9	12
13	11	4	12	10	10	11	5	5	5	6	4
12	11	13	15	6	6	4	12	6	5	11	4

Het kleinste van deze 120 getallen is 3 en het grootste is 18. Het gemiddelde is $\bar{x} = 8.2$ en de mediaan is $Me = 7.5$. Zijn het gemiddelde en de mediaan hier goede kengetallen voor "de typische ligging" van je observaties? Is het waar dat de meerderheid van die leerlingen rond de acht boeken per dag meebrengt? Om dit te weten heb je niet genoeg aan een getal, zoals een gemiddelde of een mediaan. Een figuur blijft altijd nodig.

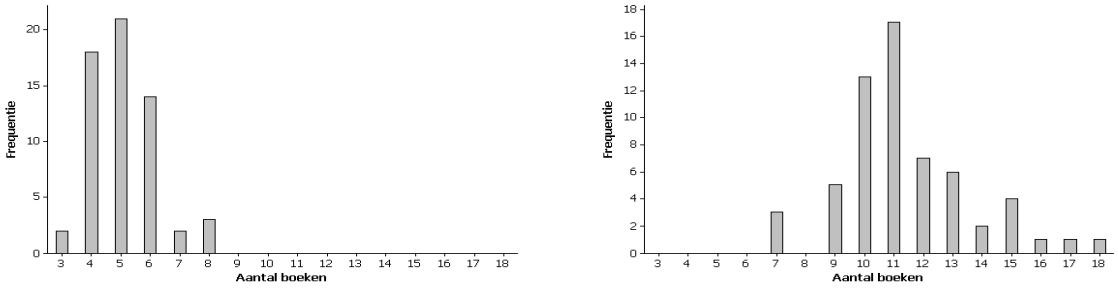


Figuur 6

Het blijkt helemaal niet waar te zijn dat een groot aantal leerlingen ongeveer 8 boeken bij heeft. Het staafdiagram vertelt een heel ander verhaal. Het lijkt wel een figuur met twee toppen. Er zijn blijkbaar 2 clusters (ophopingen), één rond 5 en één rond 11. Hoe kan dat verklaard worden?

Als je in je onderzoek een figuur ontmoet die lijkt op figuur 6, dan is de kans groot dat je te maken hebt met twee verschillende fenomenen. Probeer die te ontdekken en als je ze gevonden hebt, bestudeer dan elke situatie eens afzonderlijk. Als je niet vindt hoe je die twee clusters uit elkaar kan krijgen, beschrijf dan in je rapport dat je clusters hebt opgemerkt en geef aan dat het gemiddelde en de mediaan waarschijnlijk geen goede kengetallen zijn om je waarnemingen samen te vatten.

In het voorbeeld van de boeken kan de manier van opmeten je op een spoor brengen. De getallen werden verzameld door twee verschillende leerkrachten. Eén van de leerkrachten had die vraag gesteld op een donderdag maar de andere leerkracht had die vraag gesteld op een woensdag, en dan is er maar een halve dag les. De staafdiagrammen voor die twee groepen (elk gebaseerd op 60 opmetingen) zien er als volgt uit.



Figuur 7

Voor de eerste groep is de mediaan 5 en het gemiddelde 5.1. Voor de tweede groep is de mediaan 11 en het gemiddelde 11.4. Beide staafdiagrammen vertonen nu geen eigenaardige patronen meer en je zou ze kunnen omschrijven als "ongeveer symmetrisch en met één top".

Zowel de mediaan als het gemiddelde zijn hier goede kengetallen om, per groep, het "typisch centrum" aan te geven. De 60 leerlingen die over de middag thuis gaan eten hebben per halve dag ongeveer 5 boeken bij. De andere 60 leerlingen brengen ongeveer 11 boeken mee voor een volledige schooldag.

Als je dataset bestaat uit duidelijk verschillende groepen met elk een eigen gemiddelde en mediaan, dan is de kans groot dat je in je staafdiagram clusters ziet. Dit is iets anders dan een figuur die scheef is naar links of naar rechts.

3 Continu numerieke gegevens: gemiddelde, standaardafwijking en histogram

Sommige veranderlijken hebben uitkomsten die een continuüm bestrijken. Het echte geboortegewicht van een baby is misschien 3.49652485421154125… kilogram. Je hebt daar geen weegschaal voor en je moet ergens "afronden". Maar in feite bestrijkt "gewicht" een heel continuüm van numerieke uitkomsten. Daarom noem je zo'n veranderlijke "continu numeriek".

3.1 Het gemiddelde

Bij continue gegevens is het gemiddelde van *n* getallen $\{x_1, x_2, ..., x_n\}$ gelijk aan $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, net zoals bij discrete gegevens.

3.2 De standaardafwijking

De standaardafwijking is een maat voor de spreiding van getallen rond hun gemiddelde. Als getallen ver uiteen liggen, dan is de standaardafwijking groot. Als zij dicht tegen elkaar liggen, dan is de standaardafwijking klein.

De standaardafwijking bepaal je als volgt. Kijk eerst hoeveel elk getal afwijkt van het gemiddelde $\bar{x}$. Voor een getal x_i is die afwijking gelijk aan $x_i - \bar{x}$. Als x_i groter is dan $\bar{x}$, dan is $x_i - \bar{x}$ positief, maar als x_i kleiner is dan $\bar{x}$, dan is $x_i - \bar{x}$ negatief. Om altijd een positieve bijdrage te hebben, kan je werken met de absolute waarde $|x_i - \bar{x}|$ maar dat is ingewikkeld. Daarom neemt men gewoon het kwadraat $(x_i - \bar{x})^2$, want kwadrateren maakt ook elke uitdrukking positief.
Al die kwadratische verschillen worden dan samengeteld. Zo krijg je $\sum_{i=1}^{n}(x_i - \bar{x})^2$. Die som deel je door $(n-1)$, wat ééntje minder is dan het totale aantal getallen. Delen door $(n-1)$ heeft een goede reden in de statistiek maar daar gaan we nu nog niet op in.
Tenslotte trek je uit dat resultaat de positieve vierkantswortel. Eerst kwadrateren en later de wortel trekken zorgt ervoor dat je uitkomst terug in dezelfde eenheid kan geschreven worden als de eenheid van je oorspronkelijke opmetingen.

De notatie voor de standaardafwijking van je opmetingen is een kleine letter "s", en de formule is

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

De formule voor de standaardafwijking moet je niet uit het hoofd leren maar je moet ze wel in woorden kunnen lezen en er de bouwstenen van begrijpen.

Rekenmachines hebben dikwijls een knop voor de standaardafwijking waarbij gedeeld wordt door n en een andere knop die $(n-1)$ gebruikt. Als je wil weten welke knop de juiste is, dan is daar een handig trucje voor. Maak een lijst met de getallen $\{1, 2, 3\}$ en druk op een knop voor de standaardafwijking. Als het antwoord 1 (één) is, dan heb je de goede knop te pakken.

Soms kom je het woord "variantie" tegen. Dat is niets anders dan het kwadraat van de standaardafwijking. De notatie hiervoor is s^2, zoals verwacht.

Ook voor discreet numerieke veranderlijken kan je s berekenen. De formule is dezelfde.

3.3 Het histogram

Een histogram is de meest gebruikte figuur om het globale gedrag van continu numerieke gegevens te onderzoeken.

3.3.1 Een praktijkvoorbeeld

De volgende dataset toont de lengte van 150 scharnieren (in mm).

102.4	99.7	101.6	100.5	101.4	99.2	99.5	100.2	103.7	99.6	100.0
99.1	100.2	99.5	99.8	99.5	99.4	101.9	100.1	100.0	102.9	100.9
103.5	100.4	99.3	99.3	100.3	99.5	100.8	100.4	101.7	99.3	100.6
99.6	99.8	105.2	101.5	100.2	99.5	99.9	100.0	101.3	99.9	100.6
103.5	99.9	101.5	99.4	99.7	100.9	99.4	100.3	100.3	99.2	104.1
100.5	100.6	99.7	102.4	99.9	101.2	100.7	99.1	101.3	99.9	101.8
101.4	101.9	99.2	100.3	99.2	99.8	100.4	99.3	102.7	101.1	101.1
100.2	99.6	100.0	102.5	99.7	99.9	100.4	103.6	99.9	99.5	102.5
102.0	99.2	101.3	101.6	102.1	99.2	100.5	102.2	99.5	100.8	100.7
101.0	99.6	101.2	99.1	100.1	99.6	102.8	100.8	99.7	102.8	100.9
102.2	100.7	100.5	100.3	100.4	102.4	99.7	100.1	100.5	100.9	101.2
99.3	99.1	99.9	101.3	101.8	103.8	100.8	101.0	102.4	100.4	103.2
102.6	100.7	101.3	100.8	100.7	100.7	100.2	99.7	102.2	101.8	99.6
100.2	104.0	99.3	99.1	99.7	103.4	100.8				

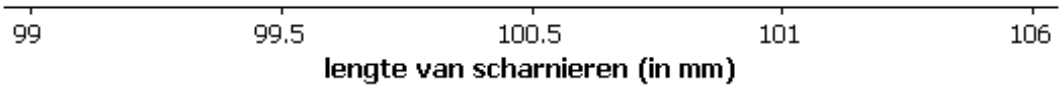
Het is de bedoeling dat de scharnieren 10 cm (= 100 mm) lang zijn, maar een afwijking (zowel te groot als te klein) tot ongeveer 0.5 millimeter is nog altijd prima. De scharnieren met een lengte tussen 99 en 99.5 mm of tussen 100.5 en 101 mm zijn ook nog bruikbaar maar moeten afzonderlijk verpakt worden. Alle scharnieren die 101 mm of meer zijn mogen niet verkocht worden. Om een zicht te krijgen op hoeveel scharnieren er in elk van die groepen zitten, heeft de bedrijfsleider al die lengtes laten samenvatten in een frequentietabel. In deze tabel lees je dat 50 van de 150 scharnieren niet verkocht mogen worden. Dat is één derde van deze scharnieren.

Frequentietabel met klassenindeling voor de lengte van scharnieren (in mm)	
Klasse	Aantal scharnieren f_i (frequentie)
[99.0 ; 99.5 [	20
[99.5 ; 100.5 [	56
[100.5 ; 101.0 [	24
[101.0 ; 106.0 [	50

De bedrijfsleider wil deze resultaten bespreken op een werkvergadering waarbij hij met een figuur de frequentietabel wil verduidelijken. Op de x-as zet hij de 4 klassen uit en daarboven tekent hij rechthoeken die verwijzen naar het aantal scharnieren per klasse.

Een eerste fout: de x-as.

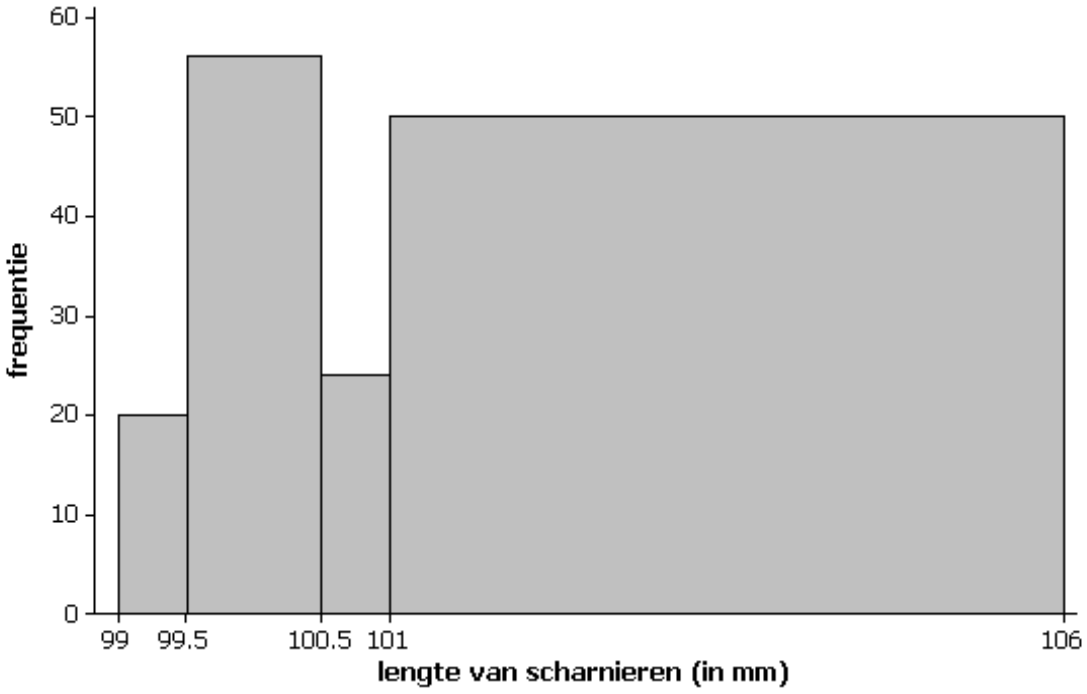
De ondergrens van de eerste klasse is 99 en de bovengrens van de laatste klasse is 106. Dat betekent dat je tussen 99 en 106 die 4 klassen moet tekenen op de x-as. Sommigen doen dit als volgt.



Als je even kijkt zie je dat dit niet juist is. Er zijn inderdaad 4 klassen, maar zij zijn niet allemaal even groot, en daar moet je natuurlijk rekening mee houden. Een juiste figuur ziet er als volgt uit.

Een tweede fout: de y-as.

Boven elke klasse moet je nu een rechthoek tekenen zodat je een goed idee krijgt van het aantal scharnieren per klasse. Sommigen nemen als hoogte van zo'n rechthoek gewoon de frequentie (dat is dus het aantal scharnieren in die klasse). Je krijgt dan de volgende figuur.



Figuur 8

In deze figuur word je overdonderd door de grote rechthoek boven de klasse met slechte scharnieren [101 ; 106 [. Bij deze "volle" figuur kijk je automatisch naar de oppervlakte. Je kan de totale oppervlakte van de eerste drie rechthoeken drie keer in de laatste rechthoek schuiven en dan heb je nog overschot! Die laatste rechthoek overheerst het totale beeld, zelfs al is hij niet de hoogste. Uit de getekende figuur zou je afleiden dat er drie keer zoveel slechte scharnieren zijn als goede.

De frequentietabel leerde ons dat er maar één derde van het totale aantal slecht is. De getekende figuur is dus verkeerd. Het is geen goed idee om op de y-as de frequenties uit te zetten. Maar wat dan wel?

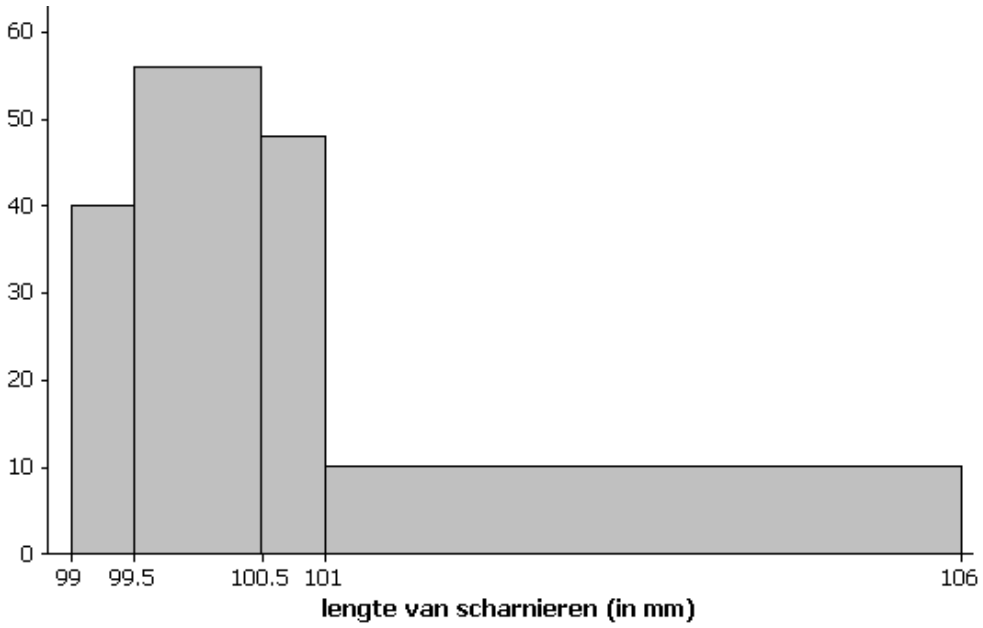
Een histogram is niet zoals een staafdiagram met losse staafjes waar de hoogte de aandacht trekt. Een histogram is een "volle" figuur van aaneensluitende rechthoeken waar de oppervlakte de aandacht trekt. Je moet er dus voor zorgen dat de oppervlakte van de rechthoeken een goede weerspiegeling is van het aantal scharnieren. De oppervlakte boven een klasse moet evenredig zijn met het aantal observaties in die klasse.

Als je bijvoorbeeld wil dat de oppervlakte van elke rechthoek gelijk is aan de frequentie (het aantal scharnieren in die klasse) dan kan je heel eenvoudig vinden wat de hoogte moet zijn. Uit $basis \times hoogte = oppervlakte$ volgt dat $klassenbreedte \times hoogte = frequentie$ of dat $b_i \times h_i = f_i$. Hierbij is b_i de breedte en f_i de frequentie van de i-de klasse. Voor de hoogte h_i van de rechthoek boven deze klasse volgt dan dat $h_i = f_i / b_i$.

Je kan de hoogte als een extra kolom aan je frequentietabel toevoegen.

Frequentietabel met klassenindeling voor de lengte van scharnieren (in mm)			
Klasse	Klassenbreedte b_i	Aantal scharnieren f_i (frequentie)	Hoogte $h_i = f_i / b_i$
[99.0 ; 99.5 [	0.5	20	40
[99.5 ; 100.5 [	1	56	56
[100.5 ; 101.0 [	0.5	24	48
[101.0 ; 106.0 [	5	50	10

Het histogram met aangepaste hoogte ziet er dan als volgt uit.



Figuur 9

Dit histogram vertelt het juiste verhaal. De lange rechthoek vertegenwoordigt nu één derde van de totale oppervlakte. Eén derde van de scharnieren is te groot. Dat bleek ook uit de frequentietabel.

Bemerk dat de y-as onbenoemd is gebleven. Dit is niet erg als je hier afspreekt dat je het histogram zo tekent dat het maatgetal van de oppervlakte gelijk is aan de frequentie. Dat is hier ook zo. De oppervlakte (zonder eenheden) van de laatste rechthoek is gelijk aan $basis \times hoogte = 5 \times 10 = 50$ en dat is inderdaad gelijk aan het aantal scharnieren dat te groot is.

3.3.2 Een histogram tekenen

De volgende dataset toont de diameter (in mm) van 160 precisiewerkstukken.

10.41 10.42 10.87 10.37 10.30 10.47 10.37 10.55 10.40 10.33 10.43 10.66 10.40
10.72 10.69 10.55 10.28 10.27 10.01 10.64 10.22 10.47 10.54 10.49 10.63 10.84
10.74 10.24 10.48 10.68 10.50 10.88 10.34 10.59 10.68 10.48 10.35 10.63 10.62
10.21 10.52 10.50 10.68 10.23 10.54 10.45 10.42 10.18 10.62 10.16 10.32 10.69
10.76 10.58 10.51 10.53 10.53 10.75 10.12 10.39 10.58 10.31 10.56 10.21 10.15
10.47 10.62 10.63 10.33 10.04 10.49 10.65 10.50 10.93 10.47 10.75 10.55 10.64
10.67 10.20 10.90 10.27 10.43 10.30 10.78 10.25 10.27 10.38 10.52 10.30 10.82
10.52 10.30 10.66 10.79 10.49 10.60 10.57 10.60 10.57 10.78 10.63 10.47 10.36
10.61 10.44 10.49 10.46 10.42 10.05 10.85 10.36 10.45 10.61 10.45 10.51 10.74
10.51 10.86 10.22 10.46 10.25 10.50 10.63 10.54 10.48 10.45 10.72 10.71 10.98
10.55 10.44 10.37 10.15 10.39 10.58 10.45 10.36 10.39 10.51 10.60 10.13 10.54
10.38 10.23 10.39 10.77 10.65 10.74 10.55 10.74 10.85 10.22 10.53 10.37 10.33
10.65 10.37 10.72 10.70

Om "met de hand" een histogram te tekenen heb je een frequentietabel met klassenindeling nodig. Hoe je die klassen maakt, mag je vrij kiezen maar je moet er natuurlijk voor zorgen dat je totale gebied groot genoeg is om al je gegevens te bevatten. Met een klein beetje speurwerk ontdek je hier dat het kleinste getal 10.01 en het grootste 10.98 is. Meestal kies je "ronde" getallen en hier zou je bijvoorbeeld kunnen starten bij 10 en lopen tot 11 met een klassenbreedte van 0.10.

Frequentietabel met klassenindeling voor diameters (in mm)			
Klasse	Klassen-breedte b_i	Frequentie f_i	Hoogte $h_i = 0.10 \times \left(f_i / b_i \right)$
[10.00 ; 10.10[	0.10	3	3
[10.10 ; 10.20[	0.10	6	6
[10.20 ; 10.30[	0.10	15	15
[10.30 ; 10.40[	0.10	25	25
[10.40 ; 10.50[	0.10	29	29
[10.50 ; 10.60[	0.10	30	30
[10.60 ; 10.70[	0.10	26	26
[10.70 ; 10.80[	0.10	16	16
[10.80 ; 10.90[	0.10	7	7
[10.90 ; 11.00[	0.10	3	3

Let erop dat de klassen beginnen met een gesloten haakje en eindigen met een open haakje. Het zijn dus "half gesloten – half open" intervallen. Het klassenmidden is het gemiddelde van de onder- en de bovengrens van de klasse. De formule voor de hoogte h_i wordt in de volgende paragraaf verklaard.

Een histogram teken je nu als volgt:

- Start met de klassen aan te duiden op de x-as. Je kan de klassengrenzen aangeven of, als alle klassen even breed zijn, enkel het klassenmidden.
- Teken dan op elk interval een rechthoek. Aangezien alle intervallen op elkaar aansluiten, liggen ook alle rechthoeken tegen elkaar.
- Hoe hoog moet die rechthoek zijn? Schrik niet, maar het antwoord hierop is: dat mag je zelf kiezen, zolang je de **basiseigenschap van een histogram** respecteert.

Basisafspraak voor het tekenen van een histogram.

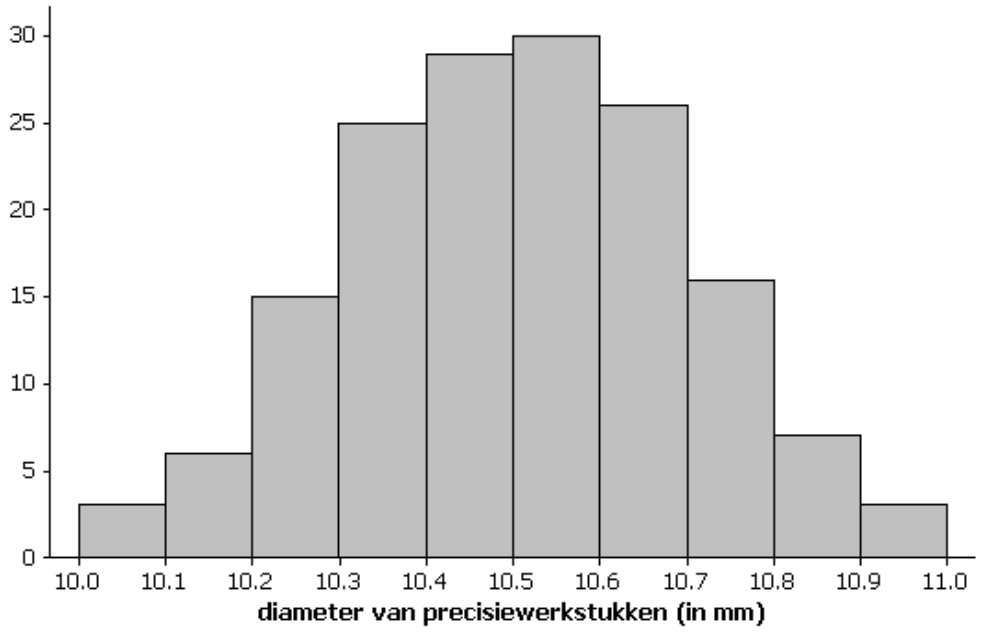
De <u>OPPERVLAKTE</u> van een rechthoek is recht evenredig met het aantal observaties in de klasse waarop die rechthoek staat

---

In de praktijk betekent dit dat, per klasse, de hoogte h_i gelijk moet zijn aan f_i / b_i, op een vrij te kiezen evenredigheidsfactor k na. De hoogte wordt dus berekend via $k \times \left(f_i / b_i \right)$.

Bij het voorbeeld met de scharnieren was de evenredigheidsfactor k gelijk aan één want je nam daar $h_i = f_i / b_i = 1 \times \left(f_i / b_i \right)$. Je hebt daar gezien dat de oppervlakte van een rechthoek dan gelijk is aan de frequentie. Als je nu de hoogte van zo'n rechthoek vermenigvuldigt met k dan wordt natuurlijk ook de oppervlakte vermenigvuldigd met hetzelfde getal k. Die oppervlakte is dan niet meer gelijk aan de frequentie maar aan k keer de frequentie.

Om het histogram voor de diameters te tekenen is als evenredigheidsfactor $k = 0.10$ gekozen. De evenredigheidsfactor is hier dus gelijk aan de klassenbreedte.

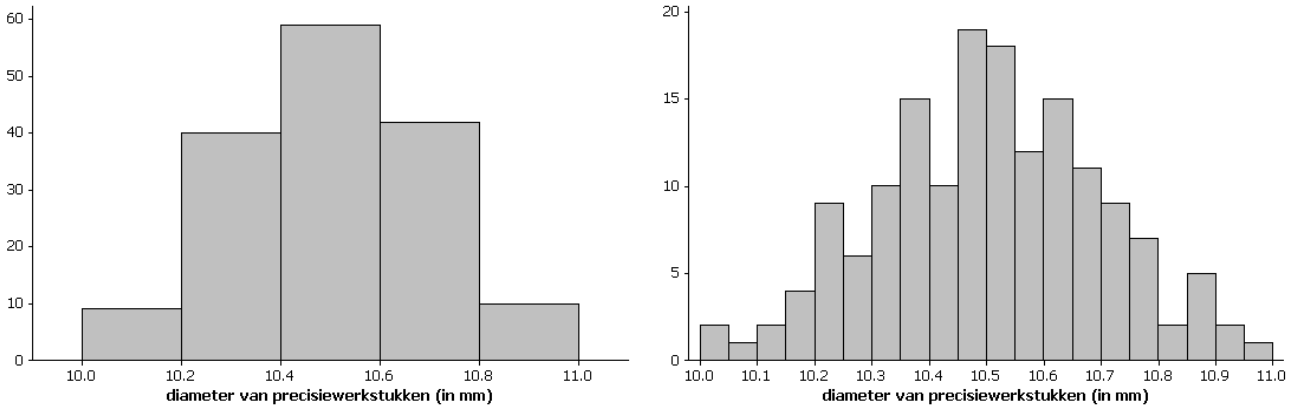


Figuur 10

Als je (zonder eenheden) de oppervlakte van de rechthoeken narekent dan vind je 0.3 voor de eerste, 0.6 voor de tweede, 1.5 voor de derde, enz. Deze getallen zijn recht evenredig met het aantal observaties per klasse. Inderdaad, het zijn de getallen die je vindt als je het product maakt: $k \times$ *frequentie* met $k = 0.10$.

Opmerking over de keuze van de klassenbreedte

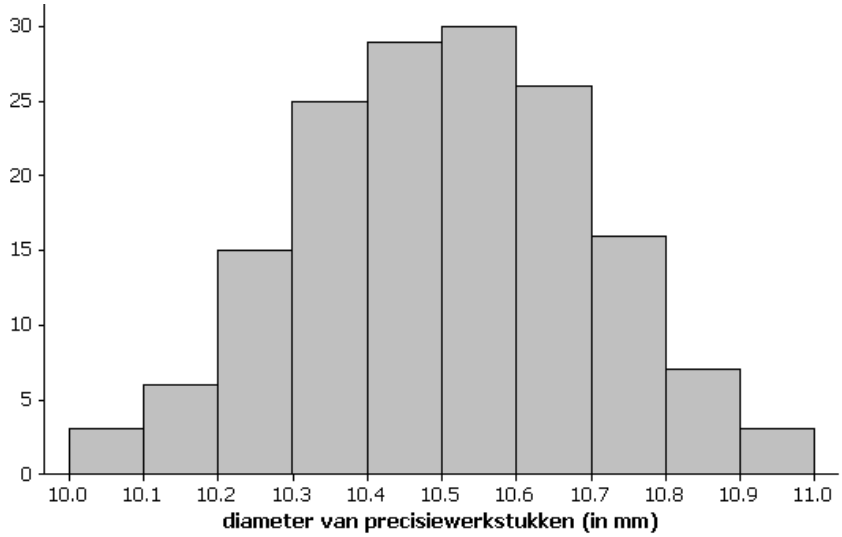
Andere keuze van klassen geven andere histogrammen. Een beetje experimenteren kan hier geen kwaad. Hieronder staan nog twee histogrammen van dezelfde dataset. Te brede klassen geven soms te weinig informatie (wat zou er hier gebeuren als je de klassenbreedte gelijk zou nemen aan één?). Te smalle klassen leiden dikwijls tot een zenuwachtige figuur.



Figuur 11

3.3.3 Een histogram interpreteren

Een histogram gebruik je om een globaal zicht te krijgen op continue data. Je kijkt naar kenmerken zoals symmetrie, scheefheid, aantal uitgesproken toppen, opvallende gaten, enz. Stap af van de drang om naar hoogtes te kijken maar laat je aandacht trekken door oppervlakten en door de "globale vorm" van de figuur.



Figuur 12

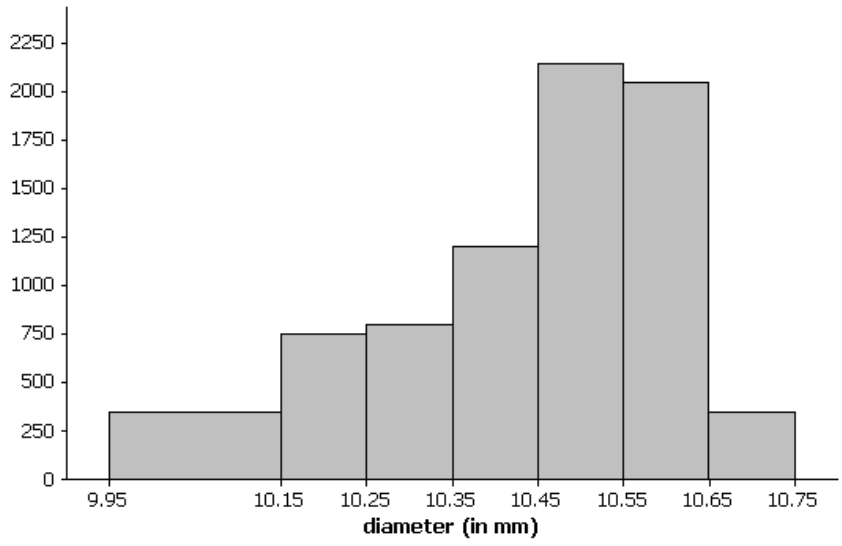
Het histogram van de diameters heeft hier één "top-gebied" waarrond de figuur vrij symmetrisch daalt naar links en naar rechts. Er zijn verder geen

eigenaardige patronen. In zo'n geval zijn het gemiddelde en de standaardafwijking goede maten voor het "centrum" en de "spreiding" van de data. Hier is $\bar{x} = 10.50\ mm$ en $s = 0.196\ mm$.

De gebieden "waar" de data liggen zie je op de x-as. Of er in bepaalde gebieden meer of minder data liggen, zie je door naar de oppervlakte boven die gebieden te kijken. Zo is de oppervlakte boven [10.50 ; 10.60[veel groter dan boven [10.80 ; 10.90[terwijl die intervallen toch even lang zijn.

Een andere dataset levert de volgende frequentietabel met klassenindeling. Voor de hoogtes in het bijhorende histogram is een evenredigheidsfactor $k = 5$ genomen, gewoon om te illustreren dat je echt een vrije keuze hebt. Je verwacht nu dat elke rechthoek een oppervlakte zal hebben die gelijk is aan 5 keer de frequentie van de klasse waarop hij staat. Dat klopt. Reken maar na!

Frequentietabel met klassenindeling voor diameters (in mm)			
Klasse	Klassen-breedte b_i	Frequentie f_i	Hoogte $h_i = 5 \times \left(f_i / b_i \right)$
[9.95 ; 10.15[	0.20	14	350
[10.15 ; 10.25[	0.10	15	750
[10.25 ; 10.35[	0.10	16	800
[10.35 ; 10.45[	0.10	24	1200
[10.45 ; 10.55[	0.10	43	2150
[10.55 ; 10.65[	0.10	41	2050
[10.65 ; 10.75[	0.10	7	350



Figuur 13

De frequentie van de eerste klasse is gelijk aan 14 en de oppervlakte van de eerste rechthoek is $0.20 \times 350 = 70$, wat gelijk is aan 5 keer 14.

Het histogram heeft één top maar is rond die top helemaal niet symmetrisch. De figuur is "scheef naar links", wat betekent dat de getallen veel verder naar links uitgespreid liggen dan naar rechts. Bij figuren die heel erg scheef zijn of die eigenaardige patronen vertonen geven het gemiddelde en de standaardafwijking je nog weinig informatie.

Dikwijls is het verstandig om naast een histogram ook een boxplot te tekenen en om naast het gemiddelde en de standaardafwijking ook te kijken naar de mediaan en de interkwartielafstand. Hierover lees je meer in het volgende hoofdstuk.

4 Continu numerieke gegevens: mediaan, interkwartielafstand en boxplot

4.1 De mediaan en de kwartielen

De mediaan en de kwartielen zijn kengetallen die je gebruikt om een bepaalde plaats aan te duiden in een geordende rij waarnemingsgetallen. Je kan die zowel bij discrete als bij continue gegevens gebruiken.

4.1.1 De mediaan

De mediaan bepaal je net zoals bij discrete veranderlijken: je ordent de getallen van klein naar groot. Bij een oneven aantal getallen neem je het middelste. Bij een even aantal neem je het gemiddelde van de 2 middelste.

4.1.2 De kwartielen

Om de kwartielen te zoeken, orden je de data van klein naar groot. Bepaal 3 plaatsen zodat je de getallenrij in vier gelijke delen verdeelt. De getallen die op die plaatsen staan, zijn de kwartielen.

"Een vierde" wordt ook wel "een kwart" genoemd en zo kan je het woord "kwartiel" gemakkelijk onthouden. Je hebt drie plaatsen nodig om een rij in vier te verdelen en eigenlijk heb je drie kwartielen: Q_1, Q_2 en Q_3. Maar het tweede kwartiel Q_2 verdeelt de rij in twee gelijke delen en is dus gelijk aan de mediaan. Daarom spreek je niet over het tweede kwartiel maar wel over de mediaan *Me*.

Als je een oneven aantal getallen hebt ga je als volgt te werk.

- Neem de geordende dataset {4, 6, 7, 8, 9, 10, 11, 12, 13, 19, 20} en deel die eerst in twee. Dat doe je door de mediaan te zoeken. Hier is de mediaan gelijk aan 10 en dat is één van de observatiegetallen zelf.
- Neem nu de eerste helft kleiner dan 10 namelijk {4, 6, 7, 8, 9} en de tweede helft groter dan 10, namelijk {11, 12, 13, 19, 20}.
- Deel deze twee helften terug in twee door telkens hun mediaan te zoeken. Het midden van de eerste helft {4, 6, 7, 8, 9} is gelijk aan 7 en dat noem je het eerste kwartiel Q_1. Het midden van de tweede helft {11, 12, 13, 19, 20} is gelijk aan 13 en dat noem je het derde kwartiel Q_3. Als je in {4, 6, 7, 8, 9, 10, 11, 12, 13, 19, 20} de getallen 7, 10 en 13 kleurt dan zie je dat zij de geordende dataset in 4 gelijke delen verdelen.

Bij een even aantal getallen is de mediaan geen observatiegetal. Je neemt dan alle getallen kleiner dan de mediaan als eerste helft en alle getallen groter dan de mediaan als tweede helft. En die twee helften deel je terug in twee door telkens hun eigen mediaan te bepalen.

Voor {4, 6, 7, 8, 9, 10, 11, 12, 13, 19} is de mediaan gelijk aan 9.5. De eerste helft is dan {4, 6, 7, 8, 9} en de mediaan daarvan is 7. De tweede helft is {10, 11, 12, 13, 19} en de mediaan daarvan is 12. Je hebt dus voor {4, 6, 7, 8, 9, 10, 11, 12, 13, 19} dat $Q_1 = 7$, $Me = 9.5$ en $Q_3 = 12$.

4.1.3 De interkwartielafstand

De interkwartielafstand is gewoon de afstand tussen de twee kwartielen Q_1 en Q_3. De interkwartielafstand heet in het Engels **i**nter-**q**uartile **r**ange en daarom wordt hij afgekort als *IQR*. De interkwartielafstand is de lengte van een gebied rond de mediaan waarbinnen de middelste helft van al je gegevens ligt. Als de *IQR* klein is, dan betekent dit dat de middelste helft van je data dicht rond de mediaan geconcentreerd ligt. Bij een grote *IQR* liggen die data verder uiteen. Daarom gebruikt men de *IQR* als een maat om aan te geven hoe groot de spreiding is van je getallen rond hun mediaan.
Voor {4, 6, 7, 8, 9, 10, 11, 12, 13, 19} is de interkwartielafstand *IQR* gelijk aan $Q_3 - Q_1 = 12 - 7 = 5$. Voor {4, 6, 7, 8, 9, 10, 11, 12, 13, 19, 20} is *IQR* = 6.

4.1.4 Uitschieters

Uitschieters zijn getallen die “uitzonderlijk groot” of “uitzonderlijk klein” zijn in vergelijking met de getallen in je dataset.

Je kan natuurlijk niet zomaar op je gevoel afgaan om te weten of een getal een uitschieter is. Daarom gebruik je in de statistiek een vuistregel.

Voor "uitzonderlijk groot" start je op het derde kwartiel Q_3 en daar tel je nog anderhalve keer de interkwartielafstand bij. Alle getallen die nog voorbij dat punt liggen zijn "uitzonderlijk groot" en worden uitschieters genoemd.
Voor "uitzonderlijk klein" doe je eigenlijk hetzelfde maar in de andere richting. Je start daar met het kleinste kwartiel Q_1 en daar trek je nog anderhalve keer de interkwartielafstand af. Alle getallen die daar nog onder liggen zijn "uitzonderlijk klein" en worden uitschieters genoemd.

Een getal dat buiten $\left[\ Q_1 - 1.5 \times IQR\ ;\ Q_3 + 1.5 \times IQR\ \right]$ valt is een uitschieter.

Voor {4, 6, 7, 8, 9, 10, 11, 12, 13, 19, 20} is $Q_1 = 7$, $Q_3 = 13$ en $IQR = 6$. Dus is $Q_1 - 1.5 \times IQR = 7 - (1.5)(6) = -2$ en $Q_3 + 1.5 \times IQR = 13 + (1.5)(6) = 22$. Er zijn geen data die buiten het interval [–2 ; 22] vallen. Hier zijn dus geen uitschieters.

Voor {4, 6, 7, 8, 9, 10, 11, 12, 13, 19, 23} is $Q_1 = 7$, $Q_3 = 13$ en $IQR = 6$. Het getal 23 is groter dan $Q_3 + 1.5 \times IQR = 13 + (1.5)(6) = 22$. Het is een uitschieter.

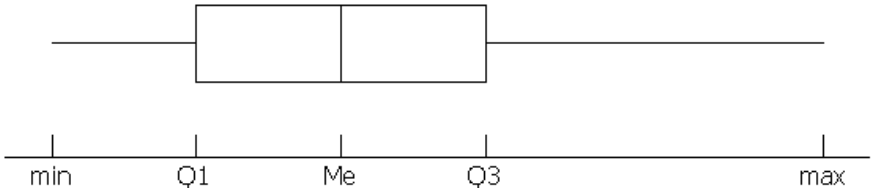
Een uitschieter beïnvloedt het gemiddelde en de standaardafwijking. De mediaan en de *IQR* veranderen niet door een uitschieter. Zij zijn in aanwezigheid van uitschieters dikwijls een betere centrum- en spreidingsmaat.

4.2 De boxplot

De mediaan en de kwartielen vertellen je al een en ander over je dataset. De boxplot is een goede figuur om die informatie nog beter te ontdekken.

4.2.1 Een boxplot tekenen

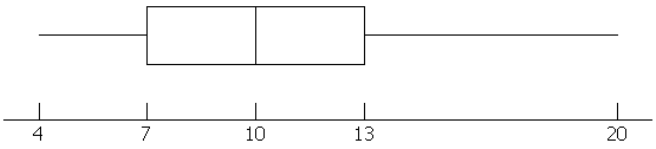
Een boxplot is een eenvoudige figuur die bestaat uit een rechthoekig doosje (een "box") waaruit langs beide zijden een staafje komt. Soms noemt men een boxplot ook wel eens een snorrendoos (een "box-and-whisker" plot).
Wanneer er geen uitschieters zijn, ziet een boxplot er als volgt uit.



Op de x-as duid je aan waar het eerste en het derde kwartiel Q_1 en Q_3 liggen en daarboven teken je de rechthoekige doos. Binnen die doos trek je een lijn op de

plaats van de mediaan *Me*. De staafjes die uit de doos komen lopen langs rechts vanaf het derde kwartiel tot aan het grootste observatiegetal en langs links vanaf het eerste kwartiel tot aan het kleinste observatiegetal.

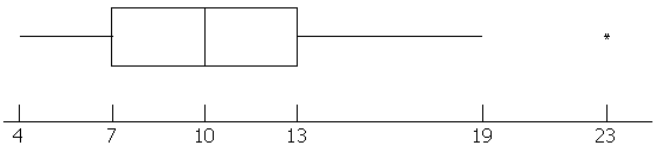
Controleer dat voor de dataset {4, 6, 7, 8, 9, 10, 11, 12, 13, 19, 20} de boxplot er als volgt uitziet.



Wanneer er uitschieters zijn in je gegevens, dan worden de staafjes anders getekend. Neem nu het grootste en het kleinste datapunt dat nog net binnen het interval $\left[Q_1 - 1.5 \times IQR \;;\; Q_3 + 1.5 \times IQR\right]$ ligt. Teken nu je staafjes tot aan dit grootste en kleinste datapunt. De getallen die er buiten vallen stel je voor door een sterretje: dat zijn uitschieters.

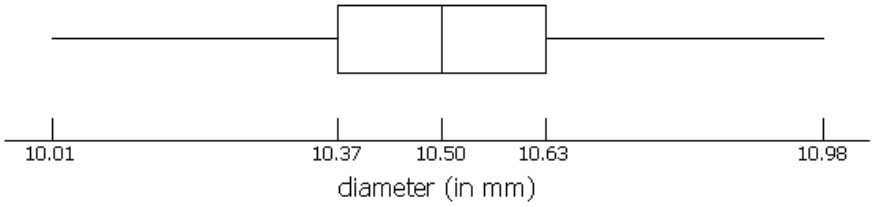
Voor {4, 6, 7, 8, 9, 10, 11, 12, 13, 19, 23} is $Q_1 = 7$, $Q_3 = 13$ en $IQR = 6$. Getallen die groter zijn dan $Q_3 + 1.5 \times IQR = 13 + (1.5)(6) = 22$ zijn uitschieters.

Een klassieke fout maak je als je het rechterstaafje laat lopen tot 22 en dan een sterretje zet op 23. Dit is niet de afspraak voor het tekenen van dat staafje. Het staafje moet lopen tot aan het grootste "gewoon" (= geen uitschieter) datapunt (= getal in je dataset). In dit voorbeeld is dat tot aan 19 want dat is het grootste getal in je dataset dat nog geen uitschieter is. Je krijgt dan de volgende boxplot.



4.2.2 Een boxplot interpreteren

Hieronder zie je de boxplot voor de diameter van die precisiewerkstukken.



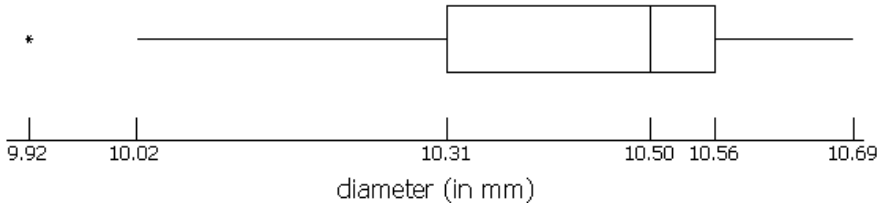
De figuur, samen met de getallen op de x-as, vertelt al heel wat over je dataset:

- alle getallen liggen tussen 10.01 (het minimum) en 10.98 (het maximum) en er zijn geen uitschieters (want er zijn geen sterretjes)
- de box toont waar de middelste helft van de geordende getallen ligt: dit is van $Q_1 = 10.37$ tot $Q_3 = 10.63$. De getallen liggen daar symmetrisch rond de mediaan $Me = 10.50$ want de mediaan ligt in het midden van de box, halfweg tussen het eerste en derde kwartiel.
- het centrale gebied tussen het eerste en derde kwartiel is veel korter dan de helft van het totale gebied (0.26 mm op een totaal van 0.97 mm). De getallen in het centrum liggen dus dichter bij elkaar. In de staarten liggen zij meer uitgespreid.
- de staafjes tonen waar het kleinste vierde en het grootste vierde van de getallen ligt. Ook hier is er symmetrie links en rechts van de box. Het eerste en laatste vierde liggen meer uitgespreid dan de box en lopen elk over een afstand van ongeveer 0.36 mm.

Een boxplot vertelt niet alles over een dataset, maar deze figuur laat toch vermoeden dat je te maken hebt met opmetingen die symmetrisch rond 10.50 liggen, waarvan de meerderheid niet te ver van dat centrum verwijderd is.

Een andere dataset ziet er als volgt uit. Ook hiervan is de boxplot getekend.

10.50 10.13 10.50 10.25 10.55 10.51 10.32 10.55 10.54 10.31 10.50
10.51 10.53 10.30 10.25 10.22 10.59 10.44 10.39 10.27 10.50 10.35
10.31 10.52 10.16 10.43 10.57 10.46 10.54 10.51 10.47 10.42 10.34
10.60 10.57 10.36 10.54 10.67 10.52 10.02 10.23 10.23 10.58 10.57
10.31 10.61 10.04 10.24 10.66 10.63 10.41 10.40 10.51 10.30 10.03
10.36 10.52 10.41 10.52 10.20 10.51 10.53 10.21 10.37 10.39 10.56
10.52 10.57 10.69 10.53 10.59 10.44 10.51 10.33 10.58 10.12 10.60
10.08 10.22 10.43 10.61 10.57 10.43 10.59 10.50 10.66 10.55 10.53
10.16 10.08 10.50 10.54 10.24 10.60 10.08 10.05 10.39 10.64 10.54
10.63 10.60 10.58 10.50 10.57 10.62 10.40 10.63 10.12 10.04 10.15
10.60 10.63 10.50 10.30 10.43 10.54 10.28 10.62 10.51 10.55 10.56
10.12 10.65 10.55 10.61 10.47 10.56 10.55 9.92 10.15 10.67 10.62
10.30 10.45 10.29 10.26 10.36 10.47 10.53 10.20 10.58 10.37 10.54
10.53 10.08 10.20 10.65 10.18 10.58 10.42 10.39 10.62 10.47



Van deze boxplot kan je meerdere dingen aflezen:

- het sterretje zegt dat 9.92 een uitschieter is. Het kleinste getal in de dataset dat geen uitschieter is, is 10.02 want vanaf dat punt begint het linkerstaafje. Op die ene uitschieter na liggen alle getallen tussen 10.02 en 10.69 (het grootste datapunt).
- de box toont waar de middelste helft van de geordende getallen ligt, dat is tussen $Q_1 = 10.31$ en $Q_3 = 10.56$. Die middelste helft is op zichzelf helemaal niet symmetrisch rond de mediaan $Me = 10.50$. Het streepje van de mediaan staat veel **dichter tegen de rechterkant** van de box dan tegen de linkerkant. Van deze middelste groep getallen zit de helft geconcentreerd tussen 10.50 en 10.56 terwijl de andere helft naar links uitgespreid is van 10.50 tot 10.31. Die middelste helft getallen is blijkbaar scheef naar **links** ten opzichte van de mediaan.
- de staafjes tonen waar het kleinste en het grootste vierde van de getallen liggen. Ook hier is er geen symmetrie. Het linkerstaafje is veel langer (0.29 mm) dan het rechterstaafje (0.13 mm). Dit betekent dat eenzelfde aantal getallen (namelijk een kwart) veel verder links uiteengespreid ligt dan rechts.
- de globale indruk van de boxplot laat vermoeden dat je hier te maken hebt met een dataset die ten opzichte van de mediaan scheef naar links is, met uiterst links zelfs een uitschieter.

Pas op! In een boxplot vertegenwoordigt de rechthoek links of rechts van de mediaan telkens een even groot aantal gegevens, namelijk een kwart. Dus hoe groter de rechthoek is, hoe meer verspreid je gegevens zijn. Bij een histogram is dat anders: hoe groter de oppervlakte van een rechthoek, hoe meer gegevens er in die klasse voorkomen.

Er is nog een ander verschil tussen boxplot en histogram: je kan maar 1 boxplot tekenen van een dataset maar er zijn meerdere histogrammen mogelijk. Bij histogrammen moet je dus zelf wat experimenteren. Meestal is het goed om zowel histogrammen als een boxplot in je rapport op te nemen bij de interpretatie van je dataset.

www.uhasselt.be/lesmateriaal-statistiek