Essential Skills Mathematics

Modulewijzer

Module S1

Statistiek

Leerdoelen en onderwerpen

Leerdoelen

Na deze module bestudeerd te hebben moet de student:

- Een percentage kunnen bepalen
- Een wel of niet gewogen rekenkundig gemiddelde kunnen bepalen
- Van een verzameling getallen de mediaan kunnen berekenen
- Van een verzameling getallen de kwartielen en de halve kwartielafstand kunnen berekenen
- Van een verzameling getallen een bijbehorend boxplot kunnen maken
- Een standaardafwijking voor een steekproef kunnen berekenen

Onderwerpen

S1.1. De bepaling van een percentage

S1.2. Het berekenen van een wel of niet gewogen rekenkundig gemiddelde

S1.3. De berekening van de mediaan

S1.4. De berekening van de kwartielen en de halve kwartielafstand van een getallenverzameling

S1.5. De constructie van een boxplot, behorend bij een bepaalde getallenverzameling

S1.6. De standaardafwijking

S1.1. De bepaling van een percentage

Voorbeeld 1
190 inwoners van Wildevanck werd naar hun mening gevraagd over een aantal door de gemeente van Wildevanck voorgestelde maatregelen. Zie voor de uitslag de onderstaande tabel:

	mee eens	geen mening	oneens	totaal
man	57	21	20	98
vrouw	23	29	40	92
totaal	80	50	60	190

We kunnen nu m.b.v. deze tabel *percentages* bepalen.

- Zo is bijvoorbeeld het percentage van de mannen dat het met de voorgestelde maatregelen eens was gelijk aan (57/98) * 100% = 58.2% .
- En als je slechts kijkt naar de groep mensen die het met de maatregelen eens waren, dan was (57/80) * 100% = 71.3 % daarvan een man.
- En van alle 92 ondervraagde vrouwen had (29/92) * 100 % = 31.5% geen mening.
- En het percentage vrouwen, dat het met de voorgestelde maatregelen eens was, als percentage van de totale steekproef, bedroeg (23/190) * 100% = 12%

Opgave S1.1.

Bij een enquete in Schiedam werden 200 personen gevraagd naar hun stemgedrag. Zie de tabel:

	PvdA	VVD	PVV	Totaal
Man	36	21	39	96
Vrouw	24	21	59	104
Totaal	60	42	98	200

Van alle ondervraagden die op de PVV stemden, is het percentage dat vrouw was gelijk aan:

(a) O 50.5 %
(b) O 35 %
(c) O 60.2 %
(d) O 70.2 %

Ga nu in Grasple aan de slag met opgaven van dit onderdeel om na te gaan of je de stof goed hebt begrepen.

S1.2. Het berekenen van een gemiddelde

Het gemiddelde van een aantal getallen berekent men door alle getallen op te tellen en te delen door het aantal. Men noemt dit gemiddelde het ***rekenkundig gemiddelde***.

Voorbeeld 1

Het gemiddelde maandinkomen van een groep van 17 pas afgestudeerde hbo-ers bedraagt € 2160 per maand. Er blijkt echter nog een vrouw bij de groep te horen. Haar maandinkomen bedraagt € 2070 per maand. Van de totale groep van 18 hbo-ers is het gemiddelde inkomen dan gelijk aan:

(a) O € 2115

(b) O € 2075

(c) O € 2155

(d) O Er zijn te weinig gegevens bschikbaar om het totale gemiddelde te bepalen

Oplossing:

Noem je de maandinkomens van de afgestudeerde hbo-ers I_1 , I_2 , ...t/m I_{17} , dan wordt het gemiddelde inkomen I_{gem} berekend door de 17 inkomens op te tellen en daarna door 17 te delen :

$$I_{gem} = (I_1 + I_2 + I_3 + ... + I_{17}) / 17 = 2160$$

Uit bovenstaande formule volgt dat $I_1 + I_2 + I_3 + ... + I_{17} = 17 * 2160 =$ € 36720.-. Met de vrouw erbij verdienen de 18 hbo-ers samen 36720 + 2070 = € 38790.- per maand. Om het gemiddelde inkomen van deze 18 hbo-ers te vinden moeten we 38790 delen door 18. Het nieuwe gemiddelde wordt 38790 : 18 = € 2155.- per maand. Antwoord (c) is correct.

Het kan zijn dat bij het berekenen van het gemiddelde sommige getallen zwaarder wegen dan andere. In dit geval spreken we dan van een ***gewogen gemiddelde***. We illustreren dit m.b.v. een voorbeeld.

Voorbeeld 2

Een docente wiskunde neemt 3 toetsen af bij haar studenten. Gezamenlijk tellen de cijfers voor die drie toetsen mee voor het eindcijfer, afgerond op 1 decimaal. De 1ste toets telt niet zwaar mee in het eindcijfer en heeft een kleine gewichtsfactor van 10%. Toets 2 telt ietsje zwaarder mee en heeft een gewichtsfactor van 20%. Maar de 3de toets is zeer belangrijk : de 3de toets heeft een gewichtsfactor van 70% . Stel dat een student op een 100 puntsschaal voor toets 1 de score 70 heeft gehaald, voor toets twee de score 63 en voor toets 3 de score 40. Wat is dan de eindscore voor deze student?

Oplossing:
Je vindt de eindscore door het resultaat van elke toets met de bij die toets behorende gewichtsfactor te vermenigvuldigen en daarna op te tellen. De gewichtsfactoren moeten dan als *fracties* genoteerd worden, dus 20% wordt 0.2 enz. De eindscore wordt:

Eindscore = 0.10 * 70 + 0.20 * 63 + 0.70 * 40 = 47.6
Het eindcijfer is dan 4.8

Je kunt gemakkelijk constateren dat de gewichtsfactor van 70% dominant is in de bepaling van de eindscore.

Voorbeeld 4
In een groep van 25 personen is van elk persoon de massa bepaald. Hieronder zie je een tabel met de gemeten massa's en het aantal keren dat de betreffende massa gemeten is:

massa (kg)	63	64	66	67	68
aantal	4	6	7	5	3

Gevraagd: bereken het gemiddelde van de 25 massa's op één decimaal nauwkeurig.

Oplossing:
Gebruiken we het idee van gewichtsfactoren dan zien we dat de massa van 63 kg dan 4 keer voor-komt, oftewel een gewichtsfactor heeft van (4/25) = 0.16. Voor 64 kg bedraagt de gewichtsfactor (6/25) = 0.24 , voor 66 kg is de gewichtsfactor (7/25) = 0.28 , voor 67 kg is de factor (5/25)= 0.20 en voor 68 kg is de gewichtsfactor gelijk aan (3/25) = 0.12.

Let op: de som van de gewichtsfactoren is altijd gelijk aan 1.
Dus ook hier: 0.16 + 0.24 + 0.28 + 0.20 + 0.12 = 1

De gemiddelde massa is nu te berekenen door elke massa met zijn bijbehorende gewichtsfactor te vermenigvuldigen en daarna op te tellen:
gemiddelde massa = 0.16 * 63 + 0.24 * 64 + 0.28 * 66 + 0.20 * 67 + 0.12 * 68 = 65.5 kg

Het is ook mogelijk het gemiddelde niet met gewichtsfactoren, maar rekenkundig uit te rekenen:

De gemiddelde massa is: (4 * 63 + 6 * 64 + 7 * 66 + 5 * 67 + 3 * 68) / 25 = 65.5 kg

Hoe betrouwbaar is het gemiddelde?
Men kan zich afvragen of het gemiddelde van een groep getallen representatief is voor die groep getallen. Anders geformuleerd: Als men een willekeurig getal pakt uit die verzameling getallen, zal het dan in de meeste gevallen ergens rond het gemiddelde liggen?
We laten met behulp van een voorbeeld zien dat dit niet zo hoeft te zijn en dat we niet altijd af kunnen gaan op het gemiddelde.

Voorbeeld
De gemiddelde leeftijd van een groep van 6 personen is 18 jaar.
Wat kun je zeggen over de leeftijd van de mensen in deze groep?
Er blijken 2 groepen te zijn met hetzelfde gemiddelde, namelijk 18 jaar:

- **groep 1:** 20; 22; 18; 16; 15; 17 gem. = 18 jr
- **groep 2:** 3 ; 5; 7; 3; 89; 1 gem. = 18 jr

In groep 1 liggen alle leeftijden redelijk mooi verdeeld rond de 18 jaar, het gemiddelde: De leeftijden liggen rond de 18 ± 4 jaar.
In groep 2 is dat zeker niet het geval. Groep 2 kan bestaan uit een oma (81 jaar) met haar kleinkinderen.

Uit dit voorbeeld zien we dat het gemiddelde gevoelig is voor uitschieters en op zichzelf niet representatief is voor een verzameling getallen.

Vandaar dat er ook andere instrumenten zijn om meer inzicht te krijgen in een verzameling getallen. Een aantal worden in de komende hoofdstukken behandeld.

Opgave S1.2.1.

Het gemiddelde maandinkomen van een groep van 23 ICT-ers is € 2840 per maand. Er blijkt nóg een persoon bij de groep te horen. Zijn maandinkomen bedraagt € 2400. Van de totale groep is het gemiddelde in-komen dan nu:

(a) O € 2829.44

(b) O € 2820

(c) O € 2821.67

Opgave S1.2.2.

Een statistiekdocent neemt drie toetsen af bij zijn studenten. Gezamenlijk leveren die toetsen een eindcijfer op, afgerond op één decimaal. De eerste toets heeft een gewicht van 40%, de tweede een gewicht van 30% en de laatste ook een gewicht van 30%. Een student behaalt op de 100-puntenschaal achter eenvolgens de scores 79, 70 en 82. Zijn eindscore afgerond op één decimaal wordt dus:

(a) 76.8

(b) 76.9

(c) 77.2

(d) 77.9

Opgave S1.2.3.

Een klein dorp heeft drie stembureaus : Centrum, Buitengebied Noord en Buitengebied Zuid. Het aantal stemgerechtigden per stembureau is respectievelijk 2000, 3000 en 3000. Bij een bepaalde verkiezing was het opkomstpercentage voor Centrum 30 %, voor Noord 30% en voor Zuid 50%. Het opkomstpercentage voor het gehele dorp zal zijn :

(a) O 36 %

(b) O 37.5 %

(c) O 39 %

(d) O 41.2 %

Ga nu in Grasple aan de slag met opgaven van dit onderdeel om na te gaan of je de stof goed hebt begrepen.

S1.3. De berekening van de mediaan

De mediaan wordt gebruikt om een indruk te geven van de ligging van het centrum van een meestal groot aantal getalswaarden van een of andere grootheid, bijvoorbeeld de mediane lengte van 110.000 dienstplichtige militairen. Van die militairen kan je dan de gemiddelde lengte bepalen, maar je kan ook de mediane lengte ofwel de mediaan bepalen.

De algemene definitie van de mediaan m_e is:
50% van alle getalswaarden is kleiner dan de mediaan en 50% van alle getalswaarden is groter dan de mediaan. Dus 50% van die 110000 militairen is groter dan de mediane lengte en de andere 50% is kleiner dan de mediane lengte. Er is natuurlijk ook een mediaan gewicht voor de dienstplichtigen en een mediaan intelligentiequotiënt, enz.
We bepalen de mediaan via 3 verschillende methoden.

Methode 1
Heb je te maken met een betrekkelijk klein *even* aantal getalswaarden , bijvoorbeeld de 10 waarden 0, 1, 6, 3, 2, 2, 1, 0, 1, en 4, dan bepaal je de mediaan door allereerst de getalswaarden in *opklimmende* grootte te rangschikken:
0 0 1 1 1 2 2 3 4 6

En dan deel je deze 10 waarden in twee gelijke blokken van 5 getallen, namelijk 0,0,1,1,1 en 2,2,3,4,6:
0 0 1 1 **1** **2** 2 3 4 6

De mediaan is nu per definitie het *gemiddelde* van de middelste 2 getallen, d.w.z. het laatste getal 1 van het eerste blok en het eerste getal 2 van het tweede blok , dus
$m_e = (1 + 2)/2 = 1.5$

Voorbeeld 1
Tien personen hebben een test afgelegd. Het aantal fouten dat ieder van hen maakte is geteld. Het resultaat is 6, 0, 2, 1, 0, 1, 4, 3, 2,1 fouten. Indien de eerste persoon niet 6 maar 17 fouten zou hebben gemaakt dan zou de mediaan:

(a) O precies 1.5 zijn
(b) O iets groter zijn dan 1.5
(c) O gelijk aan 2 worden

Oplossing:
We hebben hierboven berekend dat de mediaan van de getallen 0,0,1,1,1,2,2,3,4 en 6 gelijk is aan 1.5.

Verandert nu het getal 6 in 50, dan worden de getallen in opklimmende grootte:

0 0 1 1 1 2 2 3 4 50

En dan deel je weer deze 10 waarden in twee gelijke blokken van 5 getallen, namelijk 0,0,1,1,1 en 2,2,3,4,50 De mediaan is weer per definitie het gemiddelde van de middelste 2 getallen, d.w.z. het (vijfde) getal 1 van het eerste blok en het (zesde) getal 2 van het tweede blok , dus $m_e = (1 + 2)/2 = 1.5$. De mediaan blijft dus ongewijzigd.
We zien We zien dus dat de mediaan minder gevoelig is voor uitschieters dan het gemiddelde.

Methode 2
Je hebt nu te maken met een betrekkelijk klein *oneven* aantal getalswaarden, zoals de onderstaande 11 getallen:
0 0 1 1 1 2 2 3 4 6 13

Je deelt deze 11 waarden weer in twee gelijke blokken van 5 getallen, èn het *centrale* getal (het 6^{de} getal), namelijk 2:
0 0 1 1 1 **2** 2 3 4 6 13

De mediaan is nu per definitie het *centrale* getal (het 6^{de} getal) , dus $m_e = 2$.

Methode 3
Er bestaat een formule waarmee je behalve de mediaan, die de getallenverzameling in twee delen van 50% verdeelt, ook andere statistische grootheden snel kunt berekenen zoals *kwartielen* (delen de verzameling getallen in 4 gelijke delen van 25%), *docielen* (delen de getallenverzameling in 10 gelijke delen van 10%) en *percentielen* (delen de getallenverzameling in 100 gelijke delen van 1 %). Deze formule luidt als volgt:

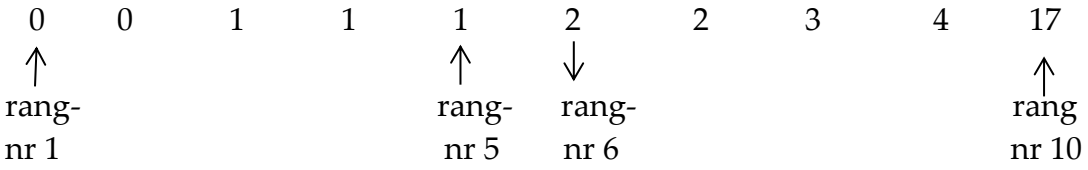
$$Rp = (n + 1) \frac{p}{100}$$

- R_p = het <u>*rangnummer*</u> van het getal dat de mediaan, dociel, kwartiel, of percentiel voorstelt.
- p = percentage, waarbij je kijkt naar de p% kleinste getallen
- n = aantal getallen.

Wat bereken je nu met deze formule?

Hebben we dus de getallen gerangschikt in een rij naar opklimmende grote. Dan berekent de formule het rangnummer in de rij, waar je kwartielen, docielen, percentielen etc. vindt.

Voorbeeld 2
Kijk terug naar voorbeeld 1. Het aantal gemaakte fouten was, naar opklimmende grootte gerangschikt, gelijk aan:



We willen de mediaan berekenen. Dan is R_P dus het rangnummer van de mediaan en p is dan 50, want 50% van de waarden is kleiner dan de mediaan. Verder is n = 10. Dan wordt $R_p = R_{50} = (10 + 1)\ (50/100) = 11 * (1/2) = 5.5$. *Dus het rangnummer van de mediaan is 5,5.* Het rangnummer van het vijfde getal is 5 en van het 6de getal is het rangnummer 6. De mediaan zit er met rangnummer 5.5 precies tussen en is dus gelijk aan 1.5.

Opgave S1.3.

a) Gegeven zijn een aantal data: 55,71,72,62,63,67,64,87,85,49,61. Bereken de mediaan.
b) Gegeven zijn een aantal data: 31,101,47,13,14,10,117,19,97,23,72,24,27,29. Bereken, afgerond op één decimaal de mediaan.

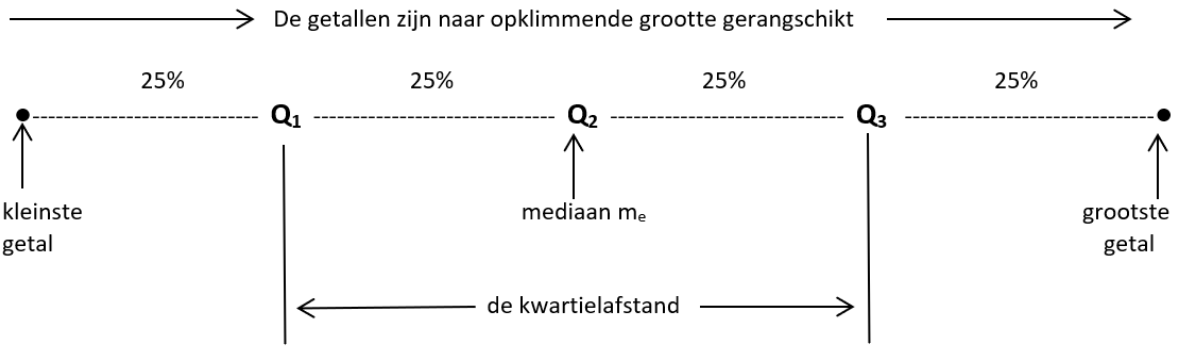
Ga nu in Grasple aan de slag met opgaven van dit onderdeel om na te gaan of je de stof goed hebt begrepen.

S1.4. De berekening van kwartielen en de halve kwartielafstand

Introductie
Kwartielen en ook de halve kwartielafstand dienen om snel een globaal overzicht te verkrijgen van een verzameling getallen. Ze worden o.a. gebruikt bij de constructie van een *boxplot* ofwel een globaal grafisch overzicht van de ligging van de waarden van een verzameling getallen.

Zoals het woord al suggereert delen *kwartielen* een getallenverzameling in *vier gelijke delen* van 25% . Om het goed te begrijpen moet je je voorstellen dat alle getallen van een getallenverzameling *naar opklimmende grootte* gerangschikt zijn.

- Kijken we dan eerst naar het eerste kwartiel Q_1 . Dit eerste kwartiel is een grenswaarde. De 25% kleinste getallen zijn kleiner dan Q_1. De overige 75 % van de getallenverzameling moet dan groter zijn dan het eerste kwartiel Q_1 .
- Voor het tweede kwartiel Q_2 geldt dat de 50% kleinste waarden kleiner dan Q_2 zijn en de overige 50% van de getallen is groter dan het tweede kwartiel Q_2. Maar dat betekent dat het tweede kwartiel Q_2 niets anders is dan de mediaan m_e . Dus de begrippen tweede kwartiel, Q_2 en mediaan m_e vallen samen.
- Dan nog het derde kwartiel Q_3. Daarvoor geldt dat 75% van de getallenverzameling kleiner is dan Q_3 en 25% is dan groter dan Q_3. Om het voorgaande nogmaals te verduidelijken is het onderstaande figuur gegeven:



Om inzicht te krijgen in een grote getallenverzameling, is men altijd geinteresseerd naar de getallen rond het midden (in dit geval de mediaan) en de spreiding daarom heen. Oftewel, men kijkt naar de 50% middelste waarnemingsgetallen en de spreiding van deze middelste getallen.
Vandaar het begrip **kwartielafstand**: de afstand tussen het eerste kwartiel Q_1 en het derde kwartiel Q_3:
Kwartielafstand = Q_3 - Q_1

Maar meestal wordt gewerkt met het begrip **halve kwartielafstand :**
Halve kwartielafstand = $(Q_3 - Q_1) / 2$

Aan de hand van een paar voorbeelden wordt nu getoond hoe men deze grootheden in de praktijk kan berekenen.

Voorbeeld 1
Gegeven zijn een aantal data: 66, 80, 48, 70, 57, 67, 60, 63, 55, 56, 75. Bereken dan:
(a) de mediaan
(b) het eerste kwartiel
(c) het derde kwartiel
(d) de halve kwartielafstand

Oplossing:
Allereerst moeten de getallen *naar opklimmende grootte* gerangschikt worden. We krijgen dan:
48 – 55 – 56 – 57 – 60 – 63 – 66 – 67 – 70 – 75 - 80

Zetten we de *rangnummers* van de getallen tussen haakjes achter de getallen dan krijgen we de volgende reeks:

48(1) – 55(2) – 56(3) – 57(4) – 60(5) – 63(6) – 66(7) – 67(8) – 70(9) – 75(10) – 80(11)

We berekenen de mediaan op twee manieren :
Methode 1
Kijk naar het *aantal* getallen. Dat is 11, dus een *oneven* aantal. Dan is de centrale waarde het 6^{de} getal. Want 5 getallen zijn kleiner dan het 6^{e} getal en 5 getallen zijn groter dan het 6^{de} getal. Dus 50% onder en 50% boven het 6^{de} getal. Het 6^{de} getal is het getal 63, dus de mediaan $m_e = 63$.

Methode 2
Gebruik de formule voor het *rangnummer* van een getal: $R_p = (n + 1) \frac{p}{100}$

De mediaan, het tweede kwartiel Q_2
In deze formule stelt p het percentage getallen voor onder R_p. Is nu R_p het rangnummer van de mediaan, dus p is gelijk aan 50. Verder is n het aantal getalswaarden en dat is 11. Substitutie van n = 11 en p = 50 geeft het rangnummer van de mediaan:
$R_{50} = (11 + 1)(50/100) = 12 * 0.5 = 6$
Dus we zoeken het getal met rangnummer 6 en dat is weer 63.

Eerste kwartielafstand Q_1
Vervolgens de bepaling van het eerste kwartiel Q_1. We gebruiken weer bovenstaande formule en substitueren wederom n = 11. Maar p is nu 25, want onder het eerste kwartiel zit 25% van de getallen.

We vinden dan : $R_{25} = (11 + 1)(25/100) = 12 * (1/4) = 3$
We zoeken het getal met rangnummer 3 en dat is 56. Dus $Q_1 = 56$.

Derde kwartiel Q_3
De bepaling van het derde kwartiel Q_3 gaat weer via de formule met n = 11 en nu p = 75. We vinden $L_{75} = 9$, dat wil zeggen het derde kwartiel Q_3 is het 9^{de} getal. Dus $Q_3 = 70$.

Kwartielafstand
De kwartielafstand is gelijk aan $Q_3 – Q_1 = 70 – 56 = 14$. Dan is de *halve kwartielafstand* gelijk aan 14/2 = 7.

Kwartielen hoeven niet altijd gehele getallen te zijn. Bij getallen die niet geheel zijn wordt de berekening iets gecompliceerder. Zie het volgende voorbeeld:

Voorbeeld 2
Gegeven zijn een aantal data: 66, 80, 48, 70, 57, 67, 60, 63, 55, 56, 75, 61, 68, 81. Bereken:
(a) de mediaan
(b) het eerste kwartiel
(c) het derde kwartiel
(d) de halve kwartielafstand

Oplossing:
Allereerst moeten de getallen naar opklimmende grootte gerangschikt worden. We krijgen dan :
48 – 55 – 56 – 57 – 60 – 61 - 63 – 66 – 67 – 68 - 70 – 75 - 80 – 81

Zetten we de rangnummers van de getallen tussen haakjes achter de getallen dan ontstaat de volgende reeks:

48(1)–55(2)–56(3)–57(4)–60(5)–61(6)–63(7)–66(8)–67(9)–68(10)–70(11)–75(12)-80(13)-81(14)

We berekenen weer de rangnummers van Q_1 , Q_2 en Q_3 .

- Voor Q_1: $R_{25} = (14+1)\ (25/100) = 3.75$
- Voor Q_2: $R_{50} = (14+1)(50/100) = 7.50$

- Voor Q_3 : R_{75} = (14+1) (75/100) = 11.25

Tweede kwartiel, de mediaan

Het tweede kwartiel Q_2, dus de mediaan, is nu het gemakkelijkst te bepalen. Het getal 63 heeft rangnummer 7 en het getal 66 heeft rangnummer 8. De mediaan ligt er met rangnummer 7.5 precies halverwege tussen, dus de mediaan wordt 63 + (66-63)/2 = 64.5. Je kan ook 63 en 66 optellen en daarna door 2 delen, dan vind je ook 63.5.

Eerste en derde kwartiel

Bij het eerste kwartiel en het derde kwartiel ligt het iets moeilijker. Het eerste kwartiel heeft rangnummer 3,75 , dus ligt qua rangnummer dichter bij het getal 57 (rangnummer 4) dan bij 56 (rangnummer 3). Gebruikelijk is nu om de afstand tussen 56 en 57 te bepalen, daarvan ¾ te nemen en het resultaat bij 56 op te tellen. Dus:

- $Q_1 = 56 + (3/4) * (57 – 56) = 56.75$. Op één decimaal afgerond: $Q_1 = 56.8$
- $Q_3 = 70 + (1/4)(75 – 70) = 71.25$. Op één decimaal afgerond: $Q_3 = 71.3$

Halve kwartielafstand

Tot slot de halve kwartielafstand: $0.5 * (Q_3 – Q_1) = 0.5 * (71.25 – 56.75) = 0.5 * 14.5 = 7.3$

Opgave S1.4.1.

Gegeven zijn een aantal data: 55,71,72,62,63,67,64,87,85,49,en 61. Bereken:
(a) het eerste kwartiel
(b) het derde kwartiel
(c) de halve kwartielafstand

Opgave S1.4.2.

Gegeven zijn een aantal data: 31,101,47,13,14,10,117,19,97,23,72,24,en 27. Bereken, afgerond op één decimaal:
(a) het eerste kwartiel
(b) het derde kwartiel
(c) de halve kwartielafstand

Ga nu in Grasple aan de slag met opgaven van dit onderdeel om na te gaan of je de stof goed hebt begrepen.

S1.5. De constructie van een boxplot, behorend bij een getallenverzameling

Zoals eerder vermeld, is men bij een getallenverzameling(bijvoorbeeld waarnemingsgetallen bij een experiment of steekproef) geïnteresseerd in de getallen rond de mediaan en de spreiding daarom heen. Om meer inzicht te krijgen, kijkt men dan naar de 50% middelste waarnemingsgetallen en hun spreiding.

Een boxplot, ook wel *snorrendoos* of *doosdiagram* genoemd, is een bepaalde grafische weergave van een getallenverzameling door middel de volgende vijf getallen :

(1) het kleinste getal oftewel het minimum.
(2) het eerste kwartiel Q_1.
(3) het tweede kwartiel Q_2 , ofwel de mediaan m_e.
(4) het derde kwartiel Q_3.
(5) het grootste getal oftewel het maximum

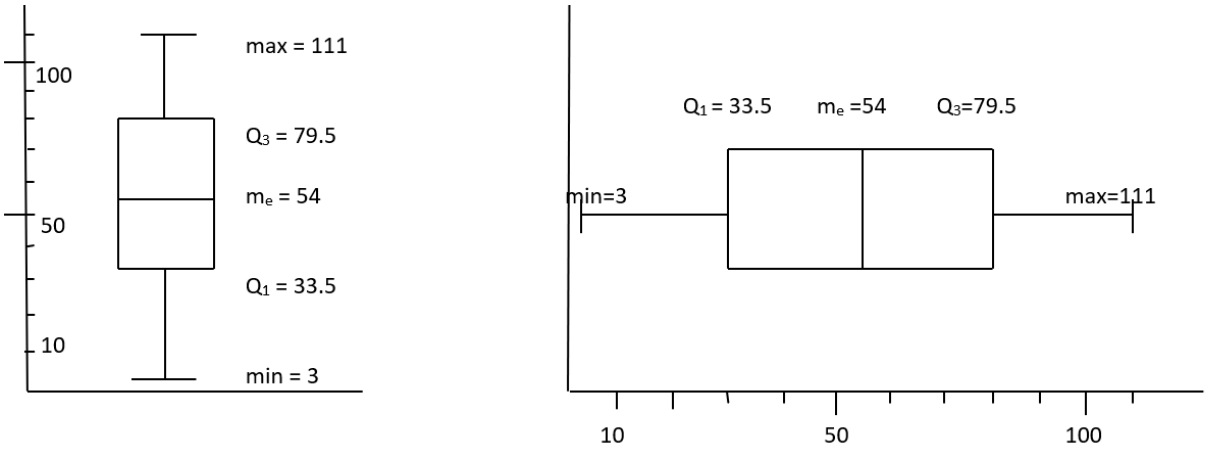
Voorbeeld 1
We maken nu een boxplot van de volgende verzameling van 37 getallen: 66, 98, 34, 3, 8, 89, 67, 65, 44, 33, 23 , 12 , 15, 16, 88, 79, 71,17, 37, 41,43,35,110,101,87, 98, 47, 46, 55, 56, 57, 54, 53, 80, 73, 24, en 111.

Oplossing:
Voor de constructie van het boxplot m.b.v. het minimum, het maximum, en de drie kwartielen Q_1 , Q_2 en Q_3 is het natuurlijk noodzakelijk om allereerst deze 37 getallen naar opklimmende grootte te rangschikken. Want dan kunnen deze 37 getallen van een rangnummer voorzien worden. De 37 getallen zijn nu in opklimmende grootte gerangschikt met tussen haakjes achter elk getal het bijbehorende rangnummer van dat getal:
3(1), 8(2), 12(3), 15(4), 16(5), 17(6), 23(7), 24(8), 33(9), 34(10), 35(11), 37(12), 41(13), 43(14), 44(15), 46(16), 47(17), 53(18), 54(19), 55(20), 56(21), 57(22), 65(23), 66(24), 67(25), 71(26), 73(27), 79(28), 80(29), 87(30), 88(31), 89(32), 98(33), 98(34), 101(35), 110(36), 111(37).

De kleinste waarde, het minimum, is dus 3. De grootste waarde, het maximum, is 111. Vervolgens is de mediaan ook vrij snel te berekenen want we kunnen deze 37 getallen verdelen in een groep van de kleinste 18 getallen, een groep met de grootste 18 getallen

en daar dan tussenin het getal met rangnummer 19, welke dan exact in het midden staat. Dus het getal met rangnummer 19 is de mediaan, d.w.z. $m_e = 54$.
Het eerste kwartiel vinden we als volgt: het rangnummer van het eerste kwartiel Q_1 vinden we door gebruik te maken van de formule voor R_p, zoals hiervoor behandeld, met n = 37 en p= 25. Het rangnummer van Q_1 wordt: $R_{25} = (37+1)(25/100) = 9.5$.
Het getal 33 heeft rangnummer 9 en het getal 34 heeft rangnummer 10. Het eerste kwartiel zit daar qua rangnummer precies halverwege tussen, dus $Q_1 = (33 + 34)/2 = 33{,}5$. Het rangnummer van het derde kwartiel Q_3 vinden we weer door gebruik te maken van de formule voor R_p met n=37 en p=75. Dan is het rangnummer van Q_1 gelijk aan $R_{75} = (37+1)(75/100) = 28.5$. Het getal 79 heeft rangnummer 28 en het getal 80 heeft rangnummer 29. Het derde kwartiel zit weer qua rangnummer precies ertussen, dus $Q_3 = (79 + 80)/2 = 79.5$.

Nu kunnen we het boxplot construeren. Er zijn twee versies: een *horizontale* versie en een *verticale* versie:



Misschien is de verticale boxplot wel het meest overzichtelijk, maar dat blijft een kwestie van persoonlijke smaak. Alle getalswaarden liggen tussen 3 en 111, oftewel in het interval [3, 111]. De kleinste 25% van de getallen liggen in het interval [3, 33.5] en de grootste 25% van de waarden ligt in het interval [79.5, 111]. De centrale 50% van alle waarden ligt in het interval [33.3 , 79.5].
Het is een interval met een breedte van 79.5 – 33.5 = 46. Deze waarde 46 wordt *de kwartielafstand* genoemd. Neem je hiervan de helft, dus 23, dan vind je de *halve kwartielafstand*.
Indien je neemt: $m_e \pm$ halve kwartielafstand, dan heb je meestal een redelijk goede indruk van de ligging van de centrale 50% van alle waarden.

Opgave S1.5.

Construeer het boxplot behorend bij de getallenverzameling : 23, 24, 24, 26, 27, 28, 29, 33, 34, 35, 37, 39, 40, 42, 44, 46, 47, 48, 49, 49, 49, 50, 51, 52, 53, 55, 56, 58, 61, 61, 62, 62, 63, 64, 67, 69.

Ga nu in Grasple aan de slag met opgaven van dit onderdeel om na te gaan of je de stof goed hebt begrepen.

S1.6. De standaardafwijking

Het gemiddelde van een aantal getallen geeft informatie over die getallen.
Eerder in S1.2. is aangetoond dat het gemiddelde ook gevoelig is voor uitschieters.
We laten wederom met een voorbeeld zien dat het gemiddelde niet altijd een nauwkeurig beeld geeft.

Voorbeeld:
Een eigenaar van twee autowinkels wil erachter komen of de managers van zijn winkels goed functioneren en altijd genoeg voorraad van de auto's in de winkels hebben. De eigenaar heeft besloten om de voorraad auto's van de laatste 6 weken te inspecteren.

De volgende resultaten werden bekend:

	week 1	week 2	week 3	week 4	week 5	week 6
Winkel 1	33	31	32	36	31	31
Winkel 2	22	34	58	52	10	21

Het berekenen van de gemiddelde van aantal auto's op voorraad binnen 6 weken voor de beide winkels heeft het volgende geleverd:

Winkel 1: $\frac{33+31+32+36+31+31}{6} = 32.3$

Winkel 2: $\frac{22+34+58+52+10+21}{6} = 32.8$

Het lijkt erop dat deze twee gemiddelde waardes van de beide winkels bijna gelijk zijn. De eigenaar zou een conclusie kunnen trekken dat beide winkels op gelijke wijze functioneren.

Maar bij verdere analyse naar de verspreiding van de getallen per week per winkel (zie boven) valt het meteen op dat winkel 2 grotere variatie in de spreiding van aantal auto's op voorraad heeft: tussen 10 en 58, terwijl het gemiddelde bijna hetzelfde is als van winkel 1.

In dit soort gevallen geeft het gemiddelde niet genoeg informatie over de data. Maar het berekenen van een standaardafwijking zal veel meer informatie bieden.

De **standaardafwijking** laat zien hoe groot de spreiding van de data is t.o.v. het gemiddelde oftewel hoever de data gemiddeld afwijkt van het gemiddelde (in dit voorbeeld: de wekelijkse voorraad).

De formule voor de standaardafwijking ziet er zo uit:

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

Met n = aantal elementen, $\bar{x}$ = gemiddelde

Deze formule wordt gebruikt bij het berekenen van een steekproef, net als in ons voorbeeld (slechts 6 weken geanalyseerd). In het geval dat het om de hele populatie gaat (dus als er alle weken geanalyseerd zouden worden vanaf de opening van de winkels), wordt onderstaande formule gebruikt:

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

Oplossing:

Dus, de standaardafwijking van winkel 1 is:

$$s = \sqrt{\frac{(33-32.3)^2 + (31-32.3)^2 + (32-32.3)^2 + (36-32.3)^2 + (31-32.3)^2 + (31-32.3)^2}{6-1}} \approx \mathbf{2.0}$$

Dan de standaardafwijking van winkel 2 is:

$$s = \sqrt{\frac{(22-32.8)^2 + (34-32.8)^2 + (58-32.8)^2 + (52-32.8)^2 + (10-32.8)^2 + (21-32.8)^2}{6-1}} \approx \mathbf{18.9}$$

Wat laten deze cijfers zien? De standaardafwijking van winkel 1 is ongeveer 2.0, dus niet heel ver van 0 en geeft aan dat de meeste waarden in de buurt van de gemiddelde waarde liggen. Hoe dichter de standaardafwijking bij 0 ligt, hoe betrouwbaarder het

gemiddelde is. Bovendien duidt de standaardafwijking dicht bij 0 op een kleine variatie van gegevens. Dat wil zeggen dat een voorraad met een standaardafwijking van 2.0 betekent dat er een stabiliteit in het verkopen van de auto's en het aanvullen van de voorraden zit.

In het geval van de tweede winkel was de standaardafwijking 18.9. Dat wil zeggen dat de voorraad per week, gemiddeld 18.9 van de gemiddelde waarde afwijkt.

Dit voorbeeld geeft dus aan dat we niet altijd alleen af kunnen gaan op het gemiddelde om een beeld te vormen over de spreiding van getallen. De standaarafwijking is hierbij dan een goed hulpmiddel.

Opgave S1.6.1.

Gegeven de gewichten van 5 vogels: 7, 17, 11, 4, 11.
Bereken de standaardafwijking.

Opgave S1.6.2.

Gegeven de leeftijden van een aantal studenten: 18, 20, 20, 21, 25, 29
Bereken de standaardafwijking.

Ga nu in Grasple aan de slag met opgaven van dit onderdeel om na te gaan of je de stof goed hebt begrepen.

Antwoorden van de opgaven

Opgave S1.1.

59 van de 98 vrouwen stemmen op de PVV. Dus 59/98 * 100% = 60.2%. Dus antwoord (c) is juist

Opgave S1.2.1.

Het gemiddelde maandinkomen van een groep van 23 ICT-ers is € 2840 per maand. In totaal verdienen deze 23 ICT-ers dus: 23 * 2480 = € 65.320.-. Inclusief het inkomen van de persoon die hier ook bij hoort geldt een totaal van: 65320 + 2400 = € 67720.-. Het gemiddelde is dan: 67720 / 24 = € 2821.67. Dus antwoord (c) is juist.

Opgave S1.2.2.

Eindscore = 0.4 * 79 + 0.3 * 70 + 0.3 * 82 = 77.2.
Antwoord (c) is juist.

Opgave S1.2.3.

Totale opkomst gehele dorp = 0.3 * 2000 + 0.3 * 3000 + 0.5 * 3000 = 3000
Totale stemgerichtigden = 2000 + 3000 + 3000 = 8000.
Opkomstpercentage gehele dorp = 3000 / 8000 * 100% = 37.5%.
Antwoord (b) is juist.

Opgave S1.3.

a) $m_e = 64$
b) $m_e = 28$

Opgave S1.4.1.

$Q_1 = 61$
$Q_3 = 72$
De halve kwartielafstand =0.5*(72-61) = 5.5

Opgave S1.4.2.

$Q_1 = 16.5$
$Q_3 = 84.5$
De halve kwartielafstand is gelijk aan 0.5 * (84.5 – 16.5) = 34

Opgave S1.5.

min=23
max= 69
rangnummer Q_1 is 9,25 , dus $Q_1 = 34 + 0.25*(35-34) = 34.25 = 34.3$.
En het rangnummer van de mediaan m_e is 18.5 dus $m_e = 48 + 0.5*(49-48) = 48.5$.
Het rangnummer van Q_3 is 27.75 , dus $Q_3 = 56 + 0.75*(58-56) = 57.5$.
Met deze gegevens is het boxplot te construeren.

Opgave S1.6.1.

s = 4.98

Opgave S1.6.2.

s = 3.72