Dataset, gemiddelde en spreiding rond het gemiddelde

Prof. dr. Herman Callaert

Statistiek voor de eerste graad

In de eerste graad komen leerlingen in contact met statistiek. Statistiek is geen onderdeel van wiskunde, het is een apart vak zoals economie of aardrijkskunde. In het secundair onderwijs komt statistiek vooral aan bod binnen de lesuren wiskunde. Onderwijsdoelen zien er daar als volgt uit:

A/B-stroom > Wiskunde – natuurwetenschappen – technologie – STEM > Onderwijsdoel 6.16 / 6.8
"De leerlingen voeren een beschrijvend statistisch onderzoek uit met 20 à *25 zelf verzamelde, niet gegroepeerde gegevens van 1 grootheid"*

In statistiek probeer je vanuit data de wereld beter te begrijpen. Dat doe je met een statistisch onderzoek waarbij je 4 grote stappen doorloopt.

1. Je formuleert een onderzoeksvraag.
2. Je verzamelt data die je moeten helpen om de onderzoeksvraag te beantwoorden.
3. Je gebruikt statistische methoden en technieken om informatie in je dataset te ontdekken.
4. Je formuleert een antwoord op de onderzoeksvraag.

Deze tekst focust op de derde stap in zo'n onderzoek. Daar gebruik je methoden en technieken om op exploratie te gaan in een dataset. Je krijgt er te maken met kengetallen voor centrum en spreiding.

Kengetallen zijn grootheden die je met een gepaste formule kan berekenen, maar dat is lang niet alles. Je kan ze ook bestuderen samen met een grafische voorstelling van de dataset. Zo krijg je extra inzicht in wat kengetallen (niet) kunnen vertellen over de dataset.

Om de aandacht toe te spitsen op deze manier van werken, houden we het super eenvoudig. We werken met een beperkt aantal opmetingen (het zijn er 7) en het zijn allemaal gehele getallen.

Inhoud

Een dataset en het gemiddelde

Eén van de meest eenvoudige grootheden die je ontmoet bij de exploratie van een dataset is het gemiddelde. Dat is een kengetal voor het centrum van een dataset.

In deze tekst bekijk je het gemiddelde in verschillende situaties. Je houdt daarbij ook de dataset heel goed in het oog. Zo ontdek je dat het gemiddelde soms "typisch" kan zijn voor de data die je hebt opgemeten, maar soms zet het gemiddelde je helemaal op het verkeerde been.

Nota.
Met het woord "gemiddelde" van een dataset bedoelt men in statistiek altijd "rekenkundig gemiddelde".

1. Een kengetal voor centrum: het gemiddelde

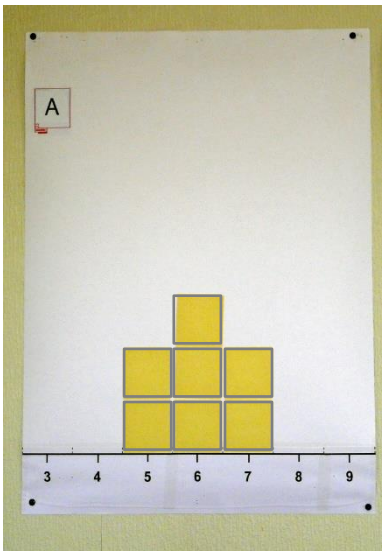
Het gemiddelde van een aantal getallen is een getal dat je als volgt bekomt: tel alle getallen samen en deel door het aantal getallen. Zo is het gemiddelde van de 7 getallen: 3 9 4 5 7 6 8 gelijk aan 6 want (3+9+4+5+7+6+8) / 7 = 6. Om het gemiddelde te berekenen heb je alleen maar zeer eenvoudige bewerkingen nodig die je al kent uit het basisonderwijs: de optelling en de deling. Natuurlijk kan je ook je rekentoestel gebruiken waar je de toets met het symbool $\bar{x}$ indrukt. Dat symbool $\bar{x}$ is de notatie voor het gemiddelde. Dat het gemiddelde van de vorige 7 getallen gelijk is aan 6 schrijf je als $\bar{x} = 6$.

Puur wiskundig bekeken is het gemiddelde gewoon "een synoniem" voor de som gedeeld door het aantal.

In statistiek wil je meer dan zomaar een wiskundige formule. Je hebt data verzameld om een onderzoeksvraag te beantwoorden. Van die data kan je het gemiddelde berekenen. Maar helpt dat gemiddelde om de dataset beter te begrijpen? Wat leer je uit het samenspel tussen een gemiddelde en een dataset? Dat ga je nu ontdekken.

2. Dataset: een aanschouwelijke voorstelling

Een dataset kan je op verschillende manieren voorstellen. Soms gebruik je een dotplot of een staafdiagram, maar soms kan je het ook heel aanschouwelijk maken met Post-it blaadjes. Op een groot blad met onderaan een getallenas kleef je boven de aangegeven getallen een Post-it blaadje voor elk getal in je dataset. Dat zou er kunnen uitzien zoals grafiek A hiernaast.



Op de grafiek zie je een getallenas met de getallen 3, 4, 5, 6, 7, 8, en 9. Boven die getallen zie je Post-it blaadjes. Er zijn geen blaadjes boven 3, 4, 8, en 9. Dat wil zeggen dat die getallen niet voorkomen in de dataset. Boven het getal 5 zijn er 2 Post-it blaadjes: het getal 5 komt twee keer voor. Verder zie je dat het getal 6 drie keer voorkomt (drie Post-it blaadjes boven 6) en het getal 7 twee keer.
Alles samen zegt de grafiek dat je dataset bestaat uit de volgende zeven getallen: 5, 5, 6, 6, 6, 7, 7.

Zodra je de volledige dataset kent, kan je het gemiddelde zoeken.

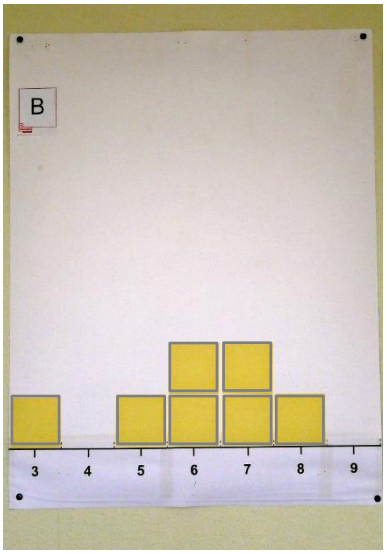
Opdracht 1

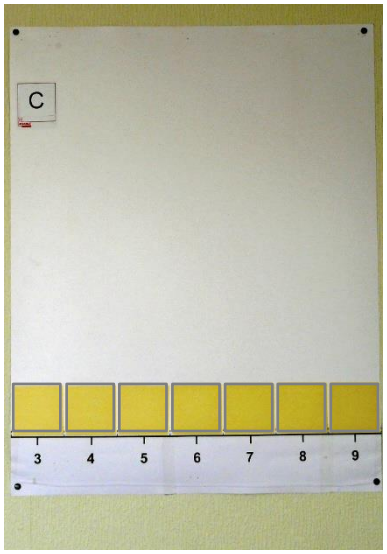
Zoek het gemiddelde van de dataset die door grafiek A wordt voorgesteld.

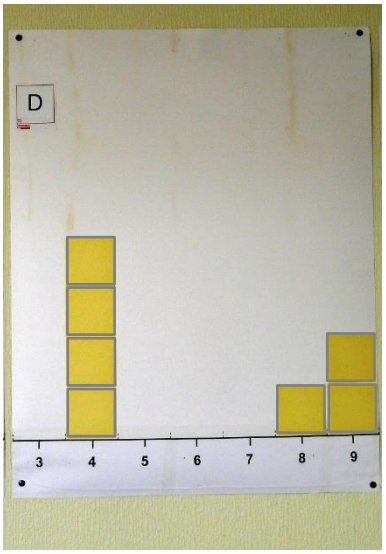
Antwoord: Grafiek A toont een dataset met gemiddelde $\bar{x} = \ldots\ldots\ldots$.

Opdracht 2

Hieronder zie je 3 grafieken die elk een dataset voorstellen. Gebruik onderstaande tabel om de volledige dataset op te schrijven voor grafiek B, grafiek C en grafiek D. Zoek ook het gemiddelde van elke dataset.







	Dataset voor deze grafiek	Het gemiddelde
Grafiek B		$\bar{x} = \ldots\ldots\ldots$
Grafiek C		$\bar{x} = \ldots\ldots\ldots$
Grafiek D		$\bar{x} = \ldots\ldots\ldots$

3. Het gemiddelde als evenwichtspunt

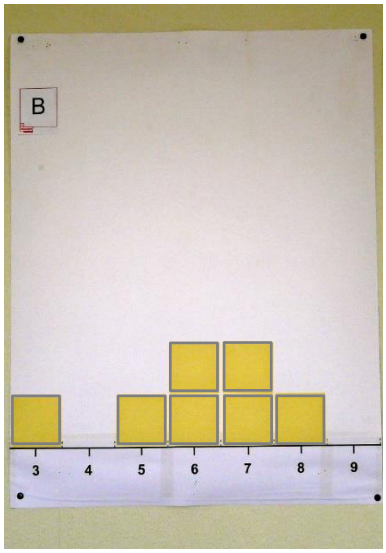
3.1. Stappen tellen

Onderstel dat je in grafiek B op de getallenas gaat staan, op de plaats van het gemiddelde. Je staat daar dan op het getal 6 (want $\bar{x} = 6$). Nu ga je op stap en je neemt telkens een stap van lengte één.

Hoe geraak je van het gemiddelde $\bar{x}$ naar de getallen in je dataset?

Het getal 7 bereik je als je vanaf het gemiddelde $\bar{x} = 6$ één stap naar rechts zet. Om bij 3 te geraken moet je vanaf het gemiddelde drie stappen naar links zetten. Of je nu naar rechts of naar links stapt, je kijkt alleen maar naar "de afgelegde afstand" = aantal stappen. Je kijkt dus niet naar "groter" (zoals +1 want 6+1=7) of kleiner (zoals –3 want 6–3=3).

Hoe je vanuit het gemiddelde op stap gaat naar alle getallen van je dataset kan je samenvatten in de volgende tabel (kijk ondertussen ook naar grafiek B).



Dataset B			
data	**stappen vanuit $\overline{x}$**	**aantal stappen**	
3	drie naar links	3	totaal aantal naar links = 4
5	één naar links	1	
6	nul stappen	0	
6	nul stappen	0	
7	één naar rechts	1	totaal aantal naar rechts = 4
7	één naar rechts	1	
8	twee naar rechts	2	

Opdracht 3

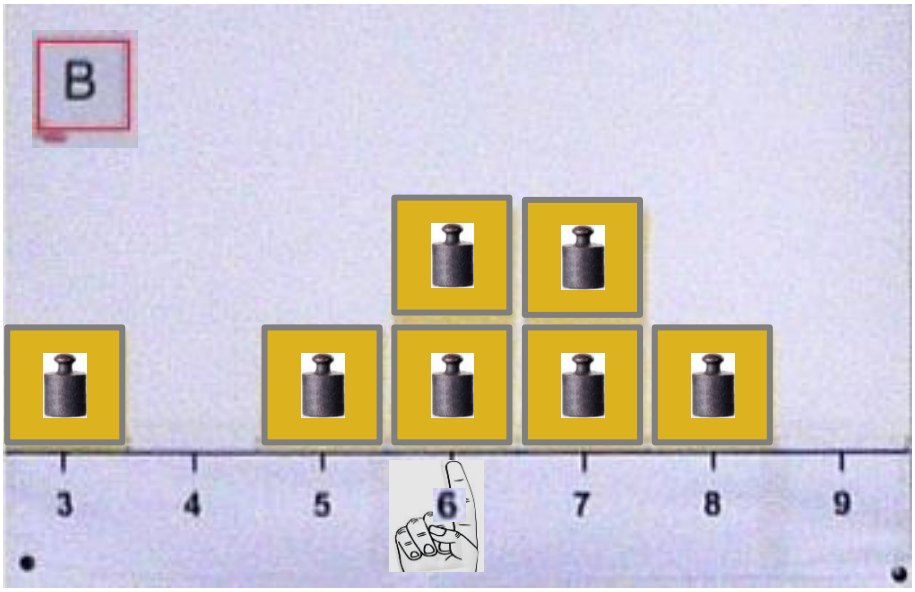
Voor grafiek D heb je de volledige dataset al opgeschreven en je hebt daar ook het gemiddelde berekend. Je kan dus ook hier een tabel maken die zegt hoe je vanuit het gemiddelde moet stappen naar alle getallen in die dataset. Vervolledig onderstaande tabel.

Dataset D			
data	**stappen vanuit $\overline{x}$**	**aantal stappen**	
			totaal aantal naar links = …..
			totaal aantal naar rechts = …..

Bij grafiek B zie je dat het gemiddelde $\overline{x}$ ergens tussen de getallen van de dataset ligt. Maar $\overline{x}$ ligt niet gewoon "zomaar ergens ertussen". Als je vanuit het gemiddelde op stap gaat naar alle getallen van de dataset dan moet je evenveel stappen naar links zetten als naar rechts. Bij grafiek B waren dat 4 stappen naar links en 4 stappen naar rechts. Bij grafiek D heb je dezelfde eigenschap ontdekt: daar ga je 8 stappen naar links en 8 stappen naar rechts. Het gemiddelde ligt dus wel op een heel speciale plaats tussen al de getallen van een dataset. Die plaats kan je zelfs "voelen". Dat zie je hieronder.

3.2. Een “fysische” interpretatie

Neem de getallenas en onderstel dat die eruitziet als een staaf. Kijk nu naar grafiek B en denk dat elk Post-it blaadje eenzelfde gewichtje voorstelt dat op de getallenas staat. Er staat dan één gewichtje op de getallen 3, 5 en 8 en er staan telkens twee gewichtjes op de getallen 6 en 7.



Neem nu die getallenas waarop de gewichtjes staan, schuif de staaf over en weer op het puntje van je vinger en zoek de plaats waar ze mooi in evenwicht blijft liggen. Die plaats is het evenwichtspunt van grafiek B en het getal waar je vinger staat is het gemiddelde.

Wat je voor grafiek B hebt ontdekt, geldt voor alle grafieken. Het is een algemene eigenschap: je kan “voelen” waar het gemiddelde ligt. Het gemiddelde is het **evenwichtspunt** bij de grafische voorstelling van een dataset.

4. Samenspel tussen gemiddelde en dataset

4.1. Van dataset naar gemiddelde

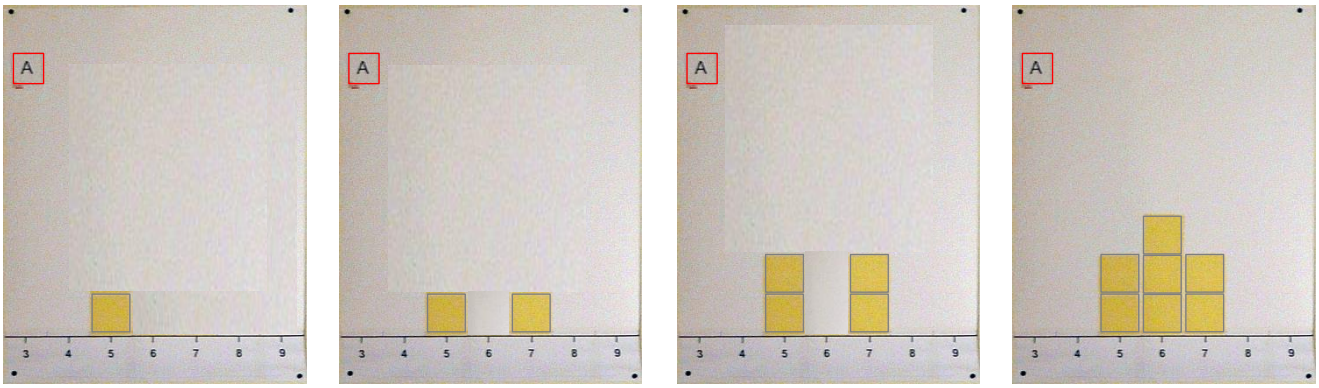
Als je de dataset volledig kent, dan ken je alle getallen en dan kan je ook tellen hoeveel getallen er zijn. Het gemiddelde ligt dan ook éénduidig vast: tel alle getallen samen en deel door het aantal. Zo is het gemiddelde van de 7 getallen: 3 9 4 5 7 6 8 gelijk aan (3+9+4+5+7+6+8) / 7 = 6. Dat wist je al, want dit is de dataset die hoort bij grafiek C.

4.2. Van gemiddelde naar dataset

Misschien is er je iets opgevallen bij de vorige grafieken. Die grafieken zien er allemaal anders uit maar zij hebben allemaal hetzelfde gemiddelde $\bar{x} = 6$.

Als je weet dat 7 getallen een gemiddelde hebben dat gelijk is aan 6, dan weet je nog helemaal niet wat die getallen zijn. Maar je kan natuurlijk wel proberen om zelf 7 getallen te zoeken waarvan het gemiddelde gelijk is aan 6. Je moet hierbij zelfs geen enkele berekening maken. Gebruik gewoon de eigenschap dat het gemiddelde het "evenwichtspunt" is. Ga grafisch te werk en zoek 7 getallen zodat je vanuit het gemiddelde $\overline{x} = 6$ evenveel stappen naar links als naar rechts moet zetten. Hieronder zie je in een voorbeeld hoe je dat zou kunnen doen.

Start met een eerste Post-it blaadje dat je bijvoorbeeld boven het getal 5 kleeft. Dat is vanuit $\overline{x} = 6$ één stap naar links. Dat compenseer je met één stap naar rechts en dus kleef je een blaadje boven het getal 7. Op dit ogenblik is de getallenas in evenwicht als je je vinger onder het getal 6 houdt.



Je zou nu terug het getal 5 kunnen kiezen en dan terug het getal 7. Zo blijft de getallenas in evenwicht. Je hebt nu nog 3 Post-it blaadjes over. Je kan daar verschillende dingen mee doen (zoals één blaadje op 3, één op 6 en één op 9), maar je kan die 3 blaadjes ook boven het getal 6 kleven. Ook op die manier blijft de getallenas in evenwicht. Alles samen heb je (zonder berekeningen) de zeven getallen: 5, 5, 6, 6, 6, 7, 7 gekozen als dataset. En inderdaad, hun gemiddelde is $\overline{x} = 6$.

Nota.
Voor de volgende twee opdrachten werken leerlingen in 5 groepjes. Elk groepje krijgt 7 Post-it blaadjes en elk groepje heeft een andere kleur (groen, rood, blauw, oranje, paars). Voorzie 5 identieke grote witte bladen met de reeds gegeven getallenas. Zo kunnen leerlingen elkaars grafieken vergelijken.

Opdracht 4

Elk groepje maakt een grafiek door de 7 gekregen Post-it blaadjes boven de getallenas te kleven. Zorg ervoor dat die grafiek een dataset met 7 getallen voorstelt waarbij het gemiddelde gelijk is aan 6. Je hebt per groepje de vrije keuze hoe je dit wil doen, maar we spreken wel af dat je geen grafiek neemt die in deze tekst al is voorgekomen.

Als alle grafieken gemaakt zijn, kan je ze samen bekijken. Zijn ze allemaal correct opgesteld? Zijn meerdere grafieken identiek of zijn ze allemaal verschillend? Als je alleen het gemiddelde kent, wat weet je dan over een dataset? Wat zie je hier als je al die grafieken bekijkt?

Opdracht 5

Nu je weet hoe je 7 getallen met gemiddelde $\bar{x} = 6$ kan voorstellen, krijg je per groepje (dus per kleur) extra voorwaarden opgelegd. Je moet de grafiek die je zopas gemaakt hebt als volgt aanpassen: het gemiddelde moet nog altijd $\bar{x} = 6$ zijn met bovendien:

- groen: juist één 4 en verder alleen oneven getallen
- rood: gebruik alle getallen minstens één keer behalve 6, 7 en 8 want die mogen niet voorkomen
- blauw: juist één 4 en verder alleen even getallen
- oranje: er moet juist één 9 voorkomen en juist één 4 en geen 3
- paars: er zijn 2 getallen meer die kleiner dan 6 zijn dan dat er groter dan 6 zijn.

Nota.
Om 7 getallen te vinden waarvan het gemiddelde gelijk is aan 6 gaan leerlingen soms anders te werk. Uit de formule: [gemiddelde = som / aantal] leiden zij af: [som = gemiddelde x aantal] = 6 x 7 = 42. Dan proberen zij experimenteel om met zeven getallen (die ook moeten voldoen aan de opgegeven voorwaarden) een som van 42 te bereiken. Daar komt soms veel "trial-and-error" bij kijken.

5. Gemiddelde, onderzoeksvraag en context

Een statistisch onderzoek start met een onderzoeksvraag. Dan verzamel je data en je probeert te weten te komen wat die data vertellen. Je gebruikt daarvoor allerlei methoden en technieken zoals kengetallen, grafieken, enz.

Als men zegt dat het gemiddelde een "kengetal voor het centrum" is, wat betekent dat dan?

- Als je bij het woord "centrum" denkt aan "evenwichtspunt" dan is het gemiddelde de perfecte grootheid. Kijk maar naar de grafieken in deze tekst: overal is het gemiddelde het evenwichtspunt.

- Als de onderzoeksvraag peilt naar "centrum" in de zin van "de standaard", "het typische", "het globale", "de norm"… dan kan het gemiddelde je soms helpen maar het kan er ook totaal naast zitten. Kijk maar eens naar de toets Nederlands waar Emma 7 op 10 haalde en het klasgemiddelde voor die toets gelijk was aan 6. Goed gewerkt, Emma … of toch niet?

  Voor meer info hierover ga je naar https://www.uhasselt.be/lesmateriaal-statistiek, klik daar in de linker kolom op "Achtergrondinformatie", klik dan op "Statistiek: een eigen manier van denken en een eigen aanpak" en ga naar de tekst "Context: een cruciale factor bij een statistisch onderzoek".

Besluit. Je kan data verzamelen en je kan een gemiddelde berekenen. Maar of je daarmee een antwoord hebt op de onderzoeksvraag hangt ook af van de context. Vraag het maar aan Emma.

Nota.
"Een grafiek" (zoals een dotplot) trekt de aandacht op eigenschappen die je niet ziet in "een getal" (zoals het gemiddelde). Teken dus altijd ook een grafiek bij een exploratief statistisch onderzoek.

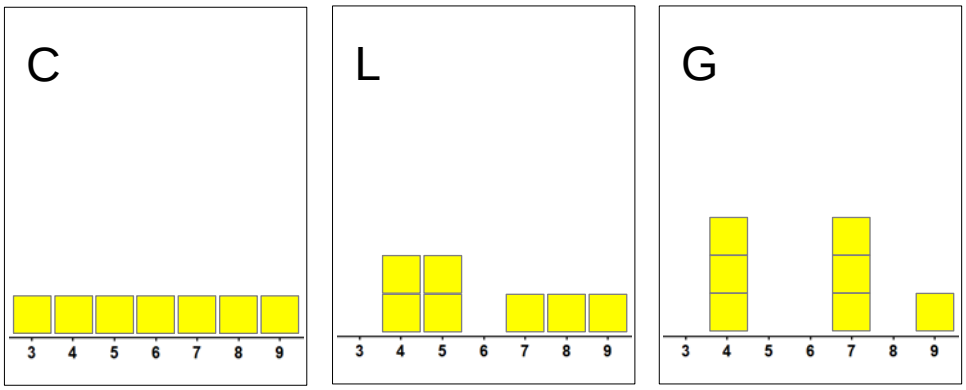
Spreiding rond het gemiddelde

6. Spreiding

Hieronder staan 3 grafieken: C, L en G. Zij stellen drie datasets voor waarbij de data 7 gehele getallen zijn met mogelijke waarden tussen 3 en 9.

6.1. Spreiding: intuïtief

Probeer eens te antwoorden op de vraag: welke dataset heeft de grootste spreiding? Je mag antwoorden gewoon op zicht, volgens je eigen aanvoelen. Zeg ook even waarom je welke dataset kiest (en doe dat vooraleer je verder leest).



Bij het woord “spreiding” vergelijk je spontaan de getallen onderling met elkaar. Maar hoe doe je dat? De data bij grafiek C zijn 3, 4, 5, 6, 7, 8, 9, bij L heb je 4, 4, 5, 5, 7, 8, 9 en bij G is dat 4, 4, 4, 7, 7, 7, 9.

Je kan nu op verschillende manieren naar de grafieken kijken.

- In grafiek C zie je dat de getallen helemaal uitgespreid liggen van 3 tot 9. Nergens vallen er data samen. Misschien zeg je daarom dat C de grootste spreiding heeft.
- Als je L en G met elkaar vergelijkt, dan zie je plaatsen waar data samenvallen. Bij L komt zowel 4 als 5 twee keer voor. Bij G is dat nog straffer, daar komen 4 en 7 elk drie keer voor. Als “veel keer dezelfde waarde” je het gevoel geeft van “weinig gespreid”, dan zeg je dat de spreiding bij G kleiner is dan bij L. Bij L heb je 5 verschillende waarden tussen je data, bij G zijn het er slechts 3.
- Zijn het misschien de “gaten” die je opvallen? Als je denkt “meer en grotere gaten” = “grotere spreiding” dan kies je voor G als dataset met de grootste spreiding.

6.2. Spreiding rond het gemiddelde

Spreiding kan je ook op een andere manier bekijken.

Wanneer je het gemiddelde als centrum neemt, dan kan je kijken of de data dicht bij dat centrum liggen of juist niet. De vraag is dan niet “liggen de data onderling ver uit elkaar?” maar wel “hoe ver liggen de data van hun gemiddelde?”. Antwoorden op deze vraag doe je niet op zicht. Voor “spreiding rond het gemiddelde” heb je een maat nodig, een “spreidingsmaat”. Die gaan we nu opstellen.

Nota.
Als je in deze tekst het woord “spreiding” ziet, dan betekent dat vanaf nu “spreiding rond het gemiddelde”.

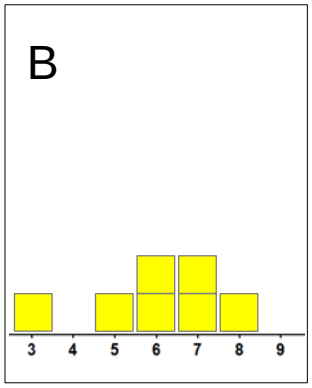
7. Een maat voor spreiding: eerste stap

We bouwen een maat voor spreiding in 2 stappen. We starten met een som.

7.1. De som van de afstanden

Kijk naar grafiek B. Die dataset ken je al. Je weet dat het gemiddelde daar gelijk is aan 6. Voor de spreiding rond het gemiddelde kijk je hoever de data van het gemiddelde liggen.

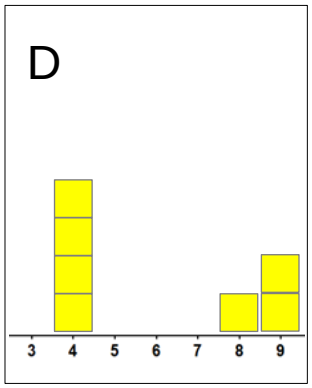
Je werkt hier met "afstanden" en niet met "verschillen". Daarom kan je gewoon "tellen" hoeveel stappen (van lengte één) je moet zetten, of het nu naar links is of naar rechts. In de tabel zie je hoe dat werkt. Voor deze dataset is ***de som van de afstanden tot het gemiddelde*** gelijk aan 8.



Dataset B		
data	aantal stappen vanaf het datapunt tot aan het gemiddelde $\bar{x} = 6$	
3	3	totaal aantal stappen = **som** van de afstanden van de data tot aan het gemiddelde **SOM = 8**
5	1	
6	0	
6	0	
7	1	
7	1	
8	2	

Opdracht 6

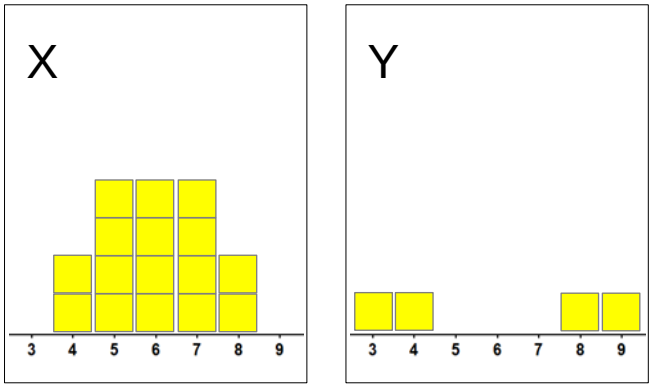
Nu je weet dat je ***de som van de afstanden tot het gemiddelde*** kan gebruiken als maat voor spreiding kijk je eens naar grafiek D. Op zicht liggen de data daar verder af van het gemiddelde $\bar{x} = 6$ dan bij grafiek B. Maar zegt onze maat voor spreiding dat ook? Gebruik de tabel om je antwoord te motiveren.



Dataset D		
data	aantal stappen vanaf het datapunt tot aan het gemiddelde $\bar{x} = 6$	
		totaal aantal stappen = **som** van de afstanden van de data tot aan het gemiddelde **SOM = …..**

7.2. Een probleem

De vorige voorbeelden tonen dat onze maat voor spreiding (de som van de afstanden tot het gemiddelde) voldoet aan wat je ervan verwacht maar … er is een probleem.



Kijk naar grafiek X en grafiek Y. Het gemiddelde is gelijk aan 6 voor beide datasets.

Bij grafiek X ligt de meerderheid van de data (met waarden 5, 6 of 7) dicht bij het gemiddelde. Enkele data liggen wat verder (2 keer 4 en 2 keer 8). Een nog grotere afstand (zoals de waarde 3 of 9) komt hier niet voor.
Grafiek Y ziet er heel anders uit. Geen enkel getal is daar 5, 6 of 7. Bij Y liggen de data globaal verder weg van het gemiddelde dan bij X. Bij Y is de spreiding (ten opzichte van het gemiddelde) groter dan bij X.

Opdracht 7

Zoek de "som van de afstanden tot het gemiddelde" bij X en bij Y. Als je wil kan je tabellen opstellen zoals bij de grafieken B en D. Dat helpt bij het tellen. De resultaten die je nu vindt, had je die verwacht? Waarom?

8. Een kengetal voor spreiding rond het gemiddelde: MAD

8.1. "Dicht bij $\bar{x}$" blijft "dicht bij $\bar{x}$", ook als je met veel bent

Bij spreiding kijk je niet naar het aantal getallen maar je kijkt hoe zij gespreid liggen rond het gemiddelde. Zomaar "de som van alle afstanden" berekenen is dan niet zo'n goed idee.

Voor de dataset X is "de som van alle afstanden tot het gemiddelde" gelijk aan 16 en er zijn 16 datapunten. ***Gemiddeld, per datapunt***, is "de afstand tot het gemiddelde" gelijk aan: som / aantal = 16 / 16 = 1.
Voor Y is "de som van alle afstanden tot het gemiddelde" gelijk aan 10. Daar zijn er maar 4 datapunten. Bij Y is ***gemiddeld, per datapunt***, "de afstand tot het gemiddelde" gelijk aan 10 / 4 = 2.5.

De nieuwe spreidingsmaat die je nu hebt opgesteld zegt hoever, ***gemiddeld***, de data van hun gemiddelde liggen. Voor elke dataset, groot of klein, is dit een goede maat die compenseert voor "het aantal" data en die dus echt kijkt naar "spreiding".

Volgens deze nieuwe spreidingsmaat liggen bij X de data gemiddeld op een afstand 1 van hun gemiddelde en bij Y liggen ze gemiddeld op een afstand 2.5. De spreiding bij X is kleiner dan bij Y.

8.2. Gemiddelde afstand tot het gemiddelde: MAD

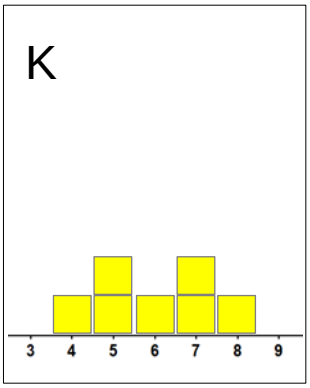
In plaats van naar de afstand, kijken sommige studies eerst naar het verschil: [datapunt – gemiddelde]. Een verschil kan positief of negatief zijn. Men schakelt dan over op de absolute waarde van het verschil. Dat is altijd positief en zo krijg je echt de "afstand" van het datapunt tot het gemiddelde.

In (internationale) teksten kom je de afkorting MAD tegen. De "D" komt van het Engelse "deviation" (afwijking of verschil). De absolute waarde van het verschil wordt dan "absolute deviation". Dat zijn de afstanden. En van al die afstanden neem je dan het gemiddelde ("mean"). Zo kom je aan **M**ean **A**bsolute **D**eviation, afgekort als **MAD**. Dit is wat wij "de gemiddelde afstand tot het gemiddelde" noemen.

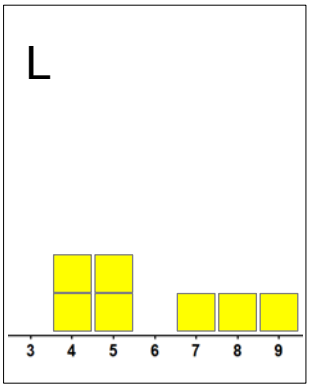
"***De gemiddelde afstand tot het gemiddelde***" is een getal dat zegt hoever, gemiddeld, de datapunten verwijderd liggen van hun gemiddelde $\bar{x}$. Dit getal gebruiken wij in deze tekst als kengetal voor spreiding en we geven het de (Engelse) afkorting "**MAD**".

Opdracht 8

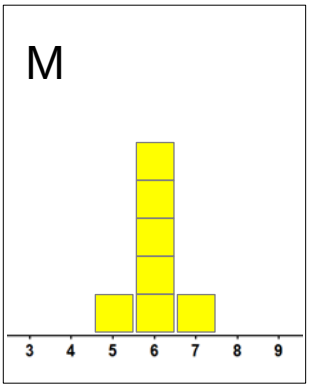
Hieronder zie je de grafieken K, L, M en N. Onder grafiek K staat een schema dat je kan gebruiken bij de berekeningen. In dat schema ontdek je dat dataset K een spreiding heeft van 1.14 want MAD = 1.14.



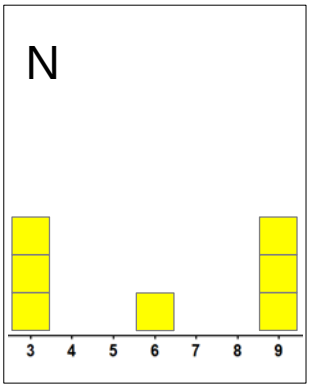
Dataset K	
aantal data n = 7	
gemiddelde $\bar{x}$ = 6	
data	afstand
4	2
5	1
5	1
6	0
7	1
7	1
8	2
som afstanden = 8	
gemiddelde afstand tot het gemiddelde: MAD = 8 / 7 = 1.14	



Dataset L	
aantal data n = …..	
gemiddelde $\bar{x}$ = …..	
data	afstand
som afstanden = …..	
gemiddelde afstand tot het gemiddelde: MAD = …/… = …..	



Dataset M	
aantal data n = …..	
gemiddelde $\bar{x}$ = …..	
data	afstand
som afstanden = …..	
gemiddelde afstand tot het gemiddelde: MAD = …/… = …..	



Dataset N	
aantal data n = …..	
gemiddelde $\bar{x}$ = …..	
data	afstand
som afstanden = …..	
gemiddelde afstand tot het gemiddelde: MAD = …/… = …..	

Het is nu jouw taak om de spreiding van de data te bestuderen bij de grafieken L, M en N.

Bedenk, vooraleer je begint te rekenen, dat "de gemiddelde afstand tot het gemiddelde" ***een spreidingsmaat*** is. Die zegt iets over de spreiding van de data rond hun gemiddelde. Kan je die spreiding ook zien in de grafieken? Probeer eens (op zicht) de datasets te rangschikken: waar liggen, gemiddeld, de data het dichtst tegen het gemiddelde $\bar{x} = 6$? Bij welke grafiek liggen de data er het verste van af? Soms is het niet gemakkelijk om spreidingen zomaar "op zicht" te vergelijken, je hebt echt een spreidingsmaat nodig.

Maak nu de berekeningen. Wat zeggen de MAD–waarden? Bij welke dataset is de spreiding het kleinst? Waar is ze het grootst? Had je dat vooraf gedacht?

Nota.
De voorbeelden in deze tekst zijn supereenvoudig. Je kan de MAD-waarden snel uitrekenen met een schema zoals in opdracht 8. Maar je kan ook gebruik maken van ICT. Daar zijn verschillende mogelijkheden voor.

- Met een TI-84 Plus zorg je dat de data in lijst [L1] staan en dan pas je het programma MAD toe. Dit programma kan je downloaden op https://www.uhasselt.be/lesmateriaal-statistiek waar je in de linker kolom klikt op "ICT-ondersteuning".
- Met Excel zet je de data in een werkblad en gebruik je de functie AVEDEV. Als de data in de eerste kolom staan (van A1 tot A7) dan werk je met f_x =AVEDEV(A1:A7).
- enz.

Opdracht 9

Bij de studie over spreiding ben je in deze tekst gestart met de grafieken C, L en G. Je wist toen nog niet dat het over spreiding rond het gemiddelde zou gaan en je hebt toen geprobeerd om, op zicht, iets te zeggen over grotere of kleinere spreiding. Je kon daarbij op verschillende manieren redeneren.
Nu weet je dat het hier gaat over spreiding rond het gemiddelde. Je moet dan eigenlijk naar de data kijken vanuit het gemiddelde. Of die data dan "ver weg" of "dichtbij" liggen, dat beslis je niet op zicht. Je gebruikt een spreidingsmaat en zoekt de MAD-waarde. Voor de dataset L heb je dat al gedaan in opdracht 8. Zoek nu ook (met ICT of met de hand) hoe groot de spreiding is bij de datasets C en G. Je kan de MAD-waarden noteren op de figuren. Misschien ben je nu overtuigd dat "intuïtie" niet altijd zo'n goed idee is.

Opdracht 10

In opdracht 5 heb je zelf datasets opgesteld met gekleurde Post-it blaadjes. Dat deed je bij de studie van het gemiddelde. Je hebt toen ontdekt dat er bij een gegeven gemiddelde ($\bar{x} = 6$) heel veel verschillende datasets mogelijk zijn. Nu kan je een stapje verder gaan en ook de spreiding rond het gemiddelde bekijken. Zoek (met ICT of met de hand) voor elke dataset die je toen hebt opgesteld "de gemiddelde afstand tot het gemiddelde" (de MAD) en noteer die MAD-waarde op de gekleurde grafiek.

9. En later?

Later, in de tweede en derde graad en ook in het hoger onderwijs, zal je nog met heel wat statistiek te maken krijgen.

De klassieke spreidingsmaat die daar gebruikt wordt, heet "de standaardafwijking". De notatie voor de standaardafwijking van een dataset is de kleine letter "**s**" en de berekening is ingewikkeld. Maar als idee zijn er toch heel wat raakvlakken met wat je nu geleerd hebt. Daarom is werken met MAD een goede aanloop om later de standaardafwijking te begrijpen (minstens intuïtief).

Nota.
De onderwijsdoelen wiskunde van de eerste graad vermelden als spreidingsmaat alleen de variatiebreedte.

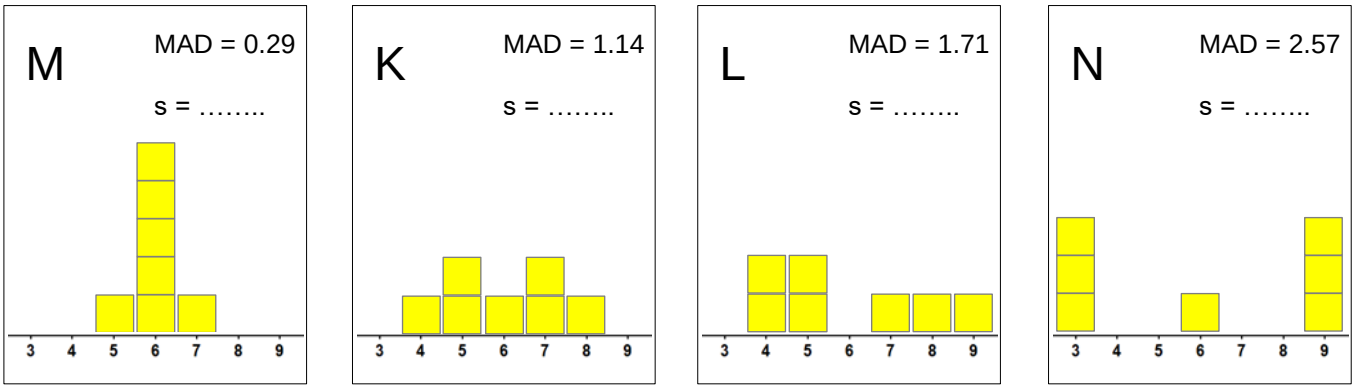
de standaardafwijking "**s**"	**de gemiddelde afstand tot het gemiddelde** "**MAD**"	**de variatiebreedte**
alle data x_i spelen een rol	alle data x_i spelen een rol	behalve min en max speelt geen enkel datapunt een rol
er is een referentiekader: het gaat over spreiding rond het gemiddelde $\overline{x}$	er is een referentiekader: het gaat over spreiding rond het gemiddelde $\overline{x}$	de variatiebreedte ($= maximum - minimum$) zegt alleen hoe groot het gebied is waarin de data terechtkomen
de bouwstenen starten met de verschillen $x_i - \overline{x}$	de bouwstenen starten met de verschillen $x_i - \overline{x}$	
schakel over op positieve getallen: kwadrateer	schakel over op positieve getallen: neem absolute waarde	
maak de som, compenseer voor het aantal data en trek dan de vierkantswortel	maak de som en compenseer voor het aantal data: deel door n	

Zonder berekeningen te maken kan je eens kijken naar de standaardafwijking als een alternatief kengetal voor spreiding rond het gemiddelde. De waarden voor "s" (de standaardafwijking) zijn niet dezelfde als de waarden voor "MAD" (de gemiddelde afstand tot het gemiddelde) want het zijn andere formules. Maar in grote lijnen ontdek je voor de spreiding toch wel eenzelfde trend.

Nota.
Op je rekentoestel (of app) zijn er soms verschillende toetsen (of functies) die standaardafwijking genoemd worden. Om zeker te zijn dat je de juiste toets (of de juiste functie) gebruikt, doe je eerst de volgende test. Zet de getallen 1, 2 en 3 in een lijst. Zoek dan met ICT de standaardafwijking voor die lijst. Als het antwoord gelijk is aan het getal 1, dan heb je de juiste toets (of de juiste functie) te pakken.

Opdracht 11

In opdracht 8 heb je de MAD gevonden voor de datasets K, L, M en N. De grafieken hieronder staan gerangschikt in volgorde van spreiding (van klein naar groot) wanneer je de MAD als spreidingsmaat gebruikt. Zoek nu voor dezelfde datasets wat hun spreiding rond het gemiddelde is als je die berekent volgens de spreidingsmaat "standaardafwijking". Gebruik ICT en vul de s-waarden in.



Opdracht 12

Je kan opdracht 11 ook toepassen op de datasets die je zelf gemaakt hebt (met de gekleurde Post-it blaadjes). Voor die datasets heb je in opdracht 10 al bepaald wat de MAD is. Schrijf die MAD-waarden op de gekleurde grafieken en orden ze van klein naar groot. Zoek nu (met ICT) de standaardafwijking en noteer die per dataset op de gepaste grafiek. Zie je hier ook, globaal, dat grotere standaardafwijkingen samengaan met grotere spreidingen?

10. Oplossingen

Opdracht 1

Grafiek A toont een dataset met gemiddelde $\bar{x} = 6$ want (5+5+6+6+6+7+7) / 7 = 6.

Opdracht 2

	Dataset voor deze grafiek	Het gemiddelde
Grafiek B	3, 5, 6, 6, 7, 7, 8	$\bar{x} = 6$
Grafiek C	3, 4, 5, 6, 7, 8, 9	$\bar{x} = 6$
Grafiek D	4, 4, 4, 4, 8, 9, 9	$\bar{x} = 6$

Opdracht 3

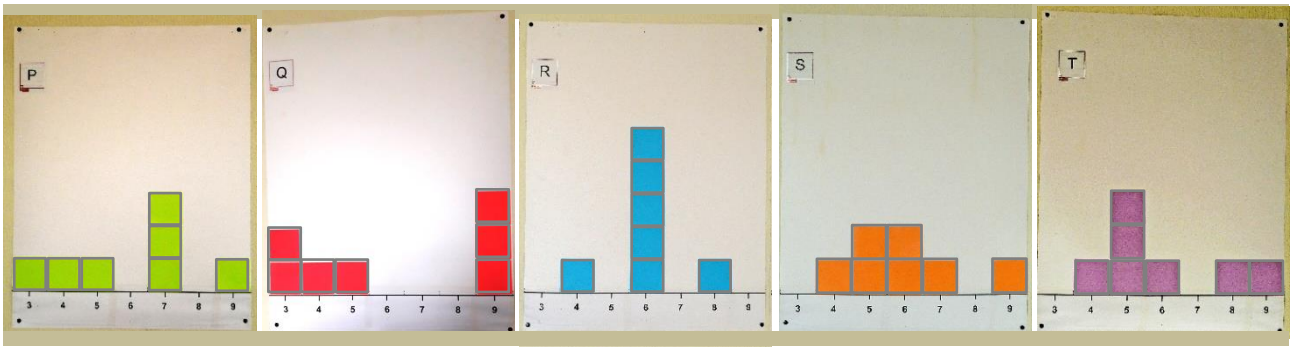
Dataset D			
data	**stappen vanuit $\bar{x}$**	**aantal stappen**	
4	twee naar links	2	totaal aantal naar links = 8
4	twee naar links	2	
4	twee naar links	2	
4	twee naar links	2	
8	twee naar rechts	2	totaal aantal naar rechts = 8
9	drie naar rechts	3	
9	drie naar rechts	3	

Opdracht 4

Verschillende juiste antwoorden zijn mogelijk (controleer bij elke grafiek of het gemiddelde $\bar{x} = 6$ is). Bemerk dat je nog niet veel weet als je alleen maar het gemiddelde kent. Er zijn dan nog heel veel verschillende datasets mogelijk. Dat zie je op de grafieken.

Opdracht 5

Per groepje kunnen er (meestal) meerdere juiste antwoorden zijn. Hieronder zie je een voorbeeld.



Nota.
Andere voorwaarden zouden kunnen zijn: “juist vier keer een 6”, “het zijn allemaal dezelfde getallen”, enz.

Opdracht 6

Ja, bij grafiek D is de spreiding groter dan bij grafiek B want nu is de som gelijk aan 16 en bij B was die 8.

Dataset D	
data	aantal stappen vanaf het datapunt tot aan het gemiddelde $\bar{x} = 6$
4	2
4	2
4	2
4	2
8	2
9	3
9	3

totaal aantal stappen
=
som van de afstanden van de data tot aan het gemiddelde
SOM = 16

Opdracht 7

Dataset X	
data	# stappen vanaf het datapunt tot aan $\bar{x}$
4	2
4	2
5	1
5	1
5	1
5	1
6	0
6	0
6	0
6	0
7	1
7	1
7	1
7	1
8	2
8	2

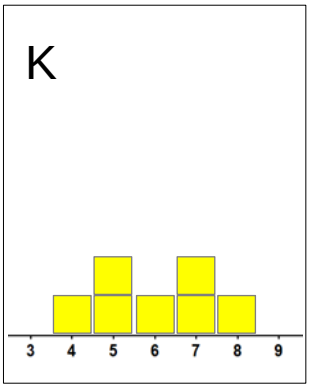
totaal aantal stappen
=
som van de afstanden van de data tot aan het gemiddelde
SOM = 16

Dataset Y	
data	# stappen vanaf het datapunt tot aan $\bar{x}$
3	3
4	2
8	2
9	3

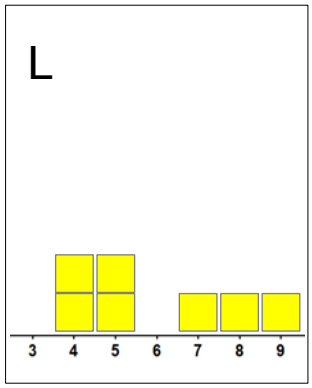
totaal aantal stappen
=
som van de afstanden van de data tot aan het gemiddelde
SOM = 10

Bij grafiek Y liggen de data globaal verder van het gemiddelde dan bij grafiek X. Je verwacht dan ook een grotere som bij Y dan bij X. Dat is hier niet het geval en je ontdekt ook waarom: bij X zijn er veel meer data dan bij Y. Je moet dus een spreidingsmaat opstellen die zich niet laat afleiden door “het aantal” data. Die maat moet ***de spreiding*** weergeven, of het nu gaat over een grote of een kleine dataset.

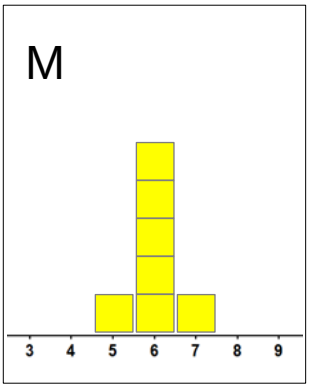
Opdracht 8



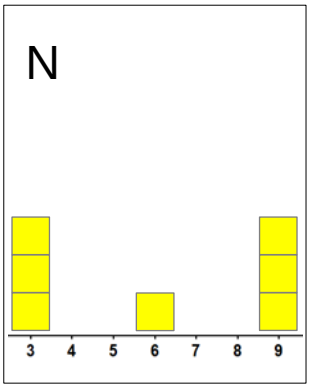
Dataset K	
aantal data n = 7	
gemiddelde $\bar{x}$ = 6	
data	afstand
4	2
5	1
5	1
6	0
7	1
7	1
8	2
som afstanden = 8	
gemiddelde afstand tot het gemiddelde: MAD = 8 / 7 = 1.14	



Dataset L	
aantal data = 7	
gemiddelde $\bar{x}$ = 6	
data	afstand
4	2
4	2
5	1
5	1
7	1
8	2
9	3
som afstanden = 12	
gemiddelde afstand tot het gemiddelde: MAD 12 / 7 = 1.71	



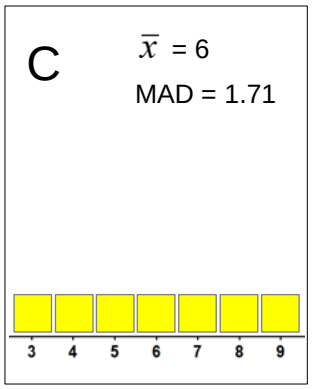
Dataset M	
aantal data = 7	
gemiddelde $\bar{x}$ = 6	
data	afstand
5	1
6	0
6	0
6	0
6	0
6	0
7	1
som afstanden = 2	
gemiddelde afstand tot het gemiddelde: MAD = 2 / 7 = 0.29	

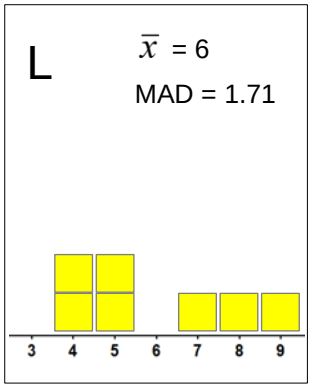


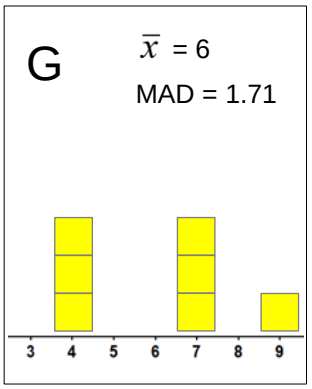
Dataset N	
aantal data = 7	
gemiddelde $\bar{x}$ = 6	
data	afstand
3	3
3	3
3	3
6	0
9	3
9	3
9	3
som afstanden = 18	
gemiddelde afstand tot het gemiddelde: MAD = 18 / 7 = 2.57	

Op zicht kan je meestal de echt kleine of echt grote spreidingen ontdekken. In deze voorbeelden zie je dat M de kleinste spreiding heeft en N de grootste. K en L zijn al moeilijker te vergelijken. Je hebt een houvast nodig die verder gaat dan “op zicht”. Die houvast is een spreidingsmaat. Hier spreek je af dat je als spreidingsmaat de MAD gebruikt. Volgens dit criterium is de spreiding het kleinste in de dataset M, dan volgt K, dan L en tenslotte N.

Opdracht 9

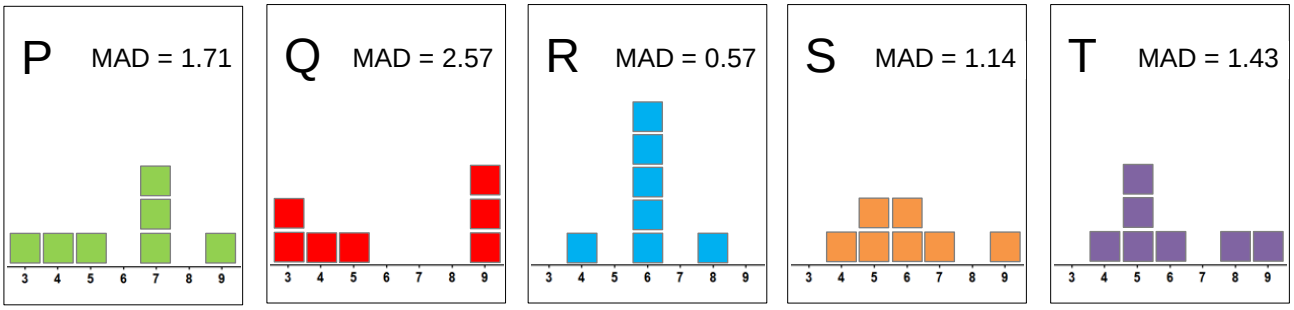






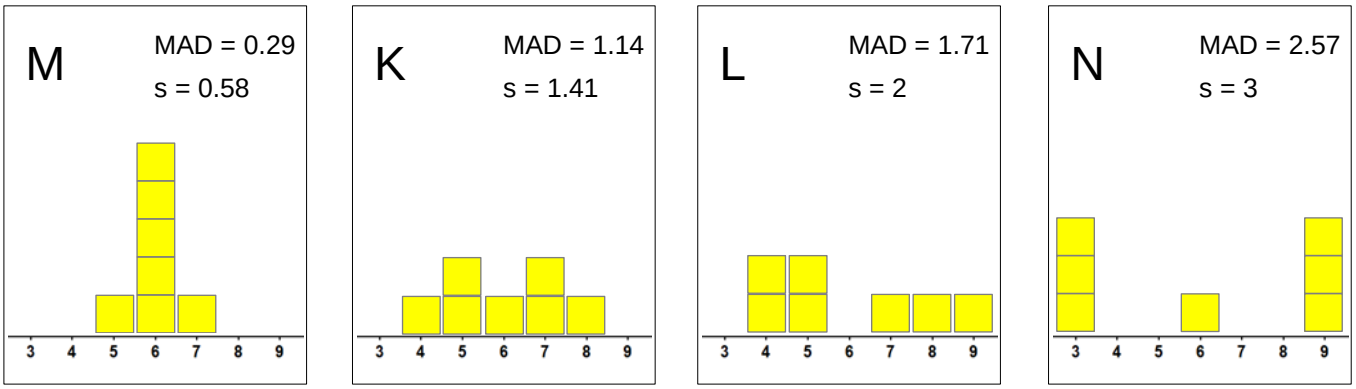
Als je “de gemiddelde afstand tot het gemiddelde” als spreidingsmaat gebruikt, dan liggen de data evenveel gespreid rond hun gemiddelde bij elk van deze grafieken (overal is MAD = 1.71).
Als men je zegt dat het gemiddelde $\bar{x}$ gelijk is aan 6 en dat de MAD-waarde gelijk is aan 1.71, dan heb je informatie over het centrum en de spreiding van de dataset. Wat weet je dan? Blijkbaar nog niet zo veel. Het kan zowel C als L als G zijn.

Opdracht 10



Opdracht 11

De datasets staan hier geordend volgens de MAD-spreidingsmaat. Als je de standaardafwijking "s" gebruikt als spreidingsmaat dan hoeft die volgorde niet identiek te zijn. In dit voorbeeld is die volgorde wel dezelfde.



Opdracht 12

Dat de spreiding groter en groter wordt, zie je ook weergegeven door de standaardafwijking die in dit voorbeeld start bij s = 1.15 en eindigt bij s = 2.89.