STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

# Correlatie: exploratieve methoden

*Werktekst voor de leerling*

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

# Inhoudstafel

**1. Veranderlijken ..... 1**
**2. Bivariate continue gegevens ..... 2**
**3. Puntenwolken ..... 3**
3.1. Een puntenwolk maken ..... 3
3.2. Een puntenwolk interpreteren ..... 6
3.3. Ellipsvormige puntenwolken ..... 7
**4. Lineaire samenhang: een grafische studie ..... 9**
4.1. Lineaire samenhang bij ellipsvormige puntenwolken ..... 9
4.2. Sterkte en zin van een lineaire samenhang ..... 11
**5. Lineaire samenhang: een numerieke studie ..... 14**
5.1. De afzonderlijke coördinaten en hun kengetallen ..... 14
5.2. Univariate en bivariate informatie ..... 17
5.3. Een aangepaste meetlat ..... 18
5.4. De typische rechte ..... 20
**6. Correlatie ..... 23**
6.1. Verstrooiing rond de typische rechte ..... 23
6.2. Gestandaardiseerde puntenwolken ..... 26
6.2.1. z–scores ..... 26
6.2.2. Standaardiseren met de GRM ..... 30
6.3. De correlatiecoëfficiënt ..... 31
6.3.1. De ideeën achter de formule ..... 31
6.3.2. De formule ..... 32
6.3.3. Eigenschappen van de correlatiecoëfficiënt ..... 33
**7. Een grafische valkuil ..... 35**
7.1. Bloemblaadjes ..... 35
7.2. Puntenwolken en hun correlatiecoëfficiënt ..... 37
**8. Een numerieke valkuil ..... 38**
8.1. Eén getal = beperkte informatie ..... 38
8.2. Uitschieters, krommen, en de voorbeelden van Anscombe ..... 39
**9. Wat kan er nog meer fout gaan? ..... 40**
9.1. Paleontologie ..... 40
9.2. Clusters ..... 41
9.3. Hoger of lager? ..... 43
9.4. De ecologische valkuil ..... 45
9.5. Oorzaak en samenhang ..... 47

Bij correlatie bestudeer je een verband tussen continu numerieke veranderlijken. Hoe zo'n veranderlijken eruit zien, bekijk je even vooraf.

# 1. Veranderlijken

Een statistische studie kan gaan over personen (baby's, leerlingen, vrouwen …) of dieren (muizen, paarden, apen …) of planten (irissen, eiken, tomaten …) of zaken (ontbijtgranen, steden, fietsen …). De dingen die je bestudeert, zijn de **elementen** in je studie.

Bij elk element ben je geïnteresseerd in bepaalde eigenschappen. Dat zijn de **veranderlijken**. Een enquête bij 500 Vlamingen kan bijvoorbeeld vragen naar het geslacht, de bloedgroep, de lengte en het gewicht. Bij elk **element** (elke ondervraagde Vlaming) worden hier 4 **veranderlijken** opgemeten.

Voor elke veranderlijke noteer je haar **waarde**.

- De **veranderlijke** "geslacht" heeft maar twee mogelijke **waarden**: mannelijk / vrouwelijk.
- De **veranderlijke** "bloedgroep" heeft vier mogelijke **waarden**: O, A, B, AB.
- De **veranderlijken** "lengte" en "gewicht" hebben heel veel mogelijke **waarden**.

De **waarden** van de veranderlijken "geslacht" en "bloedgroep" omschrijf je **met woorden** (of afkortingen), niet met getallen.

De **waarden** van de veranderlijken "lengte" en "gewicht" zijn **getallen**. Om de echte lengte of het echte gewicht te kennen zou je supergevoelige meetapparatuur moeten hebben en zelfs dan blijven er problemen. Bovendien schrijf je geen getallen op met miljoenen decimalen (je moet ergens na de komma stoppen). Het is niet omdat je de "echte" waarde niet kan opmeten of niet kan opschrijven, dat die echte waarde er niet is. "Als model" kan zo'n "echte waarde" gelijk welk getal zijn tussen bepaalde grenzen. Een veranderlijke waarbij de **waarden** *alle mogelijke getallen* zijn *tussen bepaalde grenzen* heet een **continu numerieke** veranderlijke. Voorbeelden zijn: gewicht, lengte, tijd, … Een continu numerieke veranderlijke wordt ook een continue veranderlijke genoemd.

Voorbeeld. In Vlaanderen is de gemiddelde lengte van 17-jarige meisjes 1.66 m. Bij een studie van de lengte van deze meisjes gebruik je een "min of meer" nauwkeurige meetlat en je noteert de lengte (in meter) tot op 2 decimalen. Als "model" voor de lengte van deze meisjes denk je aan een continuüm van mogelijke waarden, ergens tussen 1.20 m en 2.20 m. De naam van de veranderlijke is hier "lengte" (van 17-jarige Vlaamse meisjes) en de waarden (in m) zijn een continuüm van getallen tussen 1.20 en 2.20.

Opdracht 1

Geef een voorbeeld van een onderzoek waar je een eigenschap (van mensen, dieren of dingen) bestudeert waarbij de opgemeten veranderlijke **continu numeriek** is. Geef de **naam** van de veranderlijke en haar **waarden.**

# 2. Bivariate continue gegevens

Bij een geboorte wordt ondermeer het gewicht en de lengte van de baby opgeschreven.

De lengte (in centimeter) kan je noteren als $x$ en het gewicht (in kilogram) als $y$. Per baby kan je die informatie schrijven als een koppel: $(x, y)$ = (lengte van de baby, gewicht van de baby).

Per baby noteer je hier tegelijkertijd twee kenmerken. Dat levert een **bivariate** (bi = twee) uitkomst, waarbij de volgorde van belang is. In deze studie komt eerst de lengte en dan het gewicht als je het koppel $(x, y)$ opschrijft.

De grootheden die je opmeet (lengte en gewicht) zijn **continue** veranderlijken en dus werk je hier met **bivariate continue** gegevens.

Als je meerdere (bijvoorbeeld 10) baby's opmeet dan gebruik je een index om hun resultaat op te schrijven:

$(x_1, y_1)$ = (lengte van de 1$^{ste}$ baby, gewicht van de 1$^{ste}$ baby)

$(x_2, y_2)$ = (lengte van de 2$^{de}$ baby, gewicht van de 2$^{de}$ baby)

…….

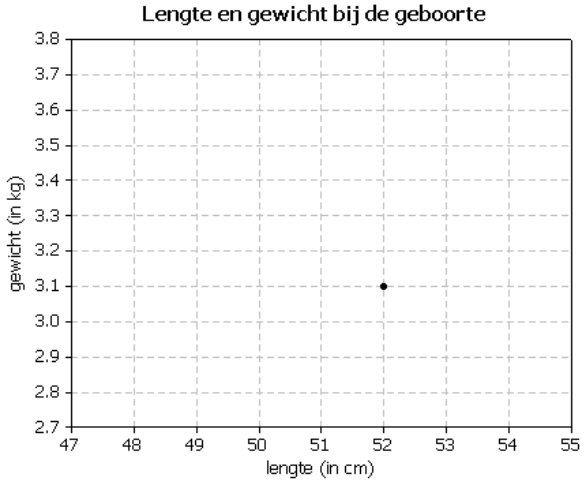
$(x_i, y_i)$ = (lengte van de i$^{de}$ baby, gewicht van de i$^{de}$ baby)

…….

$(x_{10}, y_{10})$ = (lengte van de 10$^{de}$ baby, gewicht van de 10$^{de}$ baby)

Een bivariate uitkomst kan je ook grafisch voorstellen want elk koppel bepaalt een punt in het vlak.

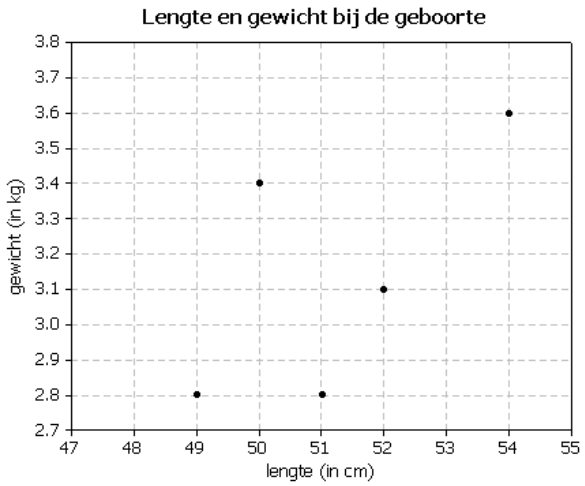
Als voor de i$^{de}$ baby $(x_i, y_i)$ = (52, 3.1) dan heb je te maken met een baby van 52 cm die 3 kilogram en 100 gram weegt. Die stel je voor door een punt in het vlak met x-coördinaat gelijk aan 52 en y-coördinaat gelijk aan 3.1. Dat zie je hiernaast.



Opdracht 2

Hieronder zie je bivariate resultaten $(x_i, y_i)$ waarbij $x_i$ de lengte en $y_i$ het gewicht is. In deze studie hebben de baby's een nieuw nummer gekregen (van klein naar groot): de kleinste baby wordt de 1ste baby genoemd en de grootste baby wordt de 10de baby genoemd. De resultaten zijn genoteerd in een tabel en zij zijn ook grafisch voorgesteld als punten in een vlak. De tabel is niet volledig, maar je kan die aanvullen met wat je ziet in de grafiek. Ook de grafiek is niet volledig, maar die kan je vervolledigen met de informatie in de tabel. Doe dat nu.

| | Lengte (in cm) | Gewicht (in kg) |
|---|---|---|
| 1ste baby | 48 | 2.9 |
| 2de baby | | |
| 3de baby | 49 | 3.1 |
| 4de baby | | |
| 5de baby | | |
| 6de baby | 51 | 3.5 |
| 7de baby | | |
| 8ste baby | 53 | 3.1 |
| 9de baby | 53 | 3.7 |
| 10de baby | | |

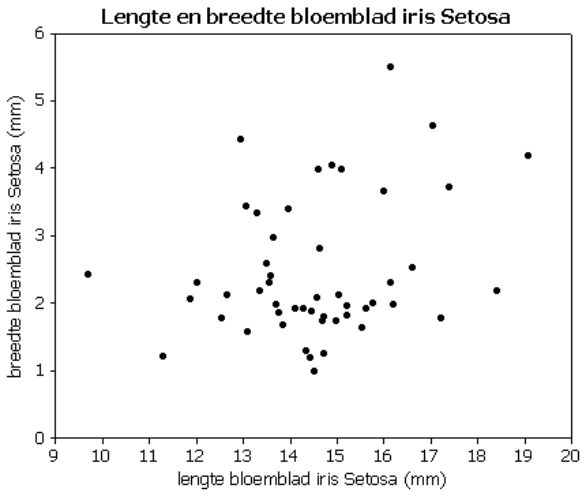



# 3. Puntenwolken

## *3.1. Een puntenwolk maken*

R. A. Fisher (1890 – 1962) was een beroemde statisticus maar hij was ook bioloog en geneticus. Hij haalde heel wat van zijn gegevens uit experimenten in de biologie.

Hiernaast zie je een grafiek waarop een deel van de vermaarde "Fisher's Iris data" wordt getoond. Het gaat hier over een bepaalde soort iris (de iris Setosa) waarbij de lengte ($x$) en de breedte ($y$) van een bloemblad is opgemeten (in mm). Dat is gebeurd voor 50 bloemblaadjes.




Bivariate gegevens $(x_i, y_i)$ grafisch voorstellen, doe je door punten in het vlak te tekenen. De figuur die je zo krijgt, heet **puntenwolk**. Een puntenwolk wordt soms ook spreidingsdiagram genoemd.

## Opdracht 3

Deze opdracht gaat over de puntenwolk van de bloemblaadjes van de iris Setosa.

1. Omcirkel het punt dat het breedste bloemblad voorstelt. Is dat bloemblad ook het langste? Waarom?

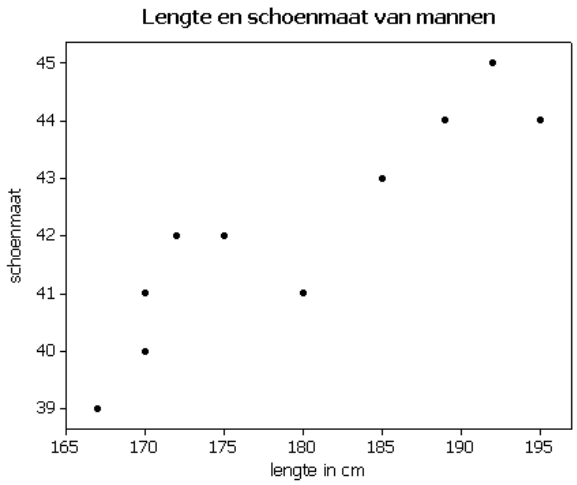
2. Hoe breed (ongeveer) is het kortste bloemblad?

## Opdracht 4

Bij 10 mannen werd de lengte ( $x$ ) en de schoenmaat ( $y$ ) genoteerd. Hieronder zie je de resultaten samen met de puntenwolk.

| Lengte in cm | 167 | 170 | 170 | 172 | 175 |
|---|---|---|---|---|---|
| Schoenmaat | 39 | 40 | 41 | 42 | 42 |

| Lengte in cm | 180 | 185 | 189 | 192 | 195 |
|---|---|---|---|---|---|
| Schoenmaat | 41 | 43 | 44 | 45 | 44 |



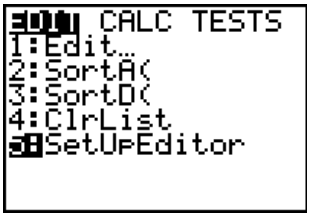

Een puntenwolk kan je tekenen met de GRM. Om te zien hoe dat werkt, volg je stap voor stap de instructies die hieronder staan. Voer die instructies niet zomaar slaafs uit, maar gebruik dit voorbeeld om te leren hoe je zelfstandig puntenwolken kan tekenen met je GRM.

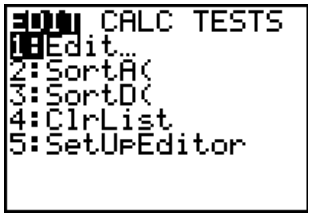
***Invoeren van de gegevens.***

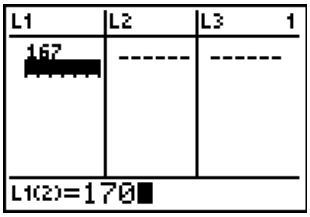
Je begint met de bivariate gegevens $(x_i, y_i)$ in te brengen in je GRM. In de lijst [L1] zet je de $x$-waarden en in de lijst [L2] de bijhorende $y$-waarden. Je kan dat op 2 manieren doen: zelf intikken of bestaande lijsten kopiëren.

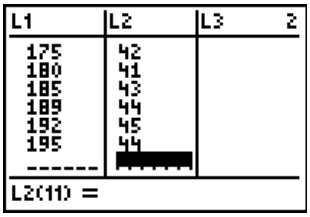
*Zelf intikken.*

De lijsten zijn hier niet lang en je verliest dus niet veel tijd als je ze zelf intikt.
Om zeker te zijn dat je start met een goede lay-out voor de in te vullen lijsten begin je als volgt: druk [STAT], loop naar 5:SetUpEditor en druk [ENTER] (of tik gewoon 5) en druk dan nog eens [ENTER].





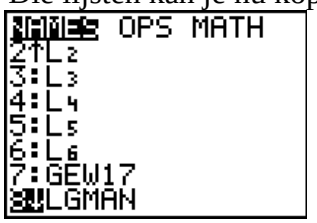


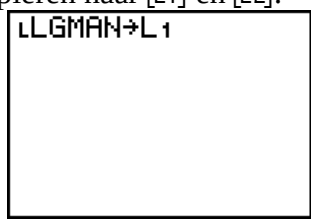


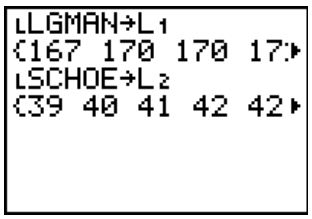

Daarna druk je [STAT] en 1:Edit en je controleert of je in de (lege) lijst [L1] staat (je kan de lijst wissen door op de naam L1 te gaan staan, op [CLEAR] te drukken en dan op [▼]). Tik daar de opeenvolgende $x_i$ – getallen, telkens gevolgd door [ENTER]. Als alle $x_i$ – getallen ingegeven zijn, dan loop je met het pijltje [▶] naar het begin van lijst [L2]. Daar vul je de $y_i$ – getallen in. Als alles ingevuld is druk je [2nd] [QUIT].

*Bestaande lijsten kopiëren.*

Op http://www.uhasselt.be/lesmateriaal-statistiek kan je alle lijsten die je in deze tekst nodig hebt vinden bij het bestand van deze werktekst. Daar staat ook uitgelegd hoe je die lijsten kan downloaden op je PC en overbrengen naar je GRM. Voor deze opdracht heb je de bestanden LGMAN.8xl (lengte man) en SCHOE.8xl (schoenmaat) gedownload en in je GRM gebracht als lijsten LGMAN en SCHOE. Die lijsten kan je nu kopiëren naar [L1] en [L2].



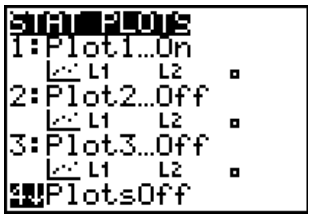




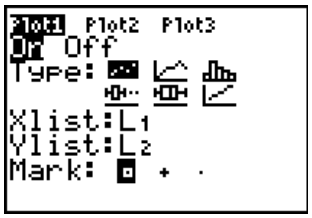

Druk [2nd] [LIST], loop naar LGMAN en druk [ENTER]. Vervolledig het commando als volgt: druk [STO▶] en [2nd] [L1] en [ENTER]. Voor de schoenmaat werk je op analoge manier. Druk [2nd] [LIST], loop naar SCHOE en druk [ENTER]. Vervolledig het commando: druk [STO▶] en [2nd] [L2] en [ENTER].

***Puntenwolk voor bivariate gegevens met x-coördinaten in [L1] en bijhorende y-coördinaten in [L2].***

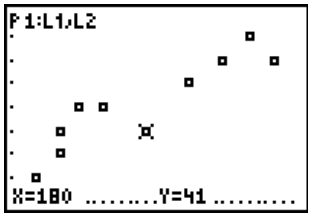
Om zeker te zijn dat je start met alle "Plots" op "Off" druk je [2nd] [STAT PLOT], dan 4:PlotsOff en dan [ENTER].




Druk nu opnieuw [2nd] [STAT PLOT] en tik dan 1. Op dit ogenblik staat Plot1 op Off. Ga op On staan en druk [ENTER]. Als type grafiek kies je het eerste type (puntenwolk). Verder is Xlist: [L1] en Ylist: [L2]. Je tekent de punten met een klein vierkantje: Mark:▫. Vergelijk met de schermafdruk hiernaast.



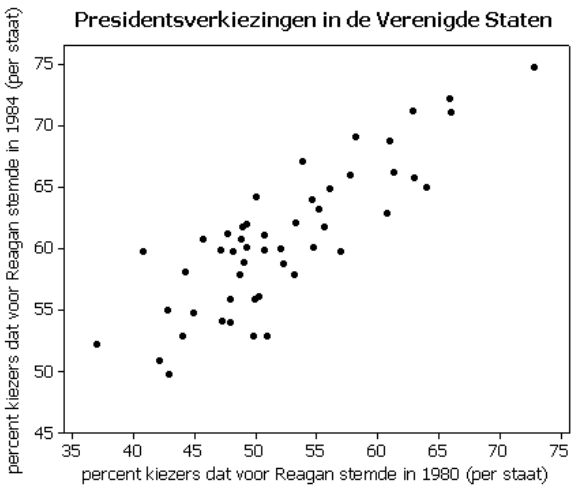

Als alles in orde is druk je [ZOOM], je loopt naar 9:ZoomStat en drukt [ENTER]. De puntenwolk verschijnt op je scherm. Druk nu [TRACE] zodat je met de pijltjes [◀] en [▶] over de grafiek kan lopen. Onderaan zie je de coördinaten van het punt waarop je staat.



## 3.2. Een puntenwolk interpreteren

In 1980 won Ronald Reagan de presidentsverkiezingen in de Verenigde Staten. In 1984 deed hij dat nog eens.

In 1980 stemde in de staat Alabama 49 % van de kiezers op Reagan en in 1984 kreeg hij daar 61 % van de stemmen. Dat noteer je (in percent) als $(x_1, y_1) = (49,61)$.
De tweede staat (in alfabetische volgorde) is Alaska. Het resultaat van die staat noteer je als $(x_2, y_2) = (54,67)$, wat betekent dat in Alaska 54 % van de kiezers op Reagan stemde in 1980 en 67 % in 1984.
Bij elk van de 50 staten hoort een resultaat $(x_i, y_i)$.
Alle verkiezingsuitslagen kan je vinden op http://uselectionatlas.org/RESULTS/index.html .




Als je geïnteresseerd bent in een globale trend, wat de staten betreft, dan krijg je een goed zicht als je de verkiezingsuitslagen grafisch voorstelt in een puntenwolk. Die staat hierboven.

Een opwaartse trend betekent dat, **globaal**, grotere $y$-waarden samengaan met grotere $x$-waarden. Bij een neerwaartse trend zie je, **globaal**, de $y$-waarden kleiner worden als de $x$-waarden vergroten.

Opdracht 5

Hieronder staan 3 uitspraken. Zeg of ze juist of fout zijn en geef ook aan waarom.

1. De puntenwolk van de presidentsverkiezingen toont een opwaartse trend.

2. Waar Reagan het hoogste percent stemmen had in 1980, daar kreeg hij ook het hoogste percent stemmen in 1984. Zoiets zie je altijd bij een opwaartse trend.

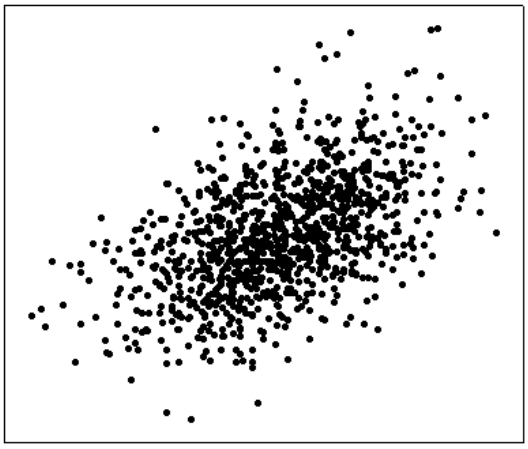
3. Neem twee willekeurige punten $(x_i, y_i)$ en $(x_j, y_j)$ in een puntenwolk die een opwaartse trend vertoont. Bij een grotere $x$-waarde hoort dan ook altijd een grotere $y$-waarde. Als dus $x_j > x_i$ dan moet $y_j > y_i$ . Staaf je reactie op deze uitspraak met een voorbeeld uit de gegeven puntenwolk (werk benaderend).

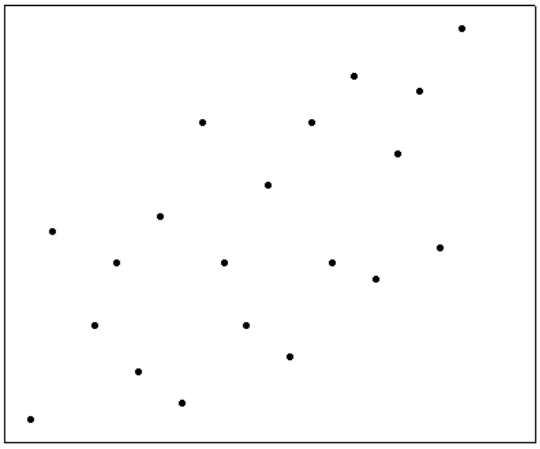
## 3.3. Ellipsvormige puntenwolken

Er zijn twee soorten puntenwolken: puntenwolken waarvan het globale uitzicht ellipsvormig is, en andere.

Als je op de afzonderlijke punten begint te letten, dan zie je in een puntenwolk meestal heel wat variabiliteit. Maar eigenlijk moet je op zoek gaan naar een “globale” vorm, zonder je vast te pinnen op enkele punten die hier en daar wat afwijken (als die afwijking tenminste niet te drastisch is).

Wanneer je besluit dat een puntenwolk er “globaal” ellipsvormig uitziet, dan betekent dit dat de grote meerderheid van de punten willekeurig verstrooid ligt binnen een ellips, zonder een ander uitgesproken patroon te vertonen.



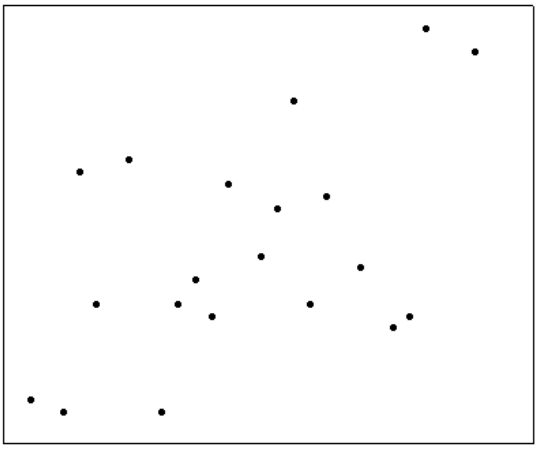


De linkse puntenwolk heeft 1078 punten en in de rechtse staan er maar 20. Toch kan je in beide gevallen zeggen dat het globale beeld ellipsvormig is.

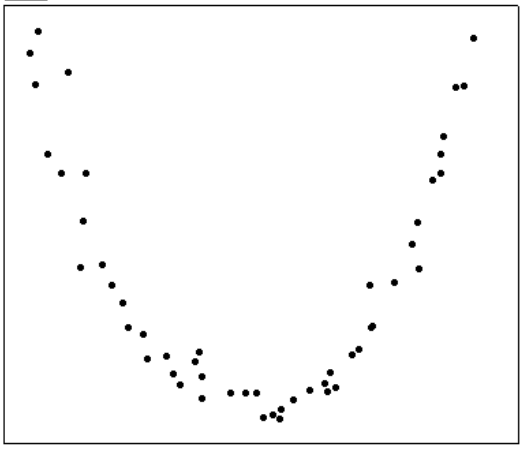
## Opdracht 6

Hieronder zie je 4 puntenwolken: a, b, c en d. Zeg bij elke puntenwolk of zij globaal ellipsvormig is of niet. Als ze niet ellipsvormig is, zeg dan ook wat er volgens jou aan de hand is.

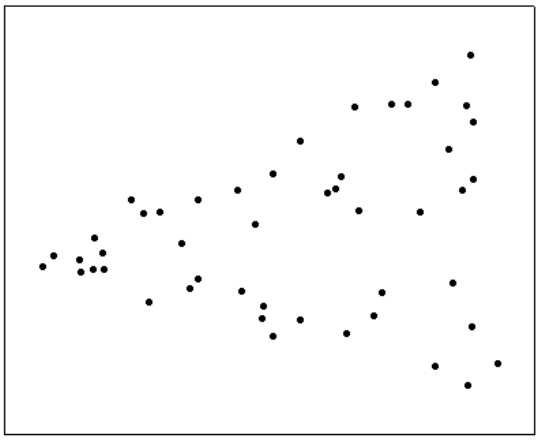
a



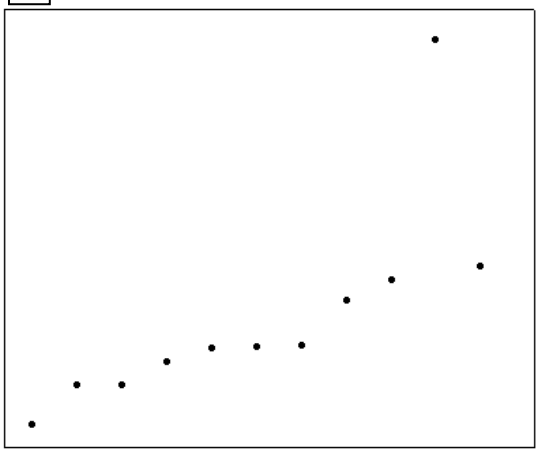
b



c



d

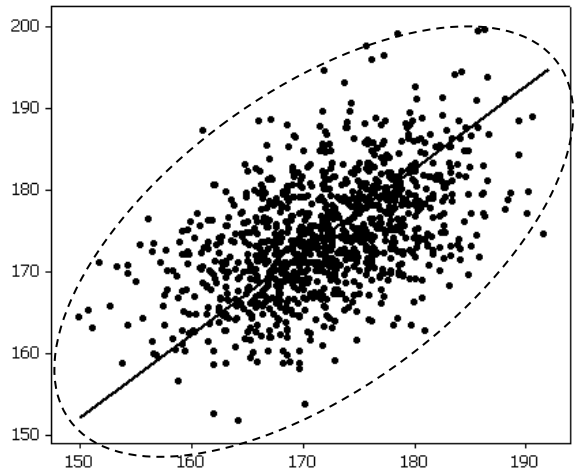


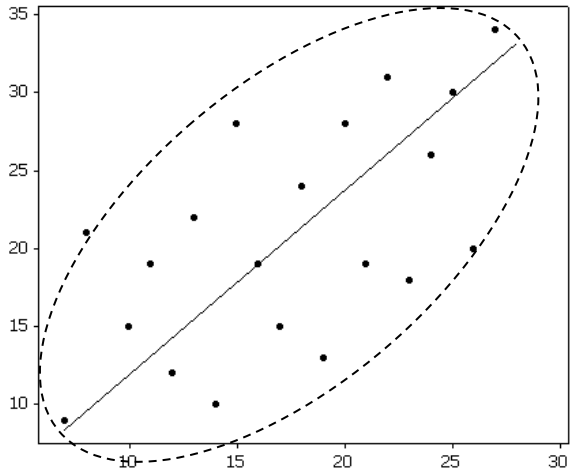
# 4. Lineaire samenhang: een grafische studie

## *4.1. Lineaire samenhang bij ellipsvormige puntenwolken*

Bij een ellipsvormige puntenwolk kan je op zoek gaan naar een rechte waarrond de punten verstrooid liggen. Je probeert dan **een rechte** te tekenen **die zo goed mogelijk aansluit bij de puntenwolk**.

*Op zicht* kom je tot figuren zoals hieronder.



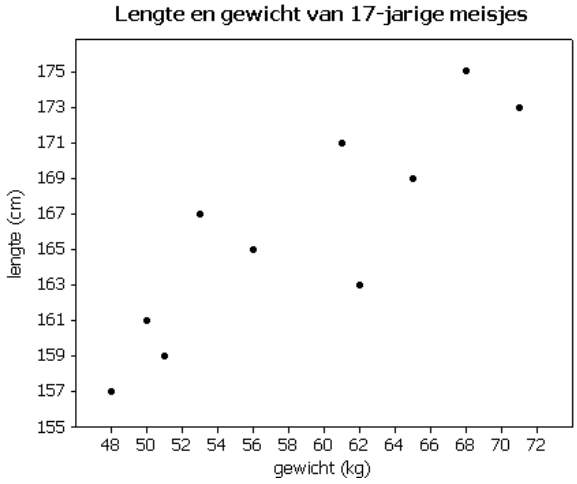




De rechte die zo goed mogelijk aansluit bij de puntenwolk heet **de typische rechte**.

## Opdracht 7

Bij ellipsvormige puntenwolken bestudeer je de manier waarop punten verstrooid zijn rond een typische rechte. Je bestudeert **de lineaire samenhang** van de lengte van moeders en de lengte van dochters, of van de lengte en het gewicht van 17-jarige meisjes. Het gaat hier niet zomaar om een samenhang tussen twee grootheden. Het woord **lineair** zegt dat het hier gaat over "**samenhang ten opzichte van een rechte**".

Teken *op zicht* de typische rechte bij de onderstaande puntenwolken.

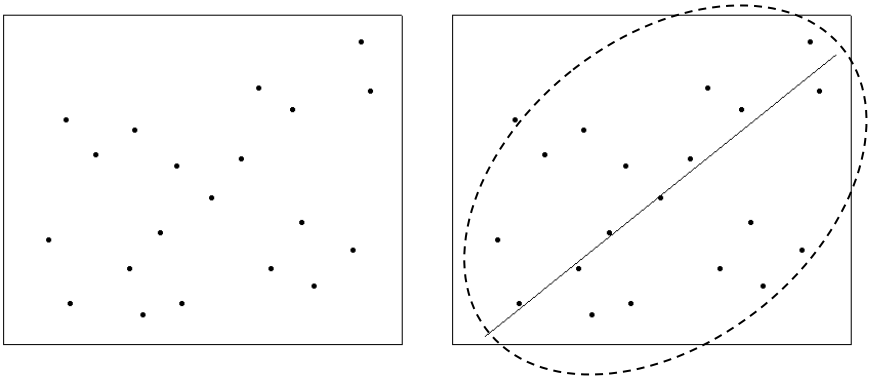



## 4.2. Sterkte en zin van een lineaire samenhang

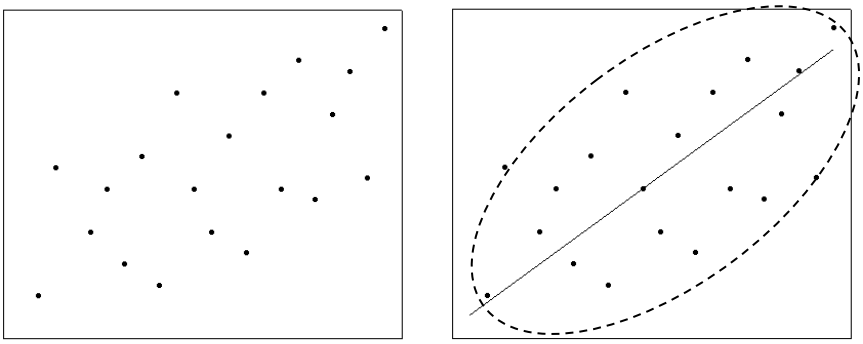
Punten kunnen op veel verschillende manieren rond een rechte verstrooid liggen. Hieronder zie je 3 voorbeelden. Links staat de gewone puntenwolk waarvan je de sterkte van de lineaire samenhang moet beoordelen. Om je te helpen zie je rechts de typische rechte waarrond de punten verstrooid zijn. Je ziet er ook een ellips die ongeveer alle punten probeert te omsluiten.

De **sterkte** van een lineaire samenhang heeft te maken met de manier waarop punten rond de rechte verstrooid liggen: dicht tegen de rechte of met grote spreiding.

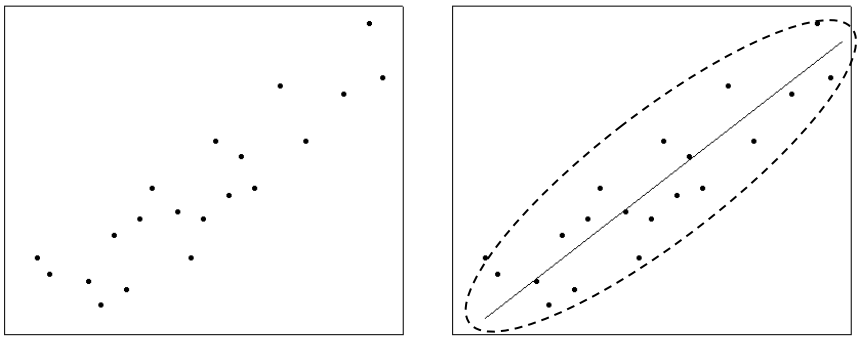
In de puntenwolk hiernaast zie je een grote spreiding en je hebt een brede ellips nodig om de punten te omvatten. Hier spreek je over een **zwakke lineaire samenhang.**



In de puntenwolk hiernaast is er al wat minder spreiding dan zopas. Ook de ellips om de punten te omvatten is smaller. Hier zeg je dat de puntenwolk wijst op een **matige lineaire samenhang**.



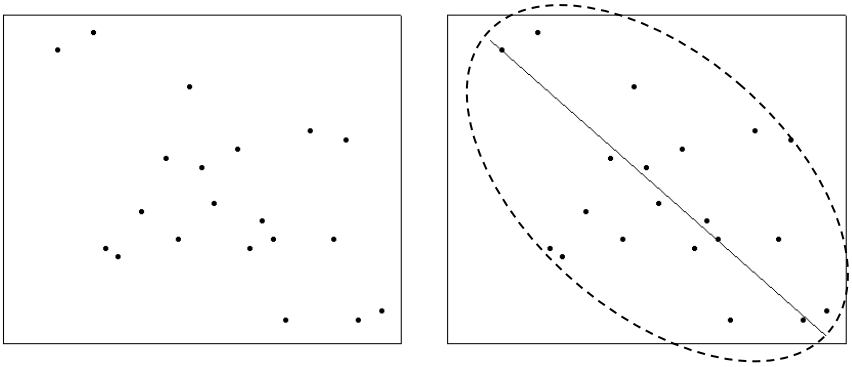
In de puntenwolk hiernaast is er weinig spreiding. De ellips die de punten omvat is smal. Hier zeg je dat de puntenwolk wijst op een **sterke lineaire samenhang**.



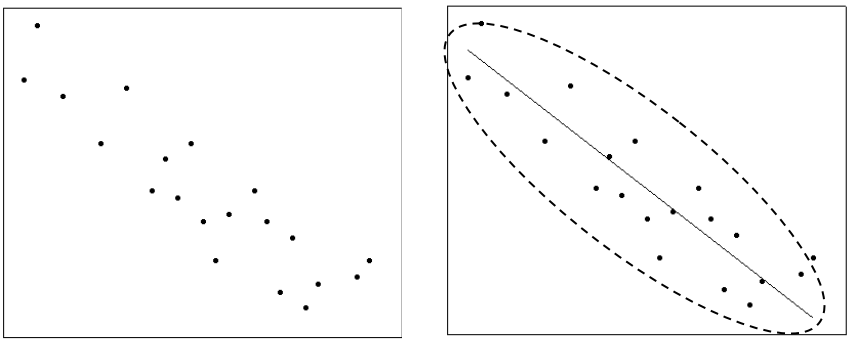
De vorige drie figuren hebben een gemeenschappelijk kenmerk: zij vertonen allemaal een opwaartse trend.

Als de ellipsvormige puntenwolk een opwaartse trend vertoont, dan spreekt men over een **positieve** lineaire samenhang. Bij een neerwaartse trend spreekt men over een **negatieve** lineaire samenhang.

In de puntenwolk hiernaast is **de lineaire samenhang negatief en matig.**

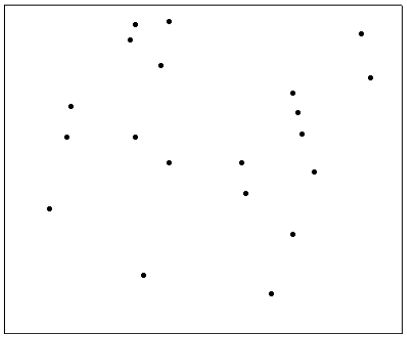


In de puntenwolk hiernaast is **de lineaire samenhang negatief en sterk.**

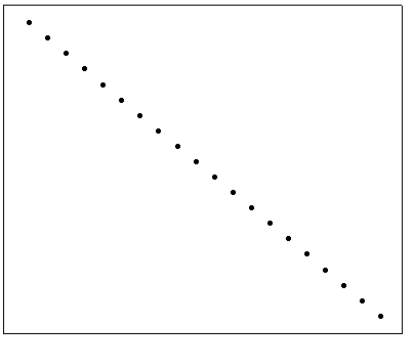


Je hebt nu zowat alle mogelijkheden bekeken voor de studie van ellipsvormige puntenwolken. Er blijven nog **twee extreme gevallen** over, die je beide als een limietsituatie kan opvatten.

Je hebt gezien dat een “zwakke” lineaire samenhang samengaat met een “dikke” ellips. Als die samenhang zwakker en zwakker wordt, dan wordt die ellips dikker en dikker. In het extreme geval wordt die ellips een cirkel en is er in de puntenwolk **geen lineaire samenhang** meer te bespeuren, zoals op het voorbeeld hiernaast.



Een “sterke” lineaire samenhang gaat samen met een “dunne” ellips. Als die samenhang sterker en sterker wordt, dan wordt die ellips dunner en dunner. In het extreme geval wordt die ellips een lijnstuk en vertoont de puntenwolk een **perfecte lineaire samenhang.** Een voorbeeld zie je hiernaast.

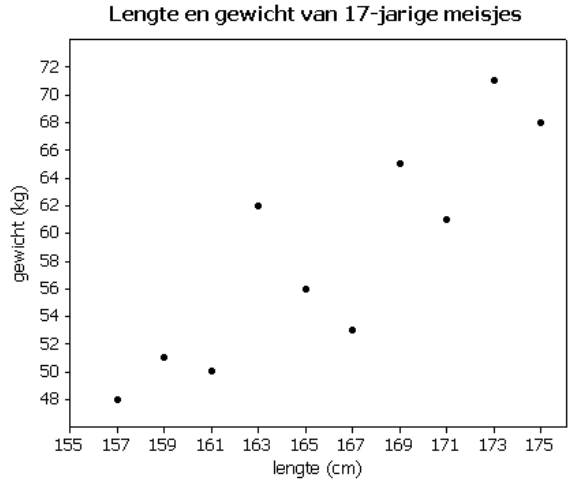


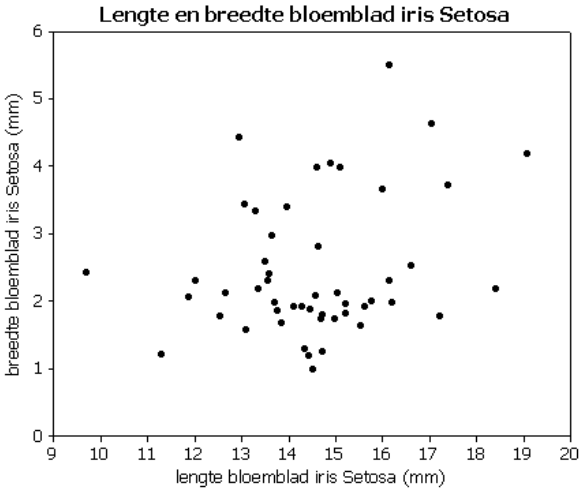
***Let op!***
De “sterkte van de samenhang” heeft niets te maken heeft met de helling (vlakker of steiler) van de typische rechte.

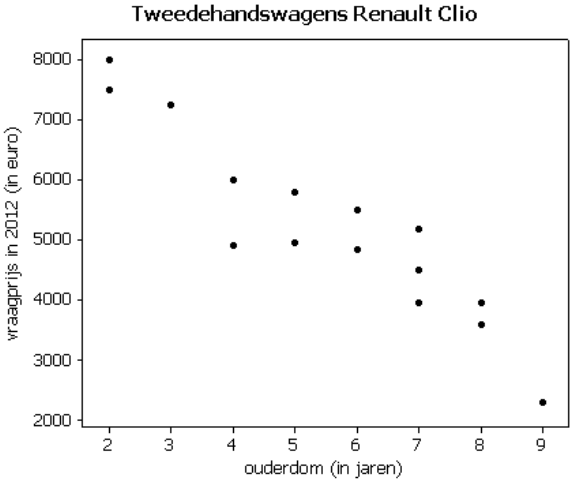
## Opdracht 8

Bij de onderstaande puntenwolken kan je ook aflezen waarover de studie gaat.
Voor elke puntenwolk doe je het volgende:

1. teken *op zicht* de typische rechte
2. teken *benaderend* een ellips die de meerderheid van de punten bevat
3. bespreek de *zin* en de *sterkte* van de puntenwolk
4. zeg in woorden welk verband getoond wordt in de context van de uitgevoerde studie.



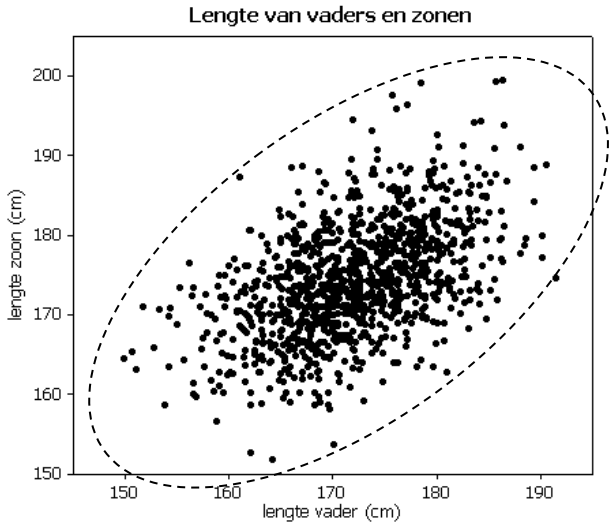



# 5. Lineaire samenhang: een numerieke studie

Bij het kijken naar puntenwolken heb je al een eerste indruk opgedaan over hun vorm en over de sterkte en zin van een lineaire samenhang. Deze grafische studie van puntenwolken is een zeer belangrijke eerste stap bij een statistisch onderzoek. In een tweede stap ga je bepaalde eigenschappen van een puntenwolk in getallen weergeven.

## *5.1. De afzonderlijke coördinaten en hun kengetallen*

In de onderstaande puntenwolk stelt elk punt de lengte voor van een vader en van zijn oudste volwassen zoon. Een deel van de dataset zie je naast de grafiek. De opmetingen zijn in cm.




| Lengte vader $x_i$ | Lengte zoon $y_i$ |
|---|---|
| 180 | 178 |
| 172 | 172 |
| 169 | 160 |
| 179 | 170 |
| 167 | 169 |
| 180 | 193 |
| 177 | 182 |
| 173 | 183 |
| 176 | 186 |
| 157 | 174 |
| …. | …. |

Er zijn 1078 punten $(x_i, y_i)$. Voor elk punt $(x_i, y_i)$ is:

$x_i$ = de lengte van de vader van het i$^{de}$ gezin.

$y_i$ = de lengte van de oudste volwassen zoon van het i$^{de}$ gezin.

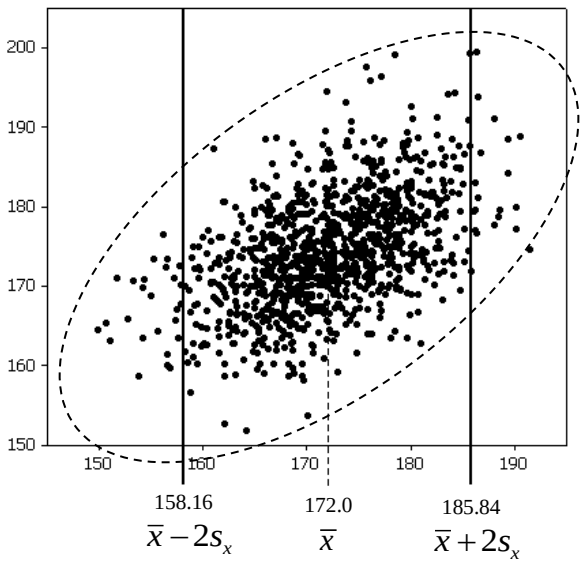
Om te starten kijk je naar de afzonderlijke coördinaten. Die kan je karakteriseren met de klassieke methoden uit de exploratieve statistiek.

Voor de $x$ - coördinaten van dit voorbeeld vind je:

- de gemiddelde lengte $\bar{x}$ van de vaders is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 172.0$ cm
- de standaardafwijking $s_x$ van de lengte van de vaders is $s_x = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2} = 6.92$ cm.

Voor elke verzameling getallen geldt dat “een meerderheid” (minstens 75 %) van die getallen niet verder dan twee standaardafwijkingen van het gemiddelde verwijderd ligt.

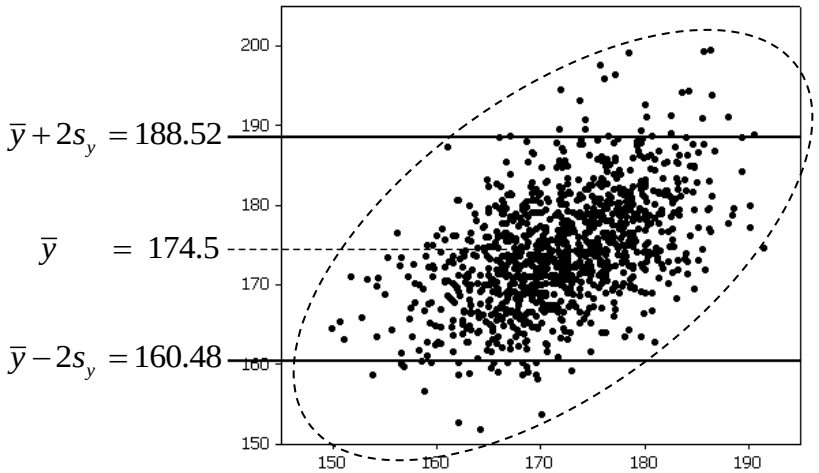
Voor de lengte van de vaders beschik je hier over 1078 $x_i$ –getallen. Bovenstaande eigenschap zegt dat een grote meerderheid van die lengtes in het interval $\left[\bar{x}-2s_x\ ;\ \bar{x}+2s_x\right]$ = [ 158.16 ; 185.84 ] terechtkomt. Je kan dat op de figuur hiernaast goed zien. Het gaat over de punten die gevangen zitten in de verticale strook.




Je kan nu op dezelfde manier tewerk gaan voor de $y$ - coördinaten (dat zijn de lengtes van de zonen).

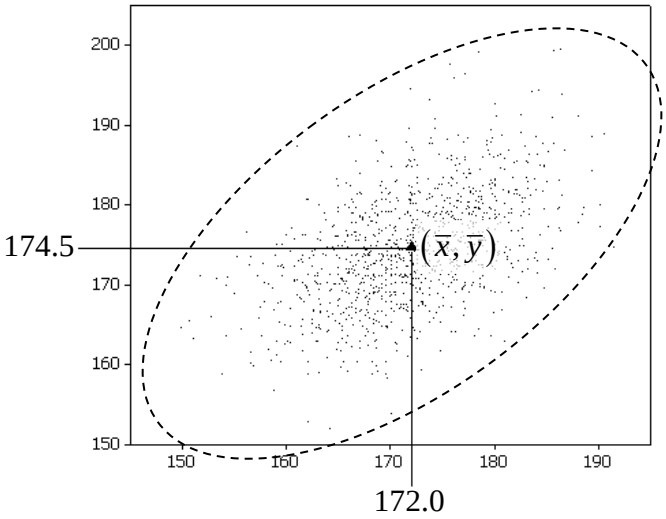
- De gemiddelde lengte $\bar{y}$ van de zonen is $\bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_i = 174.5\,\text{cm}$
- De standaardafwijking $s_y$ van de lengte van de zonen is $s_y=\sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2}=7.01$ cm.

De grote meerderheid van de 1078 $y_i$ –getallen ligt in $\left[\bar{y}-2s_y\ ;\ \bar{y}+2s_y\right]$ = [ 160.48 ; 188.52 ].
Op de figuur hieronder gaat het over alle punten van de puntenwolk die gevangen zitten in de horizontale strook.



Je weet dat $\bar{x}$ een maat is voor het centrum van de $x_i$–getallen (de lengte van de vaders) en dat $\bar{y}$ een maat is voor het centrum van de $y_i$–getallen (de lengte van de zonen). Het zal je dan waarschijnlijk niet verbazen dat $(\bar{x}, \bar{y})$ te maken heeft met het centrum van de puntenwolk. Het punt $(\bar{x}, \bar{y})$ wordt het **zwaartepunt** van de puntenwolk genoemd.

Op de figuur hiernaast is het zwaartepunt $(\bar{x}, \bar{y})$ = (172.0 , 174.5 ) aangeduid met een driehoekje.




## Opdracht 9

Bij 17-jarige meisjes is de lengte (in cm) en het gewicht (in kg) opgemeten met volgend resultaat:

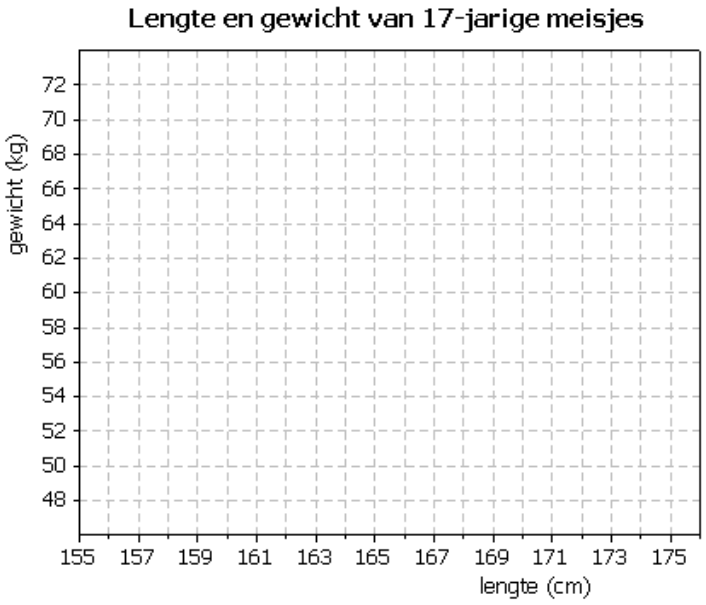
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| lengte $x_i$ | 157 | 159 | 161 | 163 | 165 | 167 | 169 | 171 | 173 | 175 |
| gewicht $y_i$ | 48 | 51 | 50 | 62 | 56 | 53 | 65 | 61 | 71 | 68 |

- Bepaal het gemiddelde en de standaardafwijking voor de lengte en voor het gewicht. Je kan werken met de GRM waarin je de lijsten LGT17 en GEW17 hebt ingebracht (of je kan die getallen met de hand intikken). Druk [STAT], loop naar CALC en druk 2:2-Var Stats. Bij Xlist: duid je LGT17 aan en bij Ylist: GEW17. Loop dan naar Calculate en druk [ENTER].

| | | |
|---|---|---|
| lengte $x_i$ | $\bar{x} =$ | $s_x =$ |
| gewicht $y_i$ | $\bar{y} =$ | $s_y =$ |

- Bij een dataset ligt minstens 75 % (en soms veel meer) van de opmetingen in het interval [ gemiddelde – twee standaardafwijkingen ; gemiddelde + twee standaardafwijkingen ]. Bepaal voor de lengtes het interval $[\bar{x} - 2s_x \,;\ \bar{x} + 2s_x]$. Hoeveel percent van de $x_i$ – getallen ligt in dat interval?

- Teken hieronder de puntenwolk. Bepaal haar zwaartepunt en teken het zwaartepunt op de grafiek met een driehoekje.




## 5.2. Univariate en bivariate informatie

Als je bij 17-jarige meisjes de lengte (in cm) en het gewicht (in kg) opmeet, dan heb je, per meisje, een bivariate opmeting: $(x_i, y_i)$ met $x_i$ de lengte en $y_i$ het gewicht van het i$^{de}$ meisje.

| lengte $x_i$ | 157 | 159 | 161 | 163 | 165 | 167 | 169 | 171 | 173 | 175 |
|---|---|---|---|---|---|---|---|---|---|---|
| gewicht $y_i$ | 48 | 51 | 50 | 62 | 56 | 53 | 65 | 61 | 71 | 68 |

Uit bivariate gegevens kan je univariate gegevens halen. Je kent de lengte van die 10 meisjes en je kent ook hun gewicht. Die univariate kenmerken kan je afzonderlijk bestuderen. Voor de lengte vond je $\bar{x} = 166$ cm, $s_x = 6.06$ cm en voor het gewicht $\bar{y} = 58.5$ kg, $s_y = 8.05$ kg.

Met univariate gegevens kan je, zonder bijkomende informatie, geen bivariate gegevens opstellen.

Als je alleen maar weet dat de opgemeten lengtes er uitzien als:

| lengte $x_i$ | 157 | 159 | 161 | 163 | 165 | 167 | 169 | 171 | 173 | 175 |
|---|---|---|---|---|---|---|---|---|---|---|

en de gewichten als

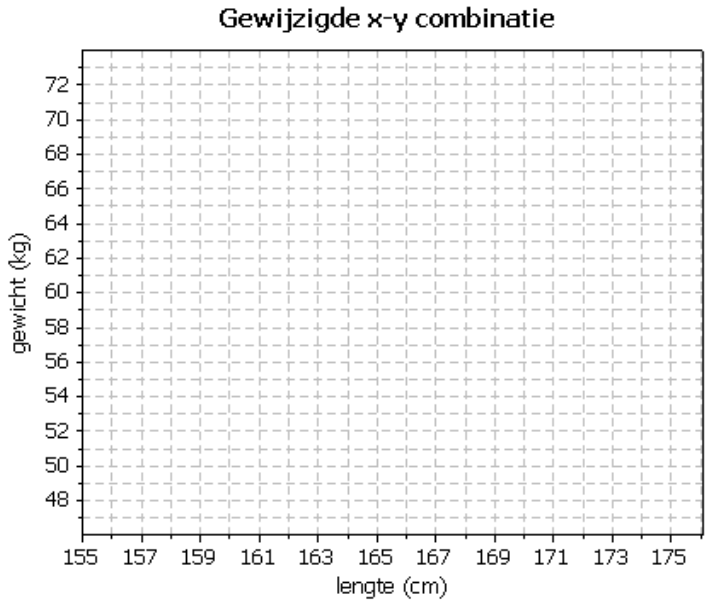
| gewicht $y_i$ | 48 | 51 | 50 | 62 | 56 | 53 | 65 | 61 | 71 | 68 |
|---|---|---|---|---|---|---|---|---|---|---|

dan is altijd $\bar{x} = 166$ cm, $s_x = 6.06$ cm en $\bar{y} = 58.5$ kg, $s_y = 8.05$ kg. Maar met de afzonderlijke $x_i$'s en $y_i$'s kan je veel verschillende combinaties $(x_i, y_i)$ maken. In de volgende opdracht werk je met dezelfde $x_i$'s en dezelfde $y_i$'s en toch krijg je een volledig andere puntenwolk.

Opdracht 10

Teken de puntenwolk voor

| lengte $x_i$ | 157 | 159 | 161 | 163 | 165 | 167 | 169 | 171 | 173 | 175 |
|---|---|---|---|---|---|---|---|---|---|---|
| gewicht $y_i$ | 61 | 56 | 65 | 48 | 71 | 50 | 68 | 62 | 51 | 53 |



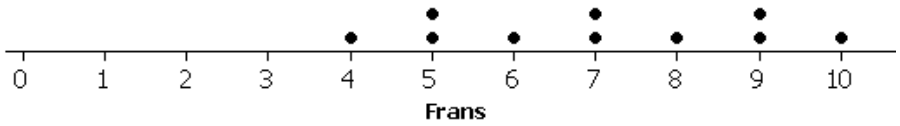

## 5.3. Een aangepaste meetlat

Opdracht 11

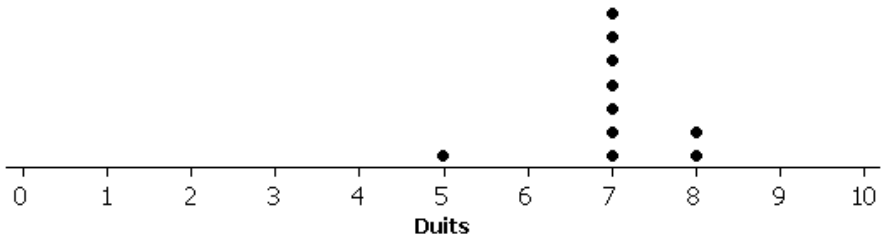
1. Pol heeft op de toets Frans 5 op 10 gehaald. Op de toets Duits haalde Pol ook 5 op 10.
   Besluit: De prestatie van Pol was twee keer dezelfde.
   Ben je akkoord met deze uitspraak? Motiveer je antwoord.

2. Pol heeft op de toets Frans 5 op 10 gehaald. Het klasgemiddelde was 7 op 10. Op de toets Duits haalde Pol ook 5 op 10 en ook op die toets was het gemiddelde van de klas 7 op 10.
   Besluit: De prestatie van Pol was twee keer dezelfde.
   Ben je akkoord met deze uitspraak? Motiveer je antwoord.

Het verhaal over de resultaten van Pol krijgt een heel andere wending als je er een eenvoudig puntendiagram bij tekent.




Bij de toets Frans liggen de punten nogal gespreid. Twee leerlingen haalden een 5, er was ook een leerling met een 4 maar er waren er ook met 9 en 10. Voor de punten van die 10 leerlingen is het gemiddelde 7 en de standaardafwijking is 2.




De toets Duits ziet er helemaal anders uit. Iedereen haalde daar een 7 of een 8, behalve…. Pol, die had een 5. Bij deze toets is het gemiddelde 7 en de standaardafwijking is 0.8.

Een getal uit een dataset zomaar vergelijken met het gemiddelde vertelt niet het hele verhaal. Soms geeft dit zelfs een verkeerd beeld. De variabiliteit rond dat gemiddelde speelt ook een rol. Bij Frans behaalde Pol een score die 1 standaardafwijking onder het gemiddelde ligt, want $5 = 7 - (1)\times(2)$. Bij Duits scoorde Pol 2.5 standaardafwijkingen onder het gemiddelde want $5 = 7 - (2.5)\times(0.8)$.

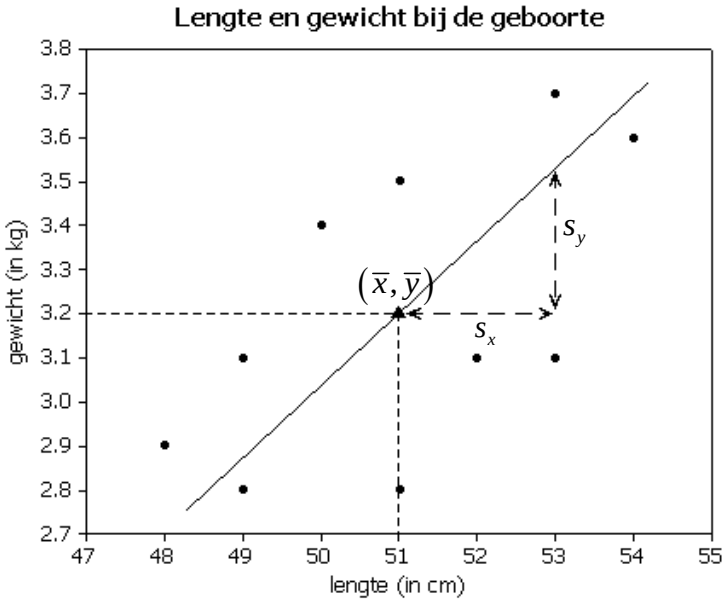
**De standaardafwijking van een dataset is dikwijls een goede meetlat om punten uit die dataset te vergelijken met hun gemiddelde**. Zo houd je ook rekening met de variabiliteit van de gegevens.

Als je de standaardafwijking als meetlat neemt dan heeft Pol "–1" op Frans en "–2.5" op Duits. In vergelijking met zijn medeleerlingen is zijn prestatie op Duits veel lager dan op Frans.

*Frans*: score van Pol = gemiddelde **– 1** standaardafwijking
*Duits*: score van Pol = gemiddelde **– 2.5** standaardafwijkingen.

## 5.4. De typische rechte

Hiernaast zie je de typische rechte waarrond de punten van een ellipsvormige puntenwolk verspreid liggen. We trachten nu de vergelijking ervan te achterhalen.




Het zal je niet verwonderen dat die rechte door **het zwaartepunt** $(\bar{x},\bar{y})$ van de puntenwolk loopt. Hiermee heb je al een eerste karakteristiek van de $x_i$'s en de $y_i$'s gebruikt: hun gemiddelde.

Nu ga je een tweede karakteristiek van de $x_i$'s en de $y_i$'s gebruiken: hun standaardafwijking. Start in het zwaartepunt. Ga in de x-richting (horizontaal) naar rechts tot aan een x-waarde die één standaardafwijking $s_x$ verwijderd ligt van het gemiddelde $\bar{x}$. Vanaf die plaats ga je in de y-richting (vertikaal) naar boven tot aan een y-waarde die één standaardafwijking $s_y$ verwijderd ligt van het gemiddelde $\bar{y}$. Het punt waar je nu staat is een tweede punt van de typische rechte.

De rechte die je zo gemaakt hebt gaat door het punt $(\bar{x},\bar{y})$ en heeft een richtingscoëfficiënt die gelijk is aan $\frac{s_y}{s_x}$. Deze manier van werken geldt voor puntenwolken waarbij de zin van de lineaire samenhang positief is. Als de zin negatief is, dan verander je het teken van de richtingscoëfficiënt.

De vergelijking van de typische rechte is:

- voor een positieve lineaire samenhang:

$$(y-\bar{y})=\frac{s_y}{s_x}(x-\bar{x}) \quad of \quad \frac{y-\bar{y}}{s_y}=\frac{x-\bar{x}}{s_x} \quad of \quad y=\frac{s_y}{s_x}x+\bar{y}-\frac{s_y}{s_x}\bar{x}$$

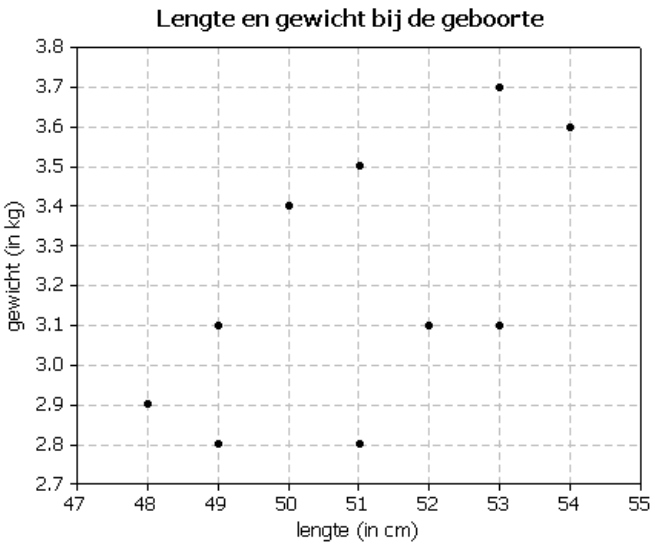
- voor een negatieve lineaire samenhang:

$$(y-\bar{y})=-\frac{s_y}{s_x}(x-\bar{x}) \quad of \quad \frac{y-\bar{y}}{s_y}=-\frac{x-\bar{x}}{s_x} \quad of \quad y=-\frac{s_y}{s_x}x+\bar{y}+\frac{s_y}{s_x}\bar{x}$$

## Opdracht 12

In opdracht 2 heb je de lengte en het gewicht van 10 baby's bepaald samen met de puntenwolk.

| Lengte (in cm) $x_i$ | Gewicht (in kg) $y_i$ |
|---|---|
| 48 | 2.9 |
| 49 | 2.8 |
| 49 | 3.1 |
| 50 | 3.4 |
| 51 | 2.8 |
| 51 | 3.5 |
| 52 | 3.1 |
| 53 | 3.1 |
| 53 | 3.7 |
| 54 | 3.6 |

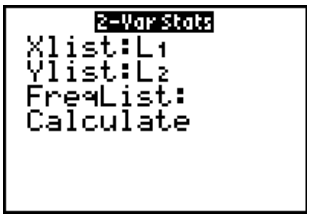



Geef de vergelijking van de typische rechte waarrond de punten verstrooid liggen en teken die rechte op de grafiek. Je kan hierbij gebruik maken van de GRM op twee manieren:

- de kengetallen die je nodig hebt, bereken je met de GRM en dan gebruik je die kengetallen in de vergelijking van de typische rechte.
- je werkt met het programma CORR en je leest de vergelijking van de typische rechte af.

In beide gevallen moet je de bestanden LBABY.8xl (lengte baby) en GBABY.8xl (gewicht baby) downloaden en in je GRM inbrengen als lijsten LBABY en GBABY. Die lijsten kan je best kopiëren naar [L1] en [L2]. Druk [2nd] [LIST], loop naar LBABY en druk [ENTER]. Vul het commando als volgt aan: druk [STO▸] en [2nd] [L1] en [ENTER]. Voor het gewicht werk je op analoge manier. Druk [2nd] [LIST], loop naar GBABY en druk [ENTER]. Vervolledig het commando: druk [STO▸] en [2nd] [L2] en [ENTER].

***Eerste manier: de vergelijking van de typische rechte opstellen.***




Om gemiddelden en standaardafwijkingen te bepalen doe je het volgende. Druk [STAT], loop naar CALC en druk 2:2-Var Stats. Bij Xlist: duid je [L1] aan en bij Ylist: [L2] . Loop dan naar Calculate en druk [ENTER]. Je kan nu alles aflezen wat je nodig hebt.

| | | |
|---|---|---|
| lengte $x_i$ | $\bar{x} =$ | $s_x =$ |
| gewicht $y_i$ | $\bar{y} =$ | $s_y =$ |

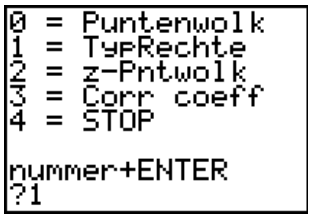
Bereken nu de vergelijking van de typische rechte (rond af op 2 decimalen). Teken de gevonden rechte op de gegeven grafiek. Heb je dezelfde figuur als hierboven?

Typische rechte: y = …

***Tweede manier: de vergelijking van de typische rechte aflezen met het programma CORR.***

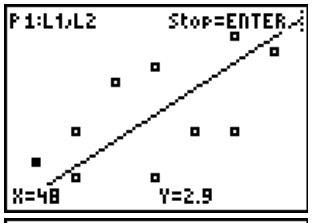
Download het bestand CORR.8xp en breng het in je GRM als programma CORR (studie van de CORRelatie). Om dit programma te kunnen gebruiken moeten de $x_i$ – getallen in de lijst [L1] staan en de bijhorende $y_i$ - getallen in de lijst [L2]. Dat heb je zopas al gedaan.

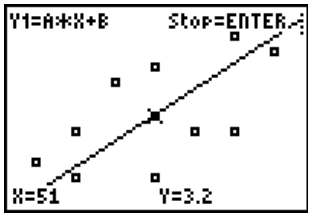
Druk [PRGM], loop naar CORR en druk 3 keer [ENTER]. Maak dan gebruik van het keuzemenu (tik het nummer 1 en druk [ENTER]). De typische rechte verschijnt onder de vorm $y = ax + b$. Je kan nu $a$ en $b$ invullen.


Schrijf nu de vergelijking van de typische rechte en teken ze op de gegeven grafiek. Heb je dezelfde figuur als hierboven?

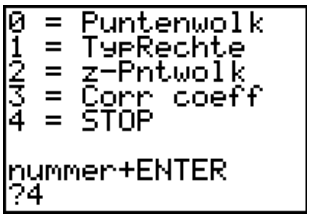
Typische rechte: y = ……




Het programma CORR tekent ook de typische rechte. Druk [ENTER] voor de grafiek. Je ziet nu de puntenwolk in de oorspronkelijke eenheden samen met de typische rechte. Met de pijltjes [▲] en [▼] verwissel je tussen puntenwolk (linksboven staat dan P1:$L_1$,$L_2$ ) en rechte (linksboven staat dan Y1=A*X+B). Op beide grafieken kan je rondlopen met [◀] en [▶]. Onderaan staan telkens de coördinaten van het punt waarop je staat. Druk [ENTER] om de figuur te verlaten.

4 = STOP: tik 4 en druk [ENTER] om het programma te verlaten.




Opdracht 13

In opdracht 9 heb je voor de lengte en het gewicht van 17-jarige meisjes een puntenwolk getekend. Teken op die figuur nu ook de typische rechte. Schrijf eerst de vergelijking van die rechte hieronder. Je kan gebruik maken van de kengetallen die je in opdracht 9 hebt berekend. Je kan anderzijds ook werken met het programma CORR als je er eerst voor zorgt dat de lijsten LGT17 en GEW17 in [L1] en [L2] staan.

Typische rechte: y = ……

# 6. Correlatie

## 6.1. *Verstrooiing rond de typische rechte*

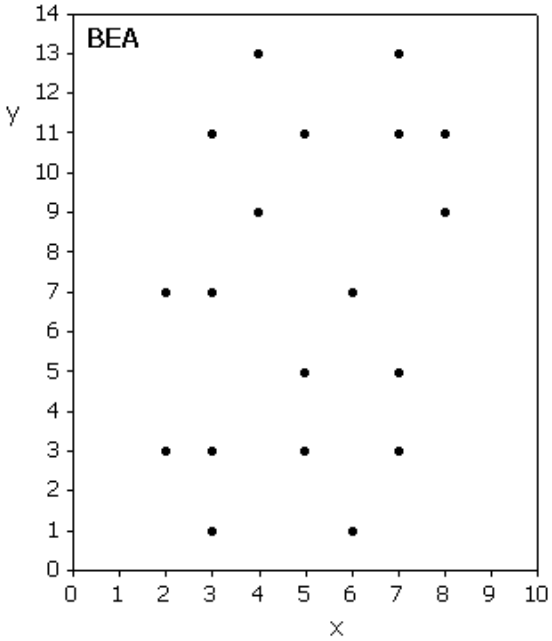
Opdracht 14

Bepaal voor de onderstaande datasets van Bea, Jan en Pol de kengetallen $\bar{x}$, $s_x$, $\bar{y}$ en $s_y$. Schrijf ook de vergelijking van de typische rechte en teken die rechte bij de gepaste puntenwolk.

Je kan de data zelf intikken in je GRM of je kan de lijsten XBEA, YBEA, XJAN, YJAN, XPOL en YPOL gebruiken.

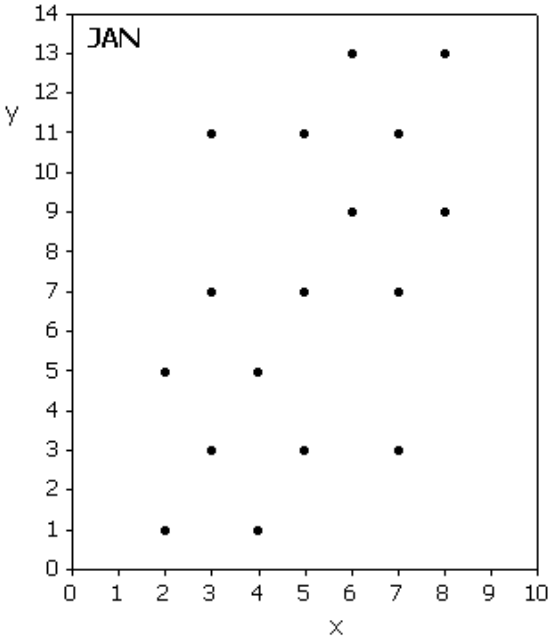
| Bea | |
|---|---|
| $x_i$ | $y_i$ |
| 2 | 3 |
| 2 | 7 |
| 3 | 1 |
| 3 | 7 |
| 3 | 11 |
| 3 | 3 |
| 4 | 13 |
| 4 | 9 |
| 5 | 3 |
| 5 | 11 |
| 5 | 5 |
| 6 | 7 |
| 6 | 1 |
| 7 | 5 |
| 7 | 3 |
| 7 | 11 |
| 7 | 13 |
| 8 | 11 |
| 8 | 9 |

| | | | |
|---|---|---|---|
| Bea | $\bar{x} =$ | $s_x =$ | Typische rechte: |
| | $\bar{y} =$ | $s_y =$ | y = |



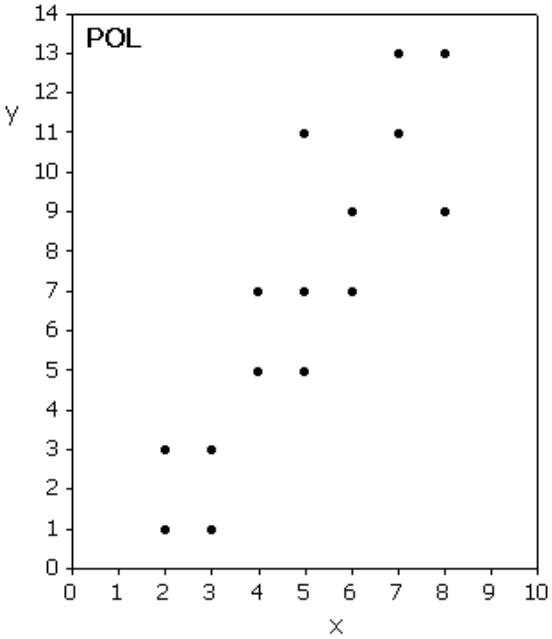
| Jan | |
|---|---|
| $x_i$ | $y_i$ |
| 2 | 5 |
| 2 | 1 |
| 3 | 7 |
| 3 | 11 |
| 3 | 3 |
| 4 | 1 |
| 4 | 5 |
| 5 | 7 |
| 5 | 3 |
| 5 | 11 |
| 6 | 9 |
| 6 | 13 |
| 7 | 7 |
| 7 | 3 |
| 7 | 11 |
| 8 | 13 |
| 8 | 9 |

| | | | |
|---|---|---|---|
| Jan | $\bar{x} =$ | $s_x =$ | Typische rechte: |
| | $\bar{y} =$ | $s_y =$ | y = |



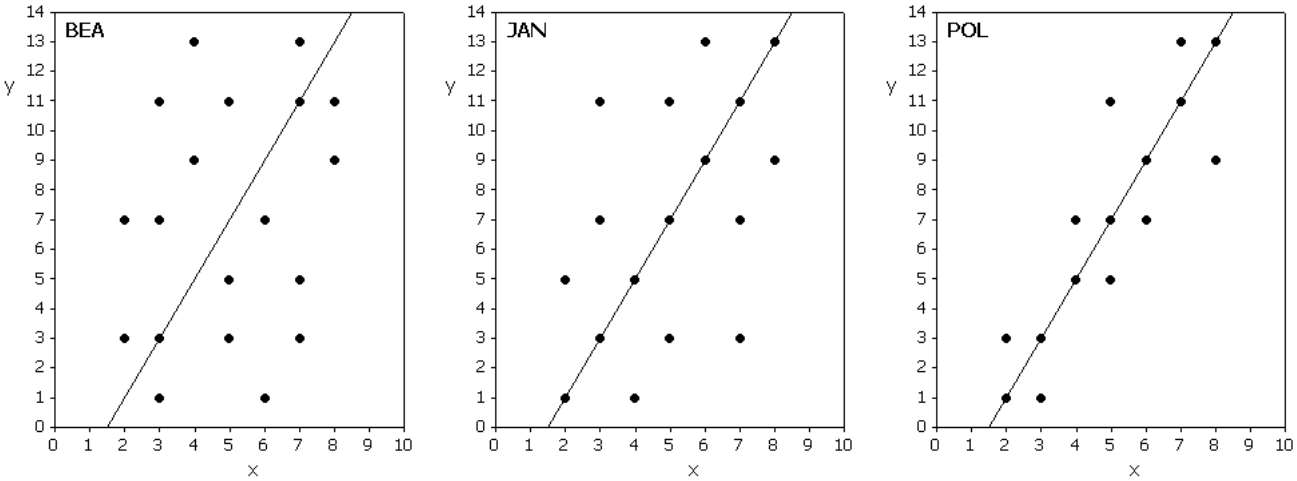

| Pol | |
|---|---|
| $x_i$ | $y_i$ |
| 2 | 3 |
| 2 | 1 |
| 3 | 1 |
| 3 | 3 |
| 4 | 7 |
| 4 | 5 |
| 5 | 7 |
| 5 | 5 |
| 5 | 11 |
| 6 | 9 |
| 6 | 7 |
| 7 | 13 |
| 7 | 11 |
| 8 | 13 |
| 8 | 9 |

| | | | |
|---|---|---|---|
| Pol | $\bar{x} =$ | $s_x =$ | Typische rechte: |
| | $\bar{y} =$ | $s_y =$ | y = |



De opmetingen van Bea, Jan en Pol zijn niet identiek maar zij hebben identieke kengetallen $\bar{x}$, $s_x$, $\bar{y}$ en $s_y$. Bovendien kan je die opmetingen voorstellen door ellipsvormige puntenwolken, waarbij de punten verstrooid liggen rond een typische rechte. Ook de typische rechte is in de drie gevallen dezelfde want de typische rechte hangt alleen af van de kengetallen.

Als je naar de puntenwolken kijkt (zij worden hieronder nog eens getoond, naast elkaar), dan zie je een verschil. Je hebt dat vroeger bij de grafische studie van de lineaire samenhang ook al ontdekt: **de sterkte** van de samenhang verschilt.




Bij Bea heb je een “dikke” ellips nodig om de punten te omvatten want zij liggen ver verstrooid rond de typische rechte. De lineaire samenhang is hier zwak.

Bij Jan liggen de punten al wat dichter tegen de typische rechte. Hier heb je een matige lineaire samenhang.

Bij Pol liggen de punten binnen een “smalle” ellips. De lineaire samenhang is hier sterk.

**Op zicht** zie je verschil tussen zwak, matig of sterk. Maar kan je **de sterkte** van de lineaire samenhang ook **in een getal** uitdrukken? Als je dat wil, dan kan je best eerst standaardiseren.

## 6.2. Gestandaardiseerde puntenwolken

Elke puntenwolk kan je transformeren naar een gestandaardiseerde puntenwolk. Dat heeft twee voordelen:

- **grafisch**: je kan puntenwolken goed *op zicht* met elkaar vergelijken omdat de keuze van de eenheden (meter, centimeter,…) geen rol meer speelt.
- **numeriek**: je kan inzien hoe men tot een getal kan komen dat de sterkte van de lineaire samenhang karakteriseert.

### 6.2.1. z–scores

De transformatie naar een gestandaardiseerde puntenwolk gaat als volgt:

– Elke puntenwolk heeft een zwaartepunt. Dat is het centrum van de puntenwolk en dat neem je als oorsprong van een nieuw assenstelsel.
– Het oorspronkelijke punt $(x_i, y_i)$ komt in het nieuwe assenstelsel terecht op het gestandaardiseerde punt $\left(\frac{x_i - \overline{x}}{s_x}, \frac{y_i - \overline{y}}{s_y}\right)$.

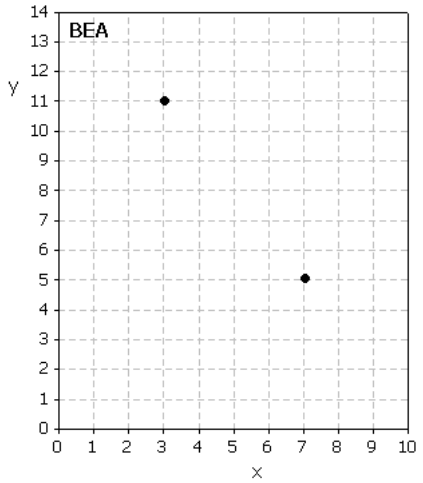
Overstappen van $x_i$ naar $\frac{x_i - \overline{x}}{s_x}$ noemt men overstappen op de z-score $z_{x_i}$ van $x_i$ met $z_{x_i} = \frac{x_i - \overline{x}}{s_x}$.

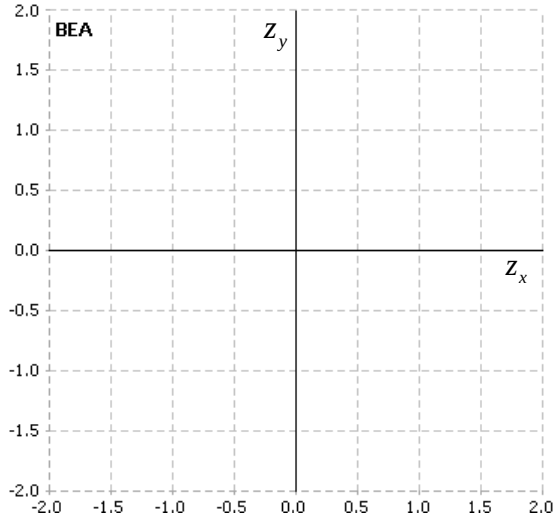
Als bijvoorbeeld $\overline{x} = 5$ en $s_x = 2$ dan heb je voor een oorspronkelijke waarde $x_i = 8$ dat de z-score van $x_i$ gelijk is aan $\frac{x_i - \overline{x}}{s_x} = \frac{8-5}{2} = 1.5$. Herken je dit? Even herschrijven geeft $x_i = \overline{x} + 1.5 s_x$. Dit betekent dat $x_i$ op 1.5 standaardafwijkingen voorbij het gemiddelde ligt. Als je dus de standaardafwijking $s_x$ als nieuwe meerlat neemt en $\overline{x}$ als oorsprong, dan krijgt de oorspronkelijke waarde $x_i = 8$ de waarde $z_{x_i}$ = 1.5 in het nieuwe assenstelsel.

Op eenzelfde manier stap je over van de oorspronkelijke $y_i$-waarden op hun z-score $z_{y_i} = \frac{y_i - \overline{y}}{s_y}$.

## Opdracht 15

Op de linkerfiguur zie je twee punten van de puntenwolk van Bea. Transformeer deze punten (bereken de z-score van hun coördinaten) en teken ze in het nieuwe gestandaardiseerde assenstelsel rechts.




| oude coördinaten | | nieuwe coördinaten | |
|---|---|---|---|
| $x_i$ | $y_i$ | $z_{x_i}$ | $z_{y_i}$ |
| | | | |
| | | | |

Bij elke $x_i$ hoort een z-score $z_{x_i}$ met $z_{x_i} = \dfrac{x_i - \overline{x}}{s_x}$. Bij elke $y_i$ hoort een z-score $z_{y_i}$ met $z_{y_i} = \dfrac{y_i - \overline{y}}{s_y}$.

Bemerk dat een z-score eenheidsloos is omdat je een quotiënt maakt waarbij de grootheid in de teller dezelfde eenheid heeft als de grootheid in de noemer. Door het quotiënt te maken vallen die eenheden weg.

**Elke bivariate dataset kan je standaardiseren door over te stappen op z-scores**. Als je dan een vaste afstand (op je papier of computerscherm) kiest van bijvoorbeeld 2 cm per eenheid, zowel op de horizontale als op de verticale as, dan krijg je "gestandaardiseerde puntenwolken" die je "fysisch" op elkaar kan leggen om ze te vergelijken. Of de oorspronkelijke opmetingen dan over gewicht, lengte, temperatuur, examenpunten of wat dan ook gaan, dat heeft allemaal geen belang. De corresponderende z-scores zijn eenheidsloos.

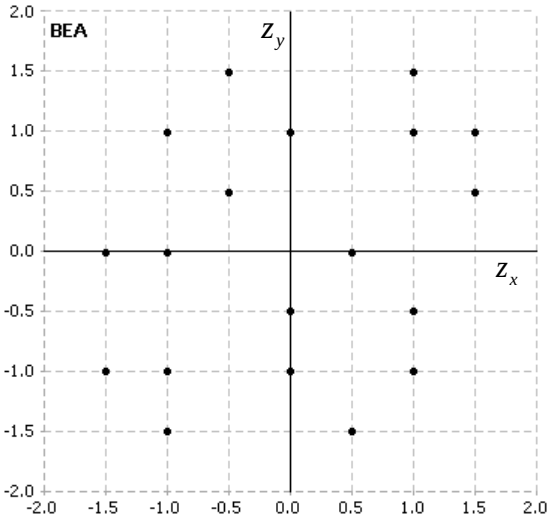
**Verband tussen oorspronkelijke gegevens en z-scores**: $z_x = \dfrac{x - \overline{x}}{s_x}$ *en* $z_y = \dfrac{y - \overline{y}}{s_y}$

zodat ook: $x = \overline{x} + z_x \cdot s_x$ *en* $y = \overline{y} + z_y \cdot s_y$

## Opdracht 16

Bereken voor de gegevens van Bea de getransformeerde typische rechte door in de oude vergelijking $y$ te vervangen door $\bar{y} + z_y \cdot s_y$ en $x$ door $\bar{x} + z_x \cdot s_x$ . Schrijf de nieuwe vergelijking expliciet op in de juiste notatie ( $z_y = f\left(z_x\right)$ ) en teken dan die typische rechte bij de gestandaardiseerde puntenwolk hieronder.

| Typische rechte | |
|---|---|
| oude coördinaten | $y$ = ..... |
| nieuwe coördinaten | ..... = .....<br><br>$z_y$ = ..... |



Voor de gestandaardiseerde puntenwolk van Bea heb je gevonden dat de punten verstrooid liggen rond de typische rechte $z_y = z_x$. Dat is de eerste bissectrice in het $(z_x, z_y)$–vlak.

Dit is geen toeval.

- Je weet dat de typische rechte door het zwaartepunt van de puntenwolk gaat. In het nieuwe assenstelsel is dat zwaartepunt de oorsprong en dus gaat de typische rechte daar altijd door de oorsprong.

- Een tweede punt van de rechte krijg je als je vanuit het zwaartepunt een afstand $s_x$ horizontaal naar rechts gaat en dan een afstand $s_y$ vertikaal naar boven. In het nieuwe assenstelsel komt een horizontale stap naar rechts van grootte 1 overeen met één standaardafwijking $s_x$ in het oude assenstelsel. Je moet daar dus een stap van grootte 1 naar rechts zetten. Op dezelfde manier komt een verticale stap naar boven van grootte 1 in het nieuwe assenstelsel overeen met één standaardafwijking $s_y$ in het oude assenstelsel. Alles samen kom je in het nieuwe assenstelsel altijd terecht op het punt $(1,1)$. De rechte door $(0,0)$ en $(1,1)$ is de eerste bissectrice.

Voor **alle gestandaardiseerde puntenwolken** in het $(z_x, z_y)$–**vlak** geldt dat de typische rechte gelijk is aan:

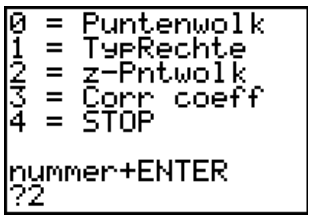
- de **eerste bissectrice** $z_y = z_x$ bij een **positieve** lineaire samenhang
- de **tweede bissectrice** $z_y = -z_x$ bij een **negatieve** lineaire samenhang

### 6.2.2. Standaardiseren met de GRM

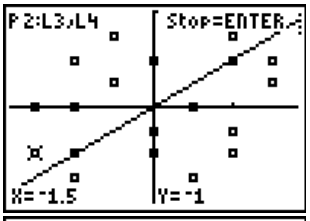
Opdracht 17

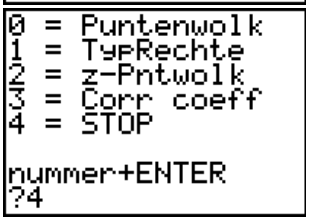
Zorg ervoor dat de lijsten XBEA en YBEA in [L1] en [L2] staan.

Druk [PRGM], loop naar CORR en druk 3 keer [ENTER] zodat je in het keuzemenu belandt. Tik nu 2 en [ENTER].



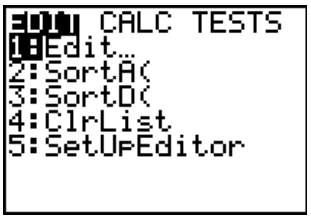

2 = z-Pntwolk: het programma tekent de gestandaardiseerde puntenwolk (z-scores) samen met de typische rechte die hier altijd een bissectrice is (de eerste bij een positieve samenhang en de tweede bij een negatieve samenhang). Met de pijltjes [▲] en [▼] verwissel je tussen puntenwolk (linksboven staat dan P2:$L_3$,$L_4$ ) en rechte (linksboven staat dan Y2=C*X). Op beide grafieken kan je rondlopen met [◄] en [►]. Onderaan staan telkens de coördinaten van het punt waarop je staat. Druk [ENTER] om de figuur te verlaten.

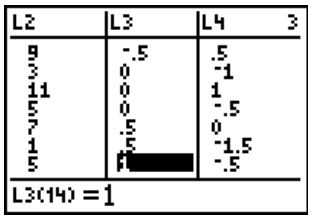



Tik 4 en druk [ENTER] om het programma te verlaten.

De z-scores staan in [L3] en [L4]. Druk [STAT] en 1:Edit… In [L3] staan de z-scores $z_{x_i}$ en in [L4] staan de z-scores $z_{y_i}$. Loop nu naar rij 14. Daar zie je de z-scores $z_{x_{14}} = 1$ en $z_{y_{14}} = -0.5$. Die heb je berekend in opdracht 15. Druk [2nd] [QUIT].

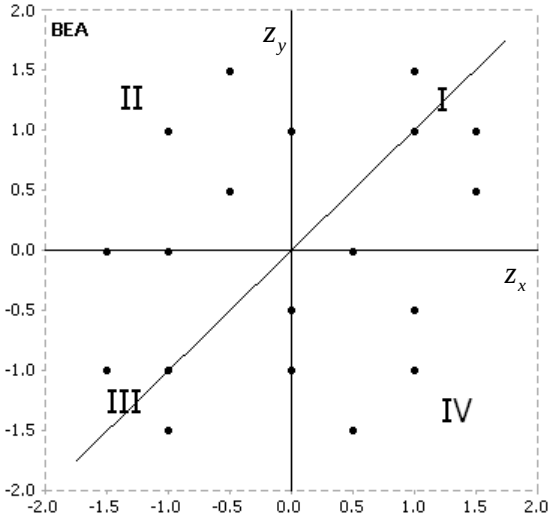


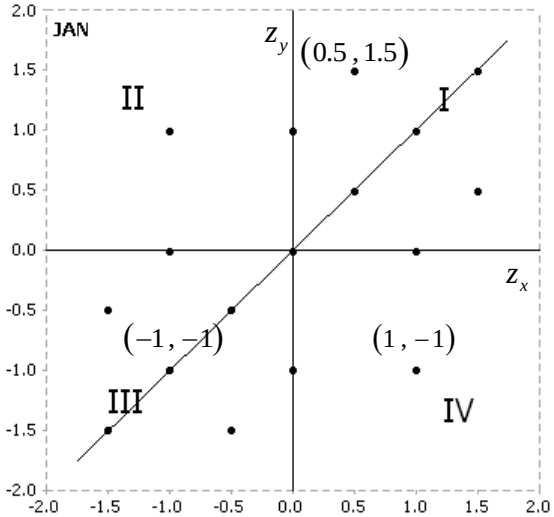



## 6.3. De correlatiecoëfficiënt

### 6.3.1. De ideeën achter de formule

Vergelijk de puntenwolk van Bea met die van Jan. Om dat goed te doen werk je gestandaardiseerd, met z–scores. Bemerk dat het hier in beide voorbeelden gaat om een **positieve** samenhang.



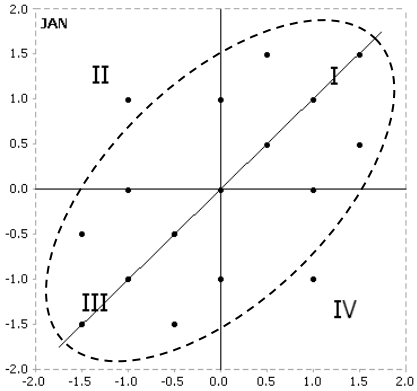




In het $(z_x\,,\,z_y)$– vlak liggen de punten verspreid in de 4 kwadranten, want de oorsprong (0 , 0) is het zwaartepunt van de puntenwolk.

- Een punt in het **eerste** kwadrant, zoals $(0.5\,,1.5)$, heeft twee positieve coördinaten. Het product van de coördinaten levert dus ook een **positief** getal.
- Een punt in het **derde** kwadrant, zoals $(-1\,,-1)$, heeft twee negatieve coördinaten. Het product van de coördinaten levert ook hier een **positief** getal.
- Een punt in het **tweede** of **vierde** kwadrant, zoals $(-1\,,1)$ of $(1\,,-1)$, heeft een positieve en een negatieve coördinaat. Het product van de coördinaten levert dan een **negatief** getal.

Punten in het eerste en derde kwadrant leveren positieve coördinatenproducten *die groter en groter zijn wanneer de punten verder van de oorsprong liggen en dicht tegen de eerste bissectrice* (= de typische rechte bij een positieve samenhang).

Als de meerderheid van de punten in een niet te brede ellips rond de typische rechte ligt (zoals bij Jan), dan liggen de punten vooral in het eerste en derde kwadrant. Dat zie je hiernaast.
Als je dan de som maakt van de coördinatenproducten van alle punten, dan zullen de positieve producten de negatieve ruim compenseren en krijg je een groot positief resultaat. Bij een "dikkere" ellips (zoals bij Bea) liggen er al wat meer punten in het tweede en vierde kwadrant en krijg je een kleinere positieve som.



Gewoon de som maken is niet echt eerlijk als je Bea met Jan wil vergelijken. Het gaat hier over “de sterkte van de samenhang rond een rechte”, niet over het aantal punten. Bij Bea zijn er 19 punten en bij Jan 17. Daarom stap je over op een soort gemiddelde en deel je de som van de coördinatenproducten $\sum_{i=1}^{n}\left(z_{x_i} \times z_{y_i}\right)$ door “het aantal punten min één” $(n-1)$.

### 6.3.2. De formule

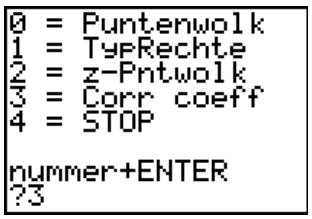
Voor een dataset van bivariate opmetingen $(x_i\ ,\ y_i)$ wordt **de correlatiecoëfficiënt** gedefinieerd als:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_{x_i} \times z_{y_i}\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\left(\frac{x_i-\bar{x}}{s_x}\right)\times\left(\frac{y_i-\bar{y}}{s_y}\right)\right)$$

De correlatiecoëfficiënt stel je voor door de letter “r”. De correlatiecoëfficiënt is een eenheidsloos getal want hij ontstaat uit producten van z-scores. Als $s_x = 0$ of $s_y = 0$, dan wordt de correlatiecoëfficiënt niet gedefinieerd.
Soms kom je als naam ook “Pearson correlatiecoëfficiënt” tegen.

Opdracht 18

Zoek de correlatiecoëfficiënt voor de puntenwolk van Bea. Gebruik je GRM en zorg ervoor dat de lijsten XBEA en YBEA in [L1] en [L2] staan. Werk dan met het programma CORR.




| | |
|---|---|
| correlatiecoëfficiënt Bea | r = |

Op zicht is er een sterkere samenhang bij Jan dan bij Bea. Wordt dit ook weergegeven in de correlatiecoëfficiënt van Jan? Hoeveel is die? Kopieer de lijsten XJAN en YJAN naar [L1] en [L2].

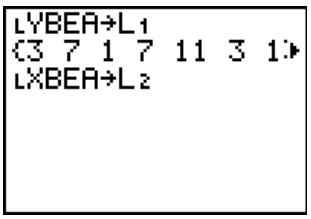
| | |
|---|---|
| correlatiecoëfficiënt Jan | r = |

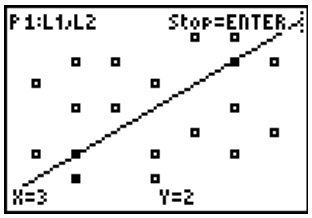
## 6.3.3. Eigenschappen van de correlatiecoëfficiënt

### ***Symmetrie in x en y***

Opdracht 19

Wat gebeurt er met de correlatiecoëfficiënt als je $x$ en $y$ omwisselt? Probeer dat voor de puntenwolk van Bea waar je YBEA in [L1] zet en XBEA in [L2].



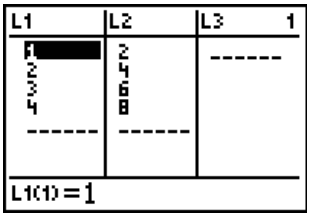

| | |
|---|---|
| correlatiecoëfficiënt Bea met $x$ en $y$ omgewisseld | r = |

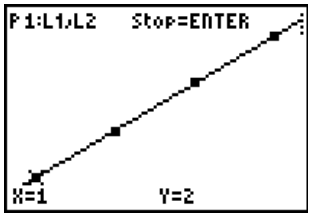
### ***Extrema***

Opdracht 20

| Perfecte positieve samenhang | |
|---|---|
| $x_i$ | $y_i$ |
| 1 | 2 |
| 2 | 4 |
| 3 | 6 |
| 4 | 8 |

Hiernaast zie je de coördinaten van punten die perfect op een rechte liggen. Tik die in je GRM in de lijsten [L1] en [L2]. Als de lijsten niet leeg zijn kan je ze eerst leeg maken. Druk [STAT] en 1:Edit... Je staat dan op het eerste getal van lijst [L1]. Druk op het pijltje [▲] zodat je op de naam L1 terechtkomt. Druk dan [CLEAR] en loop met [▼] terug naar beneden. Je hebt nu een lege lijst waar je de $x_i$ – getallen kan invullen. Ga op eenzelfde manier tewerk voor [L2] en eindig met [2nd] [QUIT].



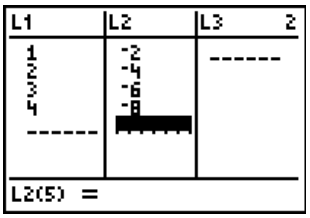



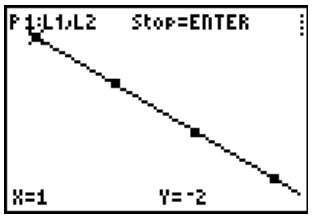

Met het programma CORR kan je controleren dat de ingebrachte punten op een rechte liggen (tik bij het keuzemenu het cijfer 1 en druk [ENTER]). Hoeveel is de correlatiecoëfficiënt hier?

| | |
|---|---|
| correlatiecoëfficiënt bij een perfecte positieve lineaire samenhang | r = |

| Perfecte negatieve samenhang | |
|---|---|
| $x_i$ | $y_i$ |
| 1 | – 2 |
| 2 | – 4 |
| 3 | – 6 |
| 4 | – 8 |

Verander nu de $y$-coördinaten in hun tegengestelde.




Hoeveel is de correlatiecoëfficiënt nu?

| | |
|---|---|
| correlatiecoëfficiënt bij een perfecte negatieve lineaire samenhang | r = |

## Opdracht 21

Plaats de lijsten XTOM en YTOM in [L1] en [L2]. Zoek de correlatiecoëfficiënt en bekijk ook de puntenwolk. Gebruik deze informatie om een uitspraak te doen over de lineaire samenhang tussen de twee veranderlijken in deze studie.

| | |
|---|---|
| correlatiecoëfficiënt Tom | r = |

Voor de correlatiecoëfficiënt gelden de volgende eigenschappen:

- de correlatiecoëfficiënt is symmetrisch: de sterkte van de lineaire samenhang tussen $y$ en $x$ is dezelfde als de sterkte van de lineaire samenhang tussen $x$ en $y$
- bij een perfecte positieve lineaire samenhang is de correlatiecoëfficiënt gelijk aan +1
- bij een perfecte negatieve lineaire samenhang is de correlatiecoëfficiënt gelijk aan –1
- de correlatiecoëfficiënt is een eenheidsloos getal tussen –1 en +1
- de correlatiecoëfficiënt is positief wanneer de zin van de lineaire samenhang positief is, en negatief wanneer de zin van de lineaire samenhang negatief is
- een correlatiecoëfficiënt die gelijk is aan nul wijst op het ontbreken van een **lineaire** samenhang.

Overzicht voor ellipsvormige puntenwolken:

*de lineaire samenhang is*

*negatief* *positief*

**–1** **0** **+1**

*perfect* *sterk* *matig* *zwak* *geen* *zwak* *matig* *sterk* *perfect*

# 7. Een grafische valkuil

Als je een puntenwolk bestudeert en je ziet dat de globale vorm lijkt op een ellips, dan kan je proberen om “op zicht” de sterkte van de lineaire samenhang te schatten. Dit geeft je dan ook een benaderend idee van de waarde van de correlatiecoëfficiënt.

## *7.1. Bloemblaadjes*

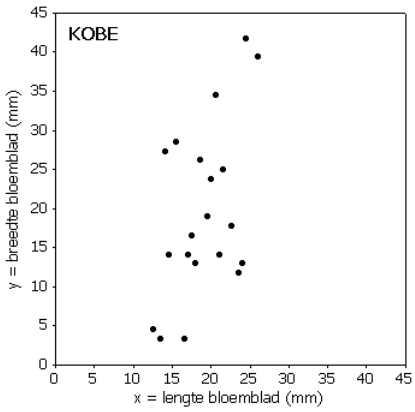
Is er een lineaire samenhang tussen de lengte en de breedte van bloemblaadjes en verschilt die samenhang volgens het soort bloem?

Om dit na te gaan deed men de volgende studie. Van 3 soorten bloemen werden heel veel bloemblaadjes verzameld en die werden bewaard in drie verschillende dozen. Drie leerlingen (Daan, Kobe en Lisa) moesten elk één doos kiezen en uit die doos lukraak 20 bloemblaadjes trekken. Daarna kregen zij gestandaardiseerde meetapparatuur om de lengte en de breedte van elk blaadje te bepalen. Op die manier hadden zij elk 20 bivariate opmetingen $(x_i, y_i)$ met $x_i$ = lengte van het i$^{de}$ bloemblad en $y_i$ = breedte van het i$^{de}$ bloemblad.

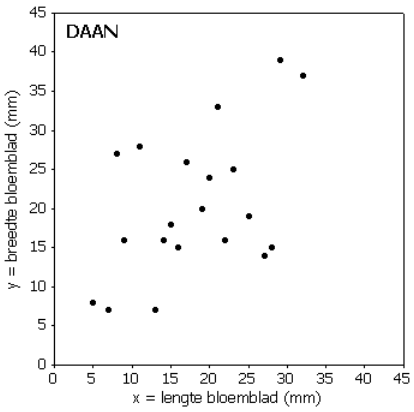
De opmetingen moesten ook grafisch worden voorgesteld. Daarvoor kregen de leerlingen elk een identiek tekenblad, waarbij de schaalverdeling op de x-as en op de y-as voor iedereen dezelfde was. Op die manier was het mogelijk om de vier grafieken letterlijk op elkaar te leggen en de 3 soorten bloemen grafisch met elkaar te vergelijken.

Opdracht 22

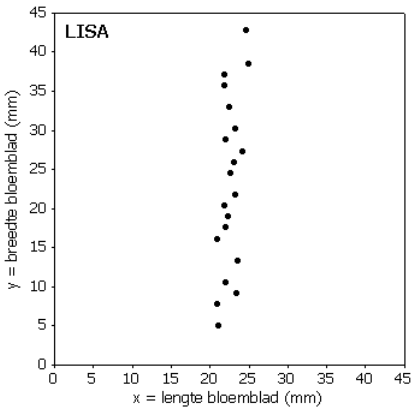
Hieronder zie je de puntenwolken van Daan, Kobe en Lisa. Schrijf bij elke puntenwolk de correlatiecoëfficiënt $r$ zoals je die op zicht schat. Kies hiervoor uit de getallen 0.50, 0.75 en 0.95.




r =

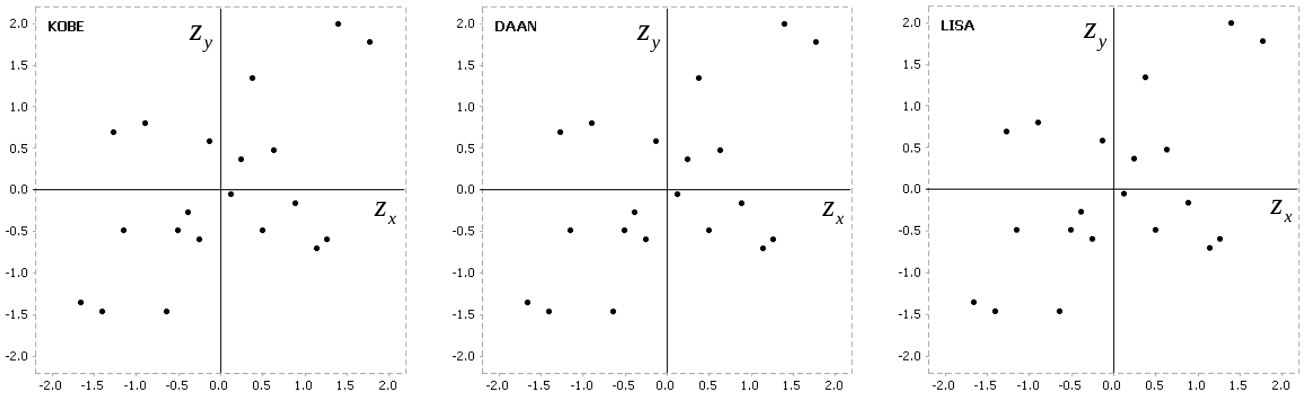



r =




r =

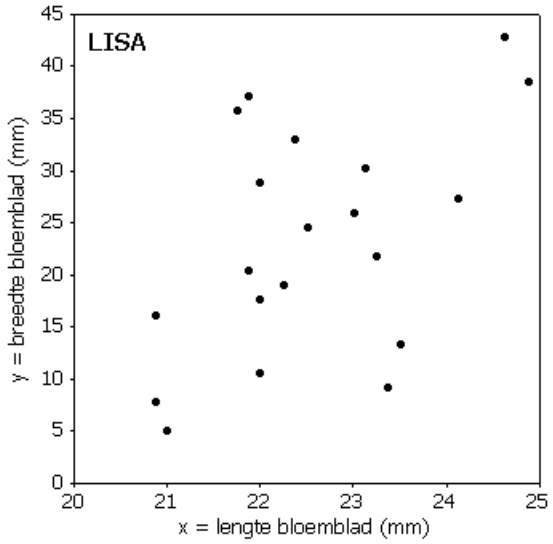
Als je niet werkt met de oorspronkelijke eenheden maar overstapt op z-scores, dan zien de gestandaardiseerde puntenwolken van Daan, Kobe en Lisa er uit zoals hieronder. Had je dat verwacht?




r = 0.50 r = 0.50 r = 0.50

Een figuur kan je soms op het verkeerde been zetten. Bij de bovenstaande studie lijkt het handig dat de 3 leerlingen eenzelfde tekenblad gebruiken. Dan is het duidelijk dat bij sommige bloemen de afmetingen van de blaadjes erg variëren (DAAN) terwijl andere bloemen blaadjes hebben waarbij de lengte bijna niet verandert (LISA).

Maar als je iets over lineaire samenhang wil zeggen, dan is het geen goed idee om in een vast assenstelsel te werken. Bij Lisa verdoezel je zo de variabiliteit in de $x_i$ – getallen (de lengtes). Als Lisa niet aan die studie had deelgenomen en zelfstandig een puntenwolk had getekend, dan zou zij zeker anders gewerkt hebben (jij zou dat ook doen en je GRM ook).



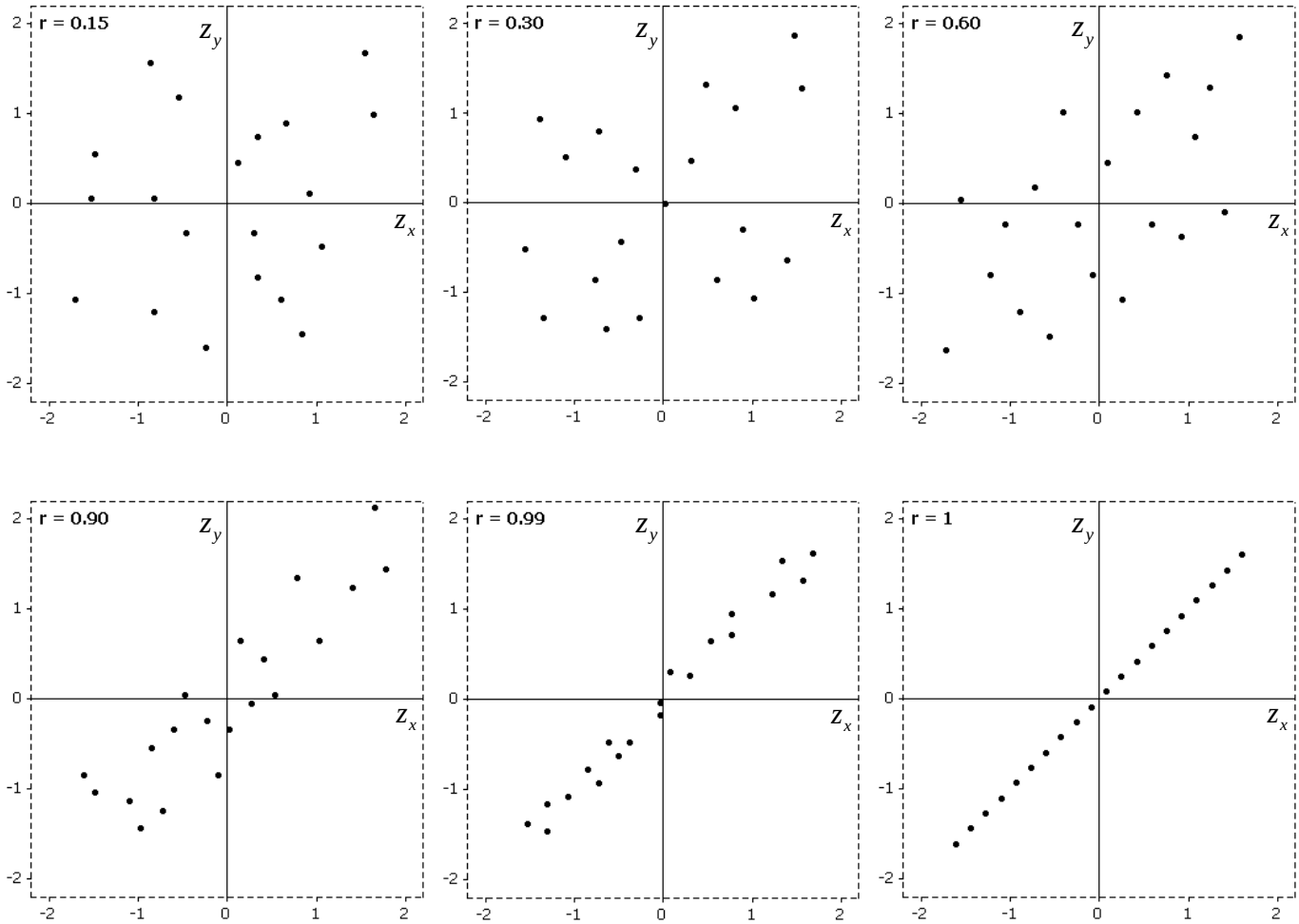

Als je bivariate opmetingen hebt zoals Lisa, waarbij alle $x_i$ – getallen liggen tussen 20 en 25, dan neem je op de x-as geen gebied dat loopt van 0 tot 45. Je zal dan heel waarschijnlijk een figuur tekenen die goed lijkt op de puntenwolk hiernaast. Bemerk dat dit exact de gegevens van Lisa zijn, gewoon met een andere keuze van de eenheid op de x-as. Bij deze figuur zal je nooit een correlatiecoëfficiënt r = 0.90 schatten.

Door de eenheid op de x-as of de y-as te veranderen krijg je een andere figuur en dus ook een andere indruk over de “sterkte” van de samenhang. De juiste indruk krijg je als je een gestandaardiseerde puntenwolk tekent.

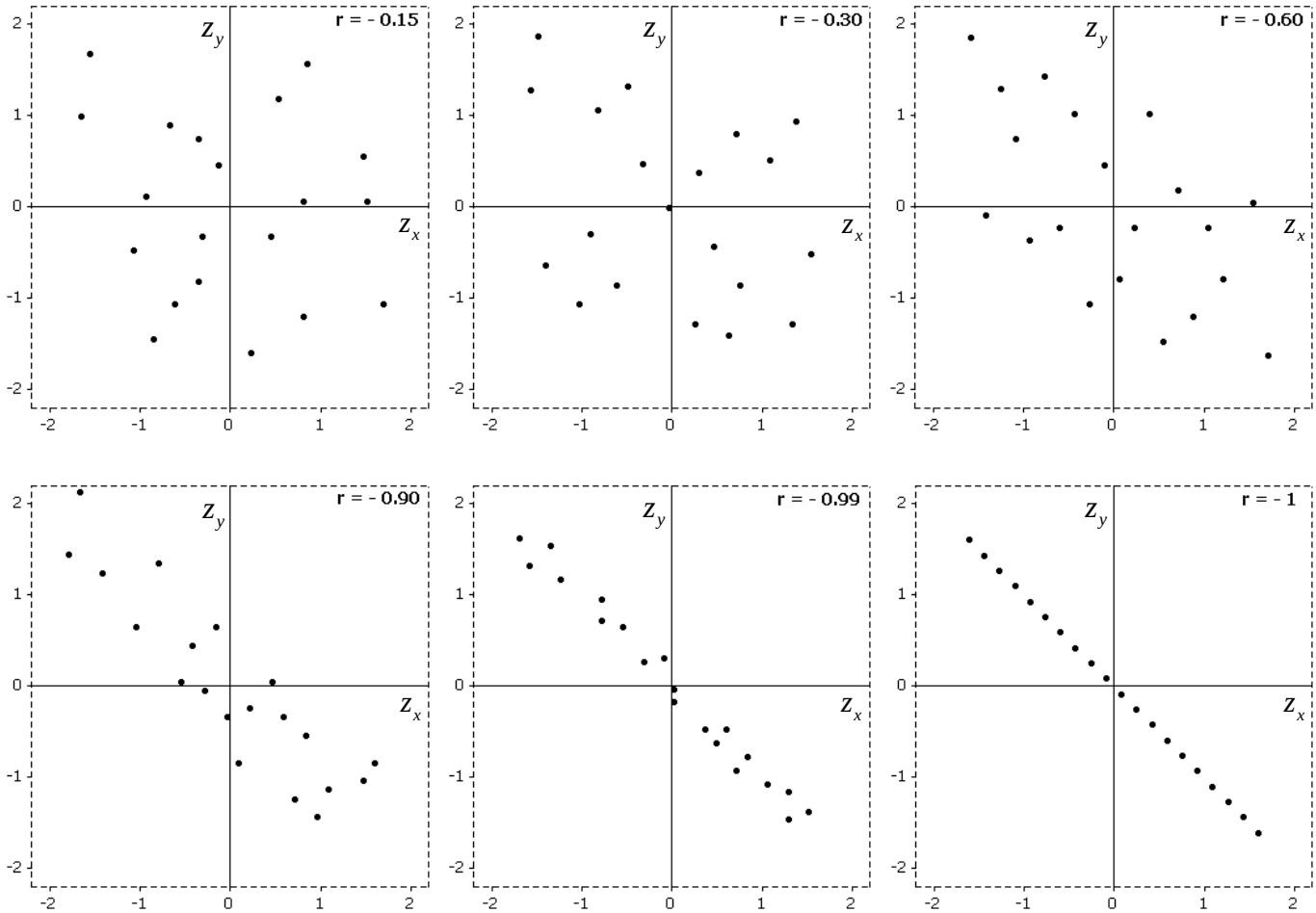
## 7.2. Puntenwolken en hun correlatiecoëfficiënt

Een grafiek, in de oorspronkelijke eenheden getekend, kan een verkeerde indruk geven. Daarom is het goed om z-scores te gebruiken als je *op zicht* de correlatiecoëfficiënt wil schatten. Jouw gestandaardiseerde puntenwolk kan je dan vergelijken met andere gestandaardiseerde puntenwolken waarvan je de correlatiecoëfficiënt kent. Hieronder zie je voorbeelden voor positieve lineaire samenhang en voor negatieve.

***Positieve lineaire samenhang***.



***Negatieve lineaire samenhang***.




# 8. Een numerieke valkuil

## *8.1. Eén getal = beperkte informatie*

Opdracht 23

"In de 4 onderstaande studies is er een ***positieve*** lineaire samenhang tussen $x$ en $y$. Die samenhang is ***redelijk sterk***. Dat volgt uit het feit dat de correlatiecoëfficiënt r ***positief*** is en ***gelijk aan 0.82***". Ben je akkoord met deze bewering? Leg uit waarom.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Studie_1 **r = 0.82** | $x_i$ | 10 | 8 | 13 | 9 | 11 | 14 | 6 | 4 | 12 | 7 | 5 |
| | $y_i$ | 8.04 | 6.95 | 7.58 | 8.81 | 8.33 | 9.96 | 7.24 | 4.26 | 10.84 | 4.82 | 5.68 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Studie _2 **r = 0.82** | $x_i$ | 10 | 8 | 13 | 9 | 11 | 14 | 6 | 4 | 12 | 7 | 5 |
| | $y_i$ | 9.14 | 8.14 | 8.74 | 8.77 | 9.26 | 8.10 | 6.13 | 3.10 | 9.13 | 7.26 | 4.74 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Studie _3 **r = 0.82** | $x_i$ | 10 | 8 | 13 | 9 | 11 | 14 | 6 | 4 | 12 | 7 | 5 |
| | $y_i$ | 7.46 | 6.77 | 12.74 | 7.11 | 7.81 | 8.84 | 6.08 | 5.39 | 8.15 | 6.42 | 5.73 |

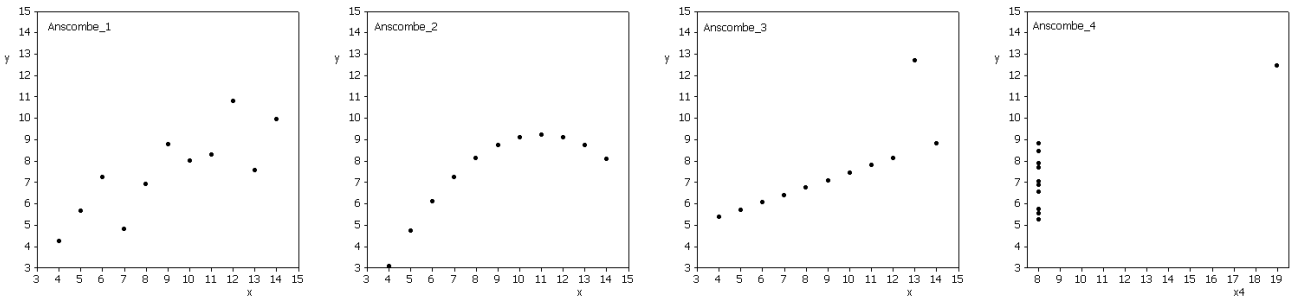
| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Studie _4 **r = 0.82** | $x_i$ | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 19 | 8 | 8 | 8 |
| | $y_i$ | 6.58 | 5.76 | 7.71 | 8.84 | 8.47 | 7.04 | 5.25 | 12.50 | 5.56 | 7.91 | 6.89 |

## 8.2. Uitschieters, krommen, en de voorbeelden van Anscombe

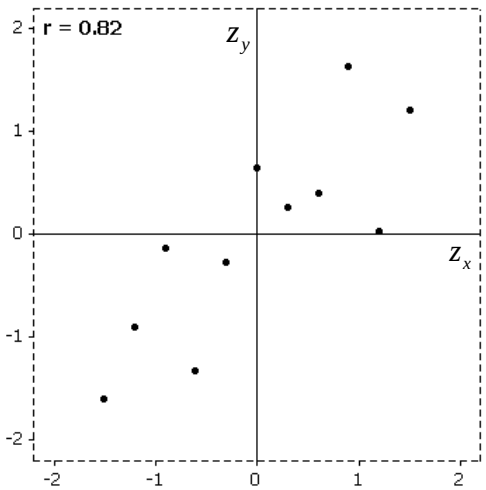
Een kengetal, zoals een correlatiecoëfficiënt, geeft informatie in een samengevatte vorm. Die informatie kan verhelderend zijn, maar soms ook misleidend.

Inzicht in een gegevensverzameling krijg je niet zomaar uit één kengetal en dikwijls zijn meerdere kengetallen zelfs niet voldoende. Bij elke statistische exploratie hoort ook een figuur. Bij ellipsvormige puntenwolken is het verstandig om met gestandaardiseerde eenheden (z-scores) te werken.

De 4 datasets die je hierboven hebt ontmoet zijn door F. J. Anscombe opgesteld. Voor elk van deze datasets is de correlatiecoëfficiënt gelijk aan 0.82. Hieruit afleiden dat er in die vier gevallen een tamelijk sterke positieve lineaire samenhang is, zou een grote vergissing zijn. Dat zie je in de bijhorende figuren hieronder. Zij zijn getekend op basis van de oorspronkelijke gegevens, zoals opgesteld door Anscombe.




De figuur Anscombe_1 is globaal ellipsvormig. Hiernaast zie je de gestandaardiseerde puntenwolk. Wat sterkte van samenhang betreft, lijkt deze puntenwolk vrij goed op de puntenwolk met r = 0.90 bij de voorbeelden.




De figuur Anscombe_2 toont een duidelijke samenhang tussen $x$ en $y$, maar die is helemaal niet lineair.

De figuren Anscombe_3 en Anscombe_4 illustreren dat de correlatiecoëfficiënt gevoelig is voor uitschieters.

# 9. Wat kan er nog meer fout gaan?

Het ergste wat er kan fout gaan, is dat je geen puntenwolk tekent.

## *9.1. Paleontologie*

In de paleontologie wordt de prehistorische mens bestudeerd. Men gebruikt ondermeer skeletten die bij opgravingen worden ontdekt. In de tabel hiernaast zie je de lengte $x_i$ en de breedte $y_i$ van een bepaald beentje, opgemeten bij skeletten van kinderen. In de tabel staat ook het identificatienummer (ID). Er is ook genoteerd of het om een meisje (M) of een jongen (J) gaat.

| ID | Sex | Lengte (cm) $x_i$ | Breedte (cm) $y_i$ |
|---|---|---|---|
| 1 | M | 10.0 | 3.0 |
| 2 | M | 11.0 | 5.0 |
| 3 | M | 11.5 | 3.5 |
| 4 | M | 12.0 | 4.0 |
| 5 | M | 12.5 | 2.5 |
| 6 | M | 13.0 | 2.0 |
| 7 | J | 13.0 | 8.0 |
| 8 | J | 13.5 | 7.5 |
| 9 | M | 14.0 | 2.5 |
| 10 | J | 14.0 | 6.5 |
| 11 | J | 15.0 | 7.0 |
| 12 | J | 15.5 | 7.5 |
| 13 | J | 16.0 | 6.0 |
| 14 | J | 16.5 | 7.0 |

Opdracht 24

De lengte van de onderzochte beenderen staat in de lijst LBEEN en de breedte in BBEEN. Plaats die lijsten in [L1] en [L2]. Druk [2nd] [LIST], loop naar LBEEN en druk [ENTER]. Druk dan [STO▸] en [2nd] [L1] en [ENTER]. Voor de breedte werk je op analoge manier. Druk [2nd] [LIST], loop naar BBEEN en druk [ENTER]. Druk dan [STO▸] en [2nd] [L2] en [ENTER].

Gebruik het programma CORR.

```
0 = Puntenwolk
1 = TypRechte
2 = z-Pntwolk
3 = Corr coeff
4 = STOP

nummer+ENTER
?
```

1. Zoek de correlatiecoëfficiënt.

| correlatiecoëfficiënt r = |
|---|

2. Bekijk de puntenwolk in de oorspronkelijke eenheden.

3. Teken op je GRM nu ook de typische rechte. Krijg je grafisch een vergelijkbare figuur als je standaardiseert (met z-Pntwolk)?

4. Gebruik nu de gevonden numerieke en grafische informatie om het juiste vakje aan te duiden:

   – de lineaire samenhang is … ☐ *positief* ☐ *negatief*
   – de lineaire samenhang is … ☐ *zwak* ☐ *matig* ☐ *sterk*

☐ *langere beenderen zijn breder*
☐ *langere beenderen zijn smaller*

5. Kijk nog eens heel goed naar de puntenwolk. Heb je bemerkingen bij deze studie? Welke?

## 9.2. Clusters

In je studie van die beenderen heb je rekening gehouden met de correlatiecoëfficiënt en met de manier waarop de puntenwolk verstrooid ligt rond de typische rechte (zowel in oorspronkelijke eenheden als gestandaardiseerd met z-scores). En toch klopt er iets niet.

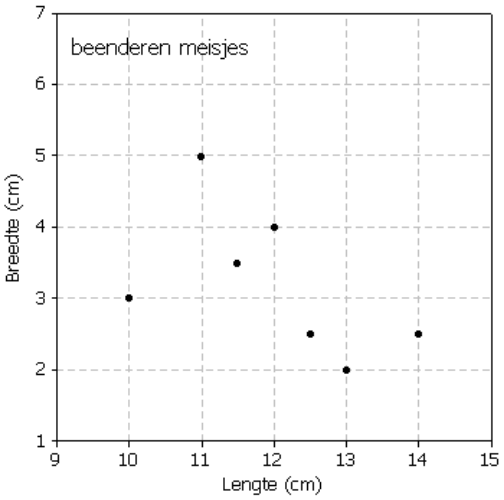
De context van de studie geeft extra informatie: het gaat over jongens en meisjes. Als je goed naar de figuur kijkt dan zie je dat de puntenwolk uit twee groepen (of twee clusters) bestaat: een groep punten links onder en een andere groep rechts boven.

Alle punten links onder zijn afkomstig van skeletten van meisjes, en alle punten rechts boven zijn opmetingen van jongens. Dat kom je ook te weten uit de dataset. Daarom is een studie van "skeletten van kinderen" niet zo verstandig hier. Kijk eens naar de twee groepen afzonderlijk.

Opdracht 25

Tik de gegevens van de meisjes in je GRM. Zoek de typische rechte en teken ze bij de puntenwolk. Zoek ook de correlatiecoëfficiënt. Welk besluit trek je op basis van de grafiek en van de correlatiecoëfficiënt?

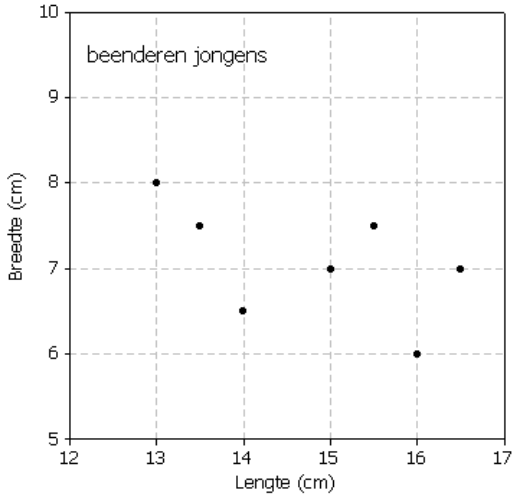
| Meisjes | | | |
|---|---|---|---|
| ID | Sex | Lengte (cm) $x_i$ | Breedte (cm) $y_i$ |
| 1 | M | 10.0 | 3.0 |
| 2 | M | 11.0 | 5.0 |
| 3 | M | 11.5 | 3.5 |
| 4 | M | 12.0 | 4.0 |
| 5 | M | 12.5 | 2.5 |
| 6 | M | 13.0 | 2.0 |
| 9 | M | 14.0 | 2.5 |




| | |
|---|---|
| Typische rechte: $y$ = | correlatiecoëfficiënt: $r$ = |
| Besluit: | |

Doe nu hetzelfde voor de jongens. Tik hun gegevens in je GRM. Zoek de typische rechte en teken ze bij de puntenwolk. Zoek ook de correlatiecoëfficiënt. Welk besluit trek je op basis van de grafiek en van de correlatiecoëfficiënt?

| Jongens | | | |
|---|---|---|---|
| ID | Sex | Lengte (cm) $x_i$ | Breedte (cm) $y_i$ |
| 7 | J | 13.0 | 8.0 |
| 8 | J | 13.5 | 7.5 |
| 10 | J | 14.0 | 6.5 |
| 11 | J | 15.0 | 7.0 |
| 12 | J | 15.5 | 7.5 |
| 13 | J | 16.0 | 6.0 |
| 14 | J | 16.5 | 7.0 |




| Typische rechte: $y$ = | correlatiecoëfficiënt: $r$ = |
|---|---|
| Besluit: | |

## 9.3. Hoger of lager?

Kleuters spelen met water. Met een bekertje scheppen zij water uit een emmer. Dan lopen zij naar de andere kant van de speeltuin waar er voor elke kleuter een holle plastieken zuil staat. Elke kleuter probeert zijn zuil zo hoog mogelijk met water te vullen.

Zeven leerlingen werken samen aan een project statistiek. Zij onderzoeken of er een samenhang is tussen de tijd dat de kleuter het waterspelletje speelt en de hoogte van het water in de zuil. Zij gaan elk één koppel $(x_i, y_i) = (\text{tijd in minuut, hoogte in centimeter})$ opmeten om daarna samen een puntenwolk te tekenen.

De leerlingen zien dat er nogal wat verschil is bij die kleuters en zij besluiten als volgt te werk te gaan. Elke leerling zal 3 kleuters observeren en telkens de tijd en de hoogte noteren. Het gemiddelde van de 3 opgemeten tijden neemt die leerling dan als "een typische tijd" en het gemiddelde van de 3 hoogtes als "een typische hoogte". Voor het gemiddelde van de tijd wordt in het tiendelig stelsel gewerkt zodat (7+10+11)/3 = 9.3 minuten en niet 9 minuten en 18 seconden.

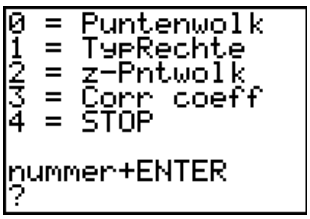
De resultaten van die 7 leerlingen zijn als volgt:

| | leerling_1 | leerling_2 | leerling_3 | leerling_4 | leerling_5 | leerling_6 | leerling_7 |
|---|---|---|---|---|---|---|---|
| $x_i$ (min) | 8.7 | 8.0 | 7.3 | 9.3 | 8.0 | 8.7 | 8.3 |
| $y_i$ (cm) | 27.1 | 32.4 | 35.9 | 23.6 | 34.2 | 25.3 | 29.7 |

Opdracht 26

Tik de $x_i$ – getallen in [L1] en de bijhorende $y_i$ – getallen in [L2].

Gebruik het programma CORR.




1. Zoek de correlatiecoëfficiënt.

| correlatiecoëfficiënt r = |
|---|

2. Bekijk de puntenwolk in de oorspronkelijke eenheden.

3. Teken op je GRM nu ook de typische rechte. Krijg je grafisch een vergelijkbare figuur als je standaardiseert (met z-Pntwolk)?

4. Gebruik nu de gevonden numerieke en grafische informatie om het juiste vakje aan te duiden:

– de lineaire samenhang is … ☐ *positief* ☐ *negatief*
– de lineaire samenhang is … ☐ *zwak* ☐ *matig* ☐ *sterk*

☐ *een langere speeltijd levert een hogere waterzuil*
☐ *een langere speeltijd levert een lagere waterzuil*

5. Had je dit resultaat verwacht? Heb je bemerkingen bij deze studie? Welke?

Als kleuters meer tijd hebben om bekertjes water in hun zuil te gieten, dan verwacht je dat er meer water in de zuil staat. Het cijfermateriaal van je vorig onderzoek spreekt deze verwachting tegen. Dat is een goede reden om dat onderzoek eens nader te bekijken.

## Opdracht 27

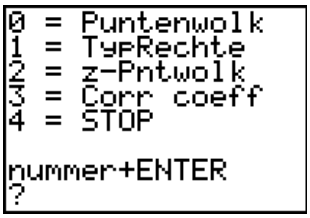
De leerlingen hebben elk 3 kleuters bestudeerd en dan het gemiddelde genomen. Dat geeft de indruk dat zij werken met "uitgebalanceerd" cijfermateriaal. Maar is dat bij een correlatiestudie wel een goed idee?

Hieronder staan alle opmetingen. In totaal zijn er 21 koppels $(x_i, y_i)$ = ( tijd, hoogte ).

| leerling 1 | | leerling 2 | | leerling 3 | | leerling 4 | | leerling 5 | | leerling 6 | | leerling 7 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | $y_i$ | $x_i$ | $y_i$ | $x_i$ | $y_i$ | $x_i$ | $y_i$ | $x_i$ | $y_i$ | $x_i$ | $y_i$ | $x_i$ | $y_i$ |
| 13 | 53.6 | 10 | 53.6 | 13 | 37.7 | 10 | 37.7 | 12 | 43.1 | 9 | 16.5 | 11 | 48.2 |
| 10 | 21.8 | 7 | 5.9 | 3 | 21.8 | 7 | 21.8 | 8 | 43.0 | 9 | 27.1 | 8 | 29.7 |
| 3 | 5.9 | 7 | 37.7 | 6 | 48.2 | 11 | 11.3 | 4 | 16.5 | 8 | 32.3 | 6 | 11.2 |
| | | | | | | | | | | | | | |
| $\bar{x}$ | $\bar{y}$ | $\bar{x}$ | $\bar{y}$ | $\bar{x}$ | $\bar{y}$ | $\bar{x}$ | $\bar{y}$ | $\bar{x}$ | $\bar{y}$ | $\bar{x}$ | $\bar{y}$ | $\bar{x}$ | $\bar{y}$ |
| 8.7 | 27.1 | 8.0 | 32.4 | 7.3 | 35.9 | 9.3 | 23.6 | 8.0 | 34.2 | 8.7 | 25.3 | 8.3 | 29.7 |

De 21 $x_i$ – getallen staan in de lijst TIJD en de bijhorende 21 $y_i$ – getallen staan in HOOG. Plaats de lijst TIJD in [L1] en de lijst HOOG in [L2].

Gebruik het programma CORR.




1. Zoek de correlatiecoëfficiënt.

| correlatiecoëfficiënt r = |
|---|

2. Bekijk de puntenwolk in de oorspronkelijke eenheden.

3. Teken op je GRM nu ook de typische rechte. Krijg je grafisch een vergelijkbare figuur als je standaardiseert (met z-Pntwolk)?

4. Gebruik nu de gevonden numerieke en grafische informatie om het juiste vakje aan te duiden:

– de lineaire samenhang is … ☐ *positief* ☐ *negatief*
– de lineaire samenhang is … ☐ *zwak* ☐ *matig* ☐ *sterk*

☐ *een langere speeltijd levert een hogere waterzuil*
☐ *een langere speeltijd levert een lagere waterzuil*

5. Je gebruikt hier dezelfde opmetingen als in de vorige opdracht en toch kom je tot een tegengesteld resultaat. Hoe zou dat komen?

## 9.4. De ecologische valkuil

De twee vorige opdrachten illustreren de **ecologische valkuil**.

De benaming "ecologische valkuil" en "ecologische correlatie" klinkt ongewoon. Dat is inderdaad zo. De naam komt van een artikel dat in 1950 gepubliceerd werd door W. Robinson: "Ecological correlations and the behaviour of individuals". Hierbij definieerde hij:

- "individuele correlatie": correlatie tussen eigenschappen **van individuen**. Hierbij bestudeer je "ondeelbare eenheden" (zoals kleuters) waarvan je eigenschappen noteert (zoals de gespeelde tijd en de hoogte van de waterzuil).
- "ecologische correlatie": correlatie tussen berekende kengetallen (zoals gemiddelde of proportie) bij **groepen**. Je berekent bijvoorbeeld de gemiddelde gespeelde tijd en de gemiddelde hoogte van de waterzuil bij groepjes van 3 kleuters.

Algemeen spreekt men over **ecologische gegevens** wanneer je te maken hebt met gegevens die zelf al groepsgewijs zijn samengevat (in gemiddelden of in proporties). Correlatie van ecologische gegevens heet **ecologische correlatie**.

Als je een sterke ecologische correlatie gevonden hebt (op groepsniveau), dan betekent dat helemaal niet dat er ook een sterke correlatie is op het niveau van de individuen. Straffer nog: niet alleen de sterkte kan wijzigen maar zelfs de zin (positief/negatief) kan omslaan.

***De ecologische valkuil kom je meer tegen dan je denkt.***

In het Europa van de 19$^{de}$ eeuw waren de zelfmoordcijfers hoger in de landen die overwegend protestants waren. Kan je daaruit besluiten dat de levenswijze opgelegd door het protestantisme de zelfmoordneiging aanwakkert?

Je hebt hier te maken met twee problemen tegelijkertijd. Het ene probleem heet "verstrengeling", wat betekent dat er nog heel wat andere factoren meespelen. Protestantse landen waren op veel punten verschillend van katholieke landen en dat was niet alleen aan de religie te wijten. Verstrengeling is uitgebreid besproken in de teksten over "Studies naar samenhang" die je kan vinden op http://www.uhasselt.be/lesmateriaal-statistiek . Op het probleem van verstrengeling gaan we in deze tekst niet dieper in.

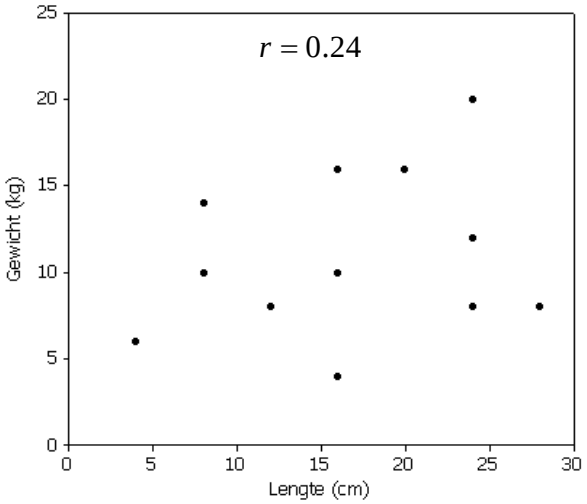
Het andere probleem heeft te maken met gegevens die per groep al samengevat zijn in een gemiddelde of in een proportie. In het voorbeeld over de zelfmoorden gaat het over gegevens "per land". Het zijn niet landen die zelfmoord plegen, maar mensen. Een typische fout bestaat er in om uit een sterke samenhang tussen gegevens per land, de conclusie te trekken dat dezelfde sterke samenhang er ook is voor de individuen in die landen. Die fout wordt de **ecologische valkuil** genoemd.

In de epidemiologie zijn er heel wat studies die "landen" vergelijken. Zo is blijkbaar het aantal borstkankers beduidend hoger in landen waar de voeding meer vetstof bevat. Maar kan je uit deze gegevens rechtstreeks besluiten dat vrouwen die meer vet eten ook een grotere kans op borstkanker hebben? Dit kan waar zijn, maar die informatie haal je niet uit gegevens "per land".

Op het terrein van het milieu en de gezondheidszorg verzamelt men soms informatie per provincie. Maar een sterke samenhang per provincie tussen pollutie en opname in ziekenhuizen, is niet voldoende om te weten te komen hoe die samenhang er uitziet voor de mensen die daar wonen.

Als je per gemeente het gemiddelde inkomen kent samen met het percent stemgerechtigden dat bij gemeenteraadsverkiezingen op “rechtse” partijen stemt, dan kan je onderzoeken of er een sterke samenhang is tussen “hoger inkomen” en “rechtser stemgedrag”. Maar uit resultaten per gemeente haal je nog helemaal niet hoe groot die samenhang is op het niveau van de individuele kiezer.

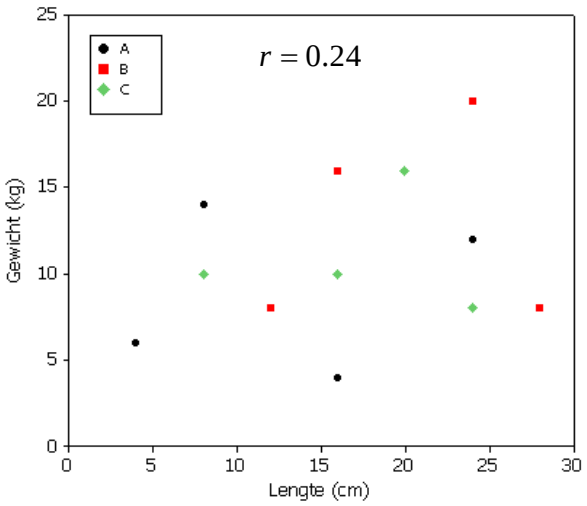
Onderstaand extreem voorbeeld (met fictieve opmetingen over lengte en gewicht van ijzeren staven) illustreert wat er meestal gebeurt bij ecologische correlaties. Bij elke opmeting $(x_i, y_i)$ is ook aangegeven uit welk land de staaf komt.

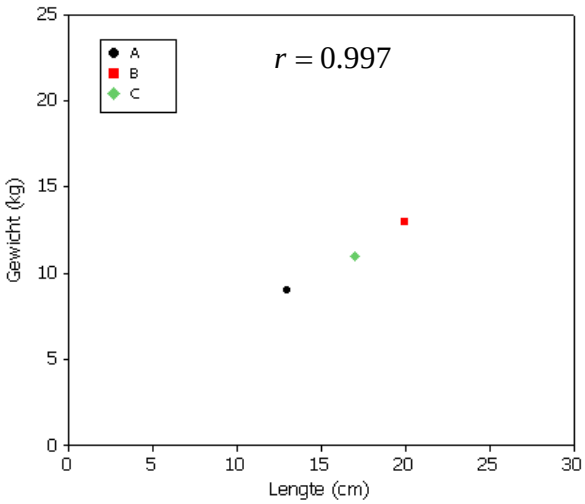



| Land | $x_i$ = lengte | $y_i$ = gewicht |
|---|---|---|
| A | 4 | 6 |
| A | 8 | 14 |
| A | 16 | 4 |
| A | 24 | 12 |
| B | 12 | 8 |
| B | 16 | 16 |
| B | 28 | 8 |
| B | 24 | 20 |
| C | 8 | 10 |
| C | 16 | 10 |
| C | 20 | 16 |
| C | 24 | 8 |

De lengte en het gewicht van die staven vertonen een positieve lineaire samenhang die zwak is. Die conclusie trek je uit de vorm van de puntenwolk samen met de waarde van de correlatiecoëfficiënt ( r = 0.24).

Hieronder links zie je de puntenwolk van de staven, gecodeerd per land. Rechts staat de puntenwolk van de “gemiddelden per land”. De correlatiecoëfficiënt rechts is r = 0.997. Dat wijst op een extreem sterke, bijna perfecte, positieve samenhang. Als je nu alleen de gegevens per land zou hebben, dan is de neiging groot om in de ecologische valkuil te trappen en te zeggen dat er een zeer sterke positieve samenhang is tussen de lengte en het gewicht van die staven.



## 9.5. Oorzaak en samenhang

Bij kinderen van de lagere school is er een sterke samenhang tussen taalvaardigheid en schoenmaat. Nochtans is het feit dat zij vlotter leren lezen er niet de oorzaak van dat hun voeten gaan groeien. Er is hier een andere factor in het spel: leeftijd.

Naar puntenwolken kijken en correlatiecoëfficiënten berekenen, behoedt je niet tegen de klassieke valkuil dat je “samenhangen” verwart met “veroorzaken”. In het voorbeeld van de schoolkinderen is het niet moeilijk om een “verstrengelende” factor (namelijk “leeftijd”) te ontdekken. Maar de meeste statistische studies zijn veel complexer en daar kan het echt moeilijk zijn om te weten te komen of er een oorzakelijk verband is. Dikwijls kan je alleen maar zeggen dat je tussen twee eigenschappen een samenhang ontdekt hebt.

Zoals boven al vermeld, kan je deze topic verder bestuderen in de teksten over “Studies naar samenhang”. Je vindt die op http://www.uhasselt.be/lesmateriaal-statistiek .