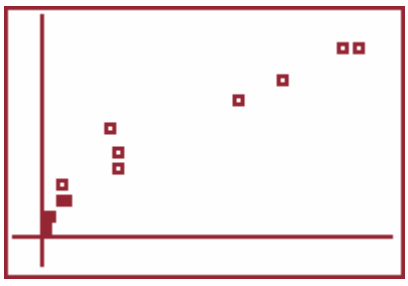
Cahiers T³ Europe
Vlaanderen nr. 3

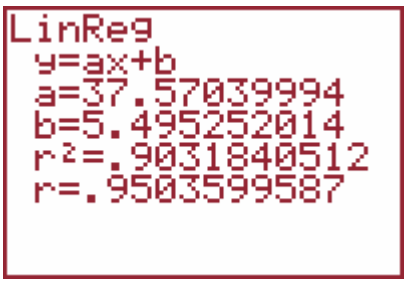
T³ EUROPE

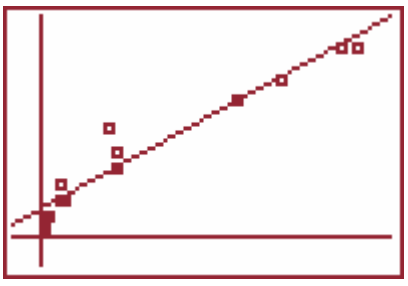
Regressie

Een eerste kennismaking

Bieke Van Deyck







KATHOLIEKE UNIVERSITEIT
LEUVEN

Regressie

Een eerste kennismaking

Bieke Van Deyck

Inhoudsopgave

HOOFDSTUK 1:
DE BIVARIATE VERDELING

A. Probleembeschrijving

Uit de beschrijvende statistiek weten we hoe we één kwantitatieve variabele X kunnen onderzoeken binnen een populatie. We nemen een representatieve ***steekproef***, onderzoeken binnen deze deelverzameling die bepaalde eigenschap en verkrijgen zo concrete waarnemingsgetallen $x_1, x_2, ..., x_n$.
Deze ordenen we in een frequentietabel, stellen ze grafisch voor en trachten ze samen te vatten door centrummaten en spreidingsmaten te berekenen (zie appendix). Op die wijze krijgen we een idee over de ***verdeling*** van de stochastische veranderlijke X.

In plaats van ons te concentreren op één enkele variabele (zoals bijvoorbeeld gewicht, lengte, punten op een examen,...) is het vaak interessant om de relatie te bestuderen tussen een koppel variabelen, wat de ***bivariate verdeling*** met zich meebrengt.
Bijvoorbeeld, van elke volwassen mens wordt zijn lengte X in cm en gewicht Y in kg gemeten. Het geordende paar (X,Y) heeft een bivariate verdeling.

We stellen ons nu de volgende vraag:
Bestaat er een ***verband*** of ***correlatie*** tussen twee kwantitatieve variabelen die betrekking hebben op dezelfde populatie?
M.a.w. bestaat er een bepaalde relatie tussen de stochastische veranderlijken X en Y van het koppel (X,Y) met hun bivariate verdeling?

☒ Voorbeelden:

- Is er een verband tussen de leeftijden van huwelijkspartners?
- Is er een verband tussen de lengte van een moeder en haar kind?
- Is er een verband tussen het gewicht en het voetoppervlak (het contactoppervlak met de grond) van Zuid-Amerikaanse slakken?
- Is er een verband tussen het aantal tewerkgestelde mannen en het aantal tewerkgestelde vrouwen van de beroepsbevolking?

We trachten een eventueel verband te beschrijven door middel van een grafische voorstelling of door middel van een getal. Dit zal in de volgende hoofdstukken uitgewerkt worden op basis van de *correlatie-* en *regressierekening*.

Opnieuw gaan we hierbij uit van een steekproef waarbij de twee variabelen worden gemeten, op die wijze verkrijgen we concrete data $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$.

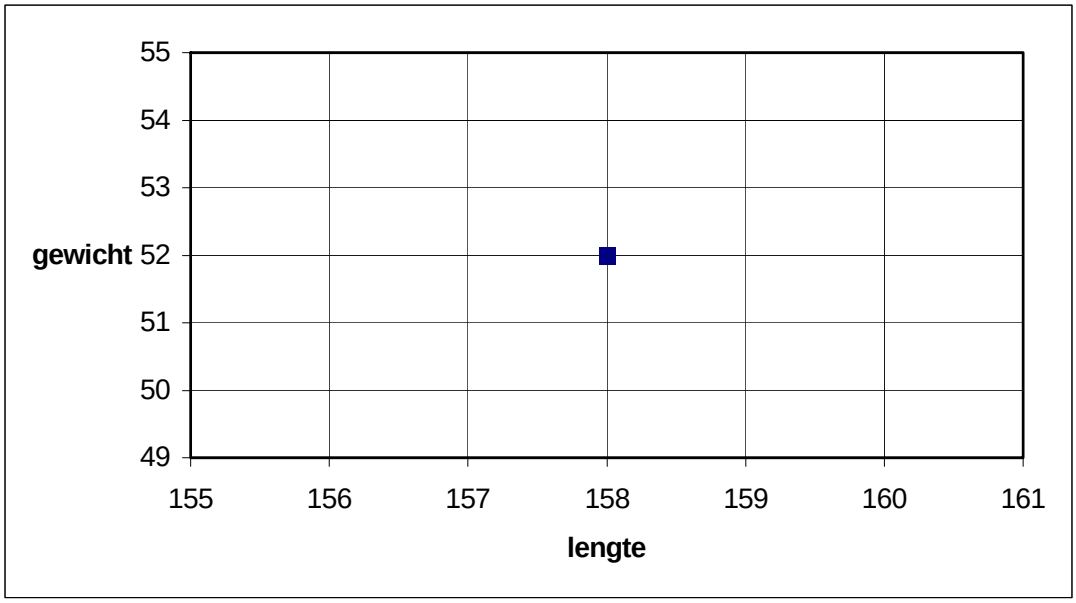
Het is belangrijk te beseffen dat we ook hier aan *beschrijvende statistiek* doen; grootheden die we zullen berekenen aan de hand van de steekproefdata (b.v. de correlatiecoëfficiënt) zullen variëren van steekproef tot steekproef.

B. Het spreidingsdiagram

De meest eenvoudige methode om twee gemeten variabelen simultaan weer te geven, is een ***spreidingsdiagram***.
Dit gebruikt een horizontale as voor één van de variabelen en een verticale as voor de andere. Er wordt een punt geplaatst voor elk observatie-paar (x_i, y_i) op de kruising van zijn twee waarden.

☒ Voorbeeld:

Een kind is 1m58 lang en weegt 52 kg.
We kiezen X = lengte (in cm)
Y = gewicht (in kg)
En we plaatsen deze observatie in het spreidingsdiagram als volgt:

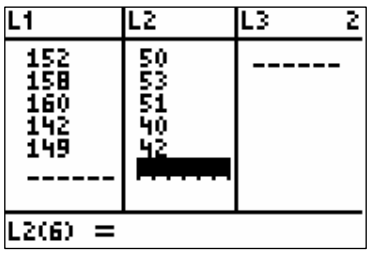


Als je de waarde van één variabele wilt gebruiken om de waarde van een andere variabele te voorspellen, is de conventie om de variabele waarmee je de voorspelling doet (=de ***onafhankelijke variabele***) op de horizontale as te plaatsen en de te voorspellen variabele (=de ***afhankelijke variabele***) op de verticale as.

💻 Hoe kunnen we zelf een spreidingsdiagram tekenen met de TI-83?

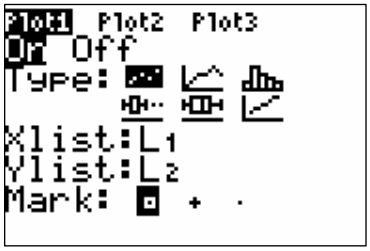
Gegeven 5 kinderen met hun lengte en gewicht:

lengte x	152	158	160	142	149
gewicht y	50	53	51	40	42



We voeren de gegevens in:

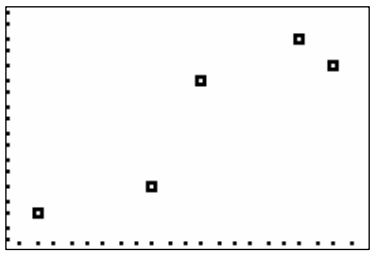
Om een spreidingsdiagram te maken, doen we het volgende:
2nd Y= (STAT PLOT) en definieer plot 1 zoals hiernaast gegeven:



Hierbij zijn:

Xlist: de gegevens die je op de horizontale as wilt
Ylist: de gegevens die je op de verticale as wilt
Mark: hoe wil je dat je punten er uit zien?

Druk dan op GRAPH en op ZOOM 9:ZoomStat en dan verschijnt je spreidingsdiagram.
Het commando Zoomstat zorgt ervoor dat alle punten op je scherm passen en dat de puntenwolk verspreid ligt over het hele scherm. Het bereik wordt zo automatisch aangepast.



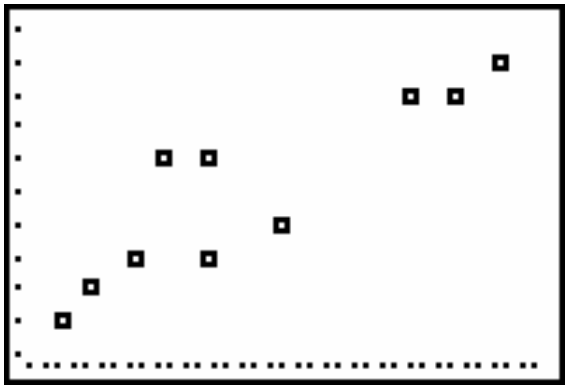
Via TRACE en de pijltjestoetsen kan je nu de coördinaten van elk punt nagaan.

☒ Voorbeeld:

Bij een keuring voor militaire dienst zijn 10 jongens aan een medisch onderzoek onderworpen. Daarbij is van elke jongen de lengte en de schoenmaat gemeten.

Lengte x	165	167	170	172	175	175	180	189	192	195
Schoenmaat y	37	38	39	42	39	42	40	44	44	45

Dit levert het volgende spreidingsdiagram:



(a) Wat is de richting van de puntenwolk?

(b) Hoe zou je de richting van de puntenwolk in woorden kunnen beschrijven? Met andere woorden hoe kan je de relatie tussen de lengte en de schoenmaat van de kandidaat-soldaten weergeven?

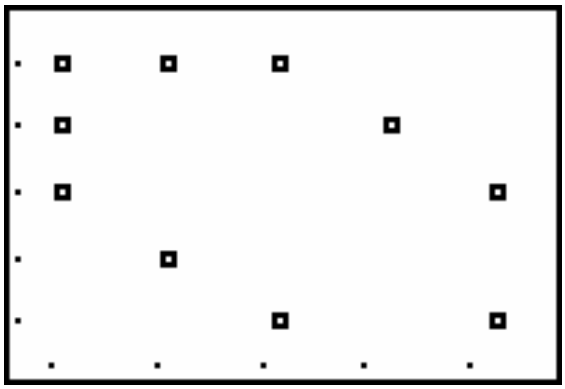
Twee variabelen worden ***positief gecorreleerd*** genoemd als grote waarden van de ene variabele overeenkomen met grote waarden voor de andere variabele.

We zien duidelijk dat de lengte en de schoenmaat positief gecorreleerd zijn. Dit zien we aan de richting van de puntenwolk: van links onder naar rechts boven.

☒ Voorbeeld:

Hieronder staat een tabel van het koffie- en theegebruik van 10 huishoudens, in koppen per persoon per dag.

koffiegebruik x	3	6	2	4	5	3	4	2	6	2
theegebruik y	6	2	4	6	5	3	2	6	4	5



(a) Wat is hier de richting van de puntenwolk?

(b) En hoe zou je in dit geval de richting van de puntenwolk in woorden kunnen beschrijven? Met andere woorden hoe kan je de relatie tussen het koffie- en theegebruik weergeven?

Omgekeerd, worden twee variabelen ***negatief gecorreleerd*** genoemd als grotere waarden van de ene variabele overeenkomen met kleinere waarden van de andere variabele.

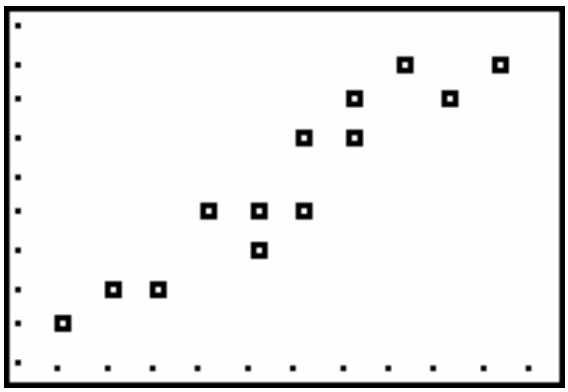
Ook hier zien we vrij duidelijk dat het koffiegebruik en het theegebruik in huishoudens negatief gecorreleerd zijn: de puntenwolk gaat van links boven naar rechts onder.

☒ Voorbeeld:

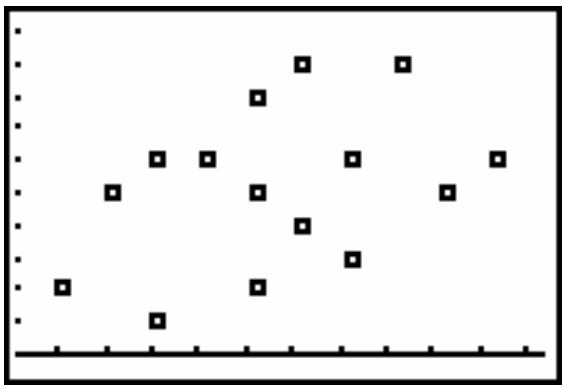
Aan een hondenshow is een wedstrijd verbonden voor de mooiste en meest verzorgde hond. Hiervoor worden drie verschillende juryleden geraadpleegd: een eigenaar van een hondenschool (jury A), een opleider van blindengeleide honden (jury B) en een hondenliefhebber en –kenner (jury C).

Er zijn 15 honden die aan de wedstrijd deelnemen en zij worden door de drie juryleden beoordeeld.

Wanneer we nu de punten van jury A (op de X-as) en jury B (op de Y-as) vergelijken bekomen we het volgende spreidingsdiagram:



Terwijl wanneer we de punten van jury A (op de X-as) en C (op de Y-as) vergelijken bekomen we dit spreidingsdiagram:



(a) Welk verschil zie je tussen de twee spreidingsdiagrammen?

(b) Hoe kan je dit verschil hier beschrijven in functie van de punten van de drie juryleden?

De ***sterkte van de correlatie*** is afhankelijk van de hoeveelheid punten die de correlatie volgen.
Bijvoorbeeld hoe meer punten de positieve correlatie volgen, hoe sterker de positieve correlatie is tussen de twee desbetreffende variabelen en hoe preciezer er een voorspelling kan gedaan worden voor de ene variabele op grond van de andere.

In het voorbeeld was er in het eerste spreidingsdiagram dus een ***zeer sterke positieve correlatie*** terwijl in het tweede spreidingsdiagram is er een ***zwakke positieve correlatie***.

☒ Voorbeeld:

Bekijken we opnieuw het voorbeeld van de kandidaat-soldaten:

Lengte x	165	167	170	172	175	175	180	189	192	195
Schoenmaat y	37	38	39	42	39	42	40	44	44	45

(a) We hadden besloten dat de lengte en de schoenmaat positief gecorreleerd waren.
Wat wilde dit juist zeggen?

(b) Bekijk de koppels (175,42) en (180,40).
Klopt dit voor deze twee koppels?

Let dus op: het concept van correlatie is slechts een ***statistische tendens***. Er kunnen dus punten zijn die niet aan die correlatie voldoen.

☒ Voorbeeld:

Een leerkracht geeft een toets maar de punten van de leerlingen zijn zo slecht, dat de leerkracht de leerlingen een tweede kans wil geven. Hij geeft een tweede toets over hetzelfde stukje leerstof.
Dit zijn de resultaten op de twee toetsen:

toets 1 x	2	2	3	5	4	6	2	5	8	5
toets 2 y	9	7	5	3	9	8	2	6	8	6

(a) Teken met je rekentoestel het spreidingsdiagram.

(b) Teken ook de rechte $y = x$ op het spreidingsdiagram.

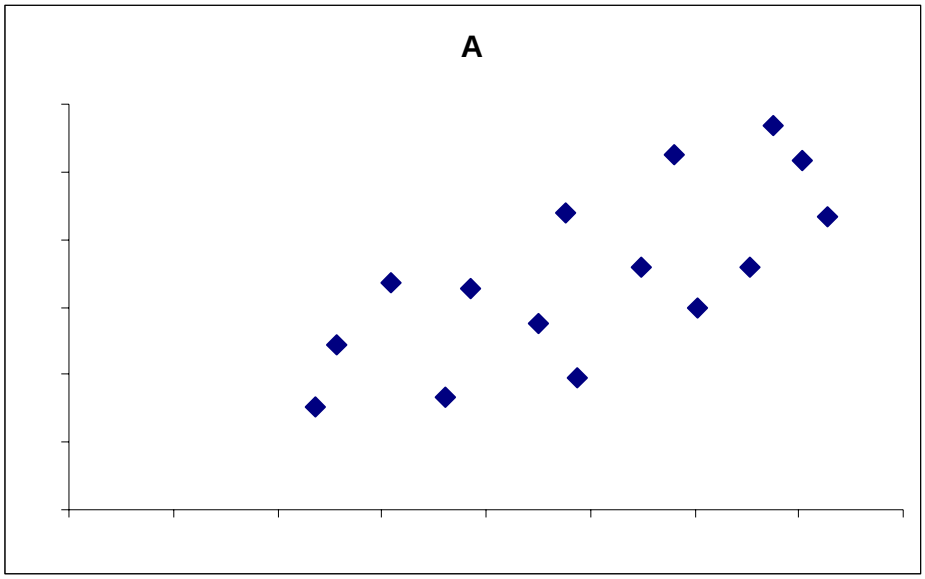
💻 **Tips voor de TI-83**:
Druk op Y= en vul in, bij de juiste plot waarin ook het spreidingsdiagram staat, Y1=X.
Druk dan op ZOOM 9:Zoomstat en je spreidingsdiagram wordt afgedrukt samen met de rechte $y = x$.

(c) Wat kun je zeggen over de punten die op het spreidingsdiagram gelegen zijn onder de rechte $y = x$? En wat betekent dit hier concreet in het voorbeeld?

(d) Als de leerkracht wil weten wie op de tweede toets meer behaalde dan op de eerste, naar wat moet hij dan juist kijken op het spreidingsdiagram?

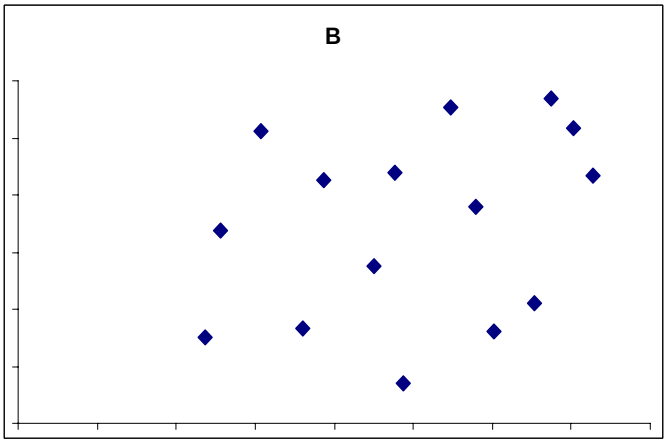
Om je wat te oefenen in het begrip “correlatie” de volgende oefeningen:

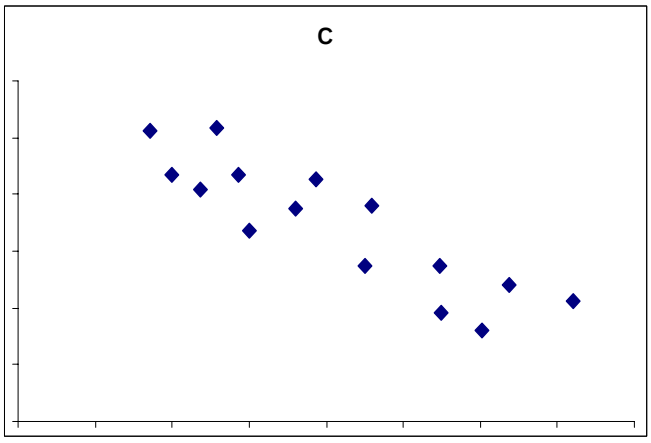
Onderzoeksopdracht 1
Beschouw het volgende spreidingsdiagram van hypothetische scores op een eerste en een tweede examen voor rijkswachtcommandanten om topfuncties bij de politie te kunnen bekleden:

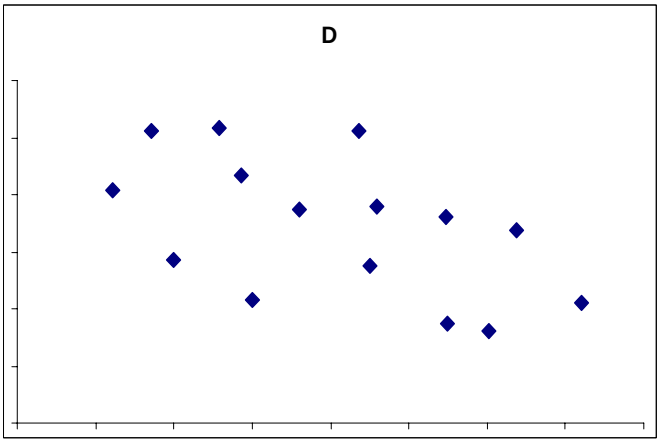


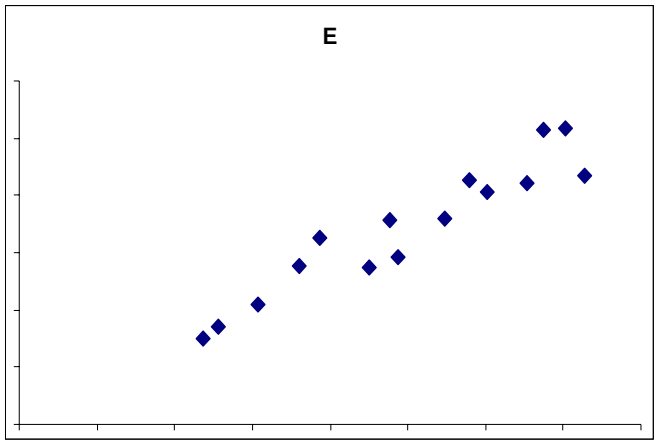
Beschrijf kort wat het spreidingsdiagram toont over de relatie tussen de scores op het eerste examen en die op het tweede examen. Met andere woorden als je de scores weet van het eerste examen, kun je dan vrij goede voorspellingen doen voor het tweede examen? Leg uit.

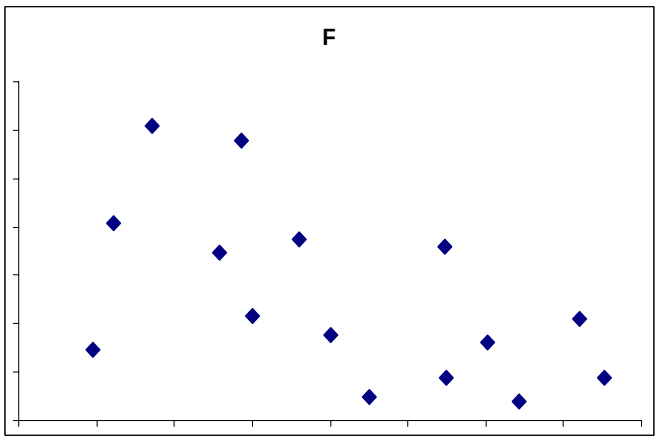
Onderzoeksopdracht 2
Hieronder staan 5 andere spreidingsdiagrammen over de hypothetische scores op de twee examens. Jouw taak is om de richting (positief of negatief) en de sterkte (sterk, gematigd of zwak) van de correlatie tussen de scores op het eerste examen en die op het tweede examen te onderzoeken voor elk voorbeeld.











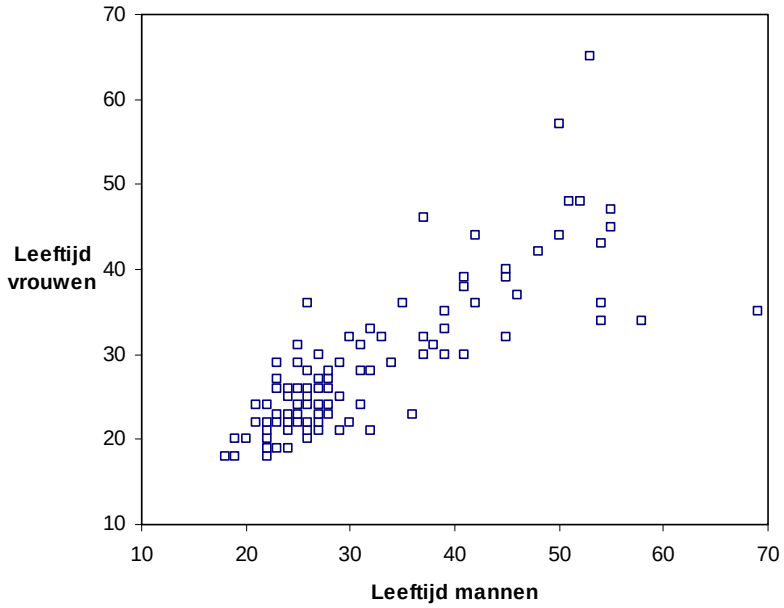
Doe dit door de bijhorende letter (A, ..., F) in te vullen in de tabel hieronder. Elke letter mag slechts één maal gebruikt worden.

	sterk	gematigd	zwak
negatief			
positief			

C. Voorbeelden

Voorbeeld 1

We vragen ons af of er een verband bestaat tussen de leeftijd van huwelijkspartners. We besluiten dit te onderzoeken d.m.v. een representatieve steekproef. Daartoe kloppen we aan bij de dienst Burgerlijke Stand van ons gemeentehuis. Hier verschaft men ons volgende gegevens: van 120 willekeurig gekozen huwelijken in het jaar 1992 noteerden we de huwelijksdatum en de geboortedata van de partners. Hieruit hebben we dan de leeftijden van de partners op hun huwelijksdag afgeleid.
Zo bekomen we 120 koppels (x_i, y_i) met x_i de huwelijksleeftijd van de man en y_i de leeftijd van zijn vrouw.
We zetten deze punten nu uit en bekomen zo het volgende spreidingsdiagram:



(a) Lijkt er een correlatie te zijn tussen de leeftijd van de twee partners? Zo ja, is deze positief of negatief? En zou je deze als sterk of als zwak karakteriseren? Leg uitgebreid uit.

(b) Zijn er veel koppels die even oud zijn? Hoe kan je dit zien op het spreidingsdiagram?

(c) Zijn er meer mannen die ouder zijn dan hun vrouw of komt het omgekeerde vaker voor? Hoe zie je dit op het spreidingsdiagram?

(d) Vat in eigen woorden samen wat je kan leren over de huwelijksleeftijd van koppels door rekening te houden met de lijn $y = x$.

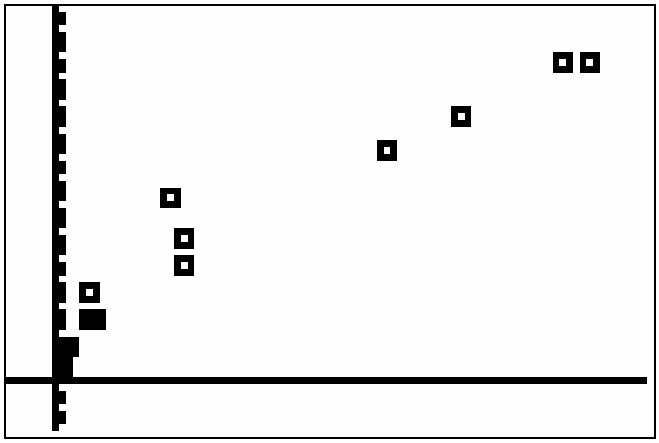
Voorbeeld 2

We bekijken de volgende data van het gewicht en het voetoppervlak van 20 Zuid-Amerikaanse slakken van de soort Biomphalaria Glabrata.

Gewicht (g)	0.64	0.21	0.85	0.53	0.02	0.01	0.21	0.18	0.06	0.20	0.07	0.01	0.05
Voetopp (mm^2)	29	16	35	25	4	1	16	20	7	13	7	3	10

Gewicht (g)	0.81	0.53	0.18	0.06	0.20	0.07	0.01
Voetopp (mm^2)	35	25	20	7	13	7	1

Wanneer we van deze gegevens het spreidingsdiagram tekenen, bekomen we de volgende grafiek:

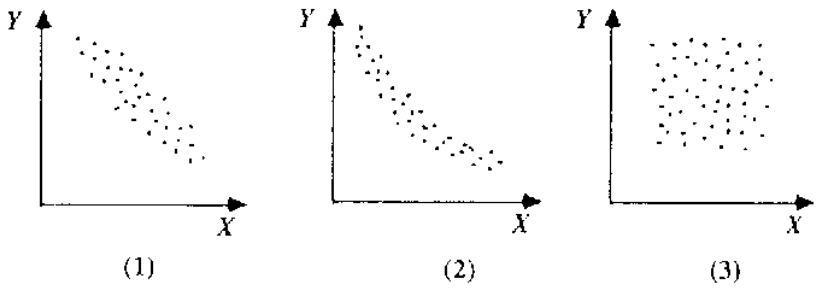


(a) Zie je of er punten samenvallen?

De punten op het spreidingsdiagram liggen duidelijk zeer dicht bij een rechte lijn. We zeggen dat er een ***benaderend lineair verband*** is tussen het gewicht en het voetoppervlak van de slakken.

(b) Welk soort correlatie is er hier?

Natuurlijk is er niet altijd een lineair verband met een positieve correlatie tussen de twee variabelen. Andere mogelijkheden zijn geïllustreerd in de volgende spreidingsdiagrammen:

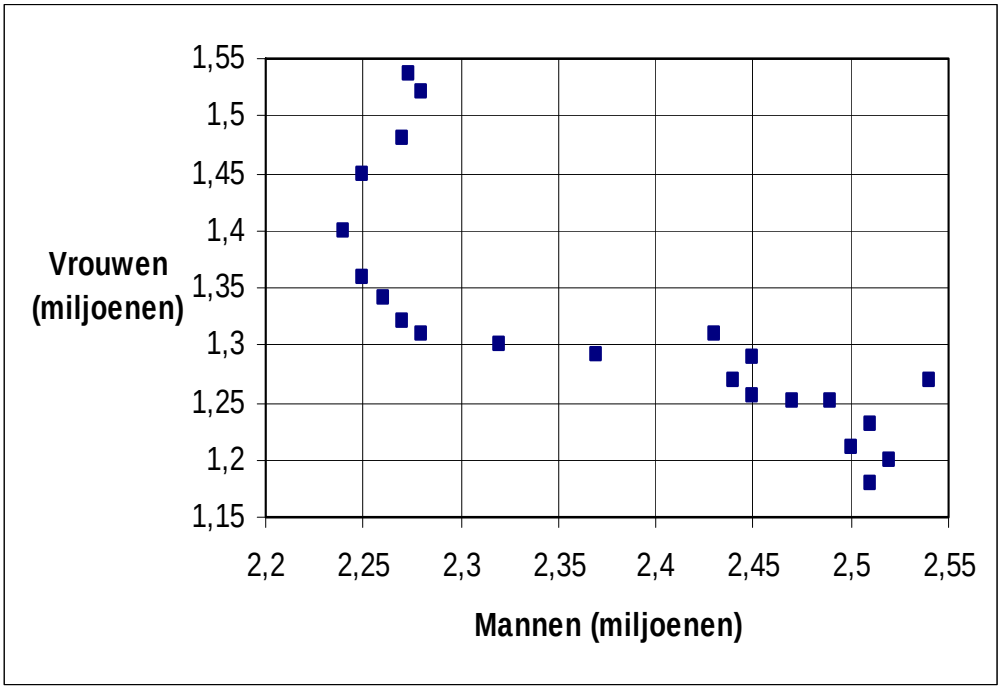


(c) Hoe zou jij, zo nauwkeurig mogelijk, het verband tussen X en Y beschrijven in deze drie gevallen?

Het doel van de komende lessen is te onderzoeken wanneer het aanvaardbaar is een lineaire benadering te gebruiken voor de relatie tussen twee variabelen. En, wanneer dit zo is, de vergelijking te vinden van die best benaderende rechte.
Het volgende voorbeeld toont ons een duidelijk niet-lineair verband.

Voorbeeld 3

Van het kabinet van het Ministerie van Tewerkstelling en Arbeid kregen we cijfers over de evolutie van de tewerkgestelde beroepsbevolking per geslacht in de periode 1970-1991. Hieruit konden we afleiden dat een groter aantal tewerkgestelde vrouwen gepaard gaat met een kleiner aantal tewerkgestelde mannen.



Kunnen we hier spreken van een lineair verband?

HOOFDSTUK 2:
COVARIANTIE ALS SPREIDINGSMAAT

A. Opbouw en interpretatie van het begrip covariantie

De bedoeling van dit hoofdstuk is een eerste maat vinden waarmee we spreiding kunnen uitdrukken bij een bivariate verdeling. We willen over spreiding kunnen spreken aan de hand van een getal en niet meer alleen op basis van het spreidingsdiagram.

Mannelijke krekels sjirpen door hun vleugels tegen elkaar te wrijven. Men wil het verband tussen de sjirpfrequentie en de temperatuur nagaan in drie verschillende landen: België, Zweden en Frankrijk.
Hieronder staan de (fictieve) gegevens voor 12 verschillende testen per land:

België		Zweden		Frankrijk	
temperatuur (°C) x	sjirpfrequentie y	temperatuur (°C) x	sjirpfrequentie y	temperatuur (°C) x	sjirpfrequentie y
17	8.4	13	7.7	24	3.5
19	8.6	14	6.1	28	4.5
20	7.8	17	9.2	31	2.8
22	7.2	18	7.9	33	2.9
25	7.1	18	10.4	34	3.4
26	6.1	18	11.3	35	2.0
28	6.5	21	11.7	35	5.0
29	5.8	22	13.0	38	3.8
31	5.4	22	14.9	41	4.4
33	4.8	23	13.9	43	2.9
33	4.2	26	14.8	48	4.3
35	4.8	26	15.8	49	3.5

(1) Teken op je rekentoestel het spreidingsdiagram voor de drie landen. Merk op dat bij één bepaalde x-waarde in een land verschillende y-waarden kunnen optreden.

Tips voor de TI-83:

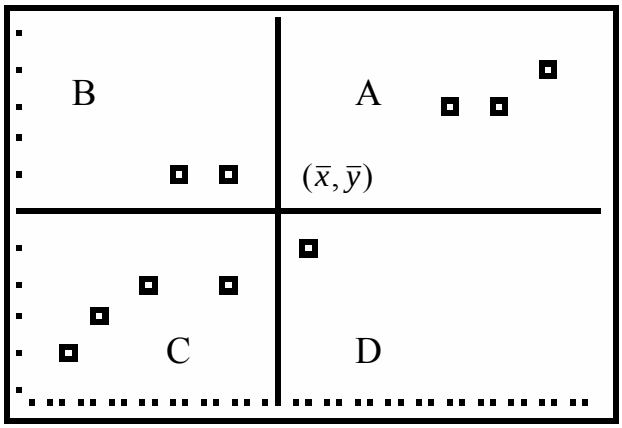
- Gebruik de lijsten L1 en L2 voor België, L3 en L4 voor Zweden en L5 en L6 voor Frankrijk.
  Pas dan ook telkens Xlist en Ylist aan, wanneer je plot1, plot2 en plot3 definieert.
- Vergeet niet plot1 en plot3 op Off te zetten wanneer je bv plot2 gaat tekenen, anders komen de spreidingsdiagrammen gewoon over elkaar.

(2) Bekijk de spreidingsdiagrammen en becommentarieer de correlaties tussen de temperatuur en de sjirpfrequentie voor elk land afzonderlijk.

Het doel van dit werkblad is correlatie nauwkeuriger te kunnen beschrijven. Tot nu toe baseerden we ons alleen op het spreidingsdiagram om te oordelen of er positieve of negatieve correlatie is en of de correlatie sterk is of zwak.
We willen nu een getal gaan ontwikkelen waarmee we de correlatie kunnen beoordelen.

Wanneer we maar één enkele variabele X hebben, geeft de variantie de spreiding weer t.o.v. het gemiddelde $\bar{x}$.
Bij een bivariate verdeling, zijn er twee variabelen X en Y van belang. We hebben hier twee gemiddelden $\bar{x}$ en $\bar{y}$ voor handen.
Om nu de spreiding van de punten in de bivariate verdeling te kunnen beschrijven, moeten we rekening houden met beide gemiddelden $\bar{x}$ en $\bar{y}$.
We tekenen dus in het spreidingsdiagram een nieuwe x- en y-as door het punt ($\bar{x}$, $\bar{y}$).

Het volgende diagram toont opnieuw het spreidingsdiagram van het voorbeeld van de kandidaat-soldaten.



We weten dat de lengte en de schoenmaat van de kandidaat-soldaten positief gecorreleerd zijn.
Vraag 3 verwijst naar dit spreidingsdiagram.

(3) Deze vraag verwijst naar het spreidingsdiagram hierboven.
   (a) In welke twee kwadranten (A, B, C of D) liggen de meeste punten? Waardoor komt dit?

   (b) Moesten de lengte en de schoenmaat negatief gecorreleerd zijn, in welke twee kwadranten zouden dan het grootst aantal punten liggen? Waarom?

   (c) En als de twee variabelen ongecorreleerd zijn, in welke kwadranten liggen dan de meeste punten?

(d) Als een punt (x, y) in kwadrant A ligt, wat is dan het teken van

- $x - \bar{x}$
- $y - \bar{y}$
- $(x - \bar{x})(y - \bar{y})$?

Beantwoord nu dezelfde vragen wanneer (x, y) in respectievelijk de kwadranten B, C en D ligt.
Vul alles in, in de volgende overzichtstabel: schrijf een "+" voor positief en een "–" voor negatief.

	A	B	C	D
$x - \bar{x}$				
$y - \bar{y}$				
$(x - \bar{x})(y - \bar{y})$				

Beschouw nu $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

(e) Wat zal het teken zijn van $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ als de twee variabelen positief gecorreleerd zijn?
Houd hiervoor rekening met (a) en de tabel in (d).
Leg uit waarom.

(f) Wat zal het teken zijn van $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ als de twee variabelen negatief gecorreleerd zijn?
Houd hiervoor rekening met (b) en de tabel in (d).
Leg ook hier uit waarom je dit denkt.

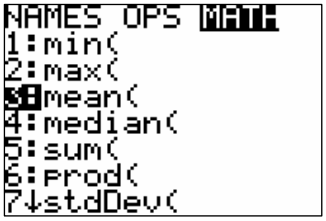
(g) Wat zal het teken zijn van $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ als de twee variabelen ongecorreleerd zijn?
Houd hiervoor rekening met (c) en de tabel in (d).
Leg ook hier uit waarom.

(h) Leg uit (m.b.v. de drie vorige vragen) waarom $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ een goede maat is voor correlatie.

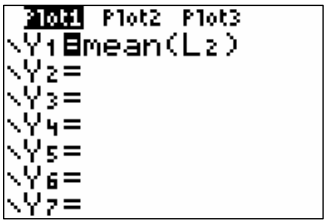
(4) Herneem nu het voorbeeld van de sjirpende krekels uit vraag 1.
Vind voor België $\bar{x}$ en $\bar{y}$.

Teken dan op het bijbehorende spreidingsdiagram de twee evenwijdige assen aan de oorspronkelijke assen, door het punt ($\bar{x}$, $\bar{y}$).

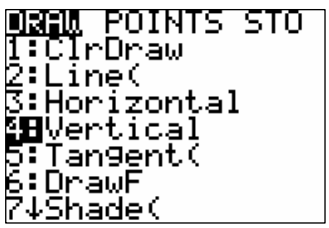
Tips voor de TI-83:



Om het gemiddelde te vinden van de elementen van een lijst, gebruik je 2nd STAT (LIST) MATH 3:mean(en dan typ je de naam van de lijst in.



Om een horizontale lijn tezamen met je spreidingsdiagram te tekenen, druk je Y=. Je vult dan in bv mean(L2) zoals het hierboven beschreven staat.
Kijk wel na of je in de juiste plot werkt: ben je bezig met de gegevens van België, dus in plot 1, moet dit aangeduid zijn bovenaan het scherm.



Om nu nog een verticale erbij te tekenen, druk je op: 2nd MODE (QUIT) om terug te keren naar het basisscherm, dan op 2nd PRGM (DRAW) DRAW 4:Vertical.
Je typt dan (mean(L1)) zoals aangeleerd hier hoger en drukt op ENTER.

(5) Bereken nu voor België $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ en leg de betekenis uit van dit resultaat. Gebruik hierbij je bevindingen in vraag 3(f) en in vraag 4.

(6) Doe nu hetzelfde voor de twee andere landen.

B. Eigenschappen en zwakheden van de covariantie als spreidingsmaat

Veronderstel dat de steekproef $(x_1, y_1),(x_2, y_2),...,(x_n, y_n)$, genomen uit de bivariate verdeling (X,Y), een steekproefgemiddelde $(\bar{x}, \bar{y})$ heeft.
Dan definiëren we de steekproefcovariantie van de data (x_i, y_i) als volgt:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Wanneer de covariantie ***dicht bij 0*** ligt, wil dit zeggen dat er weinig of geen correlatie is tussen de desbetreffende variabelen.
Een ***"grote" positieve covariantie*** wijst op een positieve correlatie tussen de twee variabelen en andersom wijst een ***"grote" negatieve covariantie*** op een negatieve correlatie tussen de twee variabelen.

We tonen nu de zwakheid aan van de covariantie als spreidingsmaat:

☒ Voorbeeld:

We hernemen het voorbeeld van de kandidaat-soldaten.
De lengte X in de bivariate verdeling (X,Y) wordt nu uitgedrukt in centimeter.
Wat zal het effect zijn op de covariantie als ze uitgedrukt zal worden in meter?

Oplossing:

Elke lengte x_i, uitgedrukt in centimeter, wordt vervangen in de berekeningen door een lengte uitgedrukt in meter, $\frac{1}{100}x_i$.
We weten dat dan ook het gemiddelde door 100 zal moeten gedeeld worden.
De covariantie verandert van $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ naar $\frac{1}{n-1}\sum_{i=1}^{n}(\frac{1}{100}x_i - \frac{1}{100}\bar{x})(y_i - \bar{y})$.
De laatste uitdrukking kan geschreven worden als:

$$\frac{1}{n-1}\cdot\frac{1}{100}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

En zo wordt de oorspronkelijke covariantie 100 keer kleiner.

Wat gebeurt er met s_{xy} als alle waarden van X vermenigvuldigd worden met 5 en alle waarden van Y met 4?

Het is duidelijk dat begrippen als "een grote covariantie" heel relatief zijn en afhangen van de grootte van de variabelen en hun eenheden.
We kunnen dus de grootte van de covariantie niet gebruiken om over spreiding te spreken.
Een nieuwe spreidingsmaat dringt zich op.

HOOFDSTUK 3:
DE CORRELATIECOËFFICIËNT ALS BETERE SPREIDINGSMAAT

A. De correlatiecoëfficiënt met zijn belangrijkste eigenschappen

In het vorige hoofdstuk zagen we dat de covariantie toch niet zo geschikt bleek te zijn als spreidingsmaat.
Daarom voeren we een nieuw begrip in: de (steekproef)correlatiecoëfficiënt *r*.

De correlatiecoëfficiënt vinden we door de covariantie te delen door het product van de standaardafwijkingen van de data x_i en y_i :

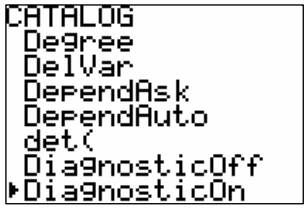
$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

⌦ Oefening:

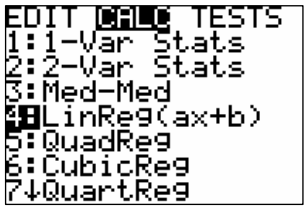
Wat gebeurt er met r als alle waarden van X vermenigvuldigd worden met 5 en alle waarden van Y met 4? Is de correlatiecoëfficiënt nog eenheidsgebonden?

💻 Met de TI-83

We kunnen met de TI-83 ook rechtstreeks de correlatiecoëfficiënt berekenen.



Plaats hiertoe je data van de stochastische veranderlijke X in de lijst L1 en die van Y in de lijst L2.
Druk op 2nd 0 (CATALOG) en ga naar DiagnosticOn, druk dan tweemaal op ENTER. Dit moet je maar één keer doen met je toestel, dit dient om extra gegevens te krijgen, waaronder de correlatiecoëfficiënt.



Druk dan op STAT CALC 4:LinReg(ax+b). En daar verschijnt dan r .
Wat a en b betekenen, zal in het volgende hoofdstuk uitgelegd worden.

⌦ Oefening:

Vind de correlatiecoëfficiënt voor de bivariate verdeling van het voorbeeld van de slakken.

Gewicht (g)	0.64	0.21	0.85	0.53	0.02	0.01	0.21	0.18	0.06	0.20	0.07	0.01	0.05
Voetopp (mm^2)	29	16	35	25	4	1	16	20	7	13	7	3	10

Gewicht (g)	0.81	0.53	0.18	0.06	0.20	0.07	0.01
Voetopp (mm^2)	35	25	20	7	13	7	1

⌦ Oefening:

Er is een verband tussen het aantal voertuigen in Nederland en het aantal verkeersongevallen per jaar.
In de jaren ’70 waren de aantallen als volgt:

jaar	‘70	‘71	‘72	‘73	‘74	‘75	‘76	‘77	‘78	‘79
Voertuigen (milj) x	2.6	3.1	3.5	3.7	4.1	4.4	4.6	4.9	5.3	5.8
Ongelukken (x 1000) y	138	163	166	153	177	201	216	208	226	238

Bereken de correlatiecoëfficiënt.

⌦ Oefening:

We weten uit de definitie dat $r = \dfrac{s_{xy}}{s_x \cdot s_y}$.

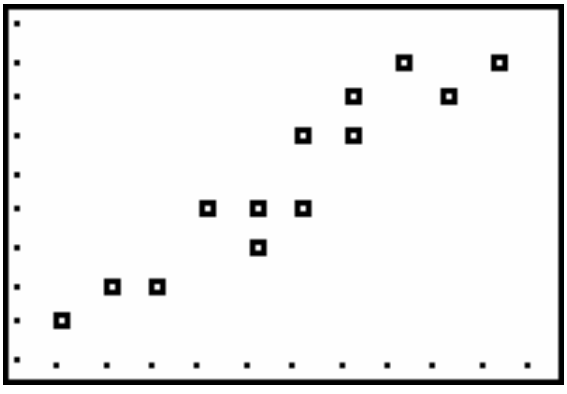
(a) Wat is het teken van s_{xy} als X en Y positief gecorreleerd zijn?

(b) En wat is het teken van s_{xy} als X en Y negatief gecorreleerd zijn?

(c) Wat is dan het teken van r in beide gevallen? Waarom?

☒ Voorbeeld:

We hernemen het voorbeeld van de schoonheidswedstrijd voor honden:
Wanneer we een spreidingsdiagram tekenen van de scores van jury A tov die van jury B, bekomen we het volgende:

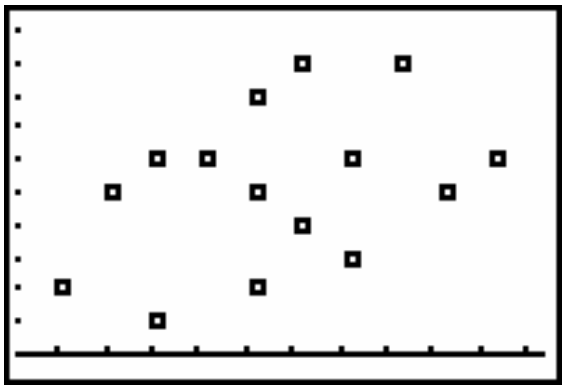


(a) Welk soort correlatie is er hier ook weer?

(b) Welk teken zou dan de correlatiecoëfficiënt moeten hebben?

Wanneer we deze met ons rekentoestel berekenen, bekomen we dat $r = 0.95$.

Wanneer we nu het spreidingsdiagram tekenen van de scores van jury A ten opzichte van die van jury C, bekomen we:



(c) Welk soort correlatie is er hier?

(d) Welk teken zou dan de correlatiecoëfficiënt moeten hebben?

Wanneer we deze met ons rekentoestel berekenen, bekomen we dat $r = 0.39$.

(e) Welk zou, denk je, de bovengrens zijn van de correlatiecoëfficiënt als er een positieve correlatie is? Wat is het verband tussen deze bovengrens en de sterkte van de correlatie?

(f) Heb je een vermoeden wat r zal zijn bij zwakke negatieve correlatie en bij sterke negatieve correlatie?

Controleer je vermoeden met het volgende voorbeeld:

☒ Voorbeeld:

We hernemen het voorbeeld van de sjirpende krekels in België. Daar waren de temperatuur en de sjirpfrequentie sterk negatief gecorreleerd.

temperatuur (°C)	17	19	20	22	25	26	28	29	31	33	33	35
sjirpfrequentie	8.4	8.6	7.8	7.2	7.1	6.1	6.5	5.8	5.4	4.8	4.2	4.8

(a) Bereken r en kijk na of je vermoeden van in de vorige vraag klopt?

(b) Welke ondergrens is er, denk je, voor r?

Tot slot, kijken we nog eens wat de correlatiecoëfficiënt gaat zijn, als de twee variabelen ongecorreleerd zijn:

☒ Voorbeeld:

We hernemen het voorbeeld van de sjirpende krekels in Frankrijk. Daar waren de temperatuur en de sjirpfrequentie ongecorreleerd.

temperatuur (°C)	24	28	31	33	34	35	35	38	41	43	48	49
sjirpfrequentie	3.5	4.5	2.8	2.9	3.4	2.0	5.0	3.8	4.4	2.9	4.3	3.5

(a) Bereken r.

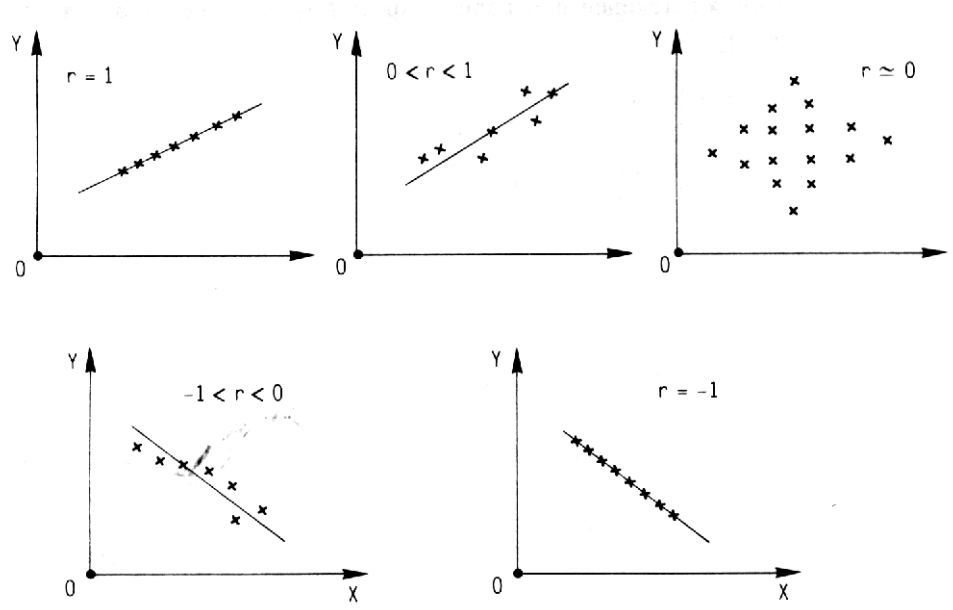
(b) Hoe zal de correlatiecoëfficiënt zijn als twee variabelen ongecorreleerd zijn?

Daar waar de covariantie ons weinig betekenis en verklaring kon geven omwille van het subjectieve idee van "groot" en "klein", kan de correlatiecoëfficiënt ons meer exacte informatie geven.

- Wanneer r dicht ligt bij 1, kan dit wijzen op een sterke tendens dat grote x_i-waarden overeenkomen met grote y_i-waarden en dat kleine x_i-waarden overeenkomen met kleine y_i-waarden.
  We spreken van een ***sterke positieve lineaire correlatie***.
- Wanneer r dicht ligt bij -1, kan dit wijzen op een sterke tendens dat kleine x_i-waarden overeenkomen met grote y_i-waarden en omgekeerd.
  We spreken van een ***sterke negatieve lineaire correlatie***.
- Wanneer r dicht ligt bij 0, kan dit erop wijzen dat er geen bepaalde tendens is onder de koppels (x_i, y_i).
  We spreken van een ***zwakke lineaire correlatie***.

Wanneer r dicht ligt bij 1, verwachten we dus dat de punten (x_i, y_i) dicht liggen bij een rechte met positieve richtingscoëfficiënt. Terwijl als r dicht ligt bij –1, verwachten we dat de punten (x_i, y_i) dicht liggen bij een rechte met negatieve richtingscoëfficiënt.

Tot slot geven de volgende figuren een overzicht van de mogelijke r-waarden.



B. Enkele oefeningen

(1) De zee- en luchttemperatuur op een stukje strand in Florida werd, gedurende 10 weken, elke maandagmiddag gemeten.
Dit leverde de volgende gegevens op:

zee x (°C)	19	22	18	19	21	22	18	18	17	16
lucht y (°C)	29	34	27	29	33	35	28	27	26	25

(a) Maak een spreidingsdiagram met je rekentoestel.
(b) Bereken r en verklaar je resultaat.

(2) Teken de volgende gegevens in één spreidingsdiagram.
Set 1:

x	1	4	5	7
y	11	5	3	-1

Set 2:

x	0	2	5	6
y	20	12	0	-4

Duid hierbij de gegevens van set 1 aan met een kruisje en die van set 2 met een vierkantje.
Bereken r voor elk set en becommentarieer het resultaat.

(3) Men mat de periode T (in seconden) van 7 staanklokken van een verschillende hoogte. Dit gaf de volgende resultaten:

H (cm)	10	20	30	40	50	60	70
T (s)	0.63	0.90	1.10	1.27	1.42	1.56	1.68

(a) Bereken r voor deze gegevens.
(b) Teken het spreidingsdiagram.
(c) Denk je dat de relatie tussen H en T lineair is? Waarom?

(4) Teken een spreidingsdiagram van de volgende vier punten: (1,1), (1,3), (3,1) en (3,3).
Bereken r en verklaar zijn waarde.

C. Enkele bijzonderheden

(1) Beschouw de volgende hypothetische scores (op 100) op twee examens:

examen 1	49	52	55	58	61	64	67	70	73	76	79	82	85	88	91
examen 2	95	81	69	59	51	45	41	39	41	45	51	59	69	81	95

(a) Teken het spreidingsdiagram met je rekentoestel.
Lijkt er een verband te zijn tussen de twee examenscores?
Zo ja, beschrijf deze relatie.

(b) Bereken de correlatiecoëfficiënt.
Verbaast dit resultaat je? Wat had je verwacht?

Dit voorbeeld illustreert dat de correlatiecoëfficiënt enkel een *lineair* verband meet tussen twee variabelen. Meer ingewikkelde relaties kunnen met r niet opgemerkt worden.
Dus kan er een verband bestaan tussen twee variabelen, zelfs als de correlatiecoëfficiënt dicht bij 0 ligt. Je moet je dus bewust zijn van deze mogelijkheid en je niet louter baseren op de waarde van r om een besluit te trekken. Onderzoek zeker ook steeds het spreidingsdiagram.

(2) Beschouw de spreidingsdiagrammen van de volgende hypothetische examenscores: teken ze met je rekentoestel.
A:

examen 1	49	52	55	58	61	64	67	70	73	76	79	82	85	88	99
examen 2	54	57	60	62	65	68	70	73	76	78	81	84	87	89	12

B:

examen 1	12	52	55	58	61	64	67	70	73	76	79	82	85	88	91
examen 2	17	62	52	73	69	71	72	80	58	60	67	52	76	69	70

(a) In klas A lijken de meeste observaties een lineair patroon te volgen. Zijn er uitzonderingen?

(b) Terwijl in klas B lijken de meeste observaties eerder willekeurig geplaatst zonder een echt patroon. Zijn er hier uitzonderingen?

(c) Bereken voor beide klassen de correlatiecoëfficiënt. Ben je verrast over één of beide resultaten? Waarom?

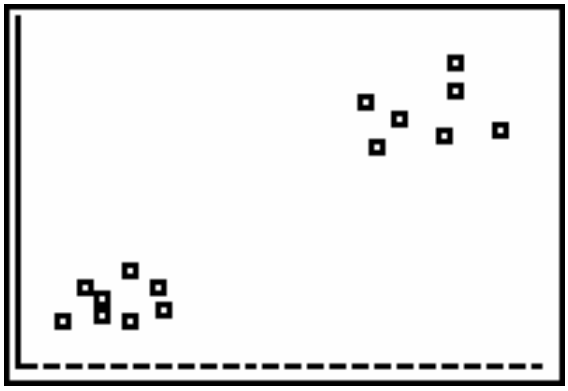
Een punt dat volledig uitspringt uit het patroon in het spreidingsdiagram, noemen we een ***uitschieter***. Dit is meestal te wijten aan een foute meting of een vergissing bij het opschrijven van de gegevens.
Maar soms zijn de gegevens juist en gaat het gewoon om een uitzondering:

- Een leerling met een dikke buis, terwijl de rest van de klas bijna het maximum haalde.
- Een slak met uitzonderlijk kleine voetjes.
- Een heel oude man die trouwt met een vrouw van 20 jaar.
- …

Uitschieters hebben vaak een grote invloed op de correlatiecoëfficiënt waardoor we soms verkeerde besluiten zouden trekken.
Het belang van het spreidingsdiagram te bekijken, is ook hier weer bewezen.
Meestal laat men voor de berekening van r de uitschieters gewoon weg.

(d) Verwijder de uitschieters uit beide klassen en bereken opnieuw de correlatiecoëfficiënt. Becommentarieer hoe deze veranderd zijn en leg uit.
Tip voor de TI-83.
Om een element in een rij (lijst) te verwijderen, druk je STAT EDIT, je gaat op het element staan en drukt DEL.
Let op: vergeet het overeenkomstige element niet te verwijderen, als het over een bivariate verdeling gaat.

(3) Beschouw het volgende spreidingsdiagram van een set hypothetische examenscores.



De gegevens zijn:

examen 1	37	44	32	37	35	41	41	45	88	72	75	81	82	71	82
examen 2	39	41	33	34	41	45	33	36	78	74	81	77	94	84	87

(a) Beschrijf wat het spreidingsdiagram jou vertelt over het verband tussen de twee examenresultaten.

(b) Bereken r. Is zijn waarde hoger dan je verwachtte?

Hier zien we dat, zelfs als er geen uitschieters zijn of geen ingewikkelder verband, de correlatiecoëfficiënt toch nog groot kan zijn, hoewel er geen lineaire relatie is tussen de twee variabelen.

(4) De correlatiecoëfficiënt wordt door onderzoekers gebruikt in veel domeinen van sociale wetenschappen tot landbouwwetenschappen. In het algemeen probeert men, aan de hand van de correlatiecoëfficiënt, te bewijzen dat de verandering van één eigenschap leidt tot de verandering van iets anders.
Bijvoorbeeld dat de stijgende werkloosheid een stijging in criminaliteit veroorzaakt.
Eigenlijk kan een resultaat waarbij $r \approx 1$ of $r \approx -1$ op drie verschillende manieren geïnterpreteerd worden:

Als y stijgt wanneer x stijgt:
- kan de stijging van x de stijging van y veroorzaakt hebben of omgekeerd, is de stijging van y de oorzaak van het stijgen van x.
- kunnen beide stijgingen een gemeenschappelijke oorzaak hebben.
- kunnen beide stijgingen totaal niets met elkaar te maken hebben.

Het is dan de taak van de onderzoeker om uit te maken op welke van de drie manieren r moet worden geïnterpreteerd.
Hieruit volgt dus dat op zich, een resultaat $r \approx 1$ of $r \approx -1$, geen informatie geeft over "het veroorzaken".

Om dit te illustreren, bekijken we het volgende voorbeeld:
De volgende tabel geeft informatie over de levensverwachting van de inwoners van 22 landen. Het geeft ook het aantal mensen per televisietoestel in elk land.

land	levensver-wachting	Mensen per TV	land	levensver-wachting	Mensen per TV
Angola	44	200	Mexico	72	6.6
Australië	76.5	2	Marokko	64.5	21
Cambodja	49.5	177	Pakistan	56.5	73
Canada	76.5	1.7	Rusland	69	3.2
China	70	8	Zuid-Afrika	64	11
Egypte	60.5	15	Sri Lanka	71.5	28
Frankrijk	78	2.6	Oeganda	51	191
Haïti	53.5	234	UK	76	3
Irak	67	18	VS	75.5	1.3
Japan	79	1.8	Vietnam	65	29
Madagaskar	52.5	92	Jemen	50	38

(a) Welk land heeft het minst aantal mensen per televisietoestel? En welk land het meest? Wat betekenen juist deze getallen?

(b) Teken met je rekentoestel het spreidingsdiagram van de levensverwachting versus het aantal mensen per televisietoestel.
Lijkt er een verband te zijn tussen de twee variabelen? Beschrijf dit verband kort.

(c) Bereken r.

(d) Omdat de correlatie zo sterk negatief is, zou men kunnen besluiten dat men, in de landen met een lagere levensverwachting, de mensen langer kan doen leven door veel televisietoestellen naar die landen te sturen.
Becommentarieer deze uitspraak.

(e) Welke van de drie hoger beschreven factoren verklaart hier de stijging van de levensverwachting in functie van de stijging van het aantal mensen per televisietoestel? Leg uit.

⌦ Oefening:

We hebben gezien dat er drie verschillende manieren bestaan om een sterke correlatie te beoordelen.

(a) In welke categorie zou je het verband tussen lengte en gewicht plaatsen?

(b) In de jaren ’80 was er een stevige toename in het aantal studenten in Sheffield. In dezelfde stad was er toen ook een ferme stijging in autodiefstallen. In welke categorie zou je dit voorbeeld plaatsen?

(c) In veel gemeenschappen vindt men een sterke positieve correlatie tussen de smaak van ijs, die in een gegeven maand het meest verkocht wordt en het aantal verdrinkingen door zelfmoord die zich die maand voordoen. Betekent dit dat ijscrème verdrinking veroorzaakt? Indien niet, kan je dan een alternatieve verklaring geven voor deze sterke correlatie?

D. Individueel project

De bedoeling bij dit individueel project is, dat je duidelijk laat merken dat je alles tot nu toe goed begrijpt. Het is dan ook een soort van controle voor jezelf: als je, tijdens het maken van deze opdracht, ergens moeilijkheden mee hebt, ga dat onderdeel dan terug bekijken in je nota's. En maak eventueel enkele oefeningen opnieuw of vraag uitleg.
Je krijgt ruim de tijd om je gegevens te verzamelen, een grondige analyse uit te voeren en je verslag te maken. Achteraf is het dan ook de bedoeling dat jullie je project kort komen presenteren voor je klasgenoten.
Zorg dus dat je alles volledig door hebt, zodat je eventuele vragen rustig kan beantwoorden. steek hier voldoende tijd in zodat je zeker bent van je analyse en je besluiten.

De opdracht zelf nu:

- Kies twee variabelen die gemeten of geteld kunnen worden en waarvan je vermoedt dat er een bepaalde relatie tussen bestaat. De variabelen kunnen zowel eigenschappen van mensen, van dieren of van dingen zijn.
  Denk hier lang genoeg over na, neem niet de eerste de beste. Probeer er twee variabelen uit te kiezen waarvan hun relatie je klasgenoten zal verbazen en interesseren.
  Misschien kun je hierover onderling wat brainstormen.
- Verzamel gegevens voor de twee gekozen variabelen en dit bij ten minste 20 verschillende bronnen.
  Bijvoorbeeld 20 verschillende mensen, dieren, testen, appelsienen, …
- Toon dat je goed begrijpt wat je twee variabelen juist betekenen, door ze uitgebreid in woorden te beschrijven.

- Leg ook uit hoe je je gegevens hebt verzameld en gemeten. Leg uit hoe je ervoor gezorgd hebt dat je gegevens representatief zijn. (bijvoorbeeld neem niet alle 20 appelsienen uit dezelfde winkel, want zo zou je steekproef onderhevig kunnen zijn aan externe factoren omdat die winkel bijvoorbeeld altijd kleinere appelsienen verkoopt)
- Maak een spreidingsdiagram voor je variabelen. Wees nauwkeurig bij je tekening. Je kan controleren met je rekentoestel.
- Zeg welke variabele je op de *X*-as hebt geplaatst en welke op de *Y*-as en leg je keuze uit.
- Geef uitleg over het verband dat er lijkt te zijn tussen jouw twee gekozen variabelen. Kan je formuleren waarom je dit verband lijkt te zien?
  Als je denkt dat er geen relatie is, verklaar dit dan ook. (en kies opnieuw twee variabelen)
- Is de correlatie positief of negatief? Zwak, gematigd of sterk?
- Toon aan dat je de relatie volledig begrijpt door ze duidelijk en volledig te beschrijven in woorden: pas hiervoor je statistische besluiten toe op je concrete voorbeeld, dus op de twee variabelen die jij hebt gekozen. Wat betekent deze relatie voor de twee variabelen?
- Kan je nu voorspellingen doen over jouw variabelen, wanneer je sommige gegevens niet hebt?
- Bereken de correlatiecoëfficiënt en geef hier wat uitleg over: wat soort correlatie geeft r aan, klopt deze correlatie met de werkelijkheid,…
- Kan je zeggen dat de ene variabele een stijging of daling in de andere variabele veroorzaakt? Of hoe interpreteer jij anders het verband tussen de twee variabelen? Is er een verborgen factor die beide variabelen beïnvloedt, of is hun verband louter toevallig?

Probeer nu aan de hand van deze vragen een samenhangende tekst te schrijven die zo goed en zo duidelijk mogelijk jouw analyse weergeeft.
Houd, terwijl je hieraan werkt, je voorblad dat op de volgende bladzijde staat, zorgvuldig bij en vul telkens de datum in wanneer je een onderdeel beëindigd hebt.
Wanneer je werkje af is, vul dan ook het evaluatieblad in en gebruik dit als “cover” voor je werkje.

Voorblad

Gebruik dit blad als cover van je werkje. Verzin ook zelf een titel voor jullie werkje. Je vindt hier ook de verschillende onderdelen die in je verslag zeker aan bod moeten komen. Schrijf de datum naast elk onderdeel wanneer je dit beëindigd hebt. Zo doe je zelf een soort van controle of je alle opdrachten wel hebt uitgevoerd.

Naam:

……

Titel:

……

Verschillende onderdelen:

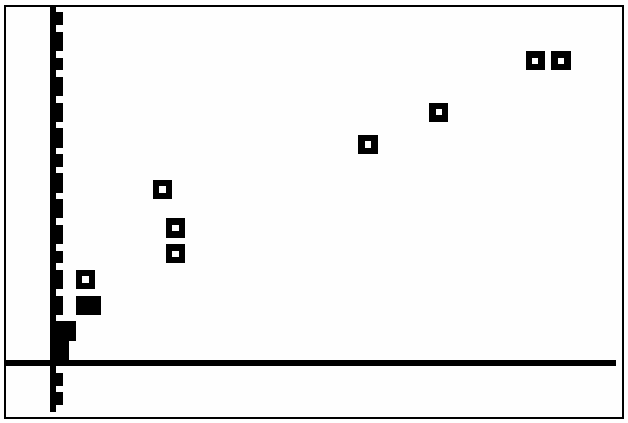
	Ik heb gegevens verzameld om te zien of de twee variabelen gecorreleerd zijn. De twee variabelen zijn: …… ……
	Ik heb duidelijk uitgelegd wat de gekozen variabelen betekenen en hoe ik ze gemeten heb.
	Het spreidingsdiagram is netjes getekend en alle eenheden liggen op elke as even ver uit elkaar.
	Ik kan de soort correlatie weergeven en verklaren.
	Ik heb een geschreven uitleg gemaakt over het feit of er een oorzaak is waardoor de ene variabele de andere beïnvloedt of of er een externe factor is.
	We hebben een overzichtelijk verslag gemaakt van al onze resultaten, als voorbereiding op de presentatie.
	We denken dat ons werk nu compleet is. We vinden zelf dat ons werk (omcirkel) • een grondige studie is • voldoende informatie bevat • nog niet voldoende onderzoek toont

HOOFDSTUK 4:
DE REGRESSIERECHTE

A. Probleemschets

In veel voorbeelden, zoals het voorbeeld van de slakken, zagen we dat de punten in het spreidingsdiagram duidelijk zeer dicht bij een rechte lijn liggen. We spraken dan van een ***benaderend lineair verband*** tussen de twee variabelen.
Eens we vermoeden dat er een lineair verband zou kunnen zijn tussen de twee variabelen, moeten we proberen de vergelijking te vinden van de rechte die het best aansluit bij de puntenwolk. Deze rechte noemen we de ***regressierechte***.
De techniek die gebruikt wordt om wiskundig de vergelijking van de regressierechte te vinden, noemt men dan ***regressie***. De regressierechte wordt o.a. gebruikt om voorspellingen te kunnen doen.

Bekijken we opnieuw het spreidingsdiagram van de slakken:



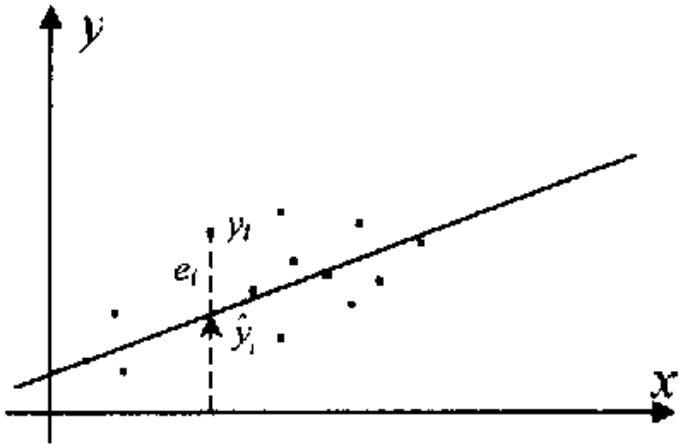
(a) Beschrijf in je eigen woorden het verband tussen het gewicht en het voetoppervlak bij de slakken.

(b) Schets op het spreidingsdiagram de regressierechte.
Vergelijk daarna je resultaat met de anderen. Heeft iedereen dezelfde rechte? Waarom is de jouwe beter/slechter?

(c) Kan je voorspellen hoe groot het voetoppervlak ongeveer zal zijn van een slak die 0.4 g weegt? Leg uit. Hoe doe je dit?

(d) Kan je voorspellen hoeveel een slak, met voetoppervlak 31 mm^2 , ongeveer weegt?

Wanneer men gegevens verzamelt van een bivariate verdeling, zijn de x-gegevens meestal de gegevens die onder controle zijn van de persoon die het experiment uitvoert. De y-waarden daarentegen zullen afhangen van deze x-waarden.
Veronderstel dat je als model een lineair verband $y = ax + b$ suggereert tussen de variabelen x en y. We spreken dan over "***lineaire regressie van y op x***".
Hierbij bekijken we het verschil tussen de ***geobserveerde*** waarde van Y (y_i) en de ***voorspelde*** waarde van Y ($\hat{y}_i$), die we uit de vergelijking van de rechte halen.



Gegeven is een puntenwolk van n punten: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, die min of meer een lineaire trend vertonen.
Beschouw dan de rechte $y = ax + b$ "door" deze punten waarmee we de grootheid y wensen te voorspellen bij gegeven x.
We definiëren voor elk punt het ***residu*** e_i, met $\hat{y}_i = ax_i + b$, als volgt:

$$\begin{aligned} e_i &= \text{observatie - voorspelling} \\ &= y_i - \hat{y}_i \\ &= y_i - (ax_i + b) \end{aligned}$$

Merk op dat een residu positief is wanneer het punt boven de rechte gelegen is en negatief wanneer het onder de rechte gelegen is.

Om nu de "beste" rechte door de puntenwolk te zoeken, gebruiken we de ***kleinste kwadratenmethode***: hierbij moeten we a en b bepalen zodat $\sum_{i=1}^{n} e_i^2$ minimaal is.

Waarom zou de methode niet werken als we $\sum_{i=1}^{n} e_i$ minimaliseren?

B. Berekening van a en b

Als we a en b zo kiezen dat $\sum_{i=1}^{n} e_i^2$ minimaal is, komen we tot de volgende formules:

Voor de regressierechte $y = ax + b$ door de punten $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ geldt:

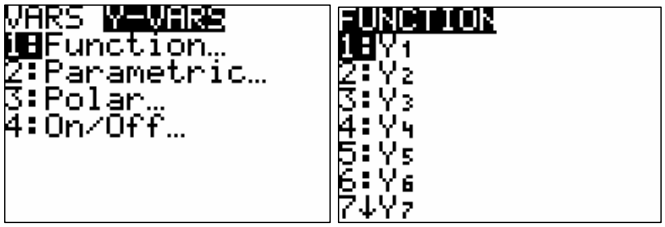
$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{{s_x}^2}$$

en $b = \bar{y} - a\bar{x}$

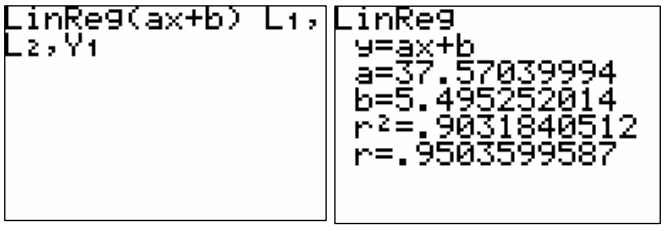
💻 Met de TI-83.

Met de TI-83 kunnen we a en b vlug berekenen. We plaatsen de X-gegevens in L1 en de Y-gegevens in L2.

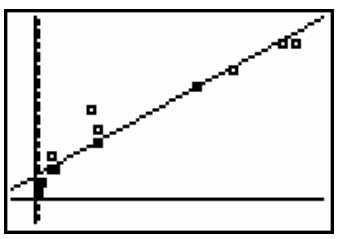
Druk dan STAT CALC 4:LinReg(ax+b).
Typ L1,L2,VARS Y-VARS 1:Function 1:Y1.
Door het toevoegen van Y1 wordt de vergelijking van de regressierechte weggeschreven in Y1.



Op je scherm zou dit nu moeten verschijnen:
Duwen we op ENTER dan verschijnen a en b op je scherm. In het voorbeeld van de slakken zien we dit scherm:



Definiëren we nu het spreidingsdiagram zoals vroeger aangeleerd dan wordt de regressierechte afgebeeld in het spreidingsdiagram.



⌦ Oefening:

We hernemen de oefening van de slakken.

Gewicht (g)	0.64	0.21	0.85	0.53	0.02	0.01	0.21	0.18	0.06	0.20	0.07	0.01	0.05
Voetopp (mm^2)	29	16	35	25	4	1	16	20	7	13	7	3	10

Gewicht (g)	0.81	0.53	0.18	0.06	0.20	0.07	0.01
Voetopp (mm^2)	35	25	20	7	13	7	1

(a) Als de regressierechte van y op x de vergelijking $y = ax + b$ heeft, bereken dan a en b.

(b) Teken de regressierechte op het spreidingsdiagram. Gebruik hiervoor je rekentoestel.
Controleer nu je resultaten die je in de vorige oefening had bekomen "op zicht".

Werk in vraag (c) en (d) eerst met de vergelijking en controleer je antwoord dan met behulp van het spreidingsdiagram.
Herinner je eraan dat je met de toets TRACE en dan met de pijltjestoetsen, de punten op je scherm krijgt met hun bijbehorende coördinaten.

(c) Gebruik de regressierechte om te voorspellen hoe groot het voetoppervlak ongeveer zal zijn van een slak die 0.4 g weegt.
Zat je voorspelling die je in de vorige oefening maakte, er dicht bij?

(d) Kan je met de regressierechte bepalen hoeveel een slak, met voetoppervlak 31 mm^2, ongeveer weegt?
Zat ook hier de voorspelling die je vroeger maakte, er dicht bij?

C. Enkele eigenschappen en bijzonderheden

(1) Een tomatenkweker gebruikt in elk van de 12 moestuintjes een verschillende hoeveelheid kunstmest.
Dit gaf hem de volgende gegevens:

hoeveelheid kunstmest (g) x	10	12	14	16	18	20	22	24	26	28	30	32
tomatenoogst (kg) y	2	2	2	3	4	3	4	3	5	6	7	9

(a) Bereken de regressierechte en teken ze in het spreidingsdiagram.

(b) Wat is het koppel $(\overline{x}, \overline{y})$?

(c) Bereken het residu van dit punt.

(d) Wat betekent deze waarde?

(e) Kunnen we dit veralgemenen? Maak je besluit hard.

(2) Beschouw de punten (0,3), (1,4), (2,7), (-1,4) en (-2,7).
(a) Teken deze punten in een spreidingsdiagram.
(b) Bereken de correlatiecoëfficiënt.

(c) Kun je hier besluiten dat er helemaal geen verband is tussen de twee variabelen?

(d) Bereken de regressierechte. Heeft deze hier zin? Kan ze bijvoorbeeld gebruikt worden om voorspellingen te doen?

We bedenken toch even dat met de geziene formules voor a en b, een regressierechte bepaald kan worden uitgaande van om het even welk spreidingsdiagram. Er kan dus – theoretisch gezien – een regressierechte beschouwd worden terwijl er helemaal geen oorzakelijk verband is tussen de twee variabelen. Dit is uiteraard niet zinvol.

Bij het opstellen van de vergelijking van de regressierechte eisen we dat $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ minimaal is.

Aangezien steeds geldt dat $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \geq 0$, is nul de kleinste waarde die $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ kan aannemen. Dit gebeurt als voor elke i geldt dat $y_i = \hat{y}_i$.

Met andere woorden alle punten van het spreidingsdiagram liggen op de regressierechte.
Men zegt dat er een ***perfecte lineaire correlatie*** is tussen x en y.

Is daarenboven a>0 (of a<0), dan spreken we van een ***perfecte positieve (negatieve) lineaire correlatie***.

In het vorige hoofdstuk zagen we dat ook de correlatiecoëfficiënt ons aanwijzingen geeft over de "goedheid" van de regressierechte.

Eigenschap:

$r = 1$	asa	de correlatie is perfect positief lineair.
$r = -1$	asa	de correlatie is perfect negatief lineair.

En hoe dichter r bij –1 of 1 ligt, hoe beter het lineaire model past bij onze punten.

In het voorbeeld waar de punten (0,3), (1,4), (2,7), (-1,4) en (-2,7) perfect op een parabool lagen, was $r = 0$ en wisten we eigenlijk zo ook al dat het lineaire verband hier niet van toepassing was.

We mogen ons echter ook niet alleen baseren op de waarde van r.

D. Oefeningen

(1) Men vermoedt zeer sterk dat de reactietijd van een persoon in verband staat met zijn hartslagritme. Elf dokters namen elk een verschillende hoeveelheid van een medicijn in, dat het hartslagritme beïnvloedt en testten zo hun vermoeden.
Dit leverde de volgende resultaten op:

hartslagritme (slagen per minuut) x	134	133	132	123	118	110	98	90	84	80	80
reactietijd (ms) y	438	455	467	505	531	557	541	562	591	603	617

(a) Is er een sterke correlatie tussen x en y?
(b) Toon de gegevens op een spreidingsdiagram.
(c) Zoek de vergelijking van de regressierechte met de kleinste kwadratenmethode voor regressie van y op x.
(d) Teken deze rechte op je spreidingsdiagram.
(e) Voorspel de reactietijd van een dokter wiens hartslagritme 95 hartslagen per minuut bedraagt.
(f) Je wordt gevraagd om de reactietijd te voorspellen van een dokter wiens hartslagritme 60 slagen per minuut bedraagt. Geef commentaar bij deze vraag.

(2) In het begin van vorige eeuw onderzocht men in enkele streken van Beieren het verband tussen kindersterfte en flessenvoeding.
Dit gaf de cijfers uit de volgende tabel:

	kindersterfte (aantal sterftes per 1000 levend geborenen) *x*	aantal kinderen met flessenvoeding (in procent) *y*
Niederbeieren	320	70
Oberfranken	170	10
Oberpfalz	300	40
Schwaben	270	60
Unterfranken	190	20
Mittelfranken	250	40

(a) Maak een spreidingsdiagram bij deze gegevens en bereken de regressierechte. Teken deze ook.

(b) In Oberbeieren kreeg 63% van de kinderen flessenvoeding en in Pfalz 15%. Doe aan de hand van de regressierechte een voorspelling voor de kindersterfte.
Ter vergelijking: de werkelijke cijfers waren respectievelijk 290 en 168.

(3) In de tabel hieronder vind je enkele Europese weerstations met hun hoogte boven de zeespiegel en gemiddelde jaartemperatuur.

station	hoogte (m)	temperatuur (°C)
Berlijn	49	9.1
Brocken	1152	2.4
Boedapest	130	10.9
Dobratsch	2140	0.1
Feuerkogel	1592	3.3
Graz	342	9.4
Innsbruck	579	8.4
Klagenfurt	448	8.1
Lugano	276	13.0
Praag	374	7.9
Salzburg	437	8.6
Säntis	2496	-2.3
Sonnblick	6106	-6.4
Wenen	203	9.1
Zugspitze	2962	-5.0

(a) Teken een spreidingsdiagram met daarop de regressierechte.
(b) Is het in het skioord Innsbruck relatief warm of relatief koud?
(c) Ukkel ligt op 100 meter boven de zeespiegel. Wat zou op basis van de regressierechte de gemiddelde jaartemperatuur in Ukkel moeten zijn?

(4) De tabel hieronder geeft telkens het gewicht van een boreling en de lengte van zijn moeder.

lengte moeder (cm) x	154	157	160	162	165	168	170	174	177	180	183
gewicht boreling (kg) y	3.10	3.05	3.10	3.10	3.25	3.25	3.25	3.30	3.35	3.40	3.40

(a) Teken een spreidingsdiagram.
(b) Bereken de correlatiecoëfficiënt.
(c) Construeer de regressierechte.
(d) Denk je dat je hier mag spreken van een lineair verband tussen de lengte van de moeder en het gewicht van de boreling?

(5) Heeft het intelligentiequotiënt (X) een invloed op het schoolresultaat (Y)?
Wat denk je?
Controleer uw vermoeden bij de volgende gegevens van 10 lukraak gekozen zesdejaars, waarbij hun eindprocent in juni gegeven is.

x	92	100	95	118	110	103	105	110	125	122
y	41	45	54	56	61	62	66	73	75	81

(6) De volgende tabel geeft de gemiddelde maandtemperatuur weer ten opzichte van het bedrag van de elektriciteitskosten die maand in een bepaald gezin.
Merk op dat er gegevens ontbreken voor 3 maanden.

maand	temperatuur (°C)	rekening (euro)	maand	temperatuur (°C)	rekening (euro)
april 91	10.5	41.7	juni 92	19	40.9
mei 91	16	42.7	juli 92	22	40.9
juni 91	23	36.6	augustus 92	22	41.4
juli 91	25	40.7	september 92	21	38.3
augustus 91	25.5	38.5	oktober 92	*	*
september 91	23	37.9	november 92	7	43.8
oktober 91	15	36	december 92	4	44.4
november 91	9	39.4	januari 93	1.5	46.3
december 91	6.5	49.7	februari 93	*	*
januari 92	1	55.5	maart 93	-1	50.8
februari 92	0	47.8	april 93	9	47.7
maart 92	5	44.4	mei 93	*	*
april 92	6	50	juni 93	20	38.7
mei 92	14	39.5	juli 93	25.5	47.5

(a) Maak een spreidingsdiagram.
Kunnen we uit deze grafiek een positieve of een negatieve correlatie besluiten of is er helemaal geen correlatie tussen de temperatuur en de elektriciteitskosten?
En als er een correlatie is, is deze sterk? Wat gebruik je om je antwoord te staven?
(b) Bereken en teken de regressierechte. Is deze goed genoeg denk je, om voorspellingen mee te doen?
(c) Gebruik de vergelijking van deze rechte om het residu te berekenen voor maart 1992.
(d) Gebruik nu de grafiek om te zien welke maanden de grootste en welke de kleinste residu's hebben. Wat wilt dit in dit voorbeeld concreet zeggen?

APPENDIX:
CENTRUMMATEN EN SPREIDINGSMATEN VAN EEN STEEKPROEF

A. Centrummaten.

Het steekproefgemiddelde.

Het steekproefgemiddelde $\bar{x}$ van de numerieke steekproefgegevens $x_1, x_2, ..., x_n$ is:

$$\bar{x} = \frac{\text{som van de waarnemingsgetallen}}{\text{aantal waarnemingsgetallen}}$$

$$= \frac{x_1 + x_2 + ... + x_n}{n}$$

$$= \frac{\sum_{i=1}^{n} x_i}{n}$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i$$

B. Spreidingsmaten

De steekproefvariantie.

De steekproefvariantie s^2 is de som van de kwadraten van de afwijkingen van de waarnemingsgetallen ten opzichte van hun rekenkundig gemiddelde, gedeeld door het aantal waarnemingen verminderd met 1.

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

De steekproefstandaardafwijking.

De steekproefstandaardafwijking s is de positieve vierkantswortel van de steekproefvariantie

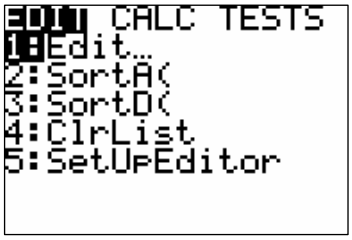
$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

C. Centrum- en spreidingsmaten met de TI-83

Voorbeeld:
Een hogeschool onderzoekt met een instaptest de kennis wiskunde van eerstejaarsstudenten.
Een steekproef van 20 studenten levert de volgende resultaten (op 20)

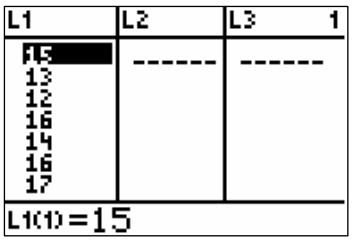
15	13	12	16	14	16	17	13	13	15	7	13	15	16	13	10	17	15	12	16

We voeren de gegevens in met behulp van lijsten (rijen).

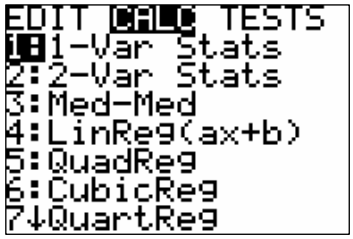


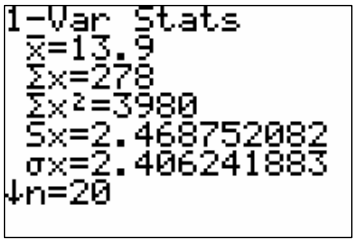
STAT EDIT 1:Edit

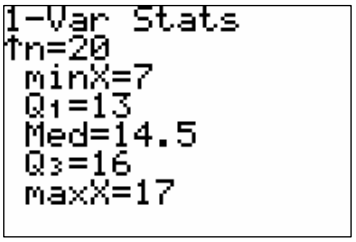
We vullen nu de 20 verschillende scores in bij lijst L1.
Dit doen we door op L1(1) te staan, de eerste score in te vullen (hier 15) en dan op ENTER te drukken. Enz.



Na het indrukken van STAT CALC 1:1-Var Stats en dan de lijst L1, verschijnen de kentallen van de lijst L1 op het basisscherm:







In de context van dit Cahier zijn vooral nuttig:

n = aantal waarnemingsgetallen
Sx = steekproefstandaardafwijking
$\bar{x}$ = gemiddelde

Bestaat er een *verband* of *correlatie* tussen twee kwantitatieve variabelen die betrekking hebben op dezelfde populatie? Om dit te onderzoeken gaan we uit van een steekproef waarbij de twee variabelen worden gemeten, dit levert concrete data $(x_1,y_1),(x_2,y_2),...,(x_n,y_n)$

Een spreidingsdiagram is een grafische voorstelling van de data, waarmee de begrippen *positieve of negatieve, sterke of zwakke correlatie* visueel worden ingevoerd.

De *correlatiecoëfficiënt* is een getal dat aangeeft in welke mate er een *lineair* verband bestaat tussen de variabelen. Eens we vermoeden dat er een lineair verband bestaat tussen twee variabelen, zoeken we de vergelijking van de rechte die het best aansluit bij de puntenwolk. Deze rechte is de *regressierechte*.

Dit cahier brengt de begrippen aan met veel voorbeelden en onderzoeksopdrachten, die door de auteur in de klas werden uitgetest.

BIEKE VAN DEYCK was praktijklector aan de faculteit Toegepaste Wetenschappen (Voorbereidend Instituut) van de K.U.Leuven. Nu is zij leerkracht wiskunde en fysica aan de Humaniora Voorzienigheid te Diest.

TEXAS INSTRUMENTS