Lengte en gewicht

vwo wiskunde keuzeonderwerp

wiskundeweb.nl versie 2004

Deze katern is bedoeld voor het keuzeonderwerp bij wiskunde voor de profielen EM en NG in het VWO. De tekst is gemaakt met Context, een typesetting-systeem van Pragma ADE in Hasselt (Nederland). Het is als PDF-bestand gratis te downloaden vanaf: www.wiskundeweb.nl. De schermafdrukken van de grafische rekenmachine zijn gemaakt met een TI-83 Plus van Texas Instruments.

Veel plezier met deze katern!

Inhoud

Inleiding

Deze katern gaat over statistisch onderzoek naar een mogelijk verband tussen twee variabelen zoals *lengte* en *gewicht* bij mensen van een bepaalde leeftijdscategorie. Vaak wordt verondersteld dat tussen deze twee variabelen een eenvoudig verband bestaat: “hoe groter iemand is, hoe zwaarder hij of zij is”. Je onderzoekt of zo’n uitspraak inderdaad geldig is en je stelt een maat vast voor de betrouwbaarheid van deze hypothese. Die maat heet de **correlatiecoëfficiënt**.

Als éénmaal is vastgesteld dat er met voldoende zekerheid kan worden aangenomen dat er een verband tussen beide variabelen bestaat, kun je proberen dit verband te beschrijven met een formule. Wat voor soort formule dat moet zijn, hangt van de situatie af. Het eenvoudigste geval is een lineaire formule. Daar ga je dan ook van uit als er geen duidelijke aanwijzingen voor een ander soort verband bestaan.

Je begint met je gegevens uit te zetten in een grafiek. Dat levert een ‘wolk’ van punten op. Om een lineaire formule te kunnen opstellen moet je daar dan zo goed mogelijk een rechte lijn in trekken. In de negentiende eeuw bedacht de beroemde wiskundige Carl Friedrich Gauss daarvoor de **methode van de kleinste kwadraten**.

Hoe die methode werkt zul je in paragraaf 3 zien. Hij wordt daar gebruikt om formules te vinden bij de zogenaamde **regressielijnen** die horen bij de puntenwolk van het verband tussen lengte en gewicht.

Er zijn natuurlijk ook andere variabelen denkbaar waartussen een verband zou kunnen bestaan. Het is de bedoeling dat je er zelf twee zoekt. Je kunt daarbij denken aan

- een verband tussen lengte en schoenmaat
- een verband tussen de cijfers voor het schoolexamen en het centraal examen
- leeftijd van overlijden en gewicht op 50-jarige leeftijd
- gasverbruik en gemiddelde temperatuur

Je moet dus zelf een statistisch onderzoek verrichten. Daarbij speelt de computer een grote rol: bij deze katern hoort een werkblad gemaakt in het rekenbladprogramma 'Excel'. Daarin vind je de opzet van het lengte-gewicht-onderzoek. Je moet dat zelf afmaken. Daarbij heb je enige kennis van statistiek nodig.
Bovendien werk je je eigen statistisch onderzoek uit in Excel.

In deze katern leer je:

- ▷ statistisch onderzoek doen naar het verband tussen twee variabelen;
- ▷ de mate waarin dit verband optreedt beschrijven met de correlatiecoëfficiënt;
- ▷ dit verband beschrijven met een lineaire formule.

Eindresultaat:

- ▷ Een uitwerking van het lengte-gewicht-onderzoek en een overzicht van de theorie die je daarbij gebruikt.
- ▷ De uitwerking van een eigen onderzoek naar het verband tussen twee variabelen.

Dit alles wordt verwerkt tot een compleet **statistisch onderzoek** naar het verband tussen twee variabelen. Daarbij werk je in het rekenbladprogramma 'Excel'. Dit onderzoek bouw je als het ware langzaam op via de laatste opgaven van elke paragraaf. Deze hebben het kopje 'Voor het werkstuk'.

1 Gegevens verzamelen

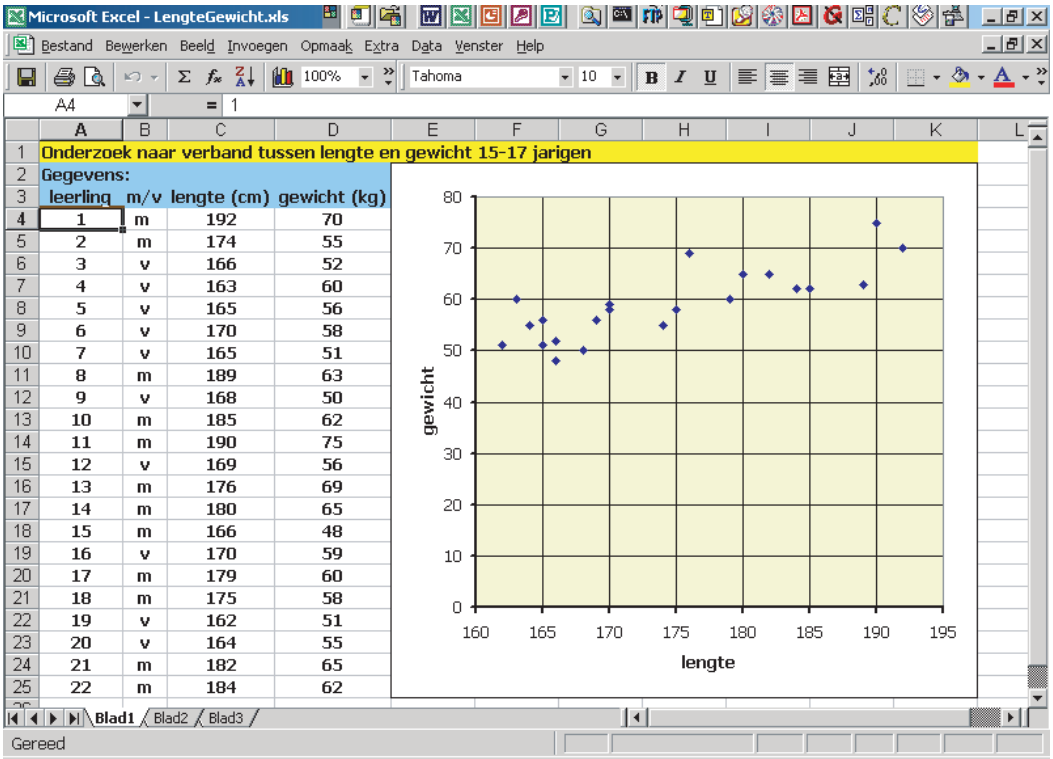
Om te onderzoeken of er een verband bestaat tussen lengte en gewicht bij mensen van 15 tot 17 jaar oud heb je gegevens nodig. Op het werkblad LengteGewicht.xls vind je de gegevens van een 4HAVO-klas van 22 leerlingen. Er zijn vier kolommen:

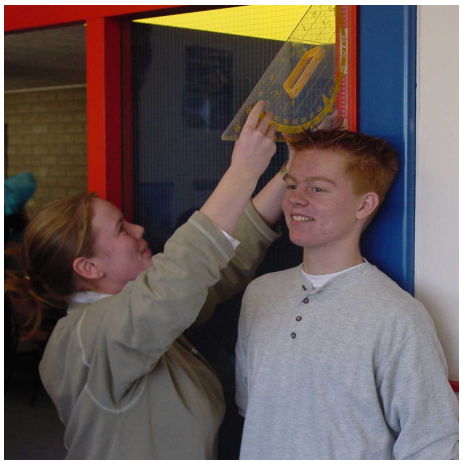
A. het nummer van de leerling op de klassenlijst;
B. het geslacht (m/v);
C. de lengte in cm;
D. het gewicht in kg.

Bij het bepalen van deze gegevens moet je zorgvuldig meten. Maak daarover bij zo'n onderzoek van tevoren goede afspraken, zoals:

- meet de lengte en het gewicht zonder schoenen aan;
- meet het gewicht steeds met dezelfde weegschaal;
- rond steeds op dezelfde manier af;

enzovoorts.





Je kunt er vanuit gaan dat de gegevens op het werkblad op zorgvuldige wijze tot stand zijn gekomen. Maar het is natuurlijk altijd beter om met eigen gegevens te werken. Bovendien kan de steekproef best wat groter zijn!

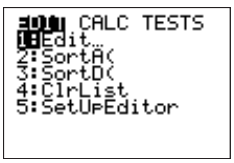
Op het eerste werkblad staan alle gegevens. Daarna worden de gegevens per variabele geordend. Het tweede werkblad gebruik je bijvoorbeeld voor alle lengtes. Je deelt daartoe je gegevens in ongeveer 10 klassen in.
Niet alleen komen daar de gegevens op volgorde, maar er wordt een staafdiagram van gemaakt en de belangrijkste centrum- en spreidingsmaten worden vermeld. Hoe meer gegevens je hebt, hoe belangrijker dat is.

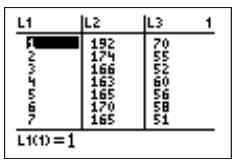
Statistische gegevens met één variabele ordenen

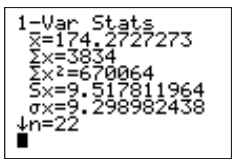
Vaak gaat het in de statistiek om één variabele en het aantal keren dat een bepaalde waarde van die variabele voorkomt. Meestal is die variabele uit te drukken in een getal.
Bij het ordenen van dergelijke statistische gegevens maak je een **frequentieverdeling**. Meestal deel je daartoe de gegevens eerst in ongeveer 10 klassen in. Van zo'n frequentieverdeling maak je dan **histogram** en/of een **frequentiepolygoon**.

Je geeft bij zo'n frequentieverdeling als centrummaat het **gemiddelde**. De bijbehorende spreidingsmaat is de **standaarddeviatie**. Daarbij reken je met de klassenmiddens. Met name bij een **normale verdeling** hebben deze maten een duidelijke betekenis!
Soms wordt ook de **spreidingsbreedte** (de klassenbreedte maal het aantal klassen) gegeven.

Al deze lijsten, grafieken en maten kun je vinden met behulp van Excel of met je grafische rekenmachine. Hoe dat in Excel gaat kun je op werkblad 2 van de werkmap LengteGewicht.xls vinden. (Klik maar op de juiste cellen!)
Hoe dat op je grafische rekenmachine gaat moet je zelf nog even nazoeken.







Lengte en gewicht 5

Opgaven

1. Gebruik de werkmap LengteGewicht.xls. Verdeel op blad 3 de gewichten in klassen zoals $45- < 50$, $50- < 55$, etc.
   Maak een frequentieverdeling, een histogram, een frequentiepolygoon.
   Bereken ook het gemiddelde, de standaarddeviatie en de spreidingsbreedte.
   Gebruik de statistische functies GEMIDDELDE en STDEVP.

2. De 22 leerlingen in de steekproef kwamen allen uit dezelfde 4HAVO-klas.
   Is deze steekproef voldoende representatief voor 15-17 jarigen?
   Motiveer je antwoord.

3. Ga met behulp van normaal waarschijnlijkheidspapier na of de lengtes van de 22 leerlingen in de voorgaande tekst ongeveer normaal verdeeld zijn.
   Doe dit ook voor de gewichten.

4. In de volgende tabel zie je de examenresultaten voor de vakken Physics (natuurkunde) en Mathematics (wiskunde) van 100 studenten van een College in de Verenigde Staten:

Physics grades \ Mathematics grades	40 - 49	50 - 59	60 - 69	70 - 79	80 - 89	90 - 99	Totals
90 - 99				2	4	4	10
80 - 89			1	4	6	5	16
70 - 79			5	10	8	1	24
60 - 69	1	4	9	5	2		21
50 - 59	3	6	6	2			17
40 - 49	3	5	4				12
Totals	7	15	25	23	20	10	100

   a. Onderzoek of de ‘Physics Grades’ normaal zijn verdeeld. Bereken het bijbehorende gemiddelde en de standaarddeviatie.
   b. Onderzoek of de ‘Mathematics Grades’ normaal zijn verdeeld. Bereken het bijbehorende gemiddelde en de standaarddeviatie.
   c. Waarom kunnen dit alleen geschatte gemiddelden en standaarddeviaties zijn?

Keuzeonderwerp wiskunde 5

d. Voor welk van beide vakken scoorden deze 100 studenten het beste? Motiveer je antwoord.
e. Kun je bij deze tabel een puntenwolk maken zoals die bij de gegevens over lengte en gewicht in de tekst? Hoe dan?

5. **Voor het werkstuk**
Kies twee variabelen waartussen vermoedelijk een bepaald verband bestaat. Bijvoorbeeld:

- ▷ lichaamslengte en armlengte
- ▷ lengte vader en lengte zoon (of moeder en dochter)
- ▷ cijfer schoolexamen en cijfer landelijk examen
- ▷ diameter en hoogte van een boom
- ▷ gemiddelde temperatuur en gasverbruik per jaar
- ▷ lengte van een draad en kracht die nodig is om hem te breken

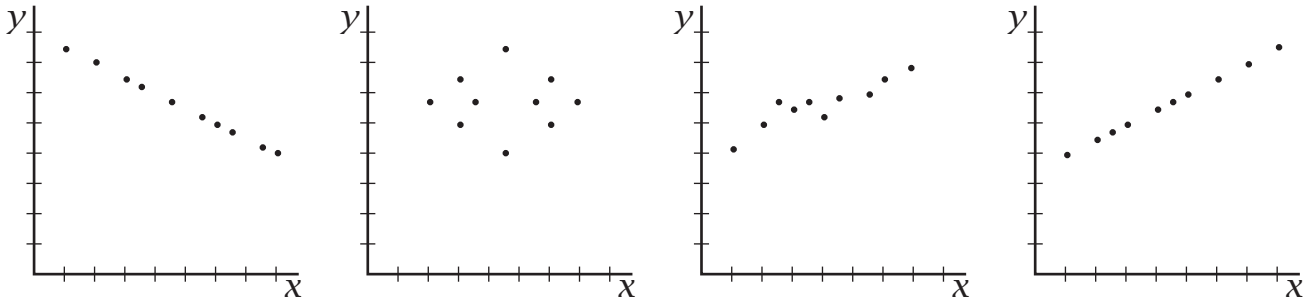
Verzamel voldoende gegevens (een puntenwolk van minstens ongeveer 50 punten!). Ontwerp een Excel-werkblad waarop je de gegevens bewaart en een puntenwolk maakt.
Maak vervolgens afzonderlijke werkbladen waarop je de gegevens per variabele statistisch verwerkt. Maak duidelijke tabellen en diagrammen. Bepaal gemiddelde en standaarddeviatie. Onderzoek of er sprake is van normale verdeling.

2 Zoeken naar correlatie

In paragraaf 1 vind je van 22 leerlingen uit een 4HAVO-klas de lengte l in cm en het gewicht g in kg. Op het bijbehorende Excel-werkblad staan l en g tegen elkaar uitgezet. Dat levert een puntenwolk op.

De punten in dit spreidingsdiagram (de puntenwolk) liggen niet netjes op een rechte lijn. Toch lijkt het er op dat er een zeker verband tussen l en g bestaat. Dat komt door de vorm van de puntenwolk. Hoe meer de punten 'op één lijn liggen', hoe sterker het vermoeden dat er een lineair verband tussen de twee variabelen bestaat.

Een maat voor de sterkte van de samenhang tussen beide variabelen is de correlatiecoëfficiënt. Dat is een getal dat kan variëren vanaf -1 tot en met 1.



$r_{xy} = -1$ $r_{xy} = 0$ $r_{xy} = 0.6$ $r_{xy} = 1$

Correlatiecoëfficiënt

Als je vermoedt dat er tussen twee variabelen x en y een lineair verband bestaat, maak je een **spreidingsdiagram** dat de vorm van een puntenwolk krijgt.
De mate waarin tussen de twee variabelen een lineair verband bestaat wordt gegeven door de **correlatiecoëfficiënt**, aangeduid door r_{xy}.

- Als $r_{xy} = 1$ dan is er een perfecte positieve correlatie tussen x en y. De punten van de puntenwolk liggen dan precies op een stijgende lijn.
- Als $r_{xy} = 0$ dan is er geen enkele correlatie tussen x en y.
- Als $r_{xy} = -1$ dan is er een perfecte negatieve correlatie tussen x en y. De punten van de puntenwolk liggen dan precies op een dalende lijn.

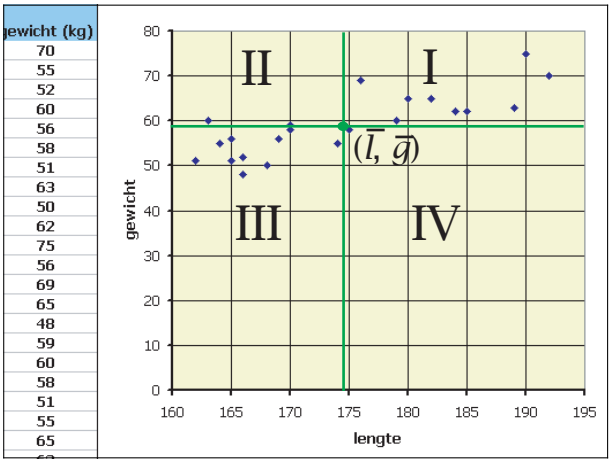
De correlatie tussen x en y wordt beter naarmate r_{xy} dichter bij 1 of -1 ligt.

Maar hoe bereken je nu die correlatiecoëfficiënt?

Bekijk daarvoor de puntenwolk die hoort bij de lengtes en de gewichten van onze 22 leerlingen nog maar eens. Je hebt de gemiddelde lengte en het gemiddelde gewicht inmiddels berekend:

$$\bar{l} = \frac{3834}{22} \approx 174{,}3 \quad \text{en} \quad \bar{g} = \frac{1300}{22} \approx 59{,}1$$

Met behulp van die gemiddelden kan het grafiekgebied in vier delen I, II, III en IV worden verdeeld (zie figuur). Je kunt nu voor elk punt (l_i, g_i) het getal $(l_i - \bar{l})(g_i - \bar{g})$ berekenen.
In de gebieden I en III is dit getal voor elk punt positief: deze punten dragen bij aan een positieve correlatie.
In de gebieden II en IV is dit getal voor elk punt juist negatief: deze punten dragen bij aan een negatieve correlatie.



Het gemiddelde van alle getallen $(l_i - \bar{l})(g_i - \bar{g})$ is een goede maat voor de correlatie. Deze maat heet de **covariantie** van de puntenwolk:

$$\text{covariantie} = \frac{\sum_{i=1}^{N}(l_i - \bar{l})(g_i - \bar{g})}{N}$$

Deze maat voor de correlatie in een puntenwolk hangt nog af van de eenheden waarin l en g zijn gemeten. Dat kun je voorkomen door telkens $(l_i - \bar{l})$ te delen door de bijbehorende standaarddeviatie σ_l en ook $(g_i - \bar{g})$ telkens te delen door σ_g. Je krijgt dan de correlatiecoëfficiënt, die niet langer afhangt van de gekozen eenheden.

De correlatiecoëfficiënt berekenen

De **correlatiecoëfficiënt** voor de variabelen x en y bereken je met de formule

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N \cdot \sigma_x \cdot \sigma_y}$$

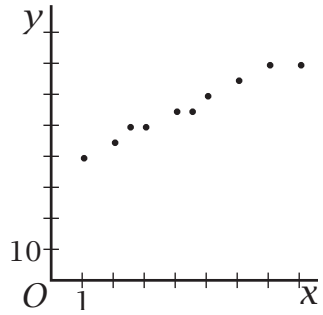
In Excel is de berekening van de correlatiecoëfficiënt niet al te moeilijk uit te voeren. Zeker niet als je de gemiddelden en de standaarddeviaties al hebt berekend met de

statistische functies. Je maakt dan een kolom voor de getallen $(l_i - \bar{l})(g_i - \bar{g})$. En daarna bereken je het gemiddelde van die kolom. Dat gemiddelde deel je nog door beide standaarddeviaties.
Overigens kent Excel ook statistische functies als COVARIANTIE en CORRELATIE.

Hoe je met de grafische rekenmachine de correlatiecoëfficiënt kunt berekenen wordt in de volgende paragraaf uit de doeken gedaan.

Opgaven

1. Bekijk dit spreidingsdiagram.



   a. Is er op het oog sprake van een goede correlatie tussen x en y?
   b. Schat de correlatiecoëfficiënt.
   c. Welke soort formule hoort er bij y als functie van x?
   d. Waarom is de schaalverdeling op de assen niet van belang voor de correlatie?

2. Bereken de correlatiecoëfficiënt bij het verband tussen de lengte en het gewicht van de 22 leerlingen. De gegevens staan op Blad 1 van de werkmap LengteGewicht.xls.
   Is er sprake van een goede correlatie tussen l en g?

3. Bekijk de gegevens van opgave 4 van de voorgaande paragraaf. Er wordt een verband verondersteld tussen het resultaat voor wiskunde m en dat voor natuurkunde p.

   a. Hoe bepaal je in dit geval (een schatting van) de correlatiecoëfficiënt?
   b. Bereken nu de (schatting van de) correlatiecoëfficiënt bijvoorbeeld met behulp van Excel. Geef een benadering in twee decimalen nauwkeurig.
   c. Is er een duidelijke correlatie tussen m en p?

4. De formule voor de correlatiecoëfficiënt is te herschrijven tot:

$$r_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sigma_x \cdot \sigma_y}$$

   Laat dat zien door in de formule in de tekst de haakjes uit te werken.

5. **Lengte van vader en zoon**
Om te onderzoeken of er enig verband bestaat tussen de lengte van een vader en die van zijn zoon zijn de lengtes van 12 vaders en die van hun oudste zoons gemeten op het moment dat die zoons volwassen werden.
De gegevens staan in deze tabel.

lengte vader v in cm	173	168	178	170	180	165	185	175	180	178	183	188
lengte zoon z in cm	180	175	180	173	183	175	180	173	188	178	180	185

a. Teken een spreidingsdiagram (een puntenwolk) bij deze gegevens.
b. Bereken de correlatiecoëfficiënt in twee decimalen nauwkeurig.
c. Kun je zeggen dat er een lineair verband bestaat tussen v en z?

6. **Vliegsnelheid en lichaamslengte**
Biologen veronderstellen op grond van metingen dat er bij vliegende dieren een verband bestaat tussen de lichaamslengte L (in cm) en de vliegsnelheid v (in cm/s).

Vliegsnelheid en lichaamslengte bij verschillende dieren

Soort		Lengte L in cm	Vlieg-snelheid v in cm/s
1.	*Drosophila melanogaster* (fruitvlieg)	0,2	190
2.	*Tabanus affinis* (paardenvlieg)	1,3	660
3.	*Archilochus colubris* (kolibriesoort)	8,1	1120
4.	*Anax* sp. (waterjuffer)	8,5	1000
5.	*Eptesicus fuscus* (grote bruine vleermuis)	11,0	690
6.	*Phylloscopus trochilus* (fitis)	11,0	1200
7.	*Apus apus* (gierzwaluw)	17,0	2550
8.	*Cypselurus cyanopterus* (vliegende vis)	34,0	1560
9.	*Numenius phaeopus* (regenwulp)	41,0	2320
10.	*Anas acuta* (pijlstaarteend)	56,0	2280
11.	*Olor columbianus bewicki* (kleine zwaan)	120,0	1880
12.	*Pelecanus onocrotalus* (witte pelikaan)	160,0	2280

a. Maak een spreidingsdiagram met v op de verticale en L op de horizontale as.
b. Bereken de correlatiecoëfficiënt. Is er sprake van een duidelijke correlatie? Bestaat er tussen v en L een verband van de vorm $v = a \cdot L + b$?
c. Maak een tabel voor $\log L$ en $\log v$ en teken een spreidingsdiagram voor deze twee variabelen.
d. Bereken de correlatiecoëfficiënt voor de variabelen $\log L$ en $\log v$.
e. Er bestaat tussen L en v dus een verband van de vorm $\log v = a \cdot \log L + b$. Laat zien dat dit betekent dat v een machtsfunctie is van L.

3 Regressielijn

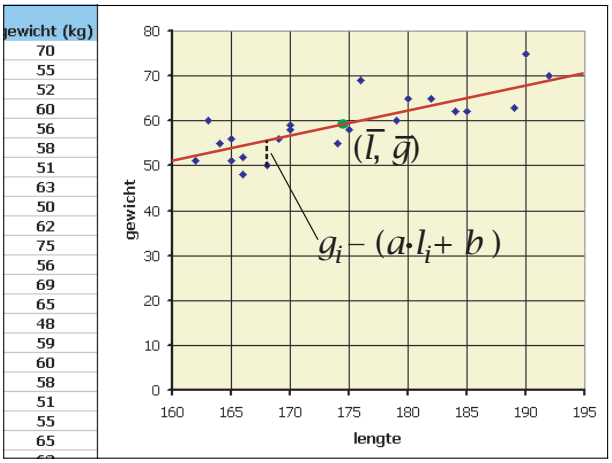
Je hebt gezien, dat er een goede correlatie is tussen de variabelen l en g die de lengtes en het gewichten voorstellen van de 22 leerlingen (zie voorgaande paragrafen).
Dat betekent dat er een lineair verband tussen l en g bestaat. Er kan dus in de puntenwolk een rechte lijn worden getrokken die het verband tussen l en g goed beschrijft. Zo'n lijn wordt een **regressielijn** genoemd.
Een regressielijn moet uiteraard door het punt $(\overline{g}, \overline{l})$ gaan. De richtingscoëfficiënt (het hellingsgetal) van die lijn kun je op dit moment echter alleen nog maar schatten.
De vraag is natuurlijk hoe je die richtingscoëfficiënt kunt berekenen.

De beroemde wiskundige Gauss bedacht daarvoor in de negentiende eeuw de **methode van de kleinste kwadraten**.
Stel je voor dat je een regressielijn wilt maken van de vorm $g = a \cdot l + b$. Je gaat dan uit van een regressielijn van g op l.
Gauss' methode houdt nu in dat de som van de kwadraten van de verticale afwijkingen van de meetpunten tot deze regressielijn zo klein mogelijk moet zijn. Dat betekent dat

$$\sum_{i=1}^{N} (g_i - (a \cdot l_i + b))^2$$

minimaal moet zijn.
Gauss vond dat dit het geval is als

$$a = \frac{\sum_{i=1}^{N} (l_i - \overline{l})(g_i - \overline{g})}{N \cdot \sigma_l^2}$$



▷ De afleiding van deze formule doe je zelf in opgave 2 van deze paragraaf.

Deze formule lijkt veel op die van de correlatiecoëfficiënt.
Dat is handig, want het betekent dat je de richtingscoëfficiënt a van de regressielijn eenvoudig kunt afleiden uit de correlatiecoëfficiënt.

Ga maar na, dat: $a = r_{lg} \cdot \frac{\sigma_g}{\sigma_l}$.

Vooral deze formule is handig bij het berekenen van de hellingsgetal bij regressie van l op g.

Regressie van y op x

Als de correlatie tussen de variabelen x en y groot genoeg is, kun je een formule van de vorm $y = ax + b$ opstellen die het verband tussen x en y weergeeft.
Deze formule heeft als grafiek een rechte lijn, de **regressielijn van y op x**.
Zo'n regressielijn gaat door het punt $(\overline{x}, \overline{y})$ en heeft als richtingscoëfficiënt (hellingsgetal):

$$a = r_{xy} \cdot \frac{\sigma_y}{\sigma_x}$$

Deze richtingscoëfficiënt heet wel de **regressiecoëfficiënt** van y op x.
Met behulp van deze regressiecoëfficiënt en het feit dat de regressielijn door $(\overline{x}, \overline{y})$ gaat, kun je de bijbehorende formule opstellen.

Als je eenmaal de correlatiecoëfficiënt hebt berekend, dan kun je dus voor onze groep van 22 leerlingen eenvoudig de regressiecoëfficiënt berekenen:

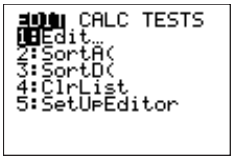
$$a = r_{lg} \cdot \frac{\sigma_g}{\sigma_l} \approx 0{,}8096 \cdot \frac{6{,}7951}{9{,}2990} \approx 0{,}59$$

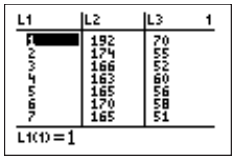
Controleer maar even of dit alles klopt met wat je zelf tot nu toe hebt gevonden.
De vergelijking van de regressielijn wordt nu: $g \approx 0{,}59 \cdot l + b$,
waarin b nog te berekenen is. Daarvoor maak je gebruik van het punt $(174{,}27; 59{,}09)$ dat op deze regressielijn moet liggen.

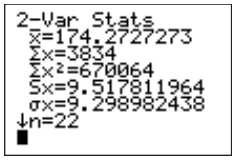
Regressie met de grafische rekenmachine

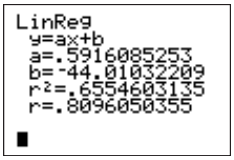
Bij het bepalen van de mate van correlatie tussen twee variabelen en het opstellen van de vergelijking van de regressielijn is de grafische rekenmachine een erg handig hulpmiddel. Tenminste, als je niet een puntenwolk hebt met heel veel meetpunten, want dan is het werken met Excel veel handiger!

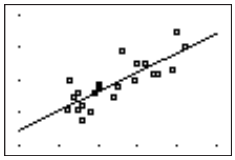
Bij de variabelen l en g (lengte en gewicht bij de groep van 22 leerlingen) voer je eerst alle meetpunten (l, g) in de grafische rekenmachine in. Je kunt dan in één klap alle gegevens betreffende gemiddelden en standaarddeviaties van beide variabelen in beeld brengen als je wilt.











Als je die gegevens hebt ingevoerd, dan kan de grafische rekenmachine de correlatiecoëfficiënt, en de waarden voor a en b van de regressielijn $g = a \cdot l + b$ direct berekenen. Vervolgens is via het statistisch tekenmenu het resultaat in beeld te brengen.
Zoek uit hoe dit op jouw grafische rekenmachine in zijn werk gaat.

Opgaven

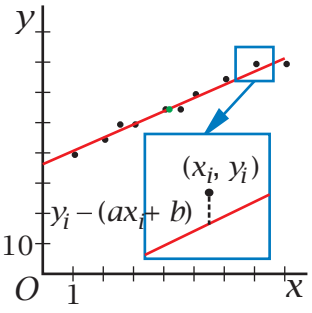
1. Kijk nog even naar opgave 1 van de voorgaande paragraaf.

   a. Maak een tabel van de 10 meetpunten. Voer deze gegevens in je grafische rekenmachine in.
   b. Bereken de coördinaten van het punt $(\overline{x}, \overline{y})$.
   c. Als je door deze punten 'op het oog' een regressielijn zou willen tekenen, hoe groot wordt dan de regressiecoëfficiënt ongeveer?
   d. Bereken nu de correlatiecoëfficiënt en de regressiecoëfficiënt.
   e. Stel een vergelijking op van de regressielijn van y op x.
   f. Welke waarde zou y moeten hebben volgens deze regressielijn als $x = 10$?

2. Neem het begin van de tekst van deze paragraaf nog even door.
   Volgens de methode van de kleinste kwadraten moet je de minimale waarde bepalen van de uitdrukking:

   $$p = \sum_{i=1}^{N} (y_i - (a \cdot x_i + b))^2$$

   als x en y de twee variabelen voorstellen.

   a. Laat door haakjes uitwerken zien, dat p een kwadratische functie van a is.
   b. Bereken voor welke waarde van a deze functie minimaal is en leidt zo de formule voor de regressiecoëfficiënt zelf af.
   c. Leg ook uit hoe je aan de formule komt waarmee je a kunt berekenen vanuit r_{xy}.

3. Stel de vergelijking op van de regressielijn van g op l bij het verband tussen lengte en gewicht van de 22 leerlingen waarvan je de gegevens in de werkmap LengteGewicht.xls aantreft.

   a. Zorg er voor dat de regressielijn in de puntenwolk zichtbaar wordt!
   b. Welke betekenis heeft deze regessielijn als je aanneemt dat de groep leerlingen voldoende representatief is voor alle 15-17 jarigen?
   c. Hoe zwaar zou iemand van 16 jaar moeten zijn als hij 180 cm lang is?

4. In opgave 5 van de voorgaande paragraaf bleek er een correlatie te bestaan tussen de lengte van een vader v en die van zijn oudste zoon z.

   a. Was er sprake van een positieve of een negatieve correlatie? Wat betekent dit in de praktijk?
   b. Stel de vergelijking op van de regressielijn van z op v. Geef benaderingen in twee decimalen!
   c. Als een bepaalde vader 1,77 m lang is, hoe lang zou dan zijn oudste zoon moeten zijn?

5. In opgave 6 van de voorgaande paragraaf ging het over het verband tussen lichaamslengte L en vliegsnelheid v van vliegende dieren.
   Stel een formule op voor de bijpassende machtsfunctie.

6. **Braadtijd van kalkoenen**
   Om het verband tussen het gewicht G (in pounds) en de braadtijd voor kalkoenen te onderzoeken, werd onder gelijke omstandigheden nagegaan hoeveel minuten t het duurde tot het binnenste van een kalkoen de temperatuur van 85°C bereikte. Er werden diverse kalkoenen aan dit onderzoek onderworpen. Ze hadden een gemiddeld gewicht van 15,24 pounds met een standaardafwijking van 6,07. Voor de waarden van t vonden de onderzoekers een gemiddelde van 205,4 minuten met een standaardafwijking van 59,1.
   De regressielijn van t op G had de vergelijking: $t = 9{,}65 \cdot G + 58{,}40$.
   Hoeveel bedroeg de correlatiecoëfficiënt?

7. **Voor het werkstuk**
   Als het goed is heb je inmiddels wel voldoende gegevens voor je werkstuk verzameld. Het wordt tijd om hierbij een correlatiecoëfficiënt en een bijpassende regressielijn te gaan berekenen.
   Mocht je op een slechte correlatie uitkomen, overleg dan met je leraar of je met twee nieuwe variabelen aan de gang zult gaan, of toch maar met deze gegevens door zult werken.

4 Regressie nader bekeken

Bij het verband tussen l en g bij de groep van 22 leerlingen heb je een regerssielijn van g op l gemaakt:

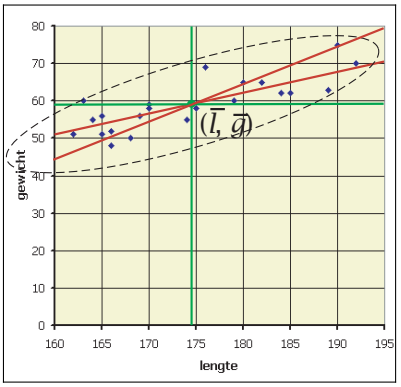
$$g = 0{,}59 \cdot l - 44{,}01$$

Er past echter ook heel goed een regressielijn van l op g bij. Ga na, dat je dan vindt:

$$l = 1{,}11 \cdot g + 108{,}80$$

Deze tweede regressielijn kun je in dezelfde figuur tekenen als de eerste.
Deze twee regressielijnen zijn verschillend!
Als je van een leerling van 15-17 jaar met een lengte van $l = 180$ cm het gewicht zou moeten voorspellen, vind je volgens de eerste regressielijn ongeveer 62,19 kg, maar volgens de tweede regressielijn hoort bij een gewicht van 62,19 kg een lengte van 177,83 cm!

Dit verschil heeft te maken met het zogenaamde **regressie-effect**.
Dat regressie-effect ontstaat doordat er geen volledige correlatie tussen g en l is, de correlatiecoëfficiënt is 'slechts' ongeveer 0,81 en dat is minder dan 1. Daarom ligt de regressielijn van g op l dichter bij de horizontale lijn door $\overline{g}$, terwijl de regressielijn van l op g juist dichter bij de verticale lijn door $\overline{l}$ komt te liggen. Dit effect wordt nog eens duidelijk zichtbaar als je kijkt naar het gewicht dat zou moeten horen bij een 15-17 jarige die precies één standaardafwijking groter is dan de gemiddelde lengte: $l = 174{,}27 + 9{,}30 = 183{,}57$.

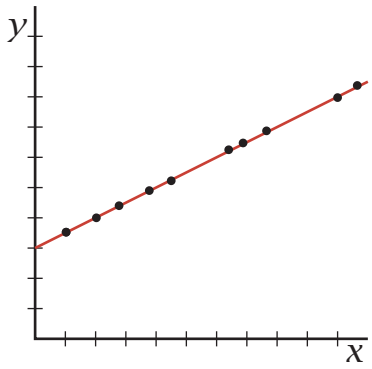


- Bij regressie van g op l voorspel je een gewicht van ongeveer 64,30 kg. Als je bij het gemiddelde gewicht echter ook één standaarddeviatie optelt, krijg je 65,89 kg. Het op grond van de regressielijn voorspelde gewicht wijkt dus minder van $\overline{g}$ af dan je zou verwachten!
  Er is sprake van een 'terugval naar het gemiddelde'.
- Bij regressie van l op g vindt je daarentegen bij een gewicht van 65,89 kg een lengte die kleiner is dan 183,57!
  Ook hier is weer sprake van een 'terugval naar het gemiddelde'.

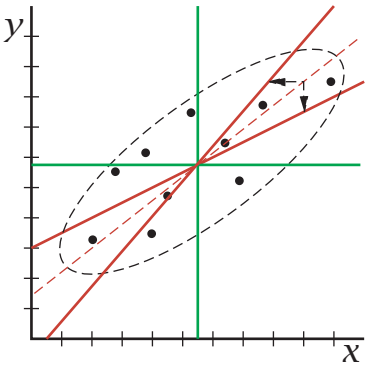
Merk nog op dat het product van de twee regressiecoëfficiënten precies het kwadraat van de correlatiecoëfficiënt is.

Het regressie-effect

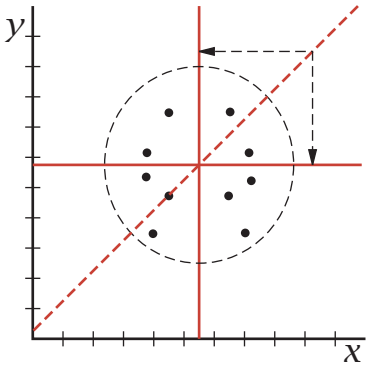
Bij correlatie tussen twee variabelen x en y kun je drie gevallen onderscheiden:



Volledige correlatie
Beide regressielijnen vallen samen. Geen regressie-effect.



Gedeeltelijke correlatie
De regressielijnen maken een hoek. Bij zwakkere correlatie is die hoek groter, dus zijn dan de regessie- effecten groter.



Geen correlatie
De regressielijn van y op x loopt horizontaal, de andere verticaal. Het regressie-effect is maximaal.

Opgaven

1. Reken alle voorbeelden in de tekst van deze paragraaf zorgvuldig na (als je dat nog niet hebt gedaan).
   Neem de gegevens in de werkmap LengteGewicht.xls er weer bij. Teken ook de regressielijn van l op g in de figuur van werkblad 1.

2. Kijk nog even naar opgave 1 van de voorgaande paragraaf.

   a. Stel een formule op voor de regressielijn van x op y.
   b. Teken zelf het spreidingsdiagram met daarin beide regressielijnen.
   c. Is er sprake van een regressie-effect? Zo ja, laat dit dan met een rekenvoorbeeld zien.

3. In opgave 5 van paragraaf 1 en opgave 4 van paragraaf 2 ging het over het verband tussen de lengtes van vader en zoon. Laat met een getallenvoorbeeld het optredende regressie-effect zien. Wat betekent dit voor de bepaling van de lengte van een zoon waarvan de vader bijvoorbeeld 2 m lang is?

4. **Bewaren van geneesmiddelen**
In een Amerikaans laboratorium heeft men proeven genomen waarbij gelet werd op het verband tussen de hoogte van de bewaartemperatuur F in graden Fahrenheit en de werkzaamheid W van een bepaald geneesmiddel. Bij temperaturen van 30°, 50°, 70° en 90° (Fahrenheit) werden drie porties van gelijk gewicht uit eenzelfde productie 20 dagen bewaard. Na deze periode werd op identieke wijze de werkzaamheid van de porties vastgesteld. De werkzaamheid werd uitgedrukt in percentages van de werkzaamheid zoals die was voor het bewaren.

Bewaartemperatuur F	30°	50°	70°	90°
Werkzaamheid W	$39, 42, 35$	$32, 26, 33$	$19, 27, 23$	$14, 19, 21$

   a. Verwerk deze gegevens in een spreidingsdiagram en bereken de correlatiecoëfficiënt. Is er sprake van een correlatie tussen W en F?
   b. Stel de vergelijking op van de regressielijn van W op F. Waarom ligt deze regressielijn meer voor de hand dan die van F op W?
   c. Het verband tussen de temperatuur in graden Fahrenheit F en die in graden Celsius wordt zoals bekend gegeven door: $F = 1{,}8C + 32$.
   Stel nu een vergelijking op van de regressielijn van W op C.
   d. Is de correlatiecoëfficiënt tussen W en C anders dan die tussen W en F? Verklaar je antwoord.
   e. Uit andere experimenten is gebleken dat de werkzaamheid bij een vaste bewaartemperatuur exponentieel afhangt van de lengte van de bewaarperiode. Schat de gemiddelde werkzaamheid van porties die 40 dagen bij een temperatuur van 20°C zijn bewaard.

5. **Exponentiële groei en correlatie**
In de tabel vind je het aantal inwoners N in een bepaalde stad.

Jaartal	1960	1970	1980	1990	2000
Aantal inwoners N	23.107	25.880	28985	32.479	36.358

   Er wordt aangenomen dat N een exponentiële functie is van t, de tijd in jaren met $t = 0$ in 1960. Voorspel met behulp van een regressielijn het aantal inwoners in 2010 en 2020. Bereken ook de bijbehorende correlatiecoëfficiënt.
   Welke gevolgen heeft een eventueel regressie-effect op je voorspelling?

6. **Voor het werkstuk**

   a. Je hebt in diverse opgaven gewerkt in de werkmap LengteGewicht.xls. Maak nu die werkmap compleet:
      - ▷ Op Blad 1 komen de gegevens van beide variabelen samen met de twee regressielijnen. Bereken daar ook de correlatiecoëfficiënt en de twee regressiecoëfficiënten.
      - ▷ Op Blad 2 worden de gegevens van de variabele l statistisch verwerkt. Daar komt een frequentietabel, een histogram, de centrum- en de spreidingsmaten voor l.
      - ▷ Op Blad 3 worden de gegevens van de variabele g statistisch verwerkt. Daar komt een frequentietabel, een histogram, de centrum- en de spreidingsmaten voor g.

      Schrijf bij die werkmap een toelichting met een duidelijke probleemstelling, een overzicht van de gegevens (centrum- en spreidingsmaten, normaal verdeeld of niet, etc.) een overzicht van de theoretische achtergronden van de correlatie tussen beide variabelen, een conclusie gebaseerd op de regressielijnen en enkele opmerkingen over het regressie-effect.

   b. Rond je eigen onderzoek af, bereken de regressielijnen en maak een onderzoeksverslag.

Index

A Antwoorden

Hier tref je de antwoorden aan bij de echte oefenopgaven. Het zijn alleen eindantwoorden zonder tekeningen.

Hoofdstuk 1

1. Gebruik zoveel mogelijk de statistische functies van Excel.
   Als het goed is vind je $\overline{g} \approx 59{,}09$ en $\sigma_g \approx 6{,}80$.
   De spreidingsbreedte wordt $75 - 48 = 27$ kg.

2. Dat is onwaarschijnlijk. Ten eerste is deze steekproef veel te klein, ten tweede is het zeer de vraag of juist HAVO-leerlingen representatief zijn voor de gehele populatie van 15-17 jarigen.

3. Maak eerst een goede klassenindeling en dan een cumulatieve frequentieverdeling. Op normaal waarschijnlijkheidspapier zet je dan de somfrequenties (omgerekend in procenten) uit tegen de *rechter* klassengrenzen.
   De verdeling zal vast niet erg normaal worden, want daarvoor is deze steekproef te klein. Bovendien kan zoiets beter voor jongens en meisjes afzonderlijk worden gedaan!

4. a. Het gemiddelde is 69,5 en de standaarddeviatie is 14,92.
   b. Het gemiddelde is 71,4 en de standaarddeviatie is 13,96.
   c. Omdat je werkt met klassenmiddens als je ze berekent met je grafische rekenmachine, of omdat je afleest van normaal waarschijnlijkheidpapier.
   d. Bij wiskunde was het gemiddelde iets hoger en waren er meer 'bovengemiddelde' scores.
   e. Ja, dat kan wel. Dan moet je alle 100 mogelijke combinaties (klassenmiddens!) onder elkaar invoeren in bijvoorbeeld Excel, net als op Blad 1 van de werkmap LengteGewicht.xls.

Hoofdstuk 2

1. a. Op het oog zeker!
   b. Waarschijnlijk wel tussen de 0,9 en de 1.
   c. Een lineaire functie, dus een functie van de vorm $y = ax + b$.
   d. Je deelt door de standaarddeviaties.

2. De correlatiecoëfficiënt is ongeveer 0,8096, dus de correlatie is behoorlijk groot.

3. Doen met behulp van Excel.

4. Haakjes uitwerken geeft:

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i \cdot y_i - x_i \cdot \overline{y} - y_i \cdot \overline{x} + \overline{x} \cdot \overline{y})}{N \cdot \sigma_x \cdot \sigma_y}$$

   Daaruit volgt:

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i \cdot y_i)}{N \cdot \sigma_x \cdot \sigma_y} - \frac{\sum_{i=1}^{N} x_i \cdot \overline{y}}{N \cdot \sigma_x \cdot \sigma_y} - \frac{\sum_{i=1}^{N} y_i \cdot \overline{x}}{N \cdot \sigma_x \cdot \sigma_y} + \frac{N \cdot \overline{x} \cdot \overline{y}}{N \cdot \sigma_x \cdot \sigma_y}$$

en dus:

$$r_{xy} = \frac{\overline{x \cdot y}}{\sigma_x \cdot \sigma_y} - \frac{\overline{x} \cdot \overline{y}}{\sigma_x \cdot \sigma_y} - \frac{\overline{y} \cdot \overline{x}}{\sigma_x \cdot \sigma_y} + \frac{\overline{x} \cdot \overline{y}}{\sigma_x \cdot \sigma_y} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sigma_x \cdot \sigma_y}$$

5. a. Gebruik je grafische rekenmachine.
   b. $r_{vz} \approx 0{,}70$
   c. Ja, hoewel de correlatie niet heel erg sterk is.

6. a. Gebruik je grafische rekenmachine.
   b. $r_{Lv} \approx 0{,}59$, dus geen erg sterke correlatie.
   c. De correlatie wordt duidelijk beter!
   d. $r \approx 0{,}90$
   e. Uit $\log v = a \cdot \log L + b$ volgt: $v = 10^b \cdot L^a$.

Hoofdstuk 3

1. a. Gebruik STAT PLOT op je grafische rekenmachine.
   b. $(4{,}3; 56)$
   c. Tussen 4 en 5 in.
   d. $r_{xy} \approx 0{,}9877$ en $a \approx 4{,}48$.
   e. $y = 4{,}48x + 36{,}72$
   f. Ongeveer 81,5.

2. a. Werk de haakjes uit en maak gebruik van $b = \overline{y} - a \cdot \overline{x}$. Je krijgt dan een nogal ingewikkelde uitdrukking in a^2 en a.
   b. Bedenk dat

$$\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N} \quad \text{en} \quad \sigma_y^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y})^2}{N}$$

   Verder is het nogal lastig geknutsel met somtekens en zo.
   c. Als je de formule voor r_{xy} vermenigvuldigt met σ_y, dan werk je de σ_y in de noemer weg. Vervolgens weer delen door σ_x en je krijgt in de noemer $\sigma_x \cdot \sigma_x$, hetgeen precies staat in de formule voor a.

3. a. $g \approx 0{,}59 \cdot l - 44{,}01$
   b. Als je van een 15-17 jarige de lengte weet, kun je met de formule voor de regressielijn het gewicht voorspellen.
   c. Ongeveer 62,5 kg.

4. a. Een positieve correlatie, dus een zoon zal over het algemeen langer zijn dan zijn vader.
   b. $z \approx 0{,}47v + 95{,}44$
   c. Ongeveer 178,6 cm.

5. De regressielijn wordt: $\log v \approx 0{,}36 \log L + 2{,}67$, dus de machtsfunctie wordt: $v \approx 10^{2{,}67} \cdot L^{0{,}36} \approx 468 \cdot L^{0{,}36}$.

6. 0,99

Hoofdstuk 4

1. Werk de voorbeelden op blz.15 zorgvuldig na!

2. a. $x \approx 0{,}22y - 7{,}88$
   b.
   c. Neem bijvoorbeeld voor x precies één keer de standaarddeviatie boven $\overline{x}$. Je zult dan voor y een uitkomst vinden die minder dan σ_y boven $\overline{y}$ zit.

3. Dat de voorspelling van de lengte van de zoon aan de lage kant zal zijn.

4. a. $r_{WF} \approx -0{,}94$, een duidelijke negatieve correlatie
   b. $W \approx -0{,}35F + 48{,}30$
   De regressielijn van W op F ligt meer voor de hand omdat gezocht wordt naar een verband waarbij de werkzaamheid afhangt van de bewaartemperatuur.
   c. $W \approx -0{,}63C + 37{,}10$
   d. Nee, want de schaalverdeling speelt geen rol bij de correlatie, het gaat alleen om de ligging van de meetpunten ten opzicht van de regressielijn.
   e. Uit de formule voor de regressielijn volgt dat de werkzaamheid in 20 dagen bij 20°C terugloopt tot ongeveer 24,5%.
   Voor een periode van 40 dagen loopt de werkzaamheid daarom terug tot $0{,}245 \cdot 24{,}5 \approx 6\%$.

5. Maak een tabel waarin je t uitzet tegen $\log N$.
   Je vindt dan een bijna perfecte correlatie ($r_{tN} = 0{,}9999\ldots$).
   De bijbehorende regressielijn is: $\log N \approx 0{,}005t + 4{,}364$, dus is: $N \approx 23.121 \cdot 10^{0{,}005}t$.
   Dat betekent voor 2010 ongeveer 41.116 inwoners en voor 2020 ongeveer 46.132 inwoners.
   Er is vrijwel geen regressie-effect.

Lengte en gewicht **25**

Keuzeonderwerp wiskunde **25**