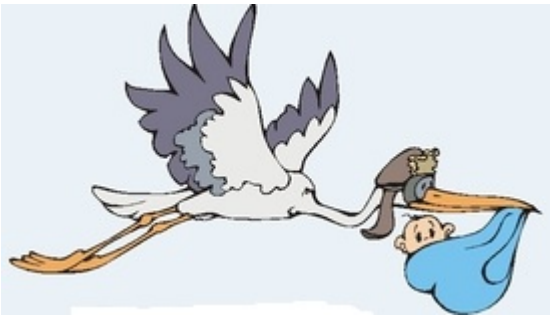
6.5 Correlatie

Inleiding

Naast het toetsen van hypothesen is in de mathematische statistiek het onderzoeken naar statistische verbanden een belangrijke tak van sport: wanneer bestaat er een verband tussen twee statistische variabelen? Bestaat er bijvoorbeeld een verband tussen het aantal overvliegende ooievaars en het aantal geboorten in een bepaalde streek? Of bestaat er een verband tussen lengte en gewicht bij scholieren?



Figuur 1

Je leert in dit onderwerp

- een puntenwolk tekenen bij twee statistische variabelen;
- de correlatiecoëfficiënt gebruiken als maat voor het statistische verband tussen beide variabelen;
- het verschil tussen statistische verbanden en oorzakelijke verbanden.

Voorkennis

- de elementaire statistische begrippen (gemiddelde, standaardafwijking, e.d.) gebruiken;
- werken met binomiale en normale toetsen.

Verkennen

Opgave V1

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 **LengteGewicht22h4.xls** vind je de gegevens van een 4HAVO-klas van 22 leerlingen.

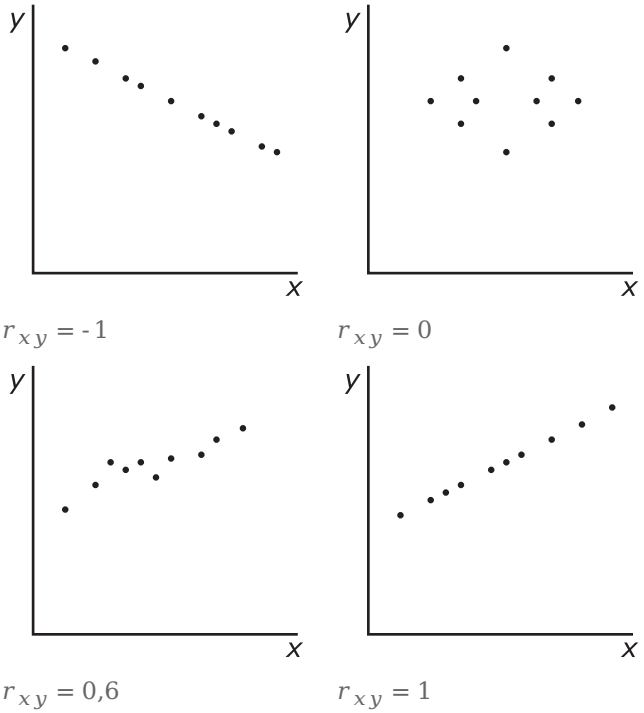
a Welke drie gegevens zijn er verzameld?

b Welke afspraken moet je maken bij het verzamelen van deze gegevens? Beschrijf er een paar. (Denk om de manier van meten!)

c Bekijk het spreidingsdiagram. Trek je op grond van de gegevens op het werkblad de conclusie dat er zo'n verband bestaat? En is dat dan uitsluitend een statistisch verband of is het ook een oorzakelijk verband. Met andere woorden wordt een groter gewicht veroorzaakt door een grotere lengte?

Uitleg

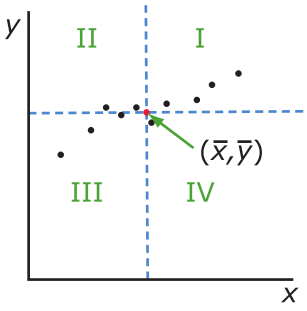
Als je vermoedt dat er tussen twee variabelen x en y een lineair verband bestaat, maak je een spreidingsdiagram dat de vorm van een puntenwolk krijgt. De mate waarin tussen de twee variabelen een lineair verband bestaat wordt gegeven door de correlatiecoëfficiënt, aangeduid door r_{xy}.



Figuur 2

De correlatie tussen x en y wordt beter naarmate r_{xy} dichter bij 1 of -1 ligt. Maar hoe bereken je nu die correlatiecoëfficiënt?

Daarbij gebruik je het punt $(\overline{x},\overline{y})$ waarin $\overline{x}$ het gemiddelde van de x-waarden en $\overline{y}$ het gemiddelde van de y-waarden is. 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 van de N punten (x_i, y_i) het getal $(x_i - \overline{x})(y_i - \overline{y})$ 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 N getallen $(x_i - \overline{x})(y_i - \overline{y})$ is een goede maat voor de correlatie.



Figuur 3

Deze maat heet de covariantie van de puntenwolk: covariantie $= \frac{\Sigma_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{N}$.

Deze maat voor de correlatie in een puntenwolk hangt nog af van de eenheden waarin x en y zijn gemeten. Dat kun je voorkomen door telkens $(x_i - \overline{x})$ te delen door de bijbehorende standaarddeviatie σ_x en ook $(y_i - \overline{y})$ telkens te delen door σ_y. Je krijgt dan de correlatiecoëfficiënt, die niet langer afhangt van de gekozen eenheden:

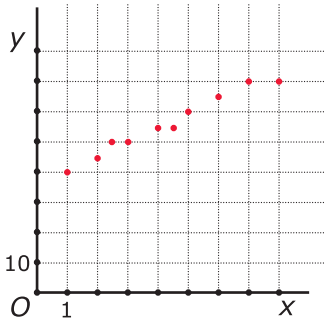
$$r_{xy} = \frac{\Sigma_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{N \cdot \sigma_x \cdot \sigma_y}.$$

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 $(x_i - \overline{x})(y_i - \overline{y})$. 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, zie het **Practi-**

cum.

Opgave 1

Bekijk dit spreidingsdiagram.



Figuur 4

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?

Opgave 2

Op het werkblad **LengteGewicht22h4.xls** vind je de gegevens van een 4HAVO-klas van 22 leerlingen.
TIP: Natuurlijk is het leuker (en beter) om met een eigen dataset van lengtes en gewichten van jouw jaargroep te werken.

a Bereken het gemiddelde, de standaarddeviatie en de spreidingsbreedte van zowel de lengtes l als de gewichten g. Gebruik de statistische functies van je grafische rekenmachine.

b Is deze steekproef voldoende representatief voor 15-17 jarigen? Motiveer je antwoord.

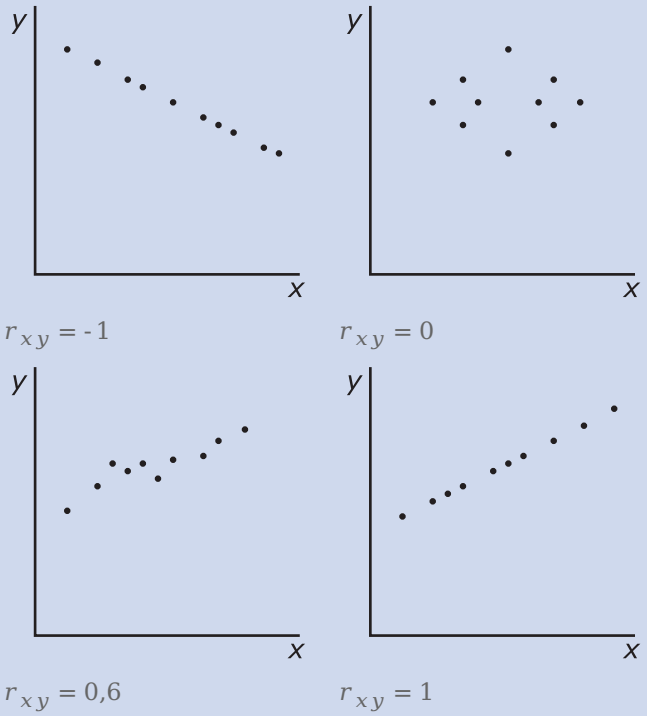
c Ga met behulp van normaal waarschijnlijkheidspapier na of de lengtes van de 22 leerlingen in de voorgaande tekst ongeveer normaal verdeeld zijn. Doe dit ook voor de gewichten.

d Bereken de correlatiecoëfficiënt bij het verband tussen de lengte en het gewicht van de 22 leerlingen. Is er sprake van een goede correlatie tussen l en g?

Theorie en voorbeelden

Om te onthouden

In een **spreidingsdiagram** van twee statistische variabelen x en y zet je alle combinaties (x,y) als een **puntenwolk** in een assenstelsel. Of er een sterk **lineair statistisch verband** bestaat tussen de variabelen wordt bepaald door de **correlatiecoëfficiënt** r_{xy}. Er geldt: $r_{xy} = \frac{\Sigma_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{N \cdot \sigma_x \cdot \sigma_y}$.



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

$r_{xy} = 0{,}6$ $r_{xy} = 1$

Figuur 5

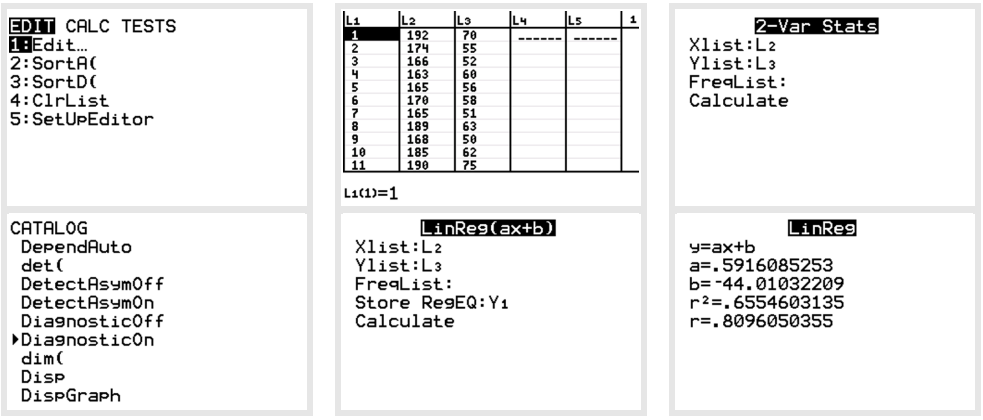
- Als $r_{xy} = 1$ dan is er een perfecte positieve correlatie tussen x en y. De punten van de puntenwolk liggen dan precies op een stijgende lijn.
- Als $r_{xy} = 0$ dan is er geen enkele correlatie tussen x en y.
- Als $r_{xy} = -1$ dan is er een perfecte negatieve correlatie tussen x en y. De punten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

Een verband waarbij de toename (of afname) van de éne variabele een gevolg is van een toename (of afname) van de andere heet een **causaal verband**: er is dan sprake van oorzaak en gevolg. Een statistisch verband tussen twee variabelen hoeft niet causaal te zijn. Andere variabelen kunnen de oorzaak zijn dat er bij twee variabelen een statistisch verband optreedt. Het is zeker niet zo, dat een grotere lengte veroorzaakt dat je daardoor automatisch ook een groter gewicht hebt.

Voorbeeld 1

Op het werkblad **LengteGewicht22h4.xls** vind je de gegevens van een 4HAVO-klas van 22 leerlingen. Je kunt deze gegevens ook in de grafische rekenmachine invoeren en die de correlatiecoëfficiënt laten berekenen. In de figuren hieronder zie je hoe dit op de TI84 in zijn werk gaat. Je hebt er het rekenalgoritme LinReg voor nodig. Dat staat voor 'lineaire regressie' en wordt in het volgende onderdeel nader bekeken. In het **Practicum** zie je hoe dit op de diverse rekenmachines gaat.



Figuur 6

Opgave 3

Voer de berekening van de correlatiecoëfficiënt bij de gegevens van de 4HAVO-klas uit **Voorbeeld 1** zelf uit met behulp van de grafische rekenmachine.

Opgave 4

De inspectie voor het onderwijs vergelijkt van een bepaalde school de cijfers voor wiskunde B van het SE (schoolexamen) en het CE (centraal examen). In de tabel vind je de gegevens van een klas van 19 leerlingen.

leerling	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
SE-cijfer	6,0	6,7	5,8	7,1	5,4	6,5	8,8	6,9	7,9	5,1	6,1	6,1	6,4	7,4	5,9	6,2	7,1	6,8	6,3
CE-cijfer	6,4	6,3	5,2	6,5	5,4	6,1	9,0	6,8	7,5	5,6	6,0	6,5	6,0	6,5	6,0	6,6	7,0	6,6	6,4

Tabel 1

Je zou kunnen onderzoeken of er een lineair statistisch verband is tussen het CE-cijfer c en het SE-cijfer s. Teken een bijpassend spreidingsdiagram en ga door berekening van de correlatiecoëfficiënt na of zo'n verband bestaat.

Verwerken

Opgave 5

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

Tabel 2

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?

Opgave 6

De formule voor de correlatiecoëfficiënt is te herleiden 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 **Uitleg** de haakjes uit te werken. (Als je de correlatiecoëfficiënt handmatig moet uitrekenen gaat dat met deze formule iets sneller.)

Opgave 7

Soms is er wel sprake van een goede correlatie tussen twee statistische variabelen, maar kun je toch je vraagtekens zetten bij het verband tussen beide.

a In een provincie neemt het aantal ooievaars en het aantal geboorten af. Het spreidingsdiagram geeft een statistisch verband te zien. Bestaat er een causaal verband tussen het aantal ooievaars en het aantal geboorten?

b Leg uit waarom er wel een statistisch verband is tussen ijsverkoop en verkoop van zonnebrillen in de zomer maar geen causaal verband.

Opgave 8

Iemand probeert aan te tonen dat de klassengrootte van invloed is op de leerprestaties. Zij vergelijkt - onder zoveel mogelijk gelijke omstandigheden - de gemiddelde cijfers voor drie wiskundetoetsen in klassen met uiteenlopende leerlingenaantallen. Hier zie je de verzamelde gegevens.

a Maak een spreidingsdiagram met c op de verticale en a op de horizontale as. Waarom is dit een logische keuze?

b Bereken de correlatiecoëfficiënt. Is er sprake van een duidelijke correlatie? Bestaat er tussen a en c een lineair statistisch verband?

c Welke conclusie zou deze onderzoekster kunnen trekken? Geef daar commentaar op.

aantal leerlingen a	gemiddelde cijfer c
30	6,1
25	6,6
32	5,5
24	7,2
18	7,4
19	6,9
30	5,2
22	7,1
29	6,0
14	7,8

Tabel 3

Opgave 9

In 1947 hielden de wiskundigen Freudenthal en Sittig een statistisch onderzoek ten behoeve van een nieuw maatsysteem voor vrouwenkleding in opdracht van het warenhuis De Bijenkorf. Zij lieten daarbij een grote verscheidenheid aan lichaamsmaten opmeten van 5001 vrouwen. In het bestand **StatFS-Bijenkorf1947.xls** vind je enkele gegevens.
Gebruik de werkbladen ‘lengte-gewicht’, ‘mouwlengte-kniehoogte’ en ‘voetlengte-breedte’.

a Waarom was De Bijenkorf geïnteresseerd in dergelijke gegevens? En waarom zijn eventuele verbanden als die tussen voetlengte en voetbreedte van belang?

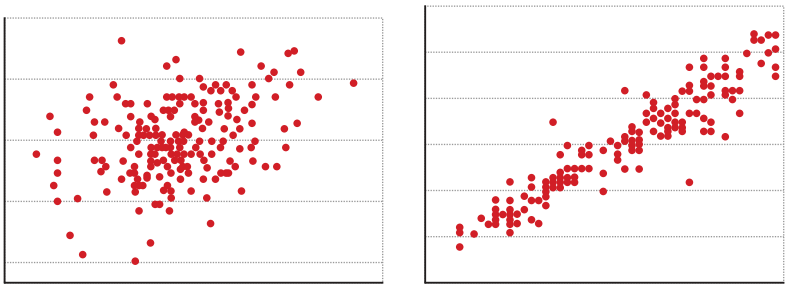
b Op de drie genoemde werkbladen is er sprake van een mogelijk verband tussen twee variabelen. Hoe zou je in dit geval een spreidingsdiagram tekenen?

c En hoe zou je een correlatiecoëfficiënt berekenen? Waarom is het hier handiger om over de oorspronkelijke ruwe meetgegevens te beschikken?

Toepassen

Opgave 10: Huwelijken

In een onderzoek onder 199 echtparen is gevraagd naar de lengte en de leeftijd van de man en de vrouw. Onder andere werd onderzocht of er bij bepaalde eigenschappen van de gehuwden sprake was van een bepaalde statistische samenhang. Dit heeft geresulteerd in de volgende twee puntenwolken:

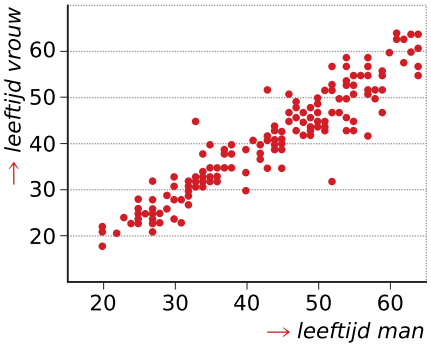


Figuur 7

Een van beide puntenwolken heeft betrekking op de leeftijden van de twee huwelijkspartners, waarbij de gegevens van de man op de horizontale as zijn uitgezet en die van de vrouw op de verticale as. De andere puntenwolk heeft betrekking op de lengte van beide partners. Ook hier zijn de gegevens van de man weer op de horizontale as uitgezet.

a Beredeneer dat, op basis van de vorm van de puntenwolk, de linker puntenwolk zeer waarschijnlijk betrekking heeft op de lengte en de rechter puntenwolk op de leeftijd.

b Bekijk de puntenwolk. Onderzoek met behulp van de puntenwolk of het in de betreffende 199 huwelijken vaker voorkomt dat de man ouder is dan de vrouw of dat het omgekeerde juist vaker voorkomt. Laat duidelijk zien hoe je tot je antwoord gekomen bent.



Figuur 8

Op basis van dergelijke puntenwolken wil men soms een schatting maken van de lengte of de leeftijd van een vrouw als men de lengte of de leeftijd van de man kent. Hoewel dit soort schattingen altijd een grote mate van onzekerheid hebben, is het toch mogelijk om aan te geven bij welk van de twee puntenwolken een dergelijke schatting het meest betrouwbaar zal zijn.

c Beredeneer bij welk van de twee puntenwolken, die met de leeftijden of die met de lengtes, een dergelijke schatting het meest betrouwbaar zal zijn.

In de tabel is een aantal kengetallen weergegeven uit het onderzoek.

	leeftijd man (jaar)	leeftijd vrouw (jaar)	lengte man (cm)	lengte vrouw (cm)
gemiddelde	42,6	40,7	173	160
minimum	20	18	156	141
maximum	64	64	195	176
standaardafwijking	11,6	11,4	6,9	6,2

Tabel 4

Ervan uitgaande dat de lengtes en de leeftijden van de huwelijkspartners nagenoeg normaal verdeeld zijn, is met behulp van deze gegevens uit te rekenen dat 95% van de lengtes van de mannen tussen de 159,2 cm en 186,8 cm zal liggen.

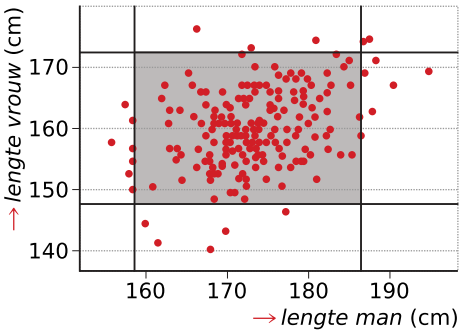
d Leg uit hoe je aan deze waarden komt.

e Bepaal tussen welke twee lengtes 95% van de vrouwen zit.

Omdat 5% van de mannen buiten de berekende grenzen zal vallen, evenals 5% van de vrouwen, concludeert de onderzoeker dat in totaal 10% van de punten uit de puntenwolk buiten de getekende rechthoek zullen vallen.

f Beargumenteer of je het met die conclusie eens bent of niet.

(bron: voorbeeldopgave Statistiek – syllabus havo A)



Figuur 9

Opgave 11: 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	Vliegsnelheid 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

Tabel 5

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.

Testen

Opgave 12

Bekijk de tabel. In een Amerikaans laboratorium heeft men proeven gedaan 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 (°F)	30	50	70	90
werkzaamheid W (%)	39, 42, 35	32, 26, 33	19, 27, 23	14, 19, 21

Tabel 6

a In hoeverre is er sprake van correlatie tussen bewaartemperatuur F en werkzaamheid W?

b Is de conclusie gerechtvaardigd dat de werkzaamheid van het geneesmiddel afneemt als de bewaartemperatuur stijgt?

Practicum

Met deze practica leer je hoe je de **de correlatiecoëfficiënt** met de grafische rekenmachine berekent.

- **Trendlijn, correlatie en de TI84**
- **Trendlijn, correlatie en de TInspire**
- **Trendlijn, correlatie en de Casio fx-CG50**
- **Trendlijn, correlatie en de HPprime**
- **Trendlijn, correlatie en de NumWorks**

Met het volgende practicum kun je zien hoe je **de correlatiecoëfficiënt in Excel** berekent. Dat is handig als je een grote set gegevens hebt. Je treft er ook in aan hoe je de **regressielijn**, dat is de meest geschikte lijn door de puntenwolk, kunt tekenen en er door Excel de vergelijking van kunt laten opstellen. In Excel heet die lijn de ‘trendlijn’. In het volgende onderdeel hoor je daar meer over.

- **Correlatie en regressie**

OPMERKING:
Natuurlijk is het veel mooier om een eigen dataset met gegevens van leerlingen in jouw jaargroep te gebruiken. Die moet je dan wel eerst zelf maken: statistisch onderzoek!