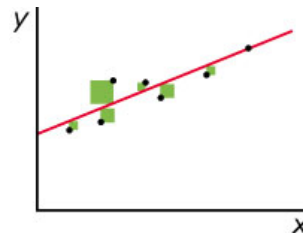


4.4 Samenhang van variabelen

Inleiding

Behalve het onderzoeken van verschillen tussen statistische variabelen is het onderzoeken naar 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

- de (statistische) samenhang tussen twee variabelen te beschrijven en te interpreteren;
- te werken met correlatiecoëfficiënt en regressielijn (trendlijn).

Voorkennis

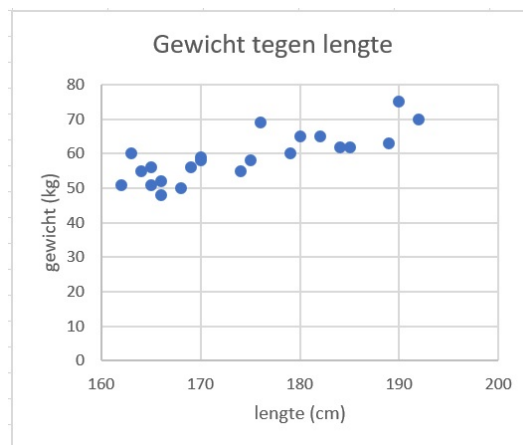
- soorten statistische variabelen herkennen;
- de begrippen onderzoek, steekproef, populatie en representatief, simulatie.

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.

- Welke gegevens zijn er verzameld?
- Welke afspraken moet je maken bij het verzamelen van deze gegevens? Beschrijf er een paar. (Denk om de manier van meten!)
- Bekijk het getekende 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, m.a.w. wordt een groter gewicht veroorzaakt door een grotere lengte?



Figuur 2

Uitleg 1

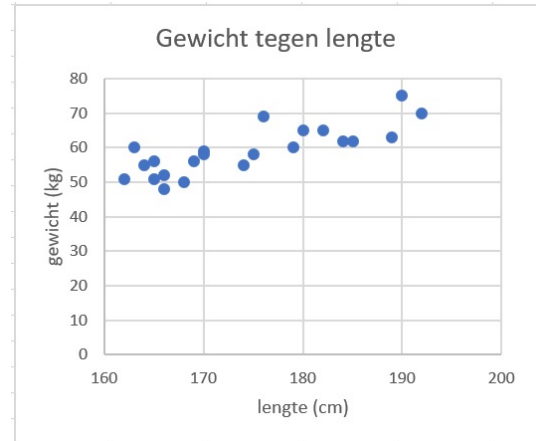
Je wilt onderzoeken of er een verband bestaat tussen lengte en gewicht bij mensen van 15 tot 17 jaar oud. Op het werkblad [LengteGewicht22H4.xls](#) vind je de gegevens van een steekproef van 22 leerlingen. Hiernaast is een spreidingsdiagram van die gegevens getekend.

Is er binnen deze groep sprake van een verband tussen lengte en gewicht?

Als je de figuur bekijkt, zie je dat bij grotere lengtes vaak ook grotere gewichten horen.

Er lijkt dus een zeker verband te zijn.

Maar de getekende punten liggen zeker niet op één lijn, dus hoe sterk is het verband?



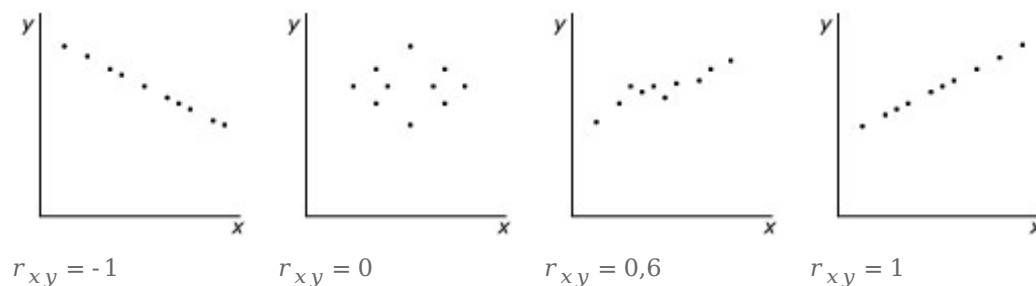
Figuur 3

De vorm van de puntenwolk, het spreidingsdiagram, zegt iets over de samenhang tussen twee variabelen. Zie daarvoor de afbeeldingen hieronder.

Een ander woord voor statistisch verband of samenhang is correlatie.

De mate van correlatie wordt uitgedrukt in de correlatiecoëfficiënt r .

Correlatie zegt iets over de statistische samenhang van twee variabelen en niets over een eventuele causale relatie (dus een oorzaak-gevolgrelatie) tussen de twee variabelen. Er is samenhang tussen de verkoop van handschoenen en de verkoop van winterbanden, maar de oorzaak is dan het mogelijke koude winterweer.



Figuur 4

Opgave 1

Gebruik het werkblad [LengteGewicht22H4.xls](#) in [Uitleg 1](#).

- Maak zelf zo'n spreidingsdiagram.
- Hoe zie je in het spreidingsdiagram dat er een zeker verband tussen l en G bestaat?
- Bereken met behulp van Excel de correlatiecoëfficiënt r_{lG} in twee decimalen nauwkeurig, bekijk het [Practicum](#).

Je hebt de waarde van r_{lG} berekend. Maar hoe weet je nu of deze correlatie genoeg is om vast te stellen dat er een verband is? Daarvoor zijn vuistregels zoals: Als $r \geq 0,7$ dan is er sprake van sterke positieve samenhang.

- Kun je vaststellen dat er een duidelijk verband is tussen lengte en gewicht van deze jongeren?

Opgave 2

Je hebt in de voorgaande opgave een verband aangetoond tussen lengte en gewicht bij mensen van 15 tot 17 jaar oud.

Daarbij gebruikte je een steekproef van 22 studenten.

- a** Geef commentaar op deze steekproef.

Stel je voor dat je met een aselechte steekproef te maken hebt. En neem ook aan dat de groep representatief is voor studenten van 15 tot 17 jaar oud. Er is een duidelijk verband gevonden.

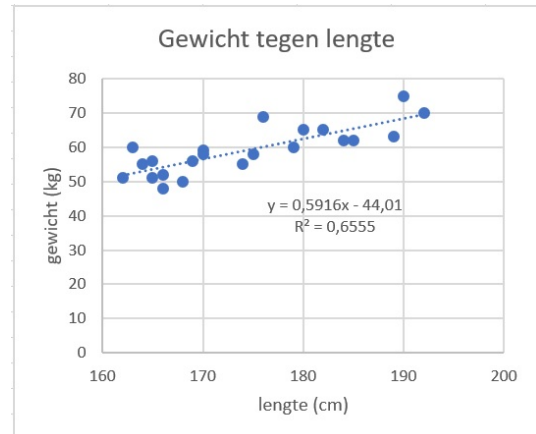
- b** Is er nu sprake van een oorzakelijk verband, dus is de lengte oorzaak van het gewicht? Of is er alleen een statistisch verband? Licht het antwoord toe.

Uitleg 2

Je wilt onderzoeken of er een verband bestaat tussen lengte en gewicht bij mensen van 15 tot 17 jaar oud. Op het werkblad [LengteGewicht22H4.xlsx](#) vind je de gegevens van een steekproef van 22 studenten. Daarnaast is een spreidingsdiagram van die gegevens getekend.

Er is binnen deze groep studenten sprake van een verband tussen lengte en gewicht, maar hoe beschrijf je dit verband?

Je ziet in de figuur hoe Excel automatisch het kwadraat van de correlatiecoëfficiënt r berekent en een best passende lijn door de puntenwolk trekt. Deze lijn heet de regressielijn of trendlijn. En r^2 heet de determinatiecoëfficiënt.



Figuur 5

Bij de trendlijn kun je ook zelf een formule opstellen. Daarvoor lees je eerst twee geschikte punten op de trendlijn af, bijvoorbeeld (160,50) en (190,68).

Met behulp van deze punten kun je de formule opstellen. Je vindt:

$$G = 0,6 \cdot l - 46$$

Hiermee kun je voorspellingen doen over bijvoorbeeld het gewicht van iemand in de doelgroep van 2 meter.

Opgave 3

Je ziet in [Uitleg 2](#) hoe Excel een lijn door de puntenwolk trekt en de bijbehorende formule berekent.

- a** Je ziet in de uitleg hoe je zelf de formule voor de trendlijn kunt berekenen.

Voer die berekening uit.

Je kunt nu met behulp van de gevonden regressielijn (trendlijn) voorspellingen doen.

- b** Hoe zwaar zou iemand van 2,00 m volgens de regressielijn moeten zijn?

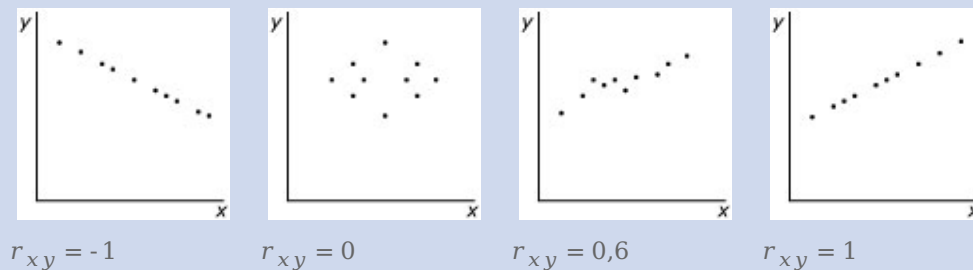
Theorie en voorbeelden

Om te onthouden

Wanneer je binnen een dataset zoekt naar een verband tussen twee statistische variabelen, gebruik je een **puntenwolk** of **spreidingsdiagram** (Engels: scatter plot).

Afhankelijk van de vorm van een puntenwolk kun je vaststellen of er een verband tussen beide variabelen is en zo ja, of dat verband sterk is. Een maat daarvoor is de **correlatiecoëfficiënt** r , een getal met waarden vanaf -1 tot en met 1. Hoe dicht r bij 1 of -1 ligt, hoe sterker het verband. In een volgend voorbeeld staan vuistregels om aan te geven of de samenhang zwak, matig of sterk is.

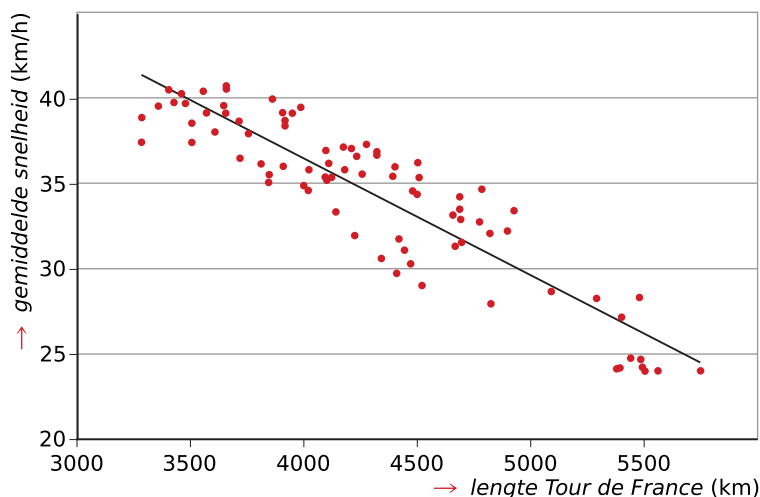
In een spreidingsdiagram kun je met een **trendlijn** de (lineaire) samenhang weergeven. Hoe je deze trendlijn en de correlatiecoëfficiënt berekent, zie je in het [Practicum](#).



Figuur 6

Let op! Je stelt alleen vast dat er een **statistische samenhang** tussen beide variabelen is. Het is de vraag of dat verband ook **causaal** is. Je kunt dus niets zeggen over oorzaak en gevolg. Er kan best een derde variabele de oorzaak zijn van de samenhang.

Voorbeeld 1



Figuur 7

In deze figuur staan twee statistische variabelen. Op de horizontale as staat de totale lengte van de Tour de France in kilometer. Op de verticale as staat de gemiddelde snelheid van de winnaar.

Welke samenhang is er tussen de lengte van de Tour de France en de gemiddelde snelheid van de winnaar?

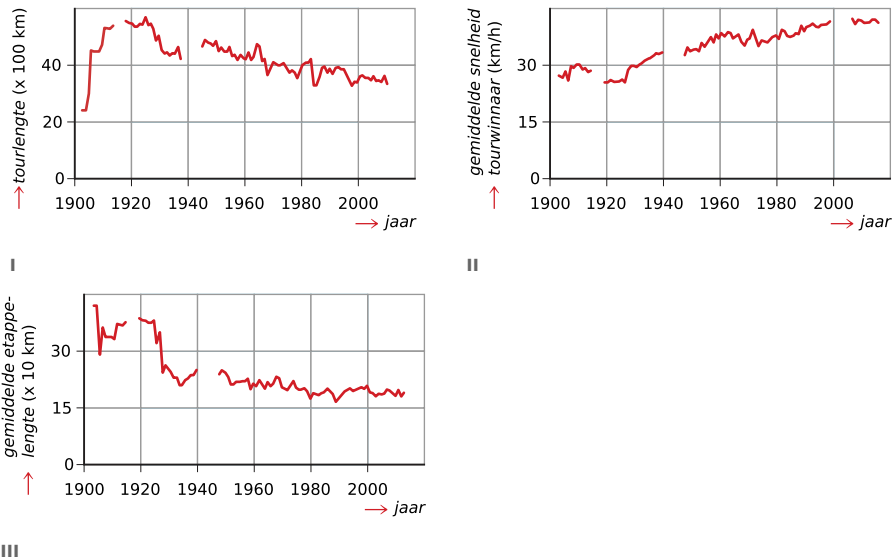
Antwoord

Als je kijkt naar de afbeelding, lijkt er een (sterke) negatieve samenhang te zijn. Hoe korter de Tour de France, hoe hoger de gemiddelde snelheid van de winnaar.

Maar pas op met oorzaak en gevolg. De oorzaak van een hogere gemiddelde snelheid van de winnaar hoeft niet de afnemende lengte van het parcours te zijn.

Opgave 4

Bekijk de drie grafieken:



Figuur 8

Twee van de grafieken kun je combineren tot een spreidingsdiagram.

- Hoe komt het dat je deze grafieken kunt combineren tot een spreidingsdiagram?
- De drie grafieken hebben allemaal twee gaten. Wat betekent dit en wat is de oorzaak?
- Welke twee (van de drie) grafieken tonen gecombineerd het spreidingsdiagram uit het voorbeeld?
- Welke samenhang wordt zichtbaar, als je steeds twee andere grafieken combineert?

Opgave 5

Het is gevaarlijk om conclusies te trekken. Het gegeven spreidingsdiagram suggereert dat de gemiddelde snelheid van de winnaar omlaag gaat, als een Tour de France langer gemaakt wordt.

- Geef drie redenen waarom de gemiddelde snelheid van de winnaar los staat van de turlengte.
- Noem nog een tweetal variabelen die vermoedelijk een sterke samenhang hebben met de gemiddelde snelheid van de winnaar.
- Noem een tweetal variabelen die geen oorzakelijk verband hebben met de gemiddelde snelheid van de winnaar.

Voorbeeld 2

De mate van samenhang is in een getal uit te drukken: de correlatiecoëfficiënt. Deze correlatiecoëfficiënt r kun je bijvoorbeeld met Excel of met de grafische rekenmachine laten uitrekenen. Zie het [Practicum](#). Vervolgens kun je met vuistregels conclusies trekken.

- Als $r \leq -0,7$ dan is er sprake van sterke negatieve samenhang.
- Als $-0,7 < r \leq -0,3$ dan is er sprake van matige negatieve samenhang.
- Als $-0,3 < r < 0$ dan is er sprake van zwakke negatieve samenhang.
- Als $0 < r < 0,3$ dan is er sprake van zwakke positieve samenhang.
- Als $0,3 \leq r < 0,7$ dan is er sprake van matige positieve samenhang.
- Als $r \geq 0,7$ dan is er sprake van sterke positieve samenhang.

Opgave 6

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 1

- Teken een spreidingsdiagram (een puntenwolk) bij deze gegevens.
- Bereken de correlatiecoëfficiënt in twee decimalen nauwkeurig.
- Bestaat er een lineair verband tussen v en z ? Zo ja, stel dan een bijpassende formule op.

Voorbeeld 3

Bekijk het spreidingsdiagram. Daarin zijn de resultaten weergegeven van een onderzoek in dertig landen. Het bnp en de ug zijn gemiddelden per inwoner.

Je kunt door deze punten een trendlijn tekenen. De bijbehorende correlatiecoëfficiënt is een maat voor de sterkte van het verband. Is de lijn dalend, dan is de correlatiecoëfficiënt negatief en is ook de richtingscoëfficiënt van de trendlijn negatief.

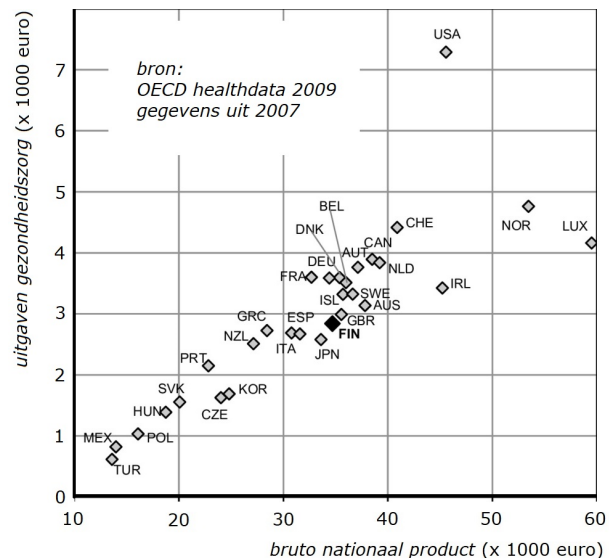
Je kunt ook een trendlijn door deze gegevens tekenen. In de praktijk heb je de gegevens vaak in een database beschikbaar. De trendlijn berekenen en tekenen je dan met de computer, zie het **Practicum**.

Welke vergelijking heeft deze trendlijn?

Antwoord

Je zoekt twee punten. Deze punten moeten niet te dicht bij elkaar liggen, bijvoorbeeld de punten voor TUR en NLD: (13; 0,5) en (39; 3,9).

De vergelijking van de trendlijn is $ug \approx 0,13 \cdot bnp - 1,20$.



Figuur 9

Opgave 7

Gebruik de gegevens uit **Voorbeeld 3**.

- Waarom moet je de punten niet te dicht bij elkaar kiezen om de vergelijking van een trendlijn op te stellen?
- Stel zelf de vergelijking van de trendlijn op.

Verwerken

Opgave 8

Bekijk het spreidingsdiagram. Op de horizontale as staat *kniehoopte* in cm en op de verticale as de *mouwlengte* in cm. De getallen stellen het aantal vrouwen voor die bij een bepaalde kniehoopte en mouwlengte horen. Je kunt hiervan een puntenwolk maken.

		Statistisch onderzoek 1947																				
		De Bijenkorf - Amsterdam																				
		5001 vrouwen met leeftijd > 18 jr.																				
	mouwlengte in cm	kniehoopte in cm																				
		34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
71																						1
70																				1		1
69																						
68																						
67																						
66																						
65																						
64																						
63																						
62																						
61																						
60																						
59																						
58																						
57																						
56																						
55																						
54																						
53																						
52																						
51																						
50																						
49																						
		1	3	19	34	76	178	313	495	632	738	675	665	466	343	195	88	45	18	12	2	3

Figuur 10

Welke mate van samenhang verwacht je op grond van deze resultaten tussen *kniehoopte* (cm) en het *mouwlengte* (cm)?

Opgave 9

Bekijk de tabel met de correlatiecoëfficiënten van een aantal variabelen van de lengte van een vrouw. Gebruik nu de vuistregels uit [Voorbeeld 2](#).

	gewicht	bovenwijdte	taille	heup	ruglengte	rugbreedte	vuistomvang	kniehoopte	voetlengte
lengte	0,2124	-0,0779	-0,1578	-0,0107	0,5933	0,0647	0,2668	0,8263	0,6737

Tabel 2

- Welke variabelen hebben een sterke samenhang met lengte?
- Welke variabelen hebben een matige samenhang met lengte?
- Welke variabelen hebben een zwakke samenhang met lengte?

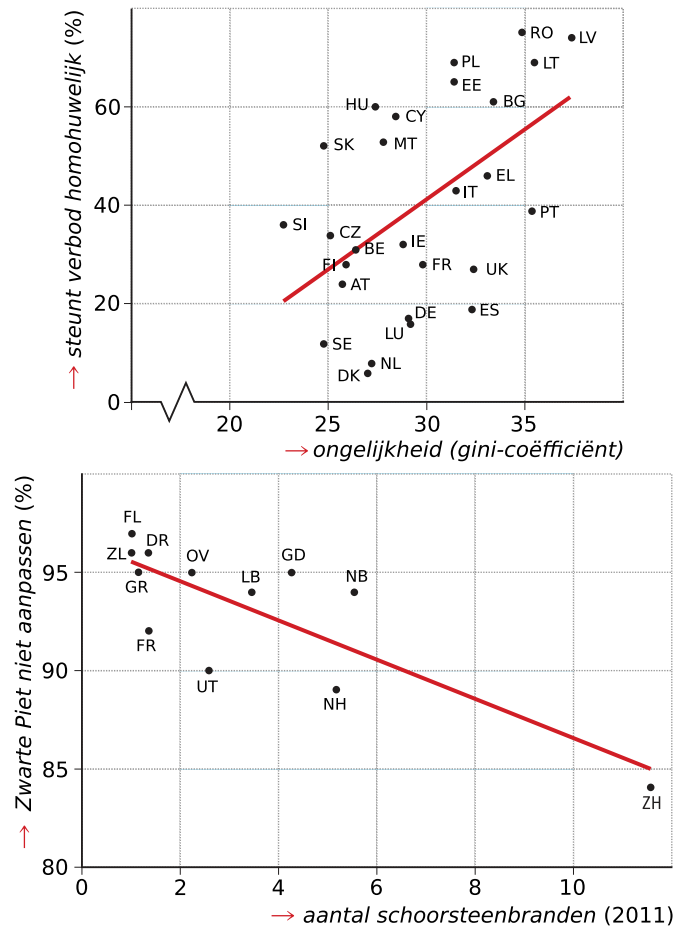
Opgave 10

Soms is er wel een samenhang tussen twee statistische variabelen, maar kun je niet spreken van een causale (oorzaak-gevolg) relatie. Leg uit dat er een samenhang is tussen ijsverkoop en de verkoop van zonnebrillen, maar dat er geen oorzakelijk verband is tussen deze twee.

Opgave 11

Bekijk deze puntenwolken.

- Wat stelt een punt voor in de eerste figuur? En in de tweede figuur?
- Tussen welke statistische variabelen wordt in de figuren gezocht naar samenhang?
- Beschrijf de statistische samenhang tussen de variabelen op grond van de puntenwolken en de getekende trendlijn.
- Is er ook sprake van een causaal verband? Geef dan een mogelijke verklaring bij de puntenwolk.



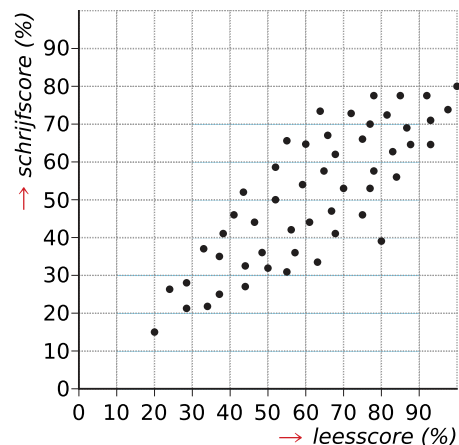
Figuur 11

Opgave 12

Een basisschool heeft een leestest en schrijftest Nederlands afgenomen bij de leerlingen in groep acht. De resultaten zijn verwerkt in een puntenwolk.

Er lijkt een verband te zijn tussen de schrijfscore S en de leesscore L .

- Stel een formule op voor de trendlijn die het verband tussen S en L weergeeft.
- Geef met behulp van de formule uit a een schatting van de schrijfscore bij een leesscore van 80%.
- Geef met behulp van de formule uit a een schatting van de leesscore bij een schrijfscore van 10%.

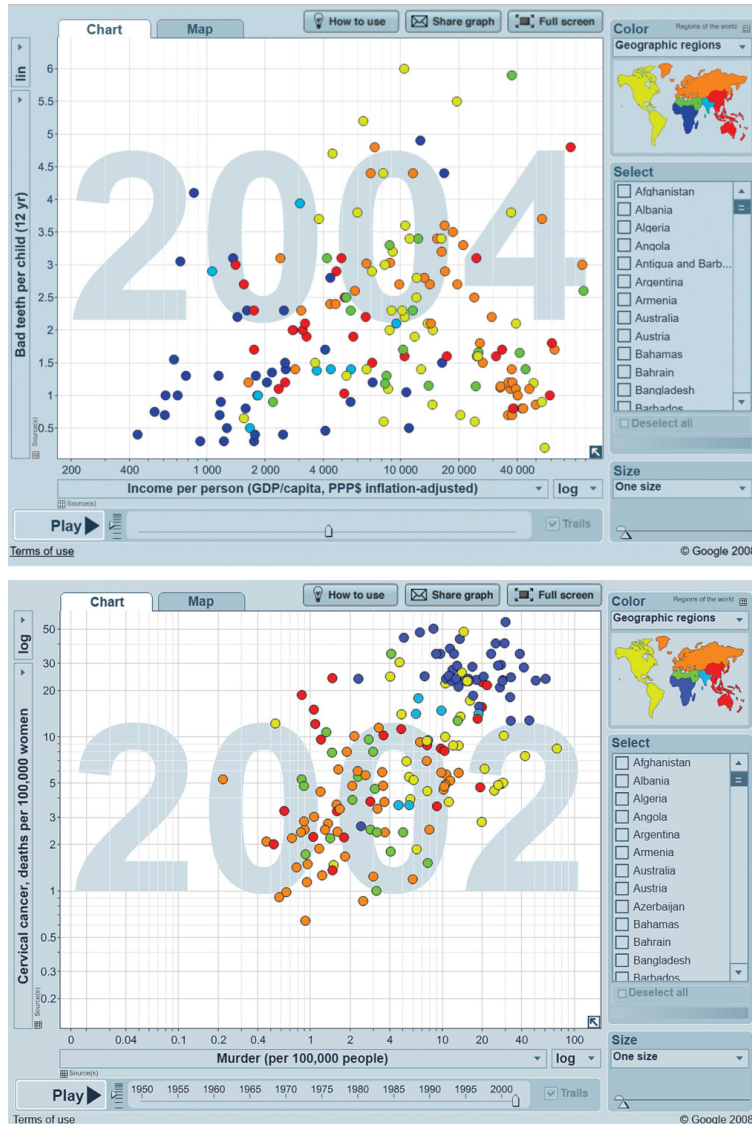


Figuur 12

Toepassen

Opgave 13: Gapminder

Met Gapminder (zie <http://www.gapminder.org/world>) kun je puntenwolken bekijken. Je kunt de statistische variabelen zelf kiezen en door op play te drukken kun je het verloop in de tijd zien. Let wel op de schaalverdelingen: lin = lineair; log = logaritmisch. Wil je de lineaire samenhang zien, dan moeten beide schalen op lin staan.

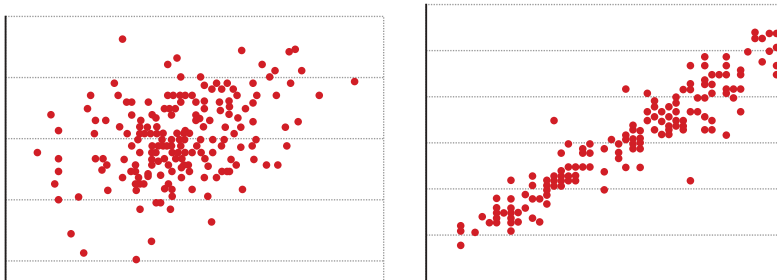


Figuur 13

- Door de vorm van de puntenwolk kun je een uitspraak doen over de lineaire samenhang. Bekijk de afbeeldingen en doe een uitspraak over de lineaire samenhang.
- Bij welke puntenwolk is er wel samenhang, maar geen lineaire samenhang?

Opgave 14: 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:

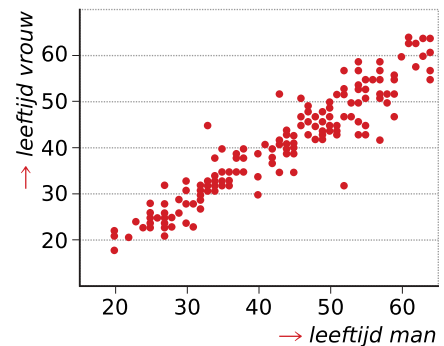


Figuur 14

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.

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.



Figuur 15

- 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 3

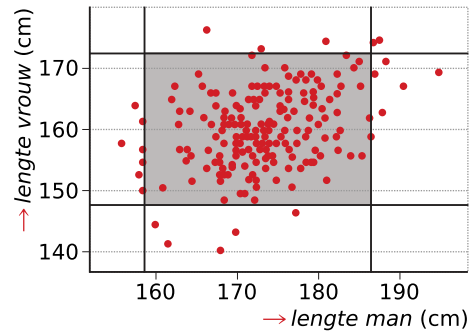
Ervan uitgaande dat de lengtes en de leeftijden van de huwelijkspartners nagenoeg normaal verdeeld zijn, is met behulp van deze gegevens uit te rekenen dat 95% van de lengtes van de mannen tussen de 159,2 cm en 186,8 cm zal liggen.

- 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)

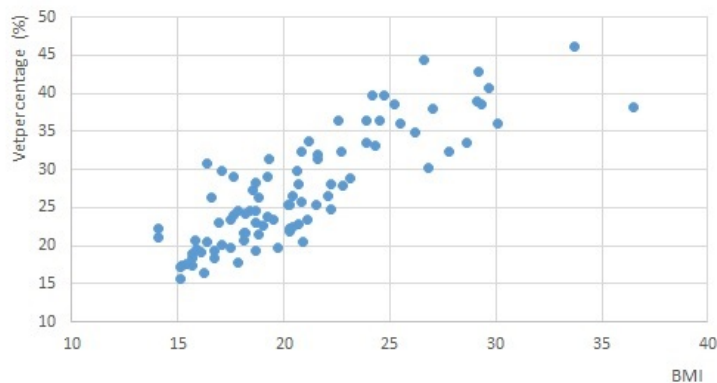


Figuur 16

Testen

Opgave 15

Bekijk de puntenwolk BMI-Vetpercentage. Daarin zijn de resultaten weergegeven van een onderzoek onder 90 jongeren. BMI is een getal dat samenhangt met lengte en gewicht, vetpercentage is het percentage van het lichaamsgewicht dat bestaat uit vet.



Figuur 17

- Is er een statistische samenhang? Geef een schatting van de correlatiecoëfficiënt.
- Is er een oorzakelijk verband?
- Bij deze puntenwolk kun je een trendlijn tekenen die door (15,17) en (30,42) gaat. Welke formule hoort bij deze trendlijn?
- Voorspel met behulp van de gevonden trendlijn het vetpercentage van iemand met een BMI van 21.

Practicum

Met deze practica leer je hoe je de **de trendlijn** met de grafische rekenmachine tekent en berekent.

- [Trendlijn, correlatie en de TI84](#)
- [Trendlijn, correlatie en de TIInspire](#)
- [Trendlijn, correlatie en de Casio](#)
- [Trendlijn, correlatie en de HPprime](#)
- [Trendlijn, correlatie en de NumWorks](#)

Met het volgende practicum kun je zien hoe je **de trendlijn en de correlatiecoëfficiënt in Excel** berekent. Dat is handig als je een grote set gegevens hebt. Je treft er ook in aan hoe je de **regressie-lijn**, 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!



© 2024

Deze paragraaf is een onderdeel van het Math4All wiskundemateriaal.

Math4All stelt het op prijs als onvolkomenheden in het materiaal worden gemeld en ideeën voor verbeteringen in de content of dienstverlening kenbaar worden gemaakt. Klik op  in de marge bij de betreffende opgave. Uw mailprogramma wordt dan geopend waarbij het emailadres en onderwerp al zijn ingevuld. U hoeft alleen uw opmerkingen nog maar in te voeren.

Email: f.spijkers@math4all.nl

Met de Math4All Foliostraat kunnen complete readers worden samengesteld en toetsen worden gegenereerd. Docenten kunnen bij a.f.otten@math4all.nl een gratis inlog voor de maatwerkdienst aanvragen.
