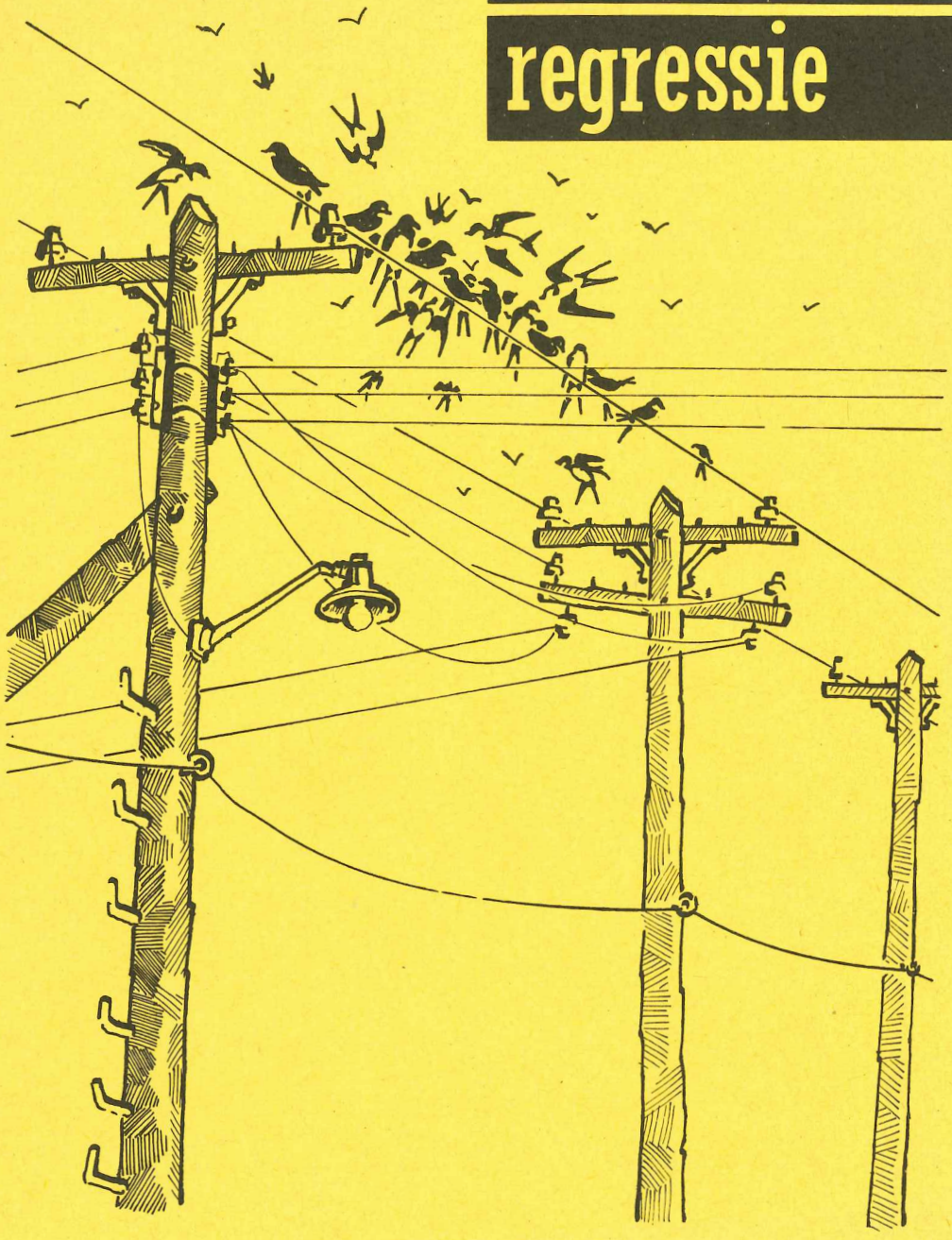
de Wageningse Methode

correlatie en regressie



inhoudsopgave

de Wageningse Methode

Auteurs: Leon v.d. Broek, Jacques Fellinger, Wim Kremers, Jan Smit, Gerard Stroomer, Stef Tijs.

Verkoopadres: Meijer & Siegers bv, Postbus 105, 6860 AC Oosterbeek.

Illustraties: Ad v.d. Broek.

1 wiskunde en muziek

Van de 102 leerlingen in de hoogste klas van een school hebben er 75 wiskunde gekozen en 27 niet. Er is dus een indeling in twee groepen: W (wiskunde) en NW (niet wiskunde).
Ook is bekend dat 34 van deze 102 leerlingen een muziekinstrument bespelen. Zij vormen de groep M. De rest, NM, speelt niet.

Nu kan de vraag gesteld worden: Is er bij deze leerlingen een verband tussen het kiezen van wiskunde en het bespelen van een muziekinstrument?

Om daar iets over te kunnen zeggen, zouden we de getallen a, b, c en d in de vakjes van de 2 x 2 - tabel moeten weten. Die getallen weten we echter niet, we kennen alleen de 'randgetallen'.
Bij de gegeven randgetallen zijn er nog allerlei mogelijkheden voor de getallen in de vier vakjes.

	W	NW	
M	a	b	34
NM	c	d	68
	75	27	102

Een paar voorbeelden:

Geen verbanden

Eén derde van alle leerlingen bespeelt een instrument. Neem even aan dat er geen verband is tussen (aanleg voor) wiskunde en muziek.

Hoeveel muzikanten verwacht je dan onder de 75 leerlingen met wiskunde? Schrijf dit getal op de juiste plaats in de tabel hiernaast.

	W	NW	
M			34
NM			68
	75	27	102

Vul ook de andere drie vakjes in.

Wat merk je op als je de verhoudingen a:c en b:d met elkaar vergelijkt?

Als de werkelijke aantallen in de vier vakjes van de 2 x 2 - tabel ongeveer gelijk zijn aan de getallen die jij invulde, dan is er dus geen aanwijzing voor een verband tussen het kiezen van wiskunde en het bespelen van een muziekinstrument.

2 wiskunde en muziek

Positief verband

Neem even aan dat er een positief verband is tussen W en M. Je verwacht dan verhoudingsgewijs veel muzikanten onder de leerlingen die wiskunde kozen.

Welke getallen in de tabel zullen dan duidelijk groter zijn dan bij 'geen verband'? Hoe groot kunnen die getallen maximaal zijn?

Schrijf die maxima in de tabel. Vul daarna ook de andere vakjes in.

	W	NW	
M			34
NM			68
	75	27	102

Negatief verband

Neem even aan dat er een negatief verband is tussen W en M.

Welke getallen zijn nu veranderd en hoe zijn ze veranderd ten opzichte van de getallen in de tabel bij 'geen verband'?

Vul de tabel in voor de extreme situatie die nu kan ontstaan.

	W	NW	
M			34
NM			68
	75	27	102

3 wiskunde en muziek

Zwak verband

	W	NW	
M	28	6	34
NM	47	21	68
	75	27	102

Zou je het verband tussen W en M in de tabel positief of negatief noemen? Waarom?

Geef zelf ook een invulling van de tabel waarbij er sprake is van een zwak verband. Zorg er daarbij voor dat het juist in de andere richting wijst dan bij de tabel daarboven.

	W	NW	
M			34
NM			68
	75	27	102

4 positief of negatief verband

WM 1

10.30
11.00
12.00
s de
soli-
une
ion!
nde:
A-
né
25
rt
r

10.00 Teletekst
10.15 Zondag op 2: met W
ner takes all; **10.25** Cham
on the wonderhorse; **10.**
Popeye and son; **11.20** Blu
Peter Omnibus; **11.50** Bos
Cat; **12.15** Boxpops.
13.00 ■□ **Anne of windy po-**
plars - Amerikaanse film uit
1940 met Anne Shirley e.a.
14.25 Darts - het w.k. in Frimley
Green.
15.15 ♫ ♫ ♫
17.10 Music in camera - Tokyo
string quartet speelt Schubert.
18.00 Rugby special
18.55 Ski sunday
19.35 The money programme
20.15 Atlantic Realm (1) -
Island Arks. Driedelige docu-
mentaire over de geschiede-
nis en natuurlijke ontwikkeling
van de Atlantische Oceaan
Voornamelijk

Bij een kijkonderzoek wordt aan n mensen gevraagd of ze de tv programma's A en/of B de afgelopen week gezien hebben.
De resultaten van de enquête worden in een 2 x 2 - tabel weergegeven. (Hierin is NA de groep die programma A niet gezien heeft.)

	A	NA	
B	a	b	
NB	c	d	
			= n

Vul, uitgaande van de getallen a, b, c en d, de randgetallen in.

Wat geeft het randgetal b+d aan? Wat geeft het getal c aan?

Er is een **positief** verband tussen A en B voor wat het kijkgedrag van de ondervraagden betreft als binnen de A-groep de verhouding B : NB hoger ligt dan binnen de NA-groep.

Leid hieruit af:

er is een **positief** verband als $\mathbf{ad - bc > 0}$.

Op dezelfde manier kun je afleiden:

er is een **negatief** verband als $\mathbf{ad - bc < 0}$ en

er is **geen** verband als $\mathbf{ad - bc = 0}$.

Geef bij elk van de tabellen hiernaast aan welk verband er tussen A en B bestaat.

	A	NA
B	183	249
NB	654	453

	A	NA
B	345	678
NB	764	876

Kies twee programma's die in de afgelopen week zijn uitgezonden en waarbij je een positief of juist een negatief verband verwacht.
Onderzoek of je vermoeden in je klas uitkomt.

5 associatie

Als we ad – bc berekend hebben, weten we of er sprake is van geen, een positief of een negatief verband.
Om een maat te krijgen die minder afhangt van de grootte van a, b, c en d, en niet afhangt van de dimensie van deze getallen, wordt ad - bc gedeeld door de wortel uit het product van de vier randgetallen a+c, b+d, c+d en a+b.
Deze maat voor het verband (de associatie, de samenhang) tussen A en B in de 2 x 2 - tabel wordt aangegeven met de letter R.

	A	NA	
B	a	b	a + b
NB	c	d	c + d
	a + c	b + d	n

Formule voor de associatiemaat R

$$R = \frac{ad - bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

Vul bij elke tabel hiernaast het passende woord (geen, positief of negatief) in.

Bereken bij elke tabel ook de associatiemaat R.

	W	NW
M	25	9
NM	50	18

_____ verband

R = _____

	W	NW
M	34	0
NM	41	27

_____ verband

R = _____

	W	NW
M	7	27
NM	68	0

_____ verband

R = _____

	W	NW
M	28	6
NM	47	21

_____ verband

R = _____

6 associatie

Bereken R ook voor de acht 2 x 2 - tabelletjes hiernaast.

2	3
3	2

R = ____

3	2
2	3

R = ____

20	30
30	20

R = ____

21	14
14	21

R = ____

1	4
4	1

R = ____

5	0
0	5

R = ____

0	5
3	0

R = ____

2	1
3	0

R = ____

Wat valt je op bij de eerste vier tabelletjes?

De associatiemaat van de tabel hiernaast is 0. Vul de ontbrekende getallen in.

	7	
20	10	

R = 0

De associatiemaat van de linkertabel is 1, die van de rechtertabel -1. Vul de ontbrekende getallen in.

20	10

R = 1

20	10

R = -1

Als je geen fouten maakte, vond je voor R steeds een uitkomst die tussen -1 en 1 ligt. R kan ook nog gelijk aan -1 of aan 1 zijn.
R is nooit kleiner dan -1 of groter dan 1.

Als R = 1, welke twee getallen (a, b, c of d) zijn dan 0 ? En welke twee als R = -1?

7 onafhankelijkheid en kansen

De 52 kaarten van een kaartspel kun je naar soort (klaveren, schoppen, harten en ruiten) of naar rang (2, 3, ..., heer, aas) in groepen verdelen.

Voorbeelden:

Ha en NHa: harten en niet harten;

Ro en NRo: rode en zwarte kaarten;

Aa en NAa: azen en niet-azen;

Pl en NPl: plaatjes en andere kaarten.

Neem je twee van die indelingen, dan heb je ook een 2 x 2 - tabel. Van die tabel kun je dan weer de associatiemaat R berekenen.
Even verderop zullen we zien dat R in verband gebracht kan worden met het begrip onafhankelijk uit de kansrekening!

Zet bij elke tabel de juiste getallen in de vakjes en langs de rand en bereken daarna de associatiemaat R.

	Ro	NRo	
Ha	13		13
NHa			
	26	26	52

R = _____

	Pl	NPl	
Aa			
NAa			
			52

R = _____

	Ro	NRo	
Pl			
NPl			
	26	26	52

R = _____

	Ha	NHa	
Aa			
NAa			
			52

R = _____

Wat valt je op bij de tabellen met R = 0, als je kijkt of er naar rang of naar soort verdeeld is?

In het kaartspel zijn rang en soort onafhankelijk. Als iemand uit een volledig spel een kaart trekt en vertelt dat het een aas is, dan weet je nog niets over de kleur.
Als één kaart ontbreekt, krijg je bijvoorbeeld bij "rood" en "plaatje" al een positief of negatief verband.

Hoe is dat verband, positief of negatief, als harten 7 ontbreekt?

onafhankelijkheid en kansen

Iemand trekt een kaart uit een volledig spel.

Hoe groot is P(Ha), de kans op een harten? Hoe groot is P(Pl), de kans op een plaatje?
Bereken ook P(Ha èn Pl), de kans op een harten plaatje.

Welk verband bestaat er tussen deze drie kansen?

Vul langs de rand van de tabel de drie nog ontbrekende kansen in. Vul ook in de vakjes de passende kansen in.

	Ha	NHa	
Pl			P(Pl) = 4/13
NPl			9/13
	P(Ha) =		

In de kansrekening heten twee gebeurtenissen A en B **onafhankelijk** als
P(A èn B) = P(A)·P(B).

Een school heeft 1000 leerlingen, 700 in de onderbouw en 300 in de bovenbouw. Van deze 1000 leerlingen zijn er 100 linkshandig, de rest is rechtshandig.

Neem aan dat er geen verband is tussen deze twee indelingen. Vul in de vakjes de te verwachten aantallen in.

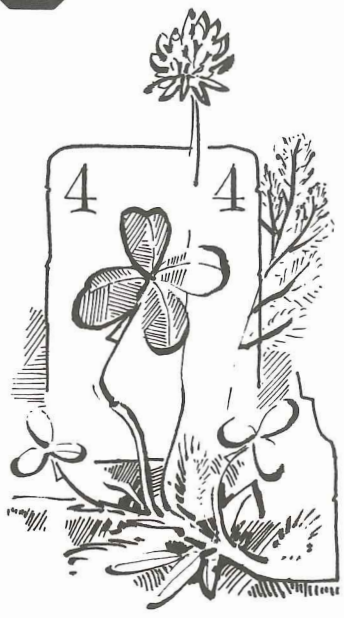
	Lh	Rh	
Ob			700
Bb			300
	100	900	1000

Aselect wordt één leerling gekozen. Vul in het schema hiernaast de passende kansen in. Neem daarbij aan dat de verdeling precies is zoals je bij 'geen verband' verwacht.

De overeenkomst tussen de twee tabellen is duidelijk!

	Lh	Rh	
Ob			P(Ob) =
Bb			
			1

9 toevallige verbanden



Helma trekt zes kaarten uit een volledig spel en let op het aantal harten en het aantal plaatjes.

Stel dat zij trekt: harten 7, ruiten boer, schoppen vrouw, schoppen 4, klaveren 8 en klaveren 9.

Vul de 2 x 2 - tabel in en bereken R.

Hoe groot zou R geweest zijn als zij twee harten plaatjes, twee andere plaatjes en twee niet-harten niet-plaatjes getrokken had?

Geef een tabel waarbij R = 1. Geef ook een tabel waarbij R = -1 en een tabel waarbij R = 0.

Zoals je ziet kan R bij zo'n trekking van zes kaarten allerlei waarden aannemen, waarden die van het toeval afhangen.

Soms kan R zelfs niet berekend worden. Geef een verdeling waarbij R onbepaald is.

Hoe groot is R voor het volledige kaartspel?

	Ha	Nha
Pl		
NPl		

R = ____

R = ____

R =1

R = -1

R = 0

R is onbepaald

Uit het bovenstaande blijkt dat een steekproef toevallig de indruk kan wekken dat er een verband is tussen twee dingen, terwijl dat verband er in de populatie zelf helemaal niet is. Vooral bij kleine steekproeven spelen zulke toevalseffecten een rol.
Hoe groter de som n van de getallen a, b, c en d in de vier vakjes, hoe kleiner de toevalseffecten.

Vuistregel:

Als er geen verband is, krijg je maar zelden een R die meer dan $2/\sqrt{n}$ van 0 afwijkt en bijna nooit een R die meer dan $3/\sqrt{n}$ van 0 afwijkt. Als je toch zo'n waarde vindt, mag je aannemen dat er verband is (of dat de steekproef niet aselect is!).

10 aspirine en hartinfarct

In januari 1987 stond dit bericht in de krant. Lees het even.

Aspirinegebruik verlaagt kans op hartinfarct met 45%

ROTTERDAM, 28 jan — Het om de dag innemen van een aspirientje verlaagt het risico op een hartinfarct bijna tot de helft. Dat is de conclusie van een grootschalig Amerikaans onderzoek dat vandaag in het medische vaktijdschrift *New England Journal of Medicine* is gepubliceerd.

Bij het onderzoek, dat in 1982 begon, waren ruim 22.000 gezonde, mannelijke artsen betrokken (mannen hebben een ongeveer 8 maal zo hoog risico op een hartinfarct als vrouwen). De helft daarvan gebruikte om de dag 300 milligram aspirine, wat ongeveer gelijkstaat aan een „gewoon" aspirientje. De andere helft slikte een placebo („fopmiddel"). Van de aspirine-slikkers kregen 104 een hartinfarct, van de placebo-slikkers 189.

Door het aspirinegebruik werd het risico op een hartinfarct dus met ongeveer 45 procent verlaagd. Dat dit grote verschil aan toeval te wijten zou zijn is praktisch uitgesloten, vanwege het grote aantal mensen dat aan de studie meewerkte. Deze werking van het aspirine berust waarschijnlijk op het tegengaan van bloedklontering op de vaatwanden.

Hart- en vaatziekten vormen in de meeste Westerse landen de belangrijkste doodsoorzaak. Men verwacht dan ook dat het onderzoeksresultaat grote gevolgen zal kunnen hebben voor de preventie van hartinfarcten.

Verheugt: „Het is pas de eerste studie. Binnenkort verschijnt er een Engelse studie met een soortgelijke resultaat, maar met een minder dramatisch verschil".

Waarom werkt men bij zo'n onderzoek met foppillen?

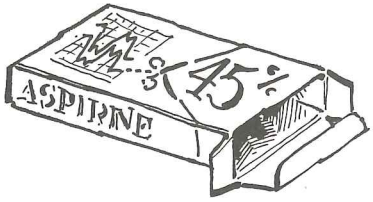
Geef de resultaten in een 2 x 2 - tabel en bereken de associatiemaat R.

	Aspirine	Placebo	
Hartinf			
Geen inf			
	11 000	11 000	22 000

R = _____

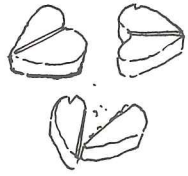
In de steekproef zijn de aantallen hartinfarcten bij "Aspirine" en "Placebo" nogal verschillend.
Wat denk je, is dit verschil toevallig? Gebruik de vuistregel van de vorige bladzijde.

11 aspirine en hartinfarct



Er is nog een tweede mogelijkheid om te bekijken of het verschil toevallig zou kunnen zijn.

Als er geen verband zou zijn, dus wanneer aspirinegebruik geen invloed zou hebben, dan kunnen we de groep van 293 hartinfarcten beschouwen als een aselecte steekproef uit de populatie van 22 000 artsen.



Natuurlijk is dit een steekproef zonder terugleggen. Waarom mogen we de steekproef toch beschouwen als een met terugleggen?

Ongeveer hoeveel procent van de steekproef zal, als er geen verband zou zijn, uit aspirineslikkers bestaan?

Het probleem dat we hier onderzoeken kunnen we dus opvatten als het 293 keer opgooien van een munt. Het aantal keren kop, noem dat X, kan ook meer of minder afwijken van de helft!

Hoe is X verdeeld en wat zijn de waarden van de parameters van die verdeling?

Bereken E(X) en SD(X).

Hoeveel keer de SD wijkt 104 van E(X) af, m.a.w. wat is de z-waarde van 104 ?

Hoe groot is de kans op deze afwijking of een afwijking die nog groter is?

Wat denk je, is het verschil tussen de aantallen hartinfarcten toevallig?

Hoeveel keer $1/\sqrt{22\,000}$ is de associatiegraad R van nul verwijderd?

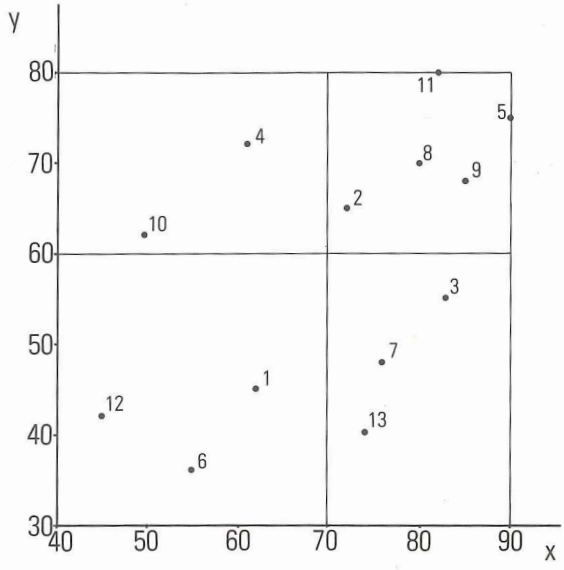
12 van 2 bij 2 naar veel bij veel

Een intelligentietest bestaat uit twee delen. Deel A test taalvaardigheid en deel B technisch inzicht.

Hieronder zie je de resultaten van 13 personen.

	i	1	2	3	4	5	6	7	8	9	10	11	12	13
A	x_i	62	72	83	61	90	55	76	80	85	53	82	45	74
B	y_i	45	65	55	72	75	36	48	70	68	62	80	42	40

De scores zijn ook nog eens als punten weergegeven in het plaatje hiernaast.



In deze **puntenwolk** zie je dat er enig verband is tussen de twee componenten van de test. Personen met een hoge (lage) x hebben meestal ook een tamelijk hoge (lage) y.

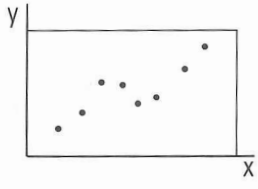
Eén manier om het verband te onderzoeken is er een 2 x 2 - tabel van te maken. We trekken ergens een verticale en een horizontale lijn, bijvoorbeeld bij x = 70 en bij y = 60.
We hebben dan twee indelingen van de 13 personen.

	$x < 70$	$x \geq 70$	
$y \geq 60$			
$y < 60$			
			13

R = ____

Vul de 2 x 2 - tabel in en bereken de associatiemaat R.

Wat is het zwakke punt in de wijze waarop de associatiemaat R is vastgelegd?



Welk verband (positief, negatief of geen) is er volgens jou tussen x en y bij de puntenwolk hiernaast?

_____ verband

R = _____

Breng met een horizontale lijn en een verticale lijn een verdeling in de puntenwolk aan, zó dat in elk vakje 2 punten komen. Bereken nu R.

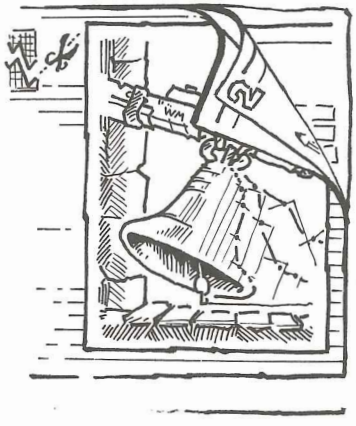
Wat is je commentaar?

13 van 2 bij 2 naar veel bij veel

Het zal duidelijk zijn dat we er niet alleen maar op moeten letten in welke hoek de punten liggen. We zullen er ook naar moeten kijken hoe ver ze in de hoeken liggen, hoe ze ten opzichte van elkaar liggen. Spreiding speelt immers ook een rol.

We bekijken eerst de spreiding in de x- en de y-waarden apart.
Als maat voor de spreiding gebruiken we de standaardafwijking. De standaardafwijking SD is de wortel uit de variantie.

In boekje 52 - normale verdeling zagen we op bladzijde 2 de volgende twee formules voor de variantie van de getallen $x_1, x_2, ..., x_n$ van een databestand:



Eerste formule voor de variantie:

$$\text{Var}(x_1, x_2, ..., x_n) = 1/n \cdot ((x_1-\bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n-\bar{x})^2).$$

In woorden: de variantie is de gemiddelde kwadratische afwijking.

Voorbeeld:
het gemiddelde van 2, 6, 6 en 8 is 5,5 en dus is de variantie gelijk aan:

$1/4 \cdot (3{,}5^2 + 0{,}5^2 + 0{,}5^2 + 2{,}5^2) = 1/4 \cdot (12{,}25 + 0{,}25 + 0{,}25 + 6{,}25) = 4{,}75.$ $\quad SD(x)=\sqrt{4{,}75} = 2{,}18.$

Tweede formule voor de variantie:

$$\text{Var}(x_1, x_2, ..., x_n) = 1/n \cdot ((x_1)^2 + (x_2)^2 + ... + (x_n)^2) - \bar{x}^2.$$

In woorden: de variantie is het gemiddelde van de kwadraten min het kwadraat van het gemiddelde.

Voorbeeld:
het gemiddelde van 2, 6, 6 en 8 is 5,5 en dus is de variantie gelijk aan:

$1/4 \cdot (2^2 + 6^2 + 6^2 + 8^2) - 5{,}5^2 = 1/4 \cdot (4 + 36 + 36 + 64) - 30{,}25 = 35 - 30{,}25 = 4{,}75.$ $\quad SD(x)=\sqrt{4{,}75} = 2{,}18.$

14 van 2 bij 2 naar veel bij veel

De resultaten van de intelligentietest waren:

	i	1	2	3	4	5	6	7	8	9	10	11	12	13
A	x_i	62	72	83	61	90	55	76	80	85	53	82	45	74
B	y_i	45	65	55	72	75	36	48	70	68	62	80	42	40

Bereken het gemiddelde en de standaarddeviatie van de x-waarden. Gebruik je rekenmachientje.

Bereken ook het gemiddelde en de standaarddeviatie van de y-waarden.

Nu gaan we nog de afwijkingen van de x- en y-waarden combineren door te letten op de producten $(x_i - \bar{x})(y_i - \bar{y})$.

Met behulp van deze producten wordt de zogenaamde **covariantie** van de punten (x_1,y_1), ...,(x_n,y_n) berekend.

Eerste formule voor de covariantie

$$Cov(x,y) = 1/n \cdot ((x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + ... + (x_n - \bar{x})(y_n - \bar{y}))$$

In woorden: de covariantie is het gemiddelde product van de afwijkingen.

Tweede formule voor de covariantie

$$Cov(x,y) = 1/n \cdot (x_1 \cdot y_1 + x_2 \cdot y_2 + ... + x_n \cdot y_n) - \bar{x} \cdot \bar{y}$$

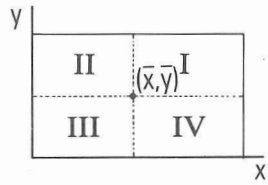
In woorden: de covariantie is het gemiddelde van de producten min het product van de gemiddelden.

Welk voordeel heeft de tweede formule boven de eerste?

Waar zijn we zoiets al eens eerder tegengekomen?

15 van 2 bij 2 naar veel bij veel

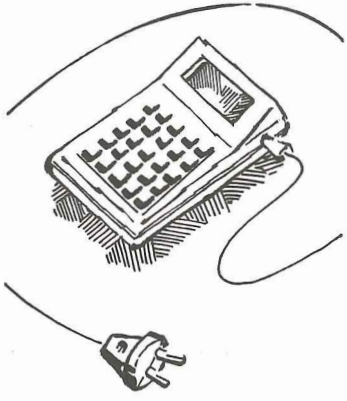
Hiernaast is bij een puntenwolk ook het punt $(\bar{x},\bar{y})$ aangegeven. De horizontale en de verticale stippellijn door $(\bar{x},\bar{y})$ verdelen het gebied waarin de puntenwolk ligt in vier gebieden.



In welke gebieden liggen de punten die een positieve bijdrage leveren aan de covariantie? Aan welke formule kun je dat het best zien?

In welke gebieden liggen de punten die een negatieve bijdrage leveren aan de covariantie?

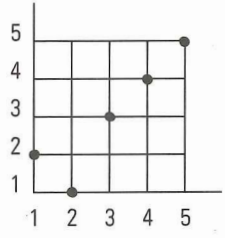
Welke bijdrage leveren de punten die op een stippellijn liggen aan de covariantie?



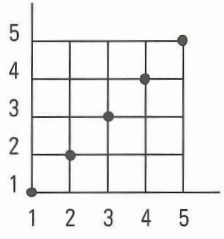
Wat denk je, is de covariantie bij de punten van de intelligentietest op blz. 12 positief óf negatief?

Bereken de covariantie van de punten (x_i,y_i) uit de intelligentietest. Gebruik je rekenmachientje.

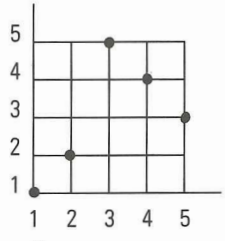
Bereken bij elk van de vier puntenwolken hiernaast de covariantie.



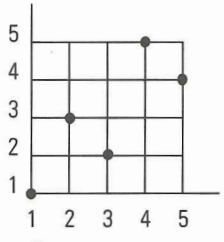
Cov = _____



Cov = _____



Cov = _____



Cov = _____

16 van 2 bij 2 naar veel bij veel

De covariantie is een maat voor de samenhang tussen de x- en y-waarden, maar nog niet zo'n beste!

Wat gebeurt er met Cov(x,y) als alle x-waarden met 10 vermenigvuldigd worden? En als alle y-waarden door 12 gedeeld worden?

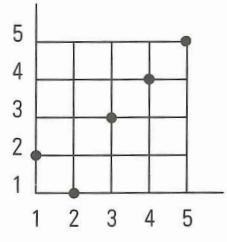
Een maat voor de samenhang die niet afhangt van de eenheden waarin x en y gemeten worden, is de **correlatiecoëfficiënt.**
Die vind je door de covariantie te delen door de standaarddeviatie van x en die van y.
De correlatiecoëfficiënt wordt meestal aangegeven met de letter r.

Eerste formule voor de correlatiecoëfficiënt

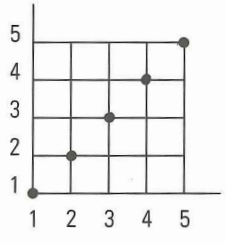
$$r = \frac{Cov(x,y)}{SD(x) \cdot SD(y)}$$

Bereken de correlatiecoëfficiënt r van de dertien punten van de intelligentietest.

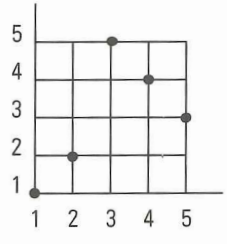
Bereken bij elk van de vier puntenwolken hiernaast de correlatiecoëfficiënt r. Op de vorige bladzijde heb je de covariantie al berekend!



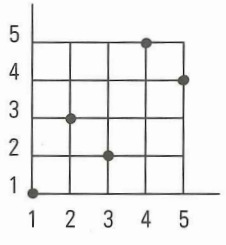
r = _____



r = _____



r = _____



r = _____

17 van 2 bij 2 naar veel bij veel

Laat zien dat de formule voor r ook zó geschreven kan worden:

$$r = \tfrac{1}{n} \cdot \left(\frac{x_1-\bar{x}}{SD(x)} \cdot \frac{y_1-\bar{y}}{SD(y)} + \ldots + \frac{x_n-\bar{x}}{SD(x)} \cdot \frac{y_n-\bar{y}}{SD(y)} \right).$$

Merk op:

$\frac{x_i - \bar{x}}{SD(x)}$ is de z-waarde van x_i, de afwijking van x_i van $\bar{x}$ gemeten in standaardafwijkingen.

Gevolg:

Tweede formule voor de correlatiecoëfficiënt

$$r = \tfrac{1}{n} \cdot (z(x_1) \cdot z(y_1) + z(x_2) \cdot z(y_2) + \ldots + z(x_n) \cdot z(y_n)),$$

waarbij $z(x_i) = \frac{x_i - \bar{x}}{SD(x)}$ de z-waarde van x_i is.

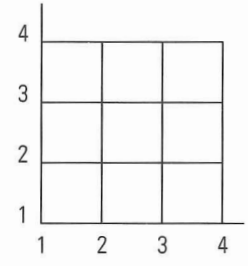
In woorden: de correlatiecoëfficiënt is het gemiddelde van de producten van de z-waarden van x_i en y_i.

Op een sportdag blijken Harrie, Tieme, Ton en Walter de beste springers. Bij het verspringen bezetten ze respectievelijk de plaatsen 1, 2, 3 en 4, bij het hoogspringen de plaatsen 2, 1, 3 en 4.

In de tabel vind je deze resultaten nog eens terug.

	x_i	y_i
Harrie	1	2
Tieme	2	1
Ton	3	3
Walter	4	4

Geef de punten (x_i, y_i) in het rooster aan. Schrijf bij elke stip de eerste letter van de naam van de bijbehorende springer.



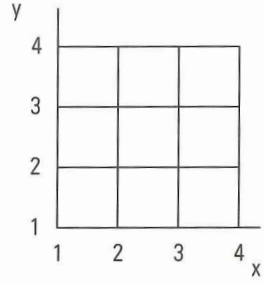
Wat denk je, is het verband tussen x en y positief of negatief?

Bereken de correlatiecoëfficiënt r van deze vier punten.

18 van 2 bij 2 naar veel bij veel

We nemen bij het hoogspringen een andere uitslag:

	x_i	y_i
Harrie	1	4
Tieme	2	3
Ton	3	1
Walter	4	2



Teken het hierbij behorende puntendiagram.

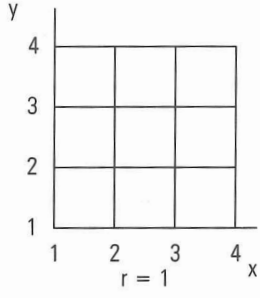
Wat is nu het verband tussen x en y, positief of negatief?

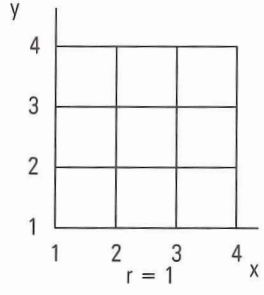
We houden steeds dezelfde uitslag bij het verspringen.

Op hoeveel verschillende manieren kunnen er rangnummers voor het hoogspringen aan gekoppeld worden?

Er is één manier waarbij r = 1. Teken het puntendiagram dat daar bij hoort.

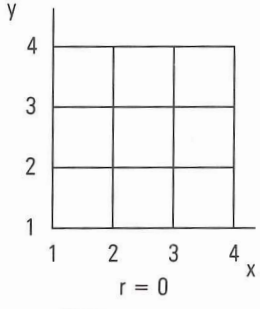
Er is één manier waarbij r = -1. Teken het puntendiagram dat daar bij hoort.

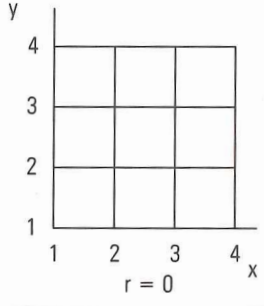




Er zijn twee manieren waarbij r = 0. Teken de twee diagrammen die daar bij horen.

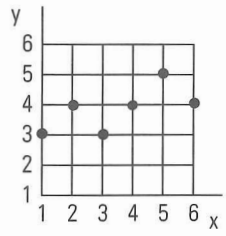
Als het je niet lukt, kijk dan nog eens naar de eerste formule voor de covariantie. En niet te snel opgeven!



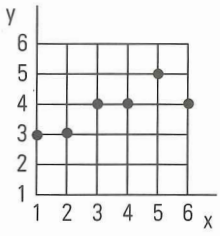


19 van 2 bij 2 naar veel bij veel

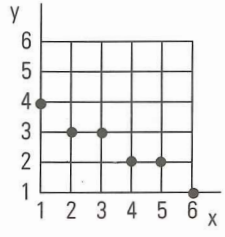
Bereken bij elk van de zes puntenwolken hiernaast de correlatiecoëfficiënt r.



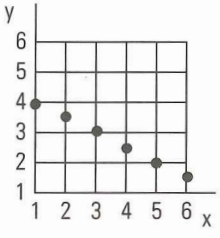
r = _____



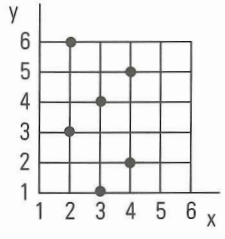
r = _____



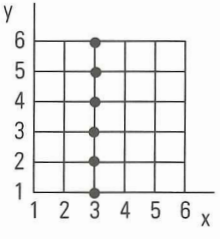
r = _____



r = _____



r = _____



r = _____

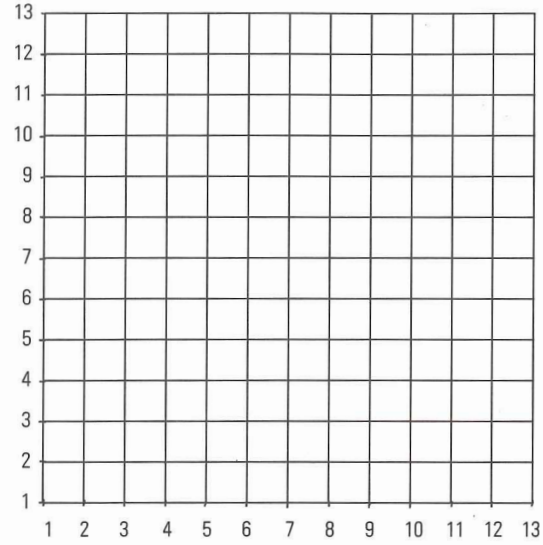
Bij het berekenen van r vind je altijd een uitkomst die tussen -1 en +1 ligt.
In extra sterk komen we daar op terug.

De correlatiecoëfficiënt is alleen dan 1 als de punten op een rechte lijn liggen met positieve richtingscoëfficiënt.
De correlatiecoëfficiënt is alleen dan -1 als de punten op een rechte lijn liggen met negatieve richtingscoëfficiënt.

20 rangnummers

Ook bij de intelligentietest (blz. 12) kunnen we de proefpersonen ordenen naar hun scores op onderdeel A en naar hun scores op onderdeel B.
Proefpersoon 12 had voor taalvaardigheid de laagste score. Die persoon geven we bij onderdeel A rangnummer $a_1 = 1$.
Proefpersoon 6 had voor technisch inzicht de laagste score. Die persoon krijgt bij onderdeel B rangnummer $b_1 = 1$, enzovoort.

	i	1	2	3	4	5	6	7	8	9	10	11	12	13
A	x_i	62	72	83	61	90	55	76	80	85	53	82	45	74
	a_i	5	6	11	4	13	3	8	9	12	2	10	1	7
B	y_i	45	65	55	72	75	36	48	70	68	62	80	42	40
	b_i	4	8	6	11	12	1	5	10	9	7	13	3	2



Teken hiernaast de wolk van de punten (a_i,b_i).
Schrijf bij elk van deze punten het nummer i van de proefpersoon.

Vergelijk deze puntenwolk met die van de punten (x_i,y_i) op blz. 12. Is er veel veranderd?

Bereken $\bar{a}$, $\bar{b}$, Var(a) en Var(b).

Bereken Cov(a,b) en de correlatiecoëfficiënt van de rangnummers r(a,b).

Welk voordeel heeft het berekenen van de correlatiecoëfficiënt van de rangnummers r(a,b) in plaats van de correlatiecoëfficiënt van de testresultaten r(x,y)?

Ga na dat de correlatiecoëfficiënt van de rangnummers r(a,b) hier weinig verschilt van de correlatiecoëfficiënt van de testresultaten r(x,y).
Overigens: dit verschil kan in sommige situaties wel aanzienlijk zijn!

21 rangnummers

De correlatiecoëfficiënt van de rangnummers noemt men wel de rangcorrelatiecoëfficiënt van Spearman naar de psycholoog Charles Spearman. Hij propageerde deze methode in het begin van deze eeuw.

In de psychologie komt het vaak voor dat men dingen kan ordenen, naar voorkeur bijvoorbeeld, maar niet met een getal kan aangeven hoe groot die voorkeur is. Het werken met rangnummers biedt dan uitkomst.

Als we n punten (x_i,y_i) hebben en alle x-waarden en alle y-waarden verschillend zijn, dan kunnen we rangnummers a_i van 1 tot en met n aan de x-waarden toekennen en rangnummers b_i, ook van 1 tot en met n, aan de y-waarden.

Bereken $\bar{a}$ en $\bar{b}$.

Met wat rekenwerk vind je $\text{Var}(a) = \frac{1}{12} \cdot (n^2-1)$ en kan de volgende formule worden afgeleid:

---

correlatiecoëfficiënt van de rangnummers:

$$r(a,b) = 1 - \frac{6}{n(n^2-1)} \cdot ((a_1-b_1)^2 + (a_2-b_2)^2 + \ldots + (a_n-b_n)^2)$$

Aan deze formule kun je eenvoudig zien aan welke voorwaarde de twee volgorden moeten voldoen wil r(a,b) gelijk aan 1 zijn. Welke voorwaarde is dat?

Bereken met deze formule nog eens de rangcorrelatiecoëfficiënt r(a,b) bij de dertien punten van de intelligentietest op de vorige bladzijde.

Bereken bij de tabel van het springen op blz. 17 de correlatiecoëfficiënt r ook nog eens met deze formule.

22 met tabelletjes werken?

In dit boekje moet je nogal eens rekenen met ingewikkelde formules. Als je eenmaal weet welke formule je wilt gebruiken, is het enige probleem rekenfouten te voorkomen. Dat kun je bereiken door het rekenwerk overzichtelijk te maken.
Tabelletjes kunnen je daarbij helpen; tabelletjes die je moet aanpassen aan de formule die je gebruikt.

Voorbeeld:
Bij een wolk van punten (x_i,y_i) wordt gevraagd de correlatiecoëfficiënt r te berekenen.

Neem aan dat je wilt werken met de formule:

$$r = \frac{\frac{1}{n}\cdot(x_1\cdot y_1 + x_2\cdot y_2 + \ldots + x_n\cdot y_n) - \bar{x}\cdot\bar{y}}{SD(x)\cdot SD(y)}$$

Het is dan handig om te werken met een tabelletje als hiernaast.

Vul die tabel verder in en bereken r(x,y).

x_i	y_i	x_i^2	y_i^2	$x_i\cdot y_i$
2,7	4,5			
1,5	1,9			
5,2	9,1			
3,9	6,3			
6,1	11,7			
19,4	33,5			+

$n =$ _____

$\bar{x} =$ _____ $\overline{x^2} =$ _____ $\bar{y} =$ _____ $\overline{y^2} =$ _____

$Var(x) = \overline{x^2} - (\bar{x})^2 =$ _____

$Var(y) = \overline{y^2} - (\bar{y})^2 =$ _____

$SD(x)\cdot SD(y) =$ _____ $\bar{x}\cdot\bar{y} =$ _____

$1/n\cdot(x_i\cdot y_i + \ldots x_n\cdot y_n) =$ _____

$1/n\cdot(x_i\cdot y_i + \ldots x_n\cdot y_n) - \bar{x}\cdot\bar{y} =$ _____

$r(x,y) =$ _____

Natuurlijk had je ook zó te werk kunnen gaan:

1. Op je rekenmachientje voer je de x-waarden in en lees je $\bar{x}$ en SD(x) af. Die uitkomsten noteer je.
2. Op je rekenmachientje voer je de y-waarden in en lees je $\bar{y}$ en SD(y) af. Die uitkomsten noteer je.
3. Je berekent $\bar{x}\cdot\bar{y}$ en noteert de uitkomst.
4. Op je rekenmachientje voer je de x•y-waarden in en lees je het gemiddelde daarvan af. Ook die uitkomst noteer je.
5. Je berekent nu $\frac{1}{n}\cdot(x_1\cdot y_1 + \ldots + x_n\cdot y_n) - \bar{x}\cdot\bar{y}$.
6. Je berekent r(x,y).

Bereken ook op deze tweede manier r(x,y). Hierbij is het van groot belang dat je alle tussenuitkomsten opschrijft. Het is zaak je berekening overzichtelijk te houden!

23 reclame

Bij een bedrijf wordt bijgehouden hoeveel reacties een advertentie oplevert. Een overzicht van de kosten x van een advertentie en het aantal reacties y vind je in de tabel.

kosten	x_i	780	515	342	1257	496	768
reacties	y_i	509	298	176	819	312	520

Geef de punten (x_i,y_i) in een puntenwolk aan.

Is er, zo op het oog, een positief verband tussen de advertentiekosten x en het aantal reacties y ?

Bereken $\bar{x}$, SD(x), $\bar{y}$ en SD(y).

Bereken de som van de producten x_iy_i en Cov(x,y).

Bereken de correlatiecoëfficiënt r(x,y).

Ken de rangnummers a_1, ..., a_6 toe aan de x-waarden en de rangnummers b_1, ..., b_6 aan de y-waarden. Geef deze rangnummers aan in de tabel hiernaast.

rangenr. kosten a_i	5					
rangenr. reacties b_i	4					

Teken de puntenwolk die bij de rangnummers hoort

Is er, zo op het oog, een positief verband tussen de rangnummers van de advertentiekosten a en de rangnummers van het aantal reacties b ?

Bereken de correlatiecoëfficiënt van de rangnummers r(a,b).

Welke correlatiecoëfficiënt geeft volgens jou het verband tussen kosten en aantal reacties het best weer r(x,y) of r(a,b)? Licht je antwoord toe.

24 van de vierde naar de vijfde klas

In het schooljaar 1986-87 begonnen 78 leerlingen uit vwo 4 van het Liemers College in vwo 5 met Wiskunde A. Met Pasen zijn hun rapportcijfers voor wiskunde A vergeleken met hun cijfers voor wiskunde een jaar eerder in vwo 4.

De resultaten zie je hiernaast in een x,y-tabel. Horizontaal x, het cijfer voor wiskunde op het paasrapport in vwo 4;
verticaal y, het cijfer voor wiskunde A een jaar later in vwo 5.

(vijfde klas)

y	2	10	33	22	10	1	
9				1			1
8				3	3	1	7
7		1	5	8	5		19
6		5	20	7			32
5	2	4	6	2	1		15
4			2	1	1		4
	4	5	6	7	8	9	

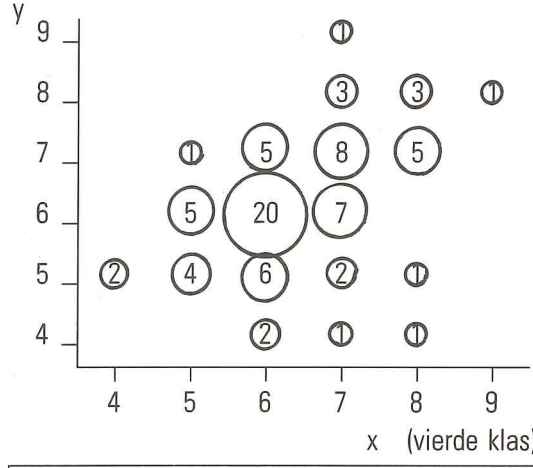
x (vierde klas)

Hoeveel leerlingen hebben hetzelfde cijfer op hun rapport gehouden?

Hoeveel leerlingen zijn er vooruit gegaan? Wat kwam het vaakst voor: vooruit gaan of achteruit gaan?

Wat is het grootste aantal punten dat een leerling achteruit is gegaan?

Hiernaast zie je dezelfde gegevens nog eens in een puntenwolk weergegeven. De grootte van de rondjes komt ongeveer overeen met het aantal punten dat er mee wordt weergegeven.



Bereken $\bar{x}$, Var(x) en SD(x). Gebruik je rekenmachientje.

25 van de vierde naar de vijfde klas

Bij hoeveel leerlingen wijkt de x-waarde meer dan 2SD(x) van $\bar{x}$ af? Is dat in overeenstemming met de vuistregel die hiervoor geldt?

Bereken $\bar{y}$, Var(y) en SD(y).

Ga met een berekening na dat Cov(x,y) = 0,4722.

Bereken de correlatiecoëfficiënt r(x,y).

We kunnen bij elke leerling ook kijken naar het aantal punten d dat hij of zij vooruit is gegaan: d = y − x.

Vul de tabel verder in.

verschil d	2	1	0	-1	-2	-3	-4
frequentie f	2						1
d•f							
d^2•f							

Bereken het gemiddelde verschil d.

Welk verband bestaat er tussen $\bar{d}$, $\bar{x}$ en $\bar{y}$?

Bereken Var(d) en SD(d) met behulp van de gegevens uit de tabel.

Waarom geldt zeker niet: Var(d) = Var(x) − Var(y)?

Waarom geldt ook niet: Var(d) = Var(x) + Var(y)?

Toch mag je een zeker verband verwachten tussen Var(d), de spreiding van de verschillen en de correlatie tussen x en y.

Ga na dat in dit geval geldt:
2 • Cov(x,y) = Var(x) + Var(y) − Var(d).

Neem even aan: alle leerlingen zijn één punt vooruit gegaan.
Hoe groot zijn dan Var(d) en r(x,y)?

van de vierde naar de vijfde klas

Algemeen geldt: hoe meer spreiding in de verschillen, des te kleiner is de correlatie.
Bewijs de volgende formule. Daarin zie je hoe het verband precies is.

$\text{Var}(d) = \text{Var}(y-x) = \text{Var}(y) + \text{Var}(x) - 2 \cdot \text{Cov}(x,y)$.

Bij een verzameling punten (x,y) heeft men berekend:
Var(x) = 169, Var(y) = 225 en Var(y−x) = 134.
Bereken Cov(x,y) en r(x,y).

Uit de formule hierboven kan de volgende formule voor de correlatiecoëfficiënt worden afgeleid:

Derde formule voor de correlatiecoëfficiënt

$$r(x,y) = \frac{\text{Var}(x) + \text{Var}(y) - \text{Var}(d)}{2 \cdot \text{SD}(x) \cdot \text{SD}(y)} \quad \text{waarbij } d = y - x.$$

Hoe leid je, met behulp van de formule boven aan de bladzijde, de derde formule voor de correlatiecoëfficiënt af?

Vijf leerlingen hebben voor twee opvolgende proefwerken voor Wiskunde A de volgende cijfers:

eerste proefwerk x_i	7,3	8,9	4,3	5,9	6,8
tweede proefwerk y_i	6,3	9,1	5,7	6,5	6,0

Bereken Var(x), Var(y), Var(d)en r(x,y).

27 associatie en correlatie

Op een middelbare school is er een leraar die als rapportcijfer altijd een zes of een zeven geeft. Een andere leraar op die school geeft alleen maar zevens en achten.

In de tabel zie je de cijfers van 50 leerlingen die van beide leraren les hebben.

	cijfer x		
y	6	7	
7	18	12	30
8	17	3	20
	35	15	50

Bereken de associatiemaat R (zie blz. 5).

Bereken ook $\bar{x}$, $\bar{y}$ en $\bar{d} = \bar{y} - \bar{x}$.

Bereken Var(x), Var (y) en Var(d).

Bereken de correlatiecoëfficiënt r.

Wat valt je op als je R met r vergelijkt?

Met behulp van de algemene tabel hiernaast kan bewezen worden dat de associatiemaat R en de correlatiecoëfficiënt r altijd gelijk zijn.

Daar komt echter nogal wat rekenwerk bij te pas. Alleen voor de echte liefhebbers!

	cijfer x		
y	x_i	x_2	
y_i	a	b	a + b
y_2	c	d	c + d
	a + c	b + d	n

28 toevallige correlatie

Aan drie toevallige voorbijgangers A, B en C, wordt hun leeftijd x en hun huisnummer y gevraagd. Het resultaat zie je in de tabel hiernaast.

	x	y
A	42	55
B	16	38
C	55	89

Bereken $\bar{x}$, $\bar{y}$ en $\bar{d} = \bar{y} - \bar{x}$.

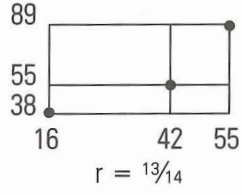
Bereken Var(x), Var(y) en Var(d).

Reken na: r(x,y)= $^{13}/_{14}$ = 0,93... .

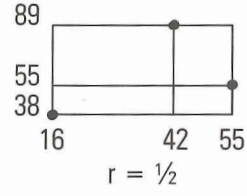
We mogen aannemen dat er in de populatie waaruit de steekproef genomen is geen enkel verband bestaat tussen leeftijd en huisnummer. Hoe kun je de grote waarde van r toch verklaren?

Op hoeveel manieren kun je drie x-waarden, bijvoorbeeld 16, 42 en 55, en drie y-waarden, bijvoorbeeld 38, 55 en 89, combineren tot drie punten (x,y) ?

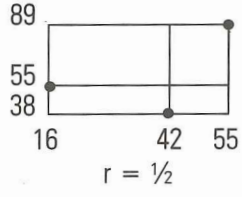
Als er geen verband is tussen x en y, dan zou elk van deze mogelijkheden dezelfde kans hebben. In het overzicht hiernaast zie je alle mogelijkheden. Bij elk van die mogelijkheden is de correlatiecoëfficiënt r berekend.



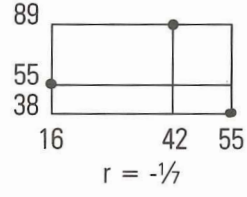
r = $^{13}/_{14}$



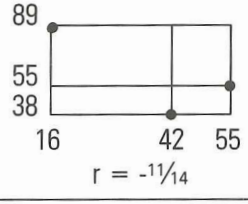
r = ½



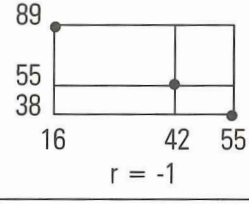
r = ½



r = -$^{1}/_{7}$



r = -$^{11}/_{14}$



r = -1

toevallige correlatie

Bereken $\bar{r}$, het gemiddelde van deze zes r-waarden.

Bereken ook Var(r).

Hoeveel mogelijke punten (x,y) zijn er als we niet bij drie maar bij n personen een x en een y genoteerd hebben? Neem aan dat er geen verband tussen x en y bestaat en dat alle x-waarden en alle y waarden verschillend zijn.

Hoeveel mogelijkheden zijn er als er twee x- en drie y-waarden gelijk zijn?

Algemeen:

In een aselecte steekproef van n personen wordt bij ieder een x en een y genoteerd.

Als er in de populatie waaruit de steekproef komt geen enkel verband is tussen x en y, dan zijn alle manieren waarop de x-waarden aan de y-waarden gekoppeld kunnen worden even waarschijnlijk.

Als alle x- en alle y-waarden verschillend zijn, dan zijn er n! mogelijkheden om die waarden te koppelen tot een wolk van n punten.

Bij elk van deze n! puntenwolken kan de correlatiecoëfficiënt r berekend worden.

Dan geldt:

het gemiddelde van deze correlatiecoëfficiënten is precies 0 en
de standaardafwijking van deze r-waarden is gelijk aan $1/\sqrt{(n-1)}$.

Dit is van groot belang voor de beoordelen van correlatiecoëfficiënten. De volgende opgaven lichten dat toe.

30 toevallige correlatie

Neem aan dat er in werkelijkheid geen verband is tussen de x- en de y-waarden, gevonden in een steekproef van tien personen.

Bereken de standaardafwijking in de bijbehorende r-waarden.

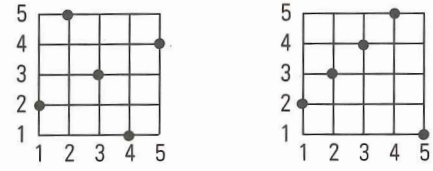
Kun je verklaren waarom een correlatiecoëfficient van 0,35 tussen de x- en de y-waarden best kan optreden maar dat r = 0,7 onwaarschijnlijk is?

Laat zien dat een correlatiecoëfficiënt van 0,35 zeer onwaarschijnlijk is als de steekproef wordt uitgebreid tot honderd personen.

Ga na dat de formule voor de standaardafwijking juist is als n = 2.

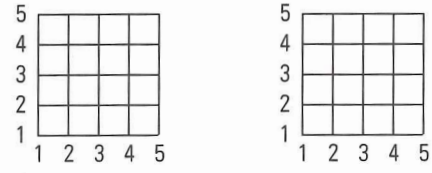
We vervangen de x-waarden x_1, x_2, ..., x_n door de rangnummers 1, 2, ..., n. Met de y-waarden doen we hetzelfde.

Hiernaast zijn bij n = 5 twee mogelijke puntenwolken getekend.



Hoeveel van deze puntenwolken zijn er mogelijk?

Teken een puntenwolk waarvan de correlatiecoëfficiënt 0 is. Ook één waarbij r = 1.

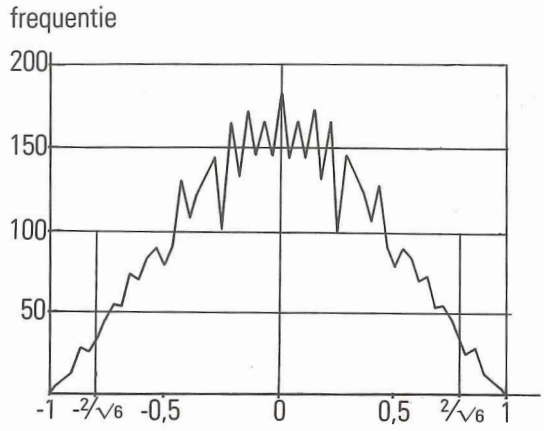


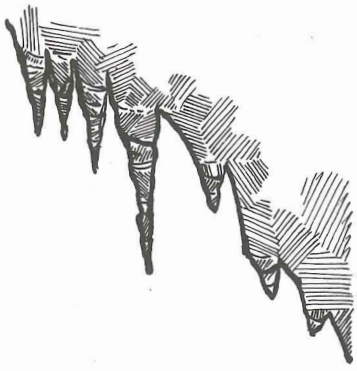
Hoe groot is de kans om toevallig r = 1 te krijgen?

31 toevallige correlatie

Hoeveel verschillende puntenwolken zijn er mogelijk bij n = 7 ?

De verdeling van de r-waarden van al deze puntenwolken kun je aflezen in de frequentiepolygoon hiernaast.





Hoe groot (ongeveer) is bij n = 7 de kans dat je een puntenwolk krijgt waarvan de correlatiecoëfficiënt 0 is?

Hoe groot is de standaardafwijking bij n = 7 ?

Hoe groot is zo ongeveer de kans dat je een r-waarde vindt die meer dan twee keer de standaardafwijking van 0 af ligt?
Tip: Let op de oppervlakte onder de grafiek. Gebruik een stukje vierkantjespapier.

voetbal en toeval

Uit het competitieoverzicht (zie ook boekje 48 - binomiale verdeling) kun je halen hoeveel doelpunten elke club in zijn thuis- en in zijn uitwedstrijden heeft gescoord.

		doelpunten thuis x	uit y
1	SC Cambuur	27	22
2	Eindhoven	35	17
3	Graafschap	24	31
4	FC den Haag	21	28
5	Heerenveen	27	20
6	Heracles	32	16
7	MVV	39	33
8	NAC	48	24
9	NEC	25	22
10	RBC	29	16
11	SVV	16	20
12	Telstar	25	27
13	FC Twente	49	33
14	Veendam	22	21
15	Vitesse	38	14
16	FC VVV	37	17
17	Wageningen	23	19

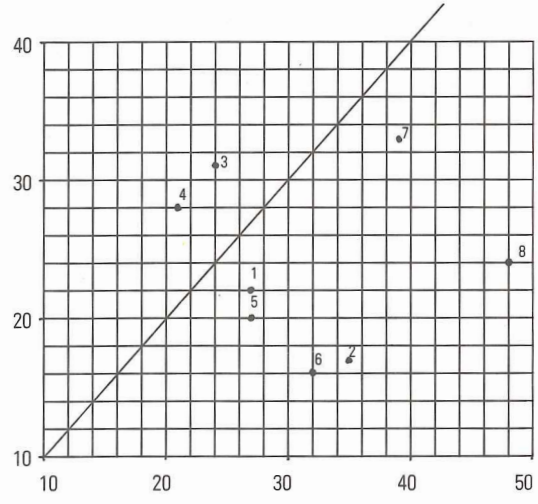
Maak de puntenwolk hiernaast af.

Wat denk je, is er positief of een negatief verband tussen x en y ?

de Volkskrant van MAANDAG 14 MEI 1984

THUIS \ UIT	SC CAMBUUR	EINDHOVEN	GRAAFSCHAP	FC DEN HAAG	HEERENVEEN	HERACLES	MVV	NAC	NEC	RBC	SVV	TELSTAR	FC TWENTE	VEENDAM	VITESSE	FC VVV	WAGENINGEN
SC CAMBUUR		3-1	1-1	1-2	2-0	1-1	1-1	0-0	1-1	2-0	4-2	2-3	1-1	3-2	1-1	3-3	1-0
EINDHOVEN	0-0		0-5	2-5	2-2	2-0	1-3	6-4	0-2	0-2	4-1	3-1	1-3	3-2	3-1	4-1	4-3
GRAAFSCHAP	3-0	2-1		2-1	0-0	2-1	2-2	6-2	0-2	2-1	1-0	0-1	0-2	2-2	1-1	0-1	1-1
FC DEN HAAG	0-1	4-0	1-3		0-3	3-0	0-0	0-1	3-1	2-1	1-0	2-1	2-4	2-0	1-0	0-1	0-1
HEERENVEEN	2-4	3-0	2-3	1-1		0-2	0-0	0-4	1-3	2-1	1-1	1-0	0-3	3-3	6-3	2-1	2-2
HERACLES	0-0	2-2	2-3	2-2	1-0		0-0	4-1	5-2	6-1	1-1	1-5	1-2	0-0	3-2	2-0	2-1
MVV	1-1	3-2	2-2	1-0	3-2	3-1		1-4	4-1	3-2	4-1	1-1	2-1	1-0	0-1	4-1	6-0
NAC	3-1	4-2	3-2	1-0	1-0	2-1	4-4		2-0	2-1	5-2	2-3	3-0	10-1	2-1	0-0	4-1
NEC	1-3	2-1	0-1	1-2	3-1	2-0	0-1	2-0		1-0	2-1	1-0	5-2	2-2	2-0	1-2	0-0
RBC	0-2	4-1	5-2	4-1	2-2	1-1	0-2	3-0	1-1		2-1	2-2	0-2	3-0	0-0	1-1	1-3
SVV	0-3	3-1	3-1	1-3	0-2	0-3	2-5	1-1	0-2	1-1		1-2	0-1	0-3	1-1	3-2	0-1
TELSTAR	1-0	1-0	1-1	1-3	5-1	1-1	1-1	0-1	3-2	2-1	4-1		0-2	1-1	2-1	2-0	1-1
FC TWENTE	3-2	7-1	2-2	2-0	2-1	1-1	1-2	2-1	3-1	3-1	1-1	3-4		5-0	5-0	2-1	7-0
VEENDAM	2-3	1-1	1-2	0-4	1-2	1-2	2-3	1-1	0-2	3-2	2-2	2-1	2-4		1-1	3-2	0-5
VITESSE	1-1	5-3	2-0	1-1	4-2	2-1	1-4	0-2	3-1	3-1	5-3	3-1	1-3	3-2		1-0	3-0
FC VVV	0-2	3-1	3-2	1-2	5-1	3-1	1-5	2-1	2-1	3-1	2-1	1-1	4-2	1-2	2-0		4-0
WAGENINGEN	1-0	0-0	1-1	1-1	1-1	3-0	1-0	2-1	0-0	5-0	2-2	1-1	1-1	2-1	0-0	2-1	

Eerste divisie



33 voetbal en toeval

Met een rekenmachientje is nagegaan:
de som van de x-waarden = 517
de som van de y-waarden = 380
de som van de x_2-waarden = 17123
de som van de y_2-waarden = 9084
de som van de xy-waarden = 11698

Bereken met behulp van deze uitkomsten:

Var(x), Var(y), Cov(x,y) en r(x,y).

Schrijf ook je berekeningen op.

Var(x):	Var(y):
Cov(x,y):	r(x,y):

Wijst deze correlatiecoëfficiënt op verband tussen x en y ? Licht je antwoord toe.

Laat het resultaat van FC Twente weg en bepaal de correlatiecoëfficiënt van de 16 overblijvende punten van je puntenwolk.

Blijf je bij je antwoord op de voorlaatste vraag?

34 rechtlijnig verband

Om een dagje uit te gaan met een stel vrienden en vriendinnen huur je een Opel Kadett.

HUURTARIEVEN 1986

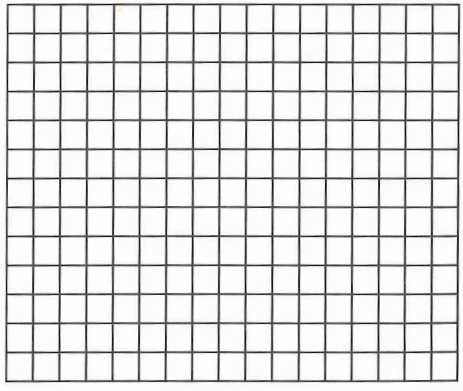
Groep	Type	deuren	zit-plaatsen		per dag 24 uur	per km
	Personenwagens					
A	Opel Corsa ★	2	4	f	26,00	0,23
B	Opel Kadett ★	2	4	f	30,00	0,26
C	Opel Record ★	4	4/5	f	42,00	0,40

Wat moet je afrekenen als je die dag 154 km gereden hebt?

Geef een formule voor de prijs p (in gld) als je op een dag x km in een Kadett gereden hebt.

Geef zo'n formule ook voor de Record.

Teken hiernaast de twee bijbehorende grafieken.



35 rechtlijnig verband

Behalve de huurprijs moet je ook de benzine betalen. Een Opel Kadett loopt 1 op 12, een Record 1 op 8. Neem aan dat een liter benzine *f* 1,60 kost.

Geef een formule voor de totale kosten y als je op een dag x km in een Kadett rijdt.

Geef die formule ook voor de Record.

Teken (rood) de bijbehorende grafieken in de figuur op de vorige bladzijde.

Er is een aantal factoren dat ervoor zorgt dat deze formules in de praktijk niet precies aangeven hoeveel de totale kosten zullen bedragen. Noem een aantal van die factoren.

Wanneer je een aantal keren zo'n auto huurt en steeds de gereden afstand x en de totale kosten y noteert, dan zullen de punten (x,y) rond de bijbehorende (rode) lijn liggen.

Die lijn geeft bij gegeven x een soort gemiddelde waarde van y. De grafiek kan dus gebruikt worden om een schatting te maken van de totale kosten als je de afstand die je moet rijden weet.

36 de lijn van de gemiddelden

Op bladzijde 24, 25 en 26 hebben we bij 78 leerlingen het verband onderzocht tussen het cijfer x voor wiskunde op het paasrapport in vwo-4 en het cijfer y voor wiskunde A in vwo-5 van dezelfde leerling een jaar later.

(vijfde klas)

y	2	10	33	22	10	1	
9				1			1
8				3	3	1	7
7		1	5	8	5		19
6		5	20	7			32
5	2	4	6	2	1		15
4			2	1	1		4
	4	5	6	7	8	9	

x (vierde klas)

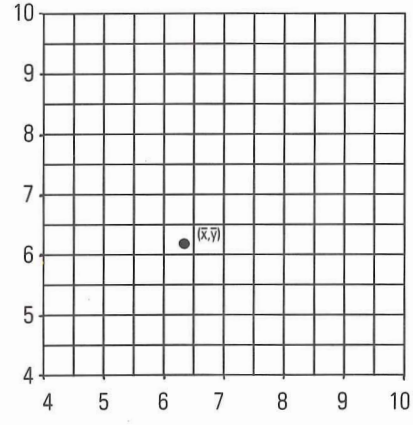
Er was duidelijk correlatie: r = 0,46. Bij een hogere x vind je gemiddeld ook een hogere y-waarde.

Deze trend kunnen we ook nog op een andere manier laten zien. We berekenen bij elke waarde van x het gemiddelde $\bar{y}(x)$ van de daarbij behorende y-waarden.

x	4	5	6	7	8	9
$\bar{y}(x)$	5,00	5,70				

Ga na dat de eerste twee waarden van $\bar{y}(x)$ in de tabel juist zijn en vul de tabel daarna verder in.

Geef de punten uit de tabel aan in de figuur hiernaast.
In die figuur is het "zwaartepunt van de puntenwolk" $(\bar{x},\bar{y})$ = (6,40;6,17) al aangegeven.



Ook deze puntenwolk maakt de tendens duidelijk: bij hogere x gemiddeld ook een hogere y.

Teken door het punt (6,40;6,17) een rechte lijn die zo ongeveer deze trend weergeeft. Bedenk daarbij dat sommige gemiddelden maar op een paar leerlingen berusten.

Hoe groot is ongeveer de richtingscoëfficiënt van de lijn die je getekend hebt?

37 de regressielijn

In plaats van de lijn van de gemiddelden zo maar op het oog te tekenen, heeft men een methode bedacht om een lijn te berekenen die "zo goed mogelijk" de stijgende of dalende trend van de gemiddelden weergeeft. Die lijn wordt de **regressielijn** genoemd. (Zie ook extra sterk, daar kijken we nog eens, op een andere manier naar deze lijn.)

Eerste formule voor de regressielijn van y op x

$$\frac{y - \bar{y}}{SD(y)} = r \cdot \frac{x - \bar{x}}{SD(x)} \qquad \rightsquigarrow z_y = r \cdot z_x$$

Tweede formule voor de regressielijn van y op x

$$y = r \cdot \frac{SD(y)}{SD(x)} \cdot x - r \cdot \frac{SD(y)}{SD(x)} \cdot \bar{x} + \bar{y}$$

Merk op:

de regressielijn gaat door het zwaartepunt $(\bar{x},\bar{y})$ van de puntenwolk

de richtingscoëfficiënt hangt af van de correlatiecoëfficiënt r (en van de standaardafwijkingen SD(x) en SD(y));

bij positieve r is de regressielijn stijgend (bij hogere x is y gemiddeld groter), bij negatieve r is de regressielijn dalend.

Ga na hoe de tweede formule, de formule in de vorm $y = ax + b$, uit de eerste volgt.

Bereken de formule voor de regressielijn die hoort bij de puntenwolk van de wiskundecijfers (zie bladzijde 24, 25 en 26).

Komt de richtingscoëfficiënt van de lijn die je zo op het oog tekende op de vorige bladzijde ongeveer overeen met de richtingscoëfficiënt van de regressielijn?

38 de regressielijn

Ga na dat de richtingscoëfficiënt a van de regressielijn ook geschreven kan worden als:

$$a = \frac{(x_1 \cdot y_1 + \ldots + x_n \cdot y_n) - n \cdot \bar{x} \cdot \bar{y}}{(x_1^2 + \ldots + x_n^2) - n \cdot \bar{x}^2}$$

In het algemeen worden statistische gegevens vaak gebruikt om voorspellingen te doen. Ook de regressielijn kan daarvoor gebruikt worden.

Als je bijvoorbeeld het cijfer x van een leerling uit de vierde klas weet kun je met de formule van de regressielijn: $y = 0{,}480x + 3{,}094$

een voorspelling doen voor het cijfer y dat deze leerling in de vijfde zal hebben.

Welk cijfer voorspel je op deze manier voor een leerling die in de vierde een 4 had?

Laat zien dat bij een positieve correlatiecoëfficiënt een x die boven het gemiddelde ligt, een voorspelling y geeft die ook boven het gemiddelde ligt.
Gebruik de eerste formule voor de regressielijn:

$$\frac{y - \bar{y}}{SD(y)} = r \cdot \frac{x - \bar{x}}{SD(x)}$$

Waarom zal (in standaardafwijkingen uitgedrukt) de afwijking $y - \bar{y}$ kleiner zijn dan $x - \bar{x}$?

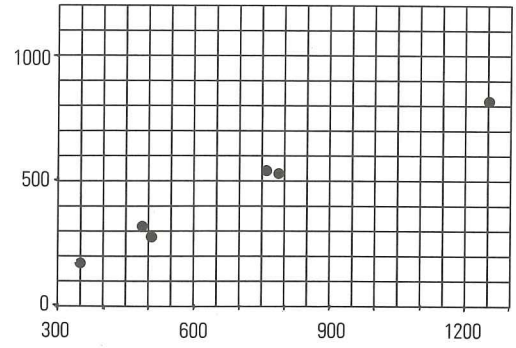
Aan deze laatste eigenschap dankt de regressielijn zijn naam: regressie: teruggaan, terugvallen. Hier naar het midden, het gemiddelde.

39 nog eens reclame

Hieronder staat nog eens de tabel van de kosten x (in guldens) van advertenties van een bedrijf en het aantal reacties y op zo'n advertentie.

kosten	x_i	780	515	342	1257	496	768
reacties	y_i	509	298	176	819	312	520

Hiernaast is de bijbehorende puntenwolk getekend.



Bereken met je rekenmachientje:

de som van de x-waarden = _____

de som van de y-waarden = _____

de som van de x^2-waarden = _____

de som van de y^2-waarden = _____

de som van de xy-waarden = _____

Bereken met deze gegevens $\bar{x}$ en $\bar{y}$.

Bereken ook de richtingscoëfficiënt a van de regressielijn.

Geef in de puntenwolk het zwaartepunt aan.
Teken de regressielijn.

Geef ook een vergelijking van deze lijn.

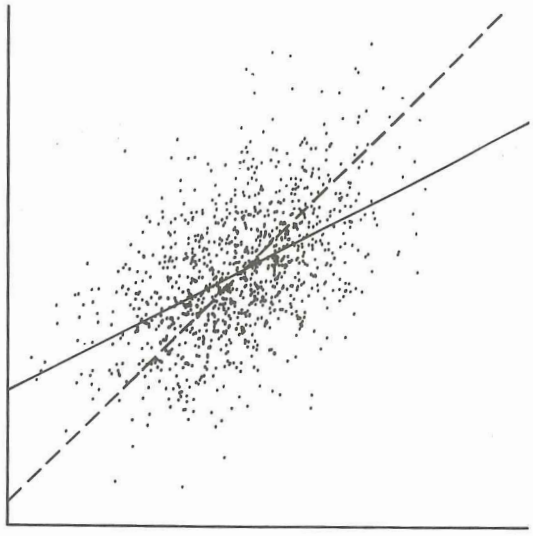
Welk aantal reacties mag je verwachten op een advertentie die 900 gulden kost?

En hoeveel reacties mag je verwachten op een advertentie die 2000 gulden kost?

Welke van deze laatste twee voorspellingen is het meest betrouwbaar? Waarom?

40 vader en zoon

Het plaatje hiernaast is gebruikt als omslag voor een boek over statistiek dat in 1978 in de Verenigde Staten is verschenen.
De puntenwolk (Engels: scatterdiagram) bestaat uit 1078 punten (x,y), waarbij x de lengte van een vader en y de lengte van diens volwassen zoon is.Het zou het resultaat zijn van een onderzoek dat omstreeks 1900 in Engeland werd gehouden. Men interesseerde zich toen voor de overerving van allerlei eigenschappen. Francis Galton (1822-1911) was een van de pioniers op dat gebied.
Het vader-zoon onderzoek waar het hier om gaat werd gepubliceerd in 1903 door Pearson en Lee in Biometrika, een bekend wetenschappelijk tijdschrift dat nu nog bestaat.



In het boek vind je de volgende informatie over de puntenwolk op de omslag:
$\bar{x}$, de gemiddelde lengte van de vaders, is 68 inch en SD(x) = 2,7 inch.
$\bar{y}$, de gemiddelde lengte van de zoons, is 69 inch en SD(y) = 2,7 inch.
De correlatiecoëfficiënt r is 0,5.

Reken deze gegevens om naar centimeters, 1 inch is 2,54 cm. Rond de gemiddelde lengten af op een geheel getal en de SD's op een getal met één cijfer na de komma.

$\bar{x}$ = _____ cm	$\bar{y}$ = _____ cm
SD(x) = _____ cm	SD(y) = _____ cm

Geef een vergelijking van de regressielijn.

Verdeel het plaatje in verticale stroken van bijvoorbeeld een halve centimeter en schat in een paar stroken het gemiddelde van de y-waarden. Geef bij elke strook het gemiddelde van de y-waarden met een stip aan.

Welke van de twee lijnen in de puntenwolk is de regressielijn?

Wat is de richtingscoëfficiënt van de andere lijn?

Wat zou er aan de hand geweest zijn als alle punten op deze andere lijn gelegen hadden?

41 omgekeerde regressie

Om een eenvoudiger plaatje te krijgen is op de omslag de schaalverdeling langs de assen weggelaten. Die schaalverdeling willen we terugvinden.

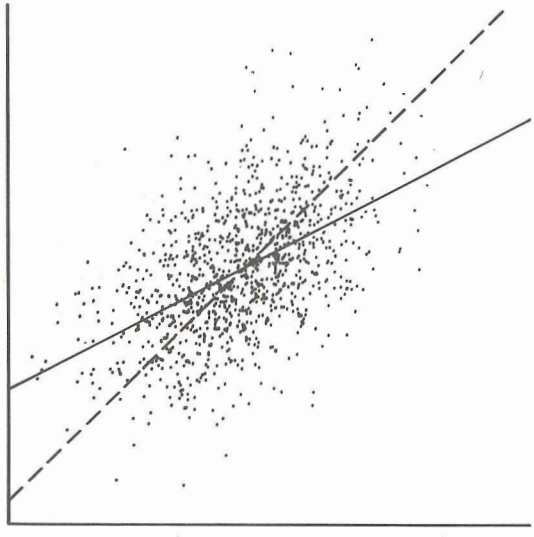
In welk punt van de puntenwolk zullen, naar je mag aannemen, de twee lijnen elkaar snijden?

Hoe zijn de x- en de y-waarden verdeeld?

Breng je verdeling (in cm) op de assen aan.

Schat, met behulp van de regressielijn, de lengte van de zoon van een vader die 180 cm lang is.

Je kunt ook de vraag stellen: Hoe groot schat je de vader als je de lengte van de zoon weet?
We moeten dan kijken naar horizontale stroken: bij een gegeven y-waarde kijken we naar het gemiddelde van de bijbehorende x-waarden.



Verdeel de puntenwolk in een stuk of tien horizontale stroken. Schat bij elke strook de gemiddelde x-waarden. Geef dat gemiddelde in het midden van de strook met een stip aan.

Teken de lijn die het best past bij je tien stippen.

Deze tweede lijn van gemiddelden noemt men de regressielijn van x op y (de andere lijn is dan de regressielijn van y op x). Deze lijn geeft dus de gemiddelde x-waarde bij een gegeven y-waarde.

Een algemene formule voor deze lijn is:

$$\frac{x - \bar{x}}{SD(x)} = r \cdot \frac{y - \bar{y}}{SD(y)} \text{ ofwel } \frac{y - \bar{y}}{SD(y)} = \frac{1}{r} \cdot \frac{x - \bar{x}}{SD(x)}$$

Geef een vergelijking van deze tweede regressielijn bij de 'vader-en-zoon-gegegens'.

Schat, met behulp van deze vergelijking, de lengte van de vader van een zoon die 180 cm lang is.

42 afhankelijke en onafhankelijke variabelen

Wanneer je het verband tussen gewicht en lengte onderzoekt bij Nederlanders van ongeveer 20 jaar, dan heeft het meer zin te kijken naar hoe het gemiddelde gewicht verandert met de lengte dan omgekeerd. Waarom is dat zo?

De regressielijn van gewicht y op lengte x geeft ons een richtlijn van wat ongeveer het normale gewicht is bij een bepaalde lengte.
De ene variabele, hier de lengte, noemt men de onafhankelijke variabele; die kan niet beïnvloed worden. De andere variabele, hier het gewicht, noemt men de afhankelijke variabele.

De onafhankelijke variabele geeft men meestal aan met de letter x, de afhankelijke variabele met de letter y.

Maar lang niet altijd is er zo'n duidelijk verschil tussen de twee variabelen als bij gewicht en lengte. Wat denk je bijvoorbeeld van de lengte x van de man en de lengte y van de vrouw bij echtparen in Nederland?

Bij een bedrijf wordt bijgehouden hoeveel reacties een advertentie oplevert. Een overzicht van de kosten x van een advertentie en het aantal reacties y staat in de tabel op bladzijde 23.

Wat is hier de afhankelijke en wat is de onafhankelijke variabele?

Bij een bedrijf wordt een product gemaakt. De productie-omvang en de productiekosten worden bijgehouden.

Wat is hier de afhankelijke en wat is de onafhankelijke variabele?

Bedenk zelf een voorbeeld waarbij het ook niet eenvoudig is aan te geven welke de afhankelijke en welke de onafhankelijke variabele is.

43 tweedehands auto's

NISSAN

SERVICE-DEALER VOOR MALDEN EN OMGEVING

Nissan Bluebird diesel, zilver		1987	31.500.-
Nissan Cherry 1.3 DX, wit	type	1983	6.750.-
Nissan Micra SDX, wit		1987	15.250.-
VW Polo, spec. uitvoering, 3-drs.		1985	14.950.-
Honda Civic Sport, 3-drs., zwart		1982	7.750.-
BMW 732i veel extra, groen		1982	16.950.-
Mitsubishi Colt GLX, blauw	type	1985	14.950.-
Opel Kadett LS 1.6i, 3.800 km, wit		1988	26.000.-
Opel Kadett LS 1.2, groen		1985	15.500.-
Opel Kadett LS, diesel		1984	15.750.-
Toyota Corolla 1.3, 4-drs., zilver		1984	11.500.-
Toyota Carina II 1.6, 5-drs.		1984	14.500.-

FINANCIERING MOGELIJK

AUTOBEDRIJF ROELOFS

Broekkant 51 - MALDEN - Telefoon 080-580334

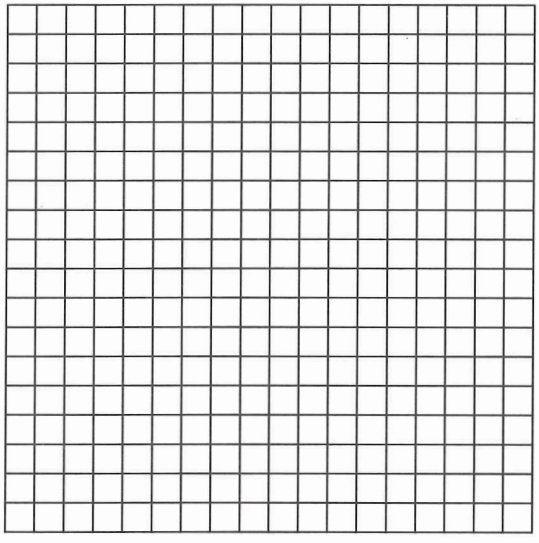
x	y
1	31,50
5	6,75

x	y

Hierboven zie je een advertentie uit de "Weekendkoerier" van 12 augustus 1988.

Vul de tabel in. Hierbij is x de leeftijd (in jaren) van de auto in 1988 en y de prijs van de auto (in duizenden guldens).

Teken de hierbij behorende puntenwolk.



Bereken $\bar{x}$, $\bar{y}$, SD(x) en SD(y).

Bereken Cov(x,y) en de correlatiecoëfficiënt r.

Geef een vergelijking van de regressielijn in de vorm: $y = ax + b$ en teken de regressielijn.

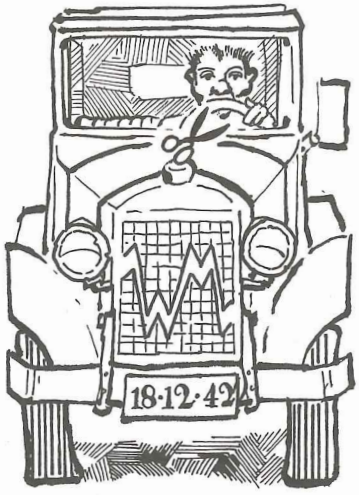
44 tweedehands auto's

Welke betekenis kun je aan a en welke aan b hechten?

Wat geeft deze regressielijn (zo ongeveer) aan?

Welke twee bedenkingen kun je maken ten aanzien van de keuze van de steekproef?

De populatie waaruit deze steekproef getrokken is, is die van alle tweedehands auto's in Nederland.



Hoe gaat de puntenwolk eruit zien als je de populatie beperkt tot een bepaald type, bijvoorbeeld Volkswagen Golf?

Wat zal er dan gebeuren met de correlatiecoëfficiënt?

De auto's uit de advertentie zijn nog tamelijk nieuw: geen enkele auto is ouder dan zes jaar.

Als je een grotere steekproef neemt met ook auto's van 10 jaar en ouder, zal de beschrijving van de puntenwolk met een rechte lijn dan nog een goed beeld geven? Licht je antwoord toe.

Rechtlijnigheid heeft zijn grenzen.

Het verband tussen leeftijd en prijs van tweedehands auto's kan redelijk beschreven worden met een rechte lijn voor de leeftijdsklasse van 0 tot 6 jaar.
Voor auto's van 6 tot 12 jaar geldt weer een andere lijn. Als je beide klassen in één figuur wilt samenvatten, is een kromme lijn misschien beter, bijvoorbeeld een parabool.

Een andere mogelijkheid kan zijn: niet rekenen met de prijs maar met de logaritme van de prijs. Als de prijzen elk jaar met ongeveer hetzelfde percentage dalen, zou dit een goede manier zijn om het verband leeftijd-prijs recht te trekken.

45 groei

Een bioloog onderzoekt de groei van zonnebloemen. Hij denkt te kunnen bewijzen dat in een bepaalde fase van de groei de hoogte h van zo'n bloem exponentiëel van de tijd t afhangt:

$h = H \cdot g^t$.

In die fase verzamelt hij de volgende gegevens:

t (in dagen)	0	4,8	11,5	16,2	22,6	30,7
h (in cm)	72	85	108	142	190	256

Teken hiernaast de wolk van de punten (t_i, y_i), waarbij $y_i = \log(h_i)$.

Mag je, zo op het oog, aannemen dat de bioloog een juiste veronderstelling maakte? Waarom?

Bereken de correlatiecoëfficiënt r(t,y).

Geef een vergelijking van de regressielijn die hoort bij de punten (t_i, y_i). Schrijf die vergelijking in de vorm $y = a \cdot t + b$.

Teken de regressielijn.

Geef met behulp van de constanten a en b een formule voor de hoogte h als functie van de tijd.

46 gemeenschappelijke belangen

Boer Jansen verbouwt aardappelen en graan, boer Pieterse aardappelen en suikerbieten.
Het inkomen van deze twee landbouwers komt voor ongeveer de helft uit de aardappelen. Een goede opbrengst en prijs van de aardappelen is dus gunstig voor beiden.
Vermoedelijk is er een duidelijke correlatie tussen de inkomsten van de twee boeren.

Een dergelijke samenhang kunnen we eenvoudig nabootsen met dobbelstenen.

Drie worpen geven X_1, X_2 en X_3 ogen.

Stel: $X = X_1 + X_2$ en $Y = X_1 + X_3$.

Door de gemeenschappelijke term X_1 zijn X en Y niet onafhankelijk.

X kan variëren van 2 tot (en met) 12, Y ook. Maar bepaalde combinaties, bijvoorbeeld $X = 3$ en $Y = 9$, zijn niet mogelijk.

Waarom kan het punt (3,9) niet optreden?

Noem nog een paar combinaties die niet kunnen optreden. Geef ook aan waarom ze niet kunnen optreden

Op hoeveel manieren kan het punt (4,5) optreden. Schrijf die mogelijkheden op.

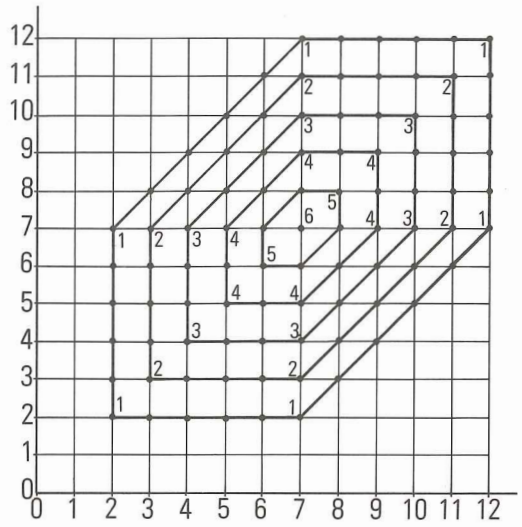
Schrijf ook alle mogelijkheden op voor het punt (7,9).

In het plaatje op de volgende bladzijde zie je welke punten (x,y) kunnen optreden en op hoeveel manieren.

Bereken het gemiddelde van de y-waarden bij $x = 3$ en bij $x = 12$.

gemeenschappelijke belangen

Geef de gemiddelde y-waarde bij x = 3 en de gemiddelde y-waarde bij x = 12 in het plaatje met rode stippen aan.



Bereken bij nog een paar waarden van x het gemiddelde van de y-waarden. Geef ook deze gemiddelden met een rode stip aan.

De gemiddelde y-waarden liggen op een rechte lijn, de regressielijn.

Hoe groot zijn de getallen a en b in de vergelijking y = ax + b van de regressielijn?

Waarom is in dit voorbeeld het getal a gelijk aan de correlatiecoëfficiënt r ?

Opmerkingen bij dit voorbeeld.

Opmerking 1.

Het plaatje geeft een volledig overzicht van alle mogelijke uitkomsten voor het paar stochasten X en Y. Je kunt ook zeggen: het geeft de gemeenschappelijke kansverdeling van (X,Y) weer.

Deze kansverdeling berust op de aanname dat alle 6^3 (= 216) mogelijke resultaten van de drie worpen met de dobbelsteen dezelfde kans hebben. De correlatiecoëfficiënt van de complete puntenwolk van de 216 punten is precies ½.

Je kunt uit deze populatie ook een steekproef nemen, bijvoorbeeld door twintig keer met drie dobbelstenen te gooien. De puntenwolk die je dan krijgt zal een correlatiecoëfficiënt hebben die waarschijnlijk niet precies ½ is. Als je echter een grotere steekproef neemt, dan zal de correlatiecoëfficiënt daarvan steeds meer gaan lijken op de correlatiecoëfficiënt van de populatie.

Opmerking 2.

We kunnen het gemeenschappelijk deel van X en Y groter of kleiner maken:

Je gooit vier keer met een dobbelsteen en stelt:

$X = X_1 + X_2 + X_3$ en $Y = X_1 + X_2 + X_4$.

Het gemeenschappelijk deel is dan groter, de correlatiecoëfficiënt is dan ⅔.

Kun je verklaren waarom de regressielijn nu gegeven wordt door de formule: $y = ⅔ \cdot x + 3½$?

Je gooit vijf keer met een dobbelsteen en stelt:

$X = X_1 + X_2 + X_3$ en $Y = X_3 + X_4 + X_5$.

Het gemeenschappelijk deel is dan kleiner.

Hoe groot is, denk je, de correlatiecoëfficiënt nu?

Wat is nu een vergelijking van de regressielijn?

Opmerking 3.

De correlatiecoëfficiënt r = ½ bij de lengtes van vader en zoon (blz. 40) is in het licht van dit voorbeeld wel aannemelijk. Waarom?

zelftoets

aantekeningen/samenvatting

De belangrijkste formules:

Eerste formule voor de covariantie

$$Cov(x,y) = 1/n \cdot ((x_1-\bar{x})(y_1-\bar{y}) + (x_2-\bar{x})(y_2-\bar{y}) + \ldots + (x_n-\bar{x})(y_n-\bar{y}))$$

In woorden: de covariantie is het gemiddelde van de producten van de afwijkingen.

Tweede formule voor de covariantie

$$Cov(x,y) = 1/n \cdot (x_1 \cdot y_1 + x_2 \cdot y_2 + \ldots + x_n \cdot y_n) - \bar{x} \cdot \bar{y}$$

In woorden: de covariantie is het gemiddelde van de producten min het product van de gemiddelden.

Eerste formule voor de correlatiecoëfficiënt

$$r = \frac{Cov(x,y)}{SD(x) \cdot SD(y)} = \frac{1/n \cdot (x_1 \cdot y_1 + x_2 \cdot y_2 + \ldots + x_n \cdot y_n) - \bar{x} \cdot \bar{y}}{SD(x) \cdot SD(y)}$$

Tweede formule voor de correlatiecoëfficiënt

$$r = 1/n \cdot (z(x_1) \cdot z(y_1) + z(x_2) \cdot z(y_2) + \ldots + z(x_n) \cdot z(y_n)),$$

waarbij $z(x_i) = \frac{x_i - \bar{x}}{SD(x)}$ de z-waarde van x_i is.

In woorden: de correlatiecoëfficiënt is het gemiddelde van de producten van de z-waarden x_i en y_i.

Derde formule voor de correlatiecoëfficiënt

$$r(x,y) = \frac{Var(x) + Var(y) - Var(d)}{2 \cdot SD(x) \cdot SD(y)}$$ waarbij $d = y - x$.

Correlatiecoëfficiënt van de rangnummers:

$$r(a,b) = 1 - \frac{6}{n(n^2 - 1)} \cdot ((a_1-b_1)^2 + (a_2-b_2)^2 + \ldots + (a_n-b_n)^2)$$

aantekeningen/samenvatting

Nog meer formules:

Eerste formule voor de regressielijn (van y op x)

$$\frac{y - \bar{y}}{SD(y)} = r \cdot \frac{x - \bar{x}}{SD(x)}$$

vorm $y = ax + b$

$a = r \cdot \frac{SD(y)}{SD(x)} = \frac{cov(x,y)}{SD(x)^2}$

door $(\bar{x}, \bar{y})$

Tweede formule voor de regressielijn

$$y = r \cdot \frac{SD(y)}{SD(x)} \cdot x - r \cdot \frac{SD(y)}{SD(x)} \cdot \bar{x} + \bar{y}$$

ofwel

$y = a \cdot x + b$ met $a = r \cdot \frac{SD(y)}{SD(x)}$ en b zó dat $(\bar{x}, \bar{y})$ op de regressielijn ligt.

Formule voor de regressielijn van x op y

$$\frac{x - \bar{x}}{SD(x)} = r \cdot \frac{y - \bar{y}}{SD(y)} \text{ ofwel } \frac{y - \bar{y}}{SD(y)} = \frac{1}{r} \cdot \frac{x - \bar{x}}{SD(x)}$$

vorm $y = ax + b$

$a = \frac{1}{r} \cdot \frac{SD(y)}{SD(x)} = \frac{SD(y)^2}{cov(x,y)}$

door $(\bar{x}, \bar{y})$

53 extra werk

Discriminatie

Bij een groot bedrijf solliciteerden het afgelopen jaar 1000 mensen: 600 mannen en 400 vrouwen.
Van deze sollicitanten werden er 400 aangenomen, evenveel mannen als vrouwen.

Vul de tabel hiernaast in.

	man	vrouw	
aangenomen			
afgewezen			

Hoeveel procent van de mannelijke sollicitanten is aangenomen? Hoeveel procent van de vrouwen heeft met succes gesolliciteerd?

Het hoofd van de afdeling personeelszaken krijgt het verwijt te horen dat ze een anti-man koers volgt.
Zij verdedigt zich hiertegen met de volgende aanvullende informatie:
Er waren 300 vacatures bij laag betaalde banen en 100 bij beter betaalde banen.
De tabel hiernaast geeft aan hoe die vacatures vervuld werden.

laag betaald	man	vrouw	
aangenomen	120	180	300
afgewezen	30	70	100
	150	250	400

Vul de tabel voor beter betaald werk in.

beter betaald	man	vrouw	
aangenomen			
afgewezen			

Hoeveel procent van de aangenomen sollicitanten was bij laag betaald werk man en hoeveel procent was vrouw? Bereken deze percentages ook bij beter betaald werk.

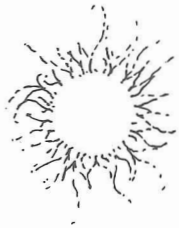
Bij welk soort werk werd een sollicitant, relatief gezien, het gemakkelijkst aangenomen?

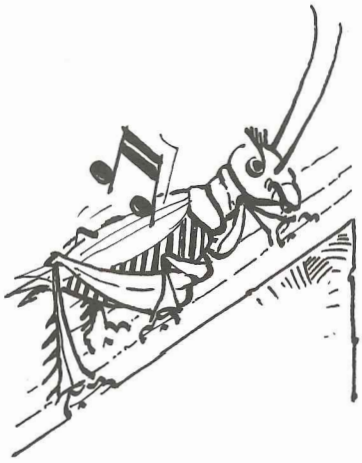
In beide groepen is van de sollicitanten die werden aangenomen het percentage mannen hoger dan het percentage vrouwen. Over de twee groepen samen genomen zijn die percentages gelijk. Hoe kan dat?

Is er sprake van discriminatie? Wie worden daar de dupe van?

54 extra werk

krekels

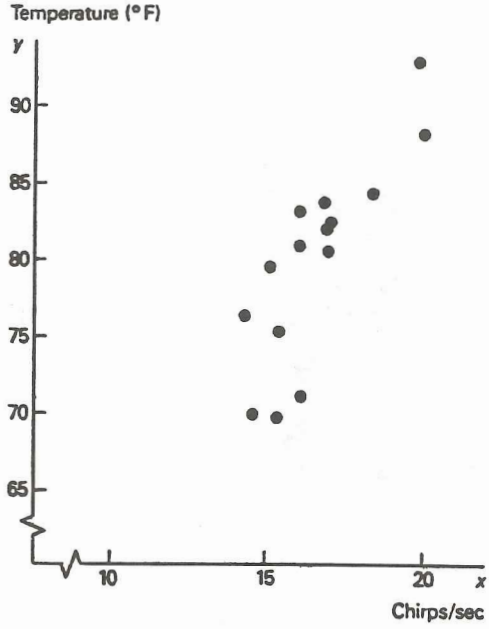




In de zomer kun je 's avonds vaak krekels horen sjirpen. De frequentie waarmee ze dat doen, is afhankelijk van de temperatuur. Van de gestreepte veldkrekel heeft George W. Pierce, professor aan de Harvard University, met behulp van speciaal ontwikkelde apparatuur een gedetailleerde studie gemaakt.

Een deel van zijn bevindingen vind je in de tabel hieronder. Hiernaast staat het bijbehorende spreidingsdiagram.
In de tabel en het diagram is x de frequentie (aantal sjirpen per seconde) en y de temperatuur (in graden Fahrenheit).
Gemakshalve staan de waarden van xy ook in de tabel.

x	y	xy
20,0	88,6	1772,0
16,0	71,6	1145,6
19,8	93,3	1847,3
18,4	84,3	1551,1
17,1	80,6	1378,3
15,5	75,2	1165,6
14,7	69,7	1024,6
17,1	82,0	1402,2
15,4	69,4	1068,8
16,2	83,3	1349,5
15,0	79,6	1194,0
17,2	82,6	1420,7
16,0	80,6	1289,6
17,0	83,5	1419,5
14,4	76,3	1098,7



Rond bij de eerste drie vragen de uitkomsten van je berekeningen af op drie cijfers na de komma.

55 extra werk

Bereken $\bar{x}$, $\bar{y}$, SD(x) en SD(y).

Bereken de correlatiecoëfficiënt r.

Geef een vergelijking van de regressielijn (van y op x). Rond de parameters af op één cijfer na de komma.

Teken deze regressielijn in het spreidingsdiagram.

Een krekel sjirpt 19,0 keer per seconde. Schat de temperatuur met behulp van je formule van de regressielijn.

Het verband tussen de temperatuur in graden Celsius (c) en in graden Fahrenheit (y) wordt gegeven door: $c = \frac{5}{9} \cdot (y - 32)$.

Stel een vergelijking op van de regressielijn waarbij de temperatuur niet in graden Fahrenheit maar in graden Celsius is gegeven.

Toon aan dat de correlatiecoëfficiënt bij gebruik van de Celsiusschaal even groot is als bij gebruik van de Fahrenheitschaal.

Hoe groot is de correlatiecoëfficiënt van x op y ?

Geef ook een vergelijking van de regressielijn van x op y. Rond de parameters af op één cijfer na de komma.

Teken ook de regressielijn van x op y in het spreidingsdiagram.

Hoe vaak schat je dat een gestreepte veldkrekel per seconde sjirpt bij een temperatuur van 85,0 graden Fahrenheit? Gebruik je formule voor de regressielijn van x op y.

Welke frequentie had je gevonden als je je formule voor de regressielijn van y op x gebruikt had?

56 extra sterk

kleinste kwadraten

Uit de verzameling getallen x_1, x_2, ..., x_n wordt zonder terugleggen n keer een getal getrokken. Vooraf moet je één getal t kiezen dat bij elk van de n trekkingen je voorspelling is.

Als x_i het getal is dat je de i^e keer trekt, dan is de fout is deze voorspelling: $t - x_i$.
Neem aan dat je nu als 'straf' $(t - x_i)^2$ gulden moet betalen.

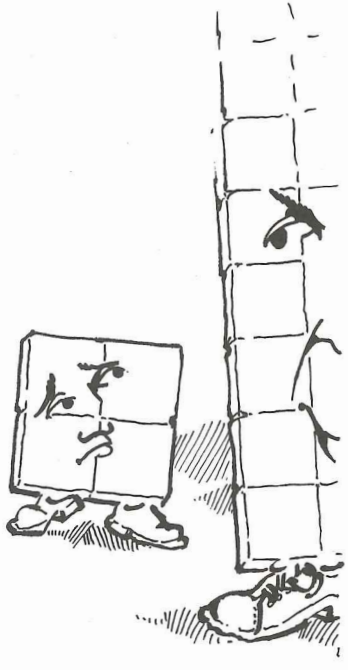
Na de n trekkingen is je totale verlies dan:

$Q(t) = (t-x_1)^2 + (t-x_2)^2 + ... + (t - x_n)^2$.

Het probleem is nu Q(t) zo klein mogelijk te krijgen.

Voorbeeld:

Laat $x_1 = 0$, $x_2 = 1$, $x_3 = 1$ en $x_4 = 4$ de getallen zijn waaruit je kiest.



Wat is het gemiddelde $\bar{x}$ van deze vier getallen?

Schrijf Q(t) als veelterm van t.

Laat met behulp van Q'(t) zien dat Q(t) minimaal is als $t = \bar{x}$.

Hoe groot is in dit voorbeeld de minimale waarde van Q ?

Behalve met differentiëren kun je nog op een tweede manier laten zien dat Q(t) minimaal is als $t = \bar{x}$.

57 extra sterk

Ga na hoe uit de twee formules voor de variantie volgt:
$(z_1-\bar{z})^2 + \ldots + (z_n-\bar{z})^2 = z_1^2 + \ldots + z_n^2 - n\cdot\bar{z}^2$

Gevolg:
$z_1^2 + \ldots + z_n^2 = (z_1-\bar{z})^2 + \ldots + (z_n-\bar{z})^2 + n\cdot\bar{z}^2$

Stel: $z_i = x_i - t$.
Laat zien: $\bar{z} = \bar{x} - t$ en $z_i - \bar{z} = x_i - \bar{x}$.

Toon nu aan:
$Q(t) = (x_1 - t)^2 + \ldots + (x_n - t)^2$

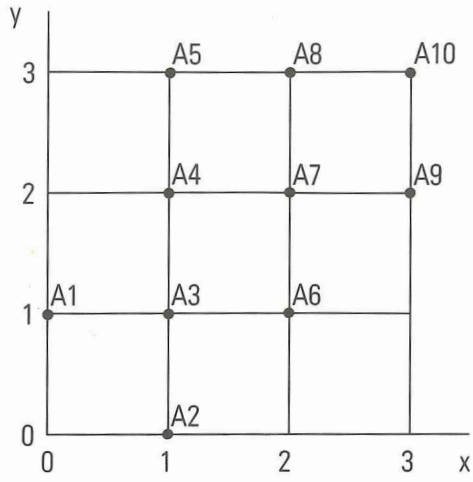
$= (x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2 + n\cdot(\bar{x} - t)^2$

Hieruit volgt direct: Q is minimaal als $t = \bar{x}$.
Hoe (precies) ?

extra sterk

de beste lijn, een bijzonder geval

Hiernaast zijn 10 punten, A_1, A_2, ..., A_{10}, in een rooster aangegeven. Door het toeval wordt één van deze punten aangewezen. Van dat punt wordt de eerste coördinaat x bekend gemaakt. Jij moet de tweede raden.
Bij gegeven eerste coördinaat x, kies je als tweede coördinaat t(x). (Als x = 0, dan kies je natuurlijk t(0) = 1. Altijd goed!)



Neem aan dat de boete voor verkeerd raden weer het kwadraat van de fout is.

Als $A_i = (x_i,y_i)$ gekozen wordt, dan wordt dus x_i bekend gemaakt. Jij voorspelt als tweede coördinaat $t(x_i)$ terwijl de juiste tweede coördinaat y_i is. Je boete wordt dus $(y_i - t(x_i))^2$.

Welk getal kies je als t(1) ? (Let op het resultaat op het eind van de vorige bladzijde.)

Bereken ook de verstandigste keus voor t(2) en voor t(3).

Ga na dat de punten $(x_i,t(x_i))$ op een rechte lijn liggen. Teken die lijn en geef daarvan vergelijking.

Merk op:
deze lijn heeft de eigenschap dat de som van de kwadraten van de afstanden $d_i = y_i - t(x_i)$ zo klein mogelijk is.

Bereken de correlatiecoëfficiënt r van deze 10 punten en laat zien dat de gevonden lijn de regressielijn is.

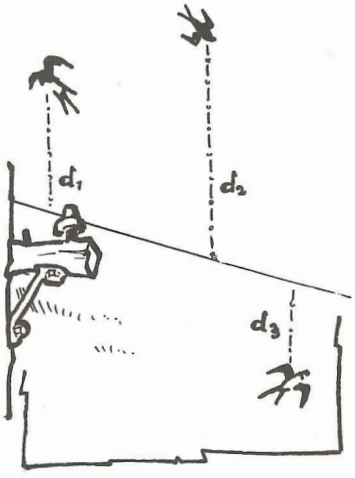
Ga na dat de te verwachten boete gegeven wordt door:

$0{,}1\cdot((y_1 - \frac{1}{2}x_1 - 1)^2 + ... + (y_{10} - \frac{1}{2}x_{10} - 1)^2)$.

Hoe groot is die te verwachten boete?

extra sterk

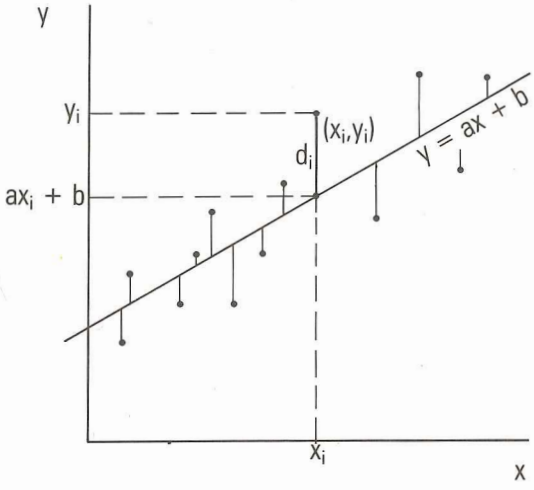
de beste lijn, het algemene geval



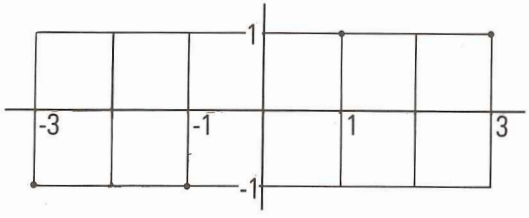
Bij een puntenwolk (x_1,y_1), ..., (x_n,y_n) zoeken we naar de lijn $y = ax + b$, zó dat de som $Q(a,b)$ van de kwadraten van de afstanden $d_i = y_i - ax_i - b$ zo klein mogelijk is.

We kiezen aselect een punt uit deze wolk en maken de eerste coördinaat x_i bekend. We proberen de tweede coördinaat te raden, te voorspellen met behulp van $ax_i + b$ (een eerstegraads functie).

$\frac{1}{n} \cdot Q(a,b)$ kunnen we dan opvatten als de gemiddelde kwadratische fout die we daarbij maken.



In het voorbeeld op de vorige bladzijde was de oplossing simpel, omdat de gemiddelde y-waarden al op een rechte lijn lagen. In het algemeen is dat natuurlijk niet zo.



Hiernaast zijn vier punten gegeven. Probeer de lijn $y = ax + b$ te bepalen waarbij $Q(a,b)$ minimaal is.

We gaan nu aantonen dat de lijn die we zoeken de regressielijn is, dat wil zeggen de lijn met richtingscoëfficiënt $r \cdot SD(y)/SD(x)$ en die door het zwaartepunt $(\bar{x},\bar{y})$ van de puntenwolk gaat.

$$Q(a,b) = (y_1 - (ax_1+b))^2 + \ldots + (y_n - (ax_n+b))^2$$

Vervang $y_i - ax_i$ in de formule voor $Q(a,b)$ door z_i en laat met de formule op bladzijde 57 zien:

$$Q(a,b) = (z_1 - \bar{z})^2 + \ldots + (z_n - \bar{z})^2 + n(\bar{z} - b)^2.$$

Ga na:

$\bar{z} = \bar{y} - a\bar{x}$ en $z_i - \bar{z} = (y_i - \bar{y}) - a(x_i - \bar{x})$.

extra sterk

Laat nu zien:

$$Q(a,b) = \sum_{i=1}^{n} ((y_i - \bar{y}) - a(x_i - \bar{x}))^2 + n(\bar{y} - a\bar{x} - b)^2$$

$$= a^2 \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2a \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + \sum_{i=1}^{n} (y_i - \bar{y})^2 + n(\bar{y} - a\bar{x} - b)^2$$

$$= a^2 \cdot Var(x) - 2a \cdot r \cdot SD(x) \cdot SD(y) + Var(y) + n(\bar{y} - a\bar{x} - b)^2$$

Ga na dat hieruit volgt dat Q(a,b) gelijk is aan:

$$(a \cdot SD(x) - r \cdot SD(y))^2 + (1 - r^2) \cdot Var(y) + n(\bar{y} - a\bar{x} - b)^2$$

En nu is ineens alles duidelijk.

Hoe kiezen we a zó dat de eerste term 0 wordt?

Hoe kiezen we b zó dat de derde term 0 wordt? Laat zien dat dit betekent dat het zwaartepunt van de puntenwolk op de lijn $y = ax + b$ ligt.

Gevolg: we kiezen a en b zó dat

$$Q(a,b) = (1 - r^2) \cdot Var(y).$$

Waarom volgt hieruit dat voor de correlatiecoëfficiënt r geldt: $-1 \leq r \leq 1$?

Waarom volgt hieruit ook dat alle punten op een lijn liggen als $r = 1$ of $r = -1$?

Als $SD(x) \neq 0 \neq SD(y)$, dan is het omgekeerde ook waar: als alle punten op een lijn liggen dan is $r = 1$ of $r = -1$. Waarom?