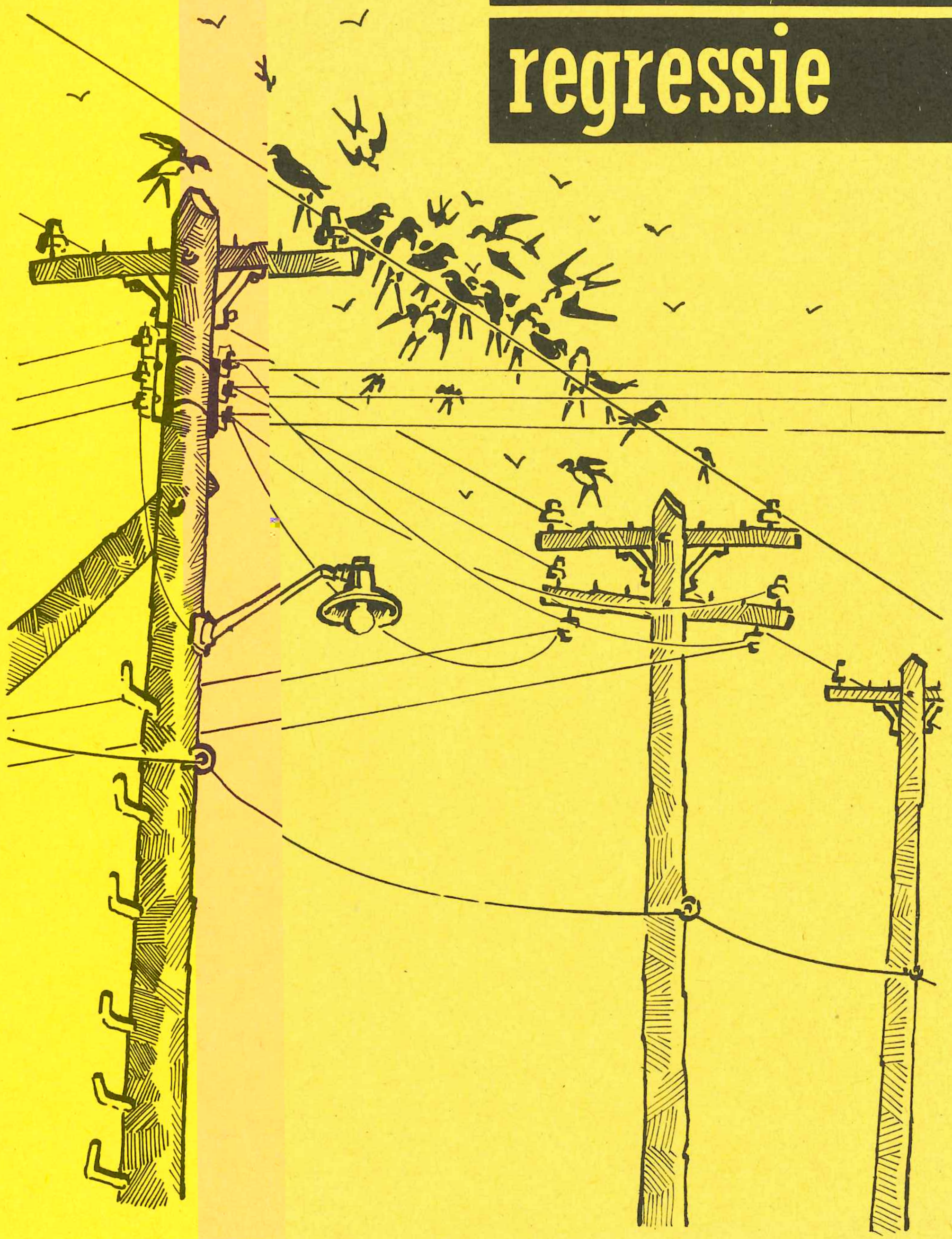
de Wageningse Methode

correlatie en regressie



...ode.

...ger, Wim Kremers, Jan Smit, Gerard Stroomer, Stef Tijs.

...tbus 105, 6860 AC Oosterbeek.

	[illegible]	[illegible]	NW	
[illegible]	[illegible]		b	34
[illegible]	[illegible]		d	68
	[illegible]		27	102

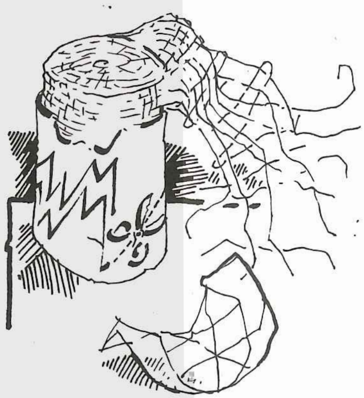
	[illegible]		NW	
[illegible]				34
[illegible]				68
	[illegible]		27	102

[redacted]

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

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 **ad − bc** > **0**.

Op dezelfde manier kun je afleiden:

er is een **negatief** verband als **ad − bc** < **0** en

er is **geen** verband als **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.

ı, weten we of
ef of een nega-

er afhangt van
et afhangt van
wordt ad - bc
product van de
en a+b.
associatie, de
2 x 2 - tabel

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

: R

ssende woord

ciatiemaat 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 = ________

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?

[redacted]

	Ro	
		13
	6	52

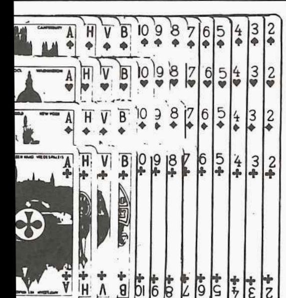
	Pl	NPl	
Aa			
NAa			
			52

R = ______

	o	
	6	52

	Ha	NHa	
Aa			
NAa			
			52

R = ______



8 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.

	Lh	Rh	
Ob			P(Ob) =
Bb			
			1

De overeenkomst tussen de twee tabellen is duidelijk!

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.

	Ha	Nha
Pl		
NPl		

R = ____

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

R = ____

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

R =1

R = -1

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.

R = 0

R is onbepaald

Hoe groot is R voor het volledige kaartspel?

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!).

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 3 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.

artinfarct

nogelijkheid om te bekijken
zou kunnen zijn.

lijn, dus wanneer aspirinege-
hebben, dan kunnen we de
cten beschouwen als een
e populatie van 22 000 art-

ekproef zonder terugleggen.
eekproef toch beschouwen
?

ent van de steekproef zal,
u zijn, uit aspirineslikkers

er onderzoeken kunnen we
13 keer opgooien van een
kop, noem dat X, kan ook
van de helft!

at zijn de waarden van de
eling?

kt 104 van E(X) af, m.a.w.
104 ?

op deze afwijking of een
is?

chil tussen de aantallen

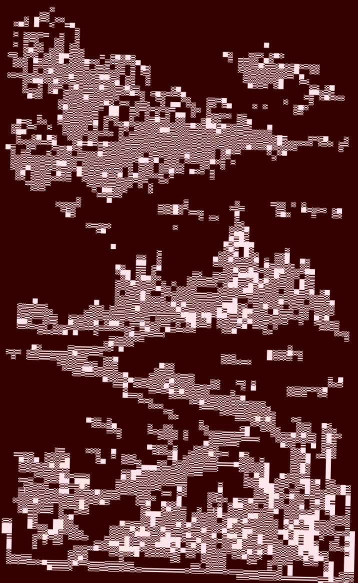
de associatiegraad R van

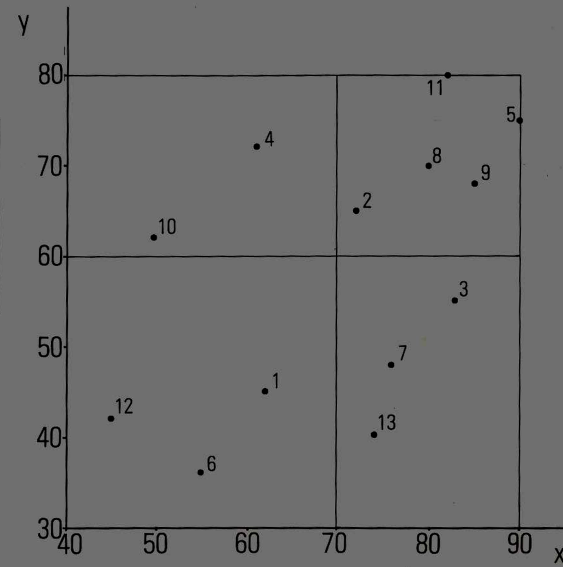
bij veel

de
len. Deel
nisch in-
pe
rsonen.

0	1		
	1	12	13
3	8		
	2	45	74
2	8		
	30	42	40

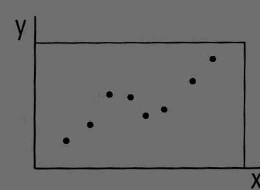
en
weerge-
g ve
tes erband is
me st. Perso-
estal ook
zoe
We ken is er
tale trekken
e lijn, bij-
n d
e 13 per-
e a
ssociatie-
e w
vaarop de
ge
ten een) is er
wolk hier-
eer
a n verticale
n an, zó dat
nu R.





	x < 70	x ⩾ 70	
y ⩾ 60			
y < 60			
			13

R = ____



____ verband

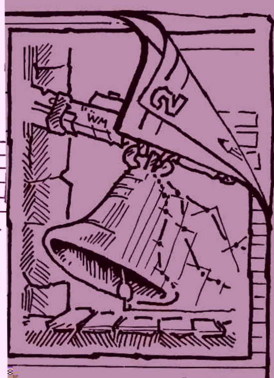
R = ____

3 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 $x_1, x_2, \ldots, x_n$ van een databestand:



Eerste formule voor de variantie:

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

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:

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

Tweede formule voor de variantie:

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

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:

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

... 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 $(x_1,y_1), \ldots,(x_n,y_n)$ berekend.

Eerste formule voor de covariantie

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

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

Tweede formule voor de covariantie

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

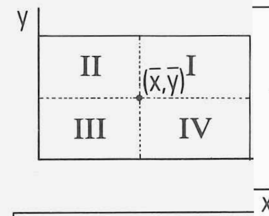
In woorden: de covariantie is het gemiddelde van de producten min het product van de gemiddelden.

Welk voordeel heeft de tweede formule boven de eerste?

Waar zijn we zoiets al eens eerder tegengekomen?

an 2 bij 2 naar veel bij veel

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



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

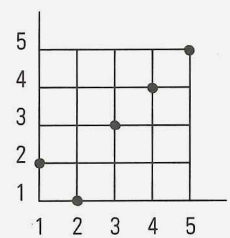
welke gebieden liggen de punten die een negatieve bijdrage leveren aan de covariantie?

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

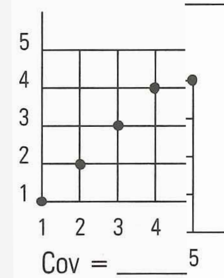
t denk je, is de covariantie bij de punten van intelligentietest op blz. 12 positief of negatief?

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

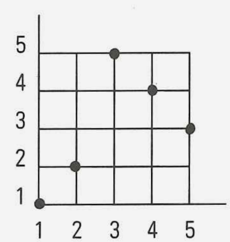
eken bij elk van de vier puntenwolken hiernaast de covariantie.



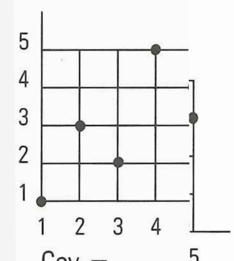
Cov = ______



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?

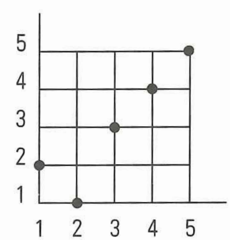
Een maat voor de samenhang die niet afhangt van de eenheden waarin x en y gemeten worden, is de **correlatiecoëfficiënt.**
Die vind je door de covariantie te delen door de standaarddeviatie van x en die van y.
De correlatiecoëfficiënt wordt meestal aangegeven met de letter r.

Eerste formule voor de correlatiecoëfficiënt

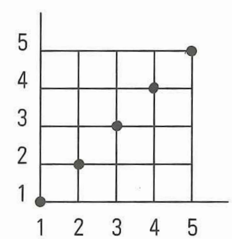
$$r = \frac{\mathrm{Cov}(x,y)}{\mathrm{SD}(x) \cdot \mathrm{SD}(y)}$$

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

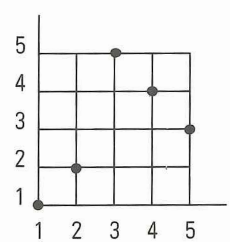
Bereken bij elk van de vier puntenwolken hiernaast de correlatiecoëfficiënt r. Op de vorige bladzijde heb je de covariantie al berekend!



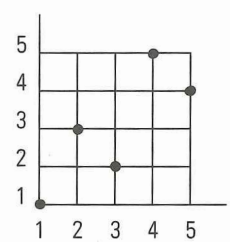
r = ______



r = ______



r = ______



r = ______

17

… bij 2 naar veel bij veel

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

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

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

Gevolg …

Tweede formule voor de correlatiecoëfficiënt

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

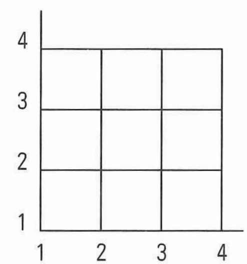
waarbij $z(x_i) = \frac{x_i - \bar{x}}{SD(x)}$ de z-waarde van x_i is.

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 zij respectievelijk de plaatsen 1, 2, 3 en 4, bij het hoogspringen de plaatsen 2, 1, 3 en 4.

In de tabel zie 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?

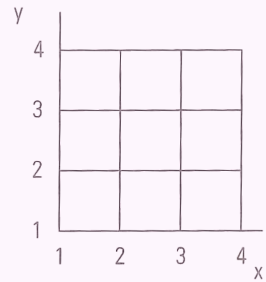
Bereken de correlatiecoëfficiënt r van deze vier punten.

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.

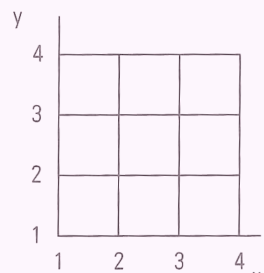
Wat is nu het verband tussen x en y, positief of negatief?

We houden steeds dezelfde uitslag bij het verspringen.

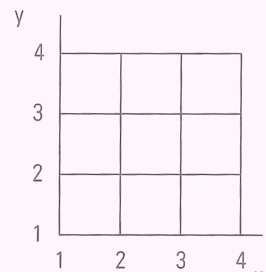
Op hoeveel verschillende manieren kunnen er rangnummers voor het hoogspringen aan gekoppeld worden?

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.



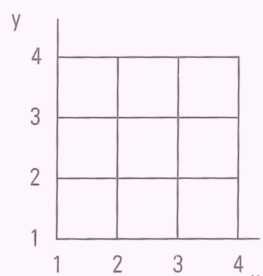
r = 1



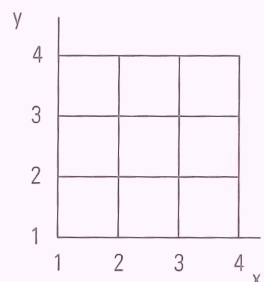
r = 1

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!



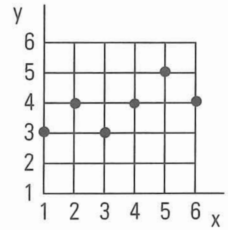
r = 0



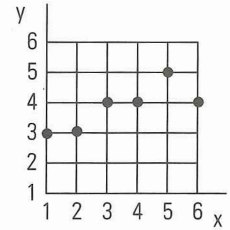
r = 0

19 van 2 bij 2 naar veel bij veel

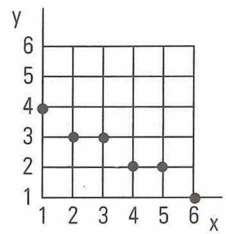
Bereken bij elk van de zes puntenwolken hiernaast de correlatiecoëfficiënt r.



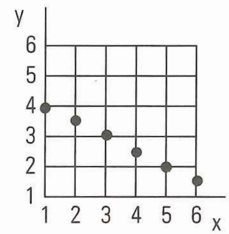
r = ______



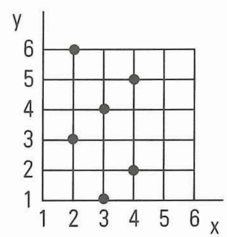
r = ______



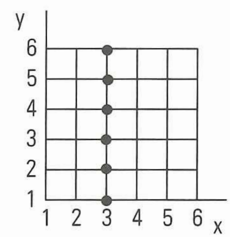
r = ______



r = ______



r = ______



r = ______

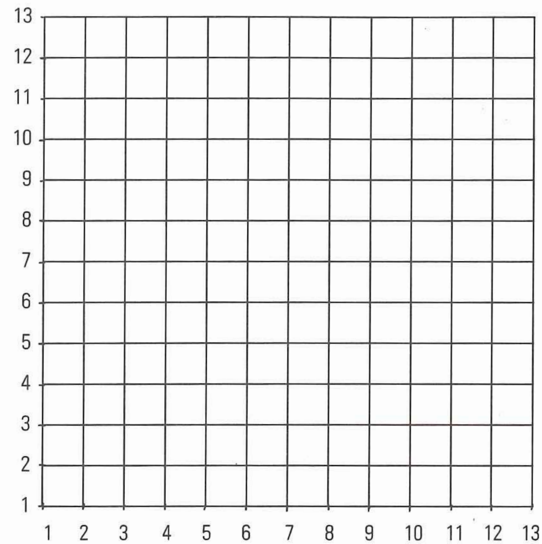
Bij het berekenen van r vind je altijd een uitkomst die tussen -1 en +1 ligt.
In extra sterk komen we daar op terug.

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 $\overline{a}$, $\overline{b}$, 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!

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 alle 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 $Var(a) = \frac{1}{12} \cdot (n^2-1)$ en kan de volgende formule worden afgeleid

correlatiecoëfficiënt van de rangnummers:

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

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 = ______

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

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

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

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

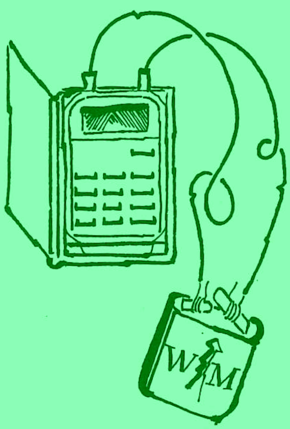
$1/n \cdot (x_i \cdot y_i + \ldots x_n \cdot y_n)$ = ______

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

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 $x_i y_i$ en Cov(x,y).

Bereken de correlatiecoëfficiënt r(x,y).

Ken de rangnummers $a_1, \ldots, a_6$ toe aan de x-waarden en de rangnummers $b_1, \ldots, b_6$ 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.

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	

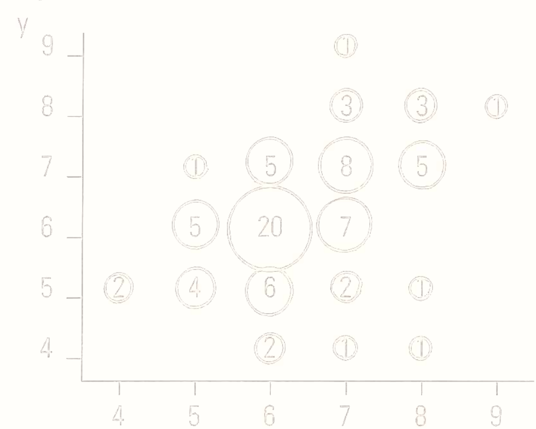
x (vierde klas)

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 $\overline{x}$, 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
$d \cdot f$							
$d^2 \cdot f$							

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

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

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.

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

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

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

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 $r(x,y)$.

...sociatie en

In de ...n middelbare schoo...
van be...rtcijfer altijd een z...
...andere leraar op ...
...zevens en achten.

Bereke...tabel zie je de cijfe...
...eide leraren les heb...

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

Bereke...

Bereke...

Bereke...en de associatiema...

Wat va...

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

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

...en de correlatiecoëf...

...alt je op als je R m...

Met be...hulp van de algem...
beweze...en worden dat de a...
correla...tiecoëfficiënt r altij...

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

Daar k...omt echter nogal w...
Alleen ...voor de echte liefh...

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 $\bar{d} = \bar{y} - \bar{x}$.

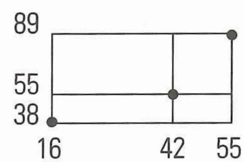
Bereken Var(x), Var(y) en Var(d).

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

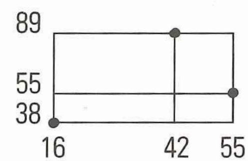
We mogen aannemen dat er in de populatie waaruit de steekproef genomen is geen enkel verband bestaat tussen leeftijd en huisnummer. Hoe kun je de grote waarde van r toch verklaren?

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) ?

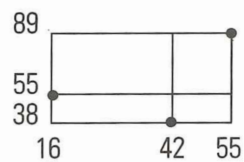
Als er geen verband is tussen x en y, dan zou elk van deze mogelijkheden dezelfde kans hebben. In het overzicht hiernaast zie je alle mogelijkheden. Bij elk van die mogelijkheden is de correlatiecoëfficiënt r berekend.



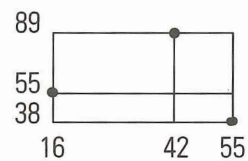
$r = {}^{13}/_{14}$



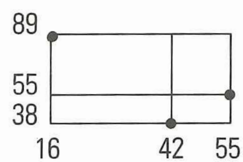
$r = {}^{1}/_{2}$



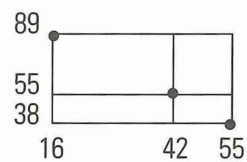
$r = {}^{1}/_{2}$



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



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



$r = -1$

ken ook Var(r).

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

eel mogelijkheden zijn er als er twee x- en
-waarden gelijk zijn?

---

meen:

aselecte steekproef van n personen wordt
der een x en een y genoteerd.

in de populatie waaruit de steekproef komt
enkel verband is tussen x en y, dan zijn alle
ren waarop de x-waarden aan de y-waar-
ekoppeld kunnen worden even waarschijn-

le x- en alle y-waarden verschillend zijn,
jn er n! mogelijkheden om die waarden te
len tot een wolk van n punten.

van deze n! puntenwolken kan de correla-
fficiënt r berekend worden.

eldt:

emiddelde van deze correlatiecoëfficiënten
ies 0 en
ndaardafwijking van deze r-waarden is
an $1/\sqrt{(n-1)}$.

an groot belang voor de beoordelen van
iecoëfficiënten. De volgende opgaven
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.

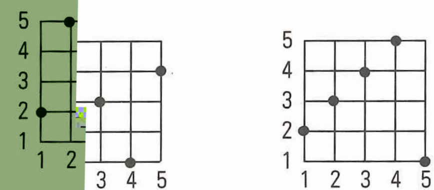
Kun je verklaren waarom een correlatiecoëfficient van 0,35 tussen de x- en de y-waarden best kan optreden maar dat r = 0,7 onwaarschijnlijk is?

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.

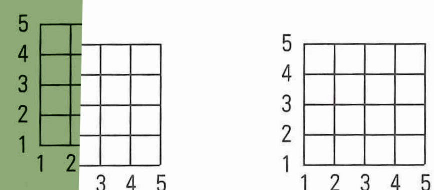
We vervangen de x-waarden $x_1, x_2, ..., x_n$ door de rangnummers 1, 2, ..., n. Met de y-waarden doen we hetzelfde.

Hiernaast zijn bij n = 5 twee mogelijke puntenwolken getekend.



Hoeveel van deze puntenwolken zijn er mogelijk?

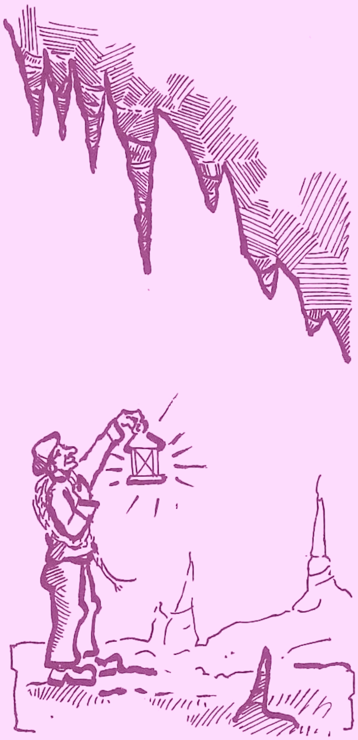
Teken een puntenwolk waarvan de correlatiecoëfficiënt 0 is. Ook één waarbij r = 1.



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

lijk

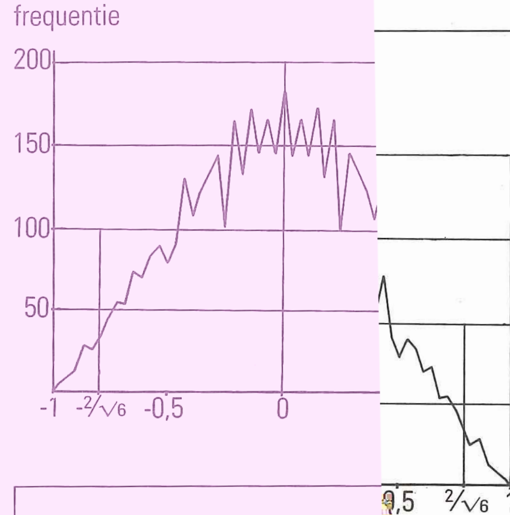
en-
er-



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.

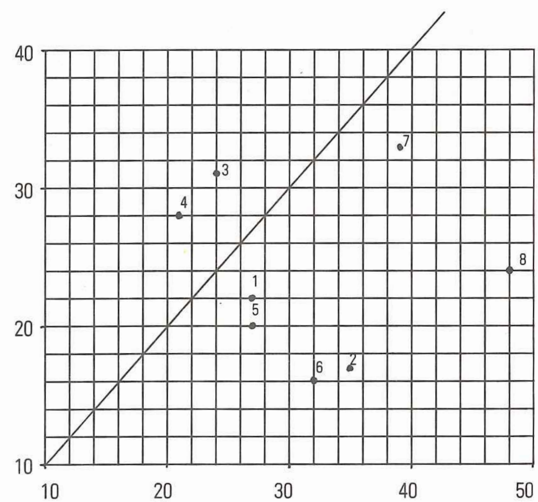
	doelpunten thuis x	doelpunten uit y
1 SC Cambuur	27	22
2 Eindhoven	35	17
3 Graafschap	24	31
4 FC den Haag	21	28
5 Heerenveen	27	20
6 Heracles	32	16
7 MVV	39	33
8 NAC	48	24
9 NEC	25	22
10 RBC	29	16
11 SVV	16	20
12 Telstar	25	27
13 FC Twente	49	33
14 Veendam	22	21
15 Vitesse	38	14
16 FC VVV	37	17
17 Wageningen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

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

Maak de puntenwolk hiernaast af.



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

en toeval

nmachientje is nagegaan:
le x-waarden = 517
le y-waarden = 380
le x_2-waarden = 17123
le y_2-waarden = 9084
le xy-waarden = 11698

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

ehulp van deze uitkomsten:

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

berekeningen op.

relatiecoëfficiënt op verband tus-
icht je antwoord toe.

aat van FC Twente weg en bepaal
ëfficiënt van de 16 overblijvende
ountenvolk.

ntwoord op de voorlaatste vraag?

vrienden

VEN 1986

ype	deuren	zit-plaatsen	f	per dag 24 uur	per km
ersonenwagens					
pel Corsa *	2	4	f	26,00	0,23
pel Kadett *	2	4	f	30,00	0,26
pel Record *	4	4/5	f	42,00	0,40

g 154 km

ld) als je
n hebt.

grafieken.

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			
8				3	3	1	1
7		1	5	8	5		7
6		5	20	7			19
5	2	4	6	2	1		32
4			2	1	1		15
	4	5	6	7	8	9	4

x (vierd

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

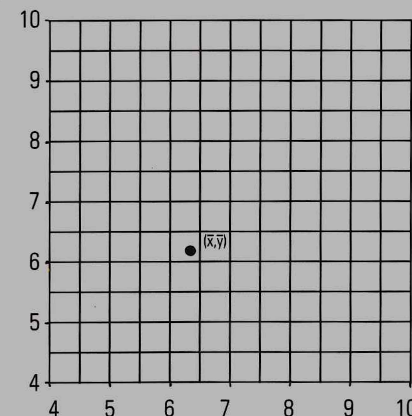
e klas)

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?

bedacht om een lijn te berekenen di
mogelijk" de stijgende of dalende tre
gemiddelden weergeeft. Die lijn wo
gressielijn genoemd. (Zie ook extra en zo maar
kijken we nog eens, op een andere m n methode
deze lijn.) e "zo goed
end van de
Eerste formule voor de regressielij rdt de **re-**
sterk, daar

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

manier naar

Tweede formule voor de regressien van y op x

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

lijn van y op x

Merk op:

de regressielijn gaat door het zwaarte

de richtingscoëfficiënt hangt af van de
en SD(y));

bij positieve r is de regressielijn stijge
regressielijn dalend. punt $(\bar{x},\bar{y})$ van de puntenwolk

correlatiecoëfficiënt r (en van de s
tandaardafwijkingen SD(x)

Ga na hoe de tweede formule, de form nd (bij hogere x is y gemiddeld gro
vorm y = ax + b, uit de eerste volgt. oter), bij negatieve r is de

Bereken de formule voor de regress
hoort bij de puntenwolk van de wisku ule in de
(zie bladzijde 24, 25 en 26).

Komt de richtingscoëfficiënt van de lijn
op het oog tekende op de vorige bladzi ielijn die
veer overeen met de richtingscoëfficië mdecijfers
regressielijn?

die je zo
jde onge-
t van de

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 $x - \bar{x}$?

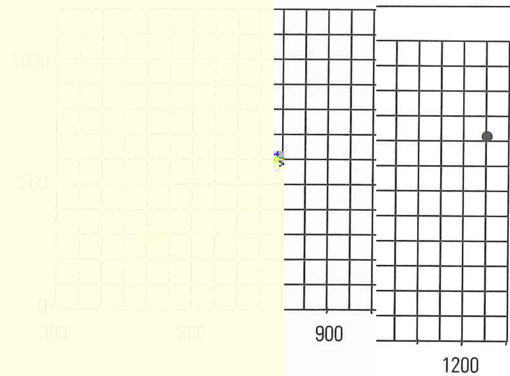
Aan deze laatste eigenschap dankt de regressielijn zijn naam: regressie: teruggaan, terugvallen. Hier naar het midden, het gemiddelde.

39 no[illegible]

Hiernaast staat nog eens de tabel van de kosten x (in guldens) van advertenties van een bedrijf en het aantal reacties y op zijn advertentie.

kosten	x_i	700	515	542	1257	406	768
reacties	y_i	509	298	176	818	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 $\overline{x}$ en $\overline{y}$.

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.

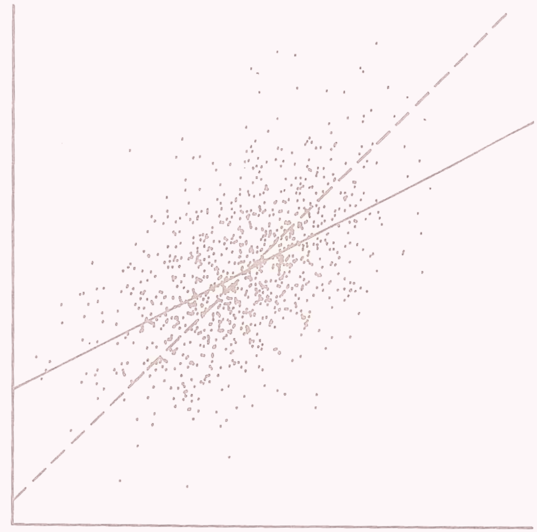
Welk aantal reacties mag je verwachten op een advertentie die 900 gulden kost?

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

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

[illegible] en zoon

[illegible] hiernaast is gebruikt als omslag voor [illegible] over statistiek dat in 1978 in de Ver[illegible]ten is verschenen.
[illegible]wolk (Engels: scatterdiagram) bestaat [illegible]nten (x,y), waarbij x de lengte van een [illegible] de lengte van diens volwassen zoon [illegible] het resultaat zijn van een onderzoek [illegible]eks 1900 in Engeland werd gehouden. [illegible]sseerde zich toen voor de overerving [illegible] eigenschappen. Francis Galton (1822-[illegible] een van de pioniers op dat gebied. [illegible]oon onderzoek waar het hier om gaat [illegible]bliceerd in 1903 door Pearson en Lee [illegible]ka, een bekend wetenschappelijk tijd[illegible]nu nog bestaat.



[illegible] vind je de volgende informatie over [illegible]olk op de omslag:
[illegible]delde lengte van de vaders, is 68 inch [illegible] 2,7 inch.
[illegible]delde lengte van de zoons, is 69 inch [illegible] 2,7 inch.
[illegible]ecoëfficiënt r is 0,5.

[illegible] gegevens om naar centimeters, 1 [illegible] cm. Rond de gemiddelde lengten af [illegible]el getal en de SD's op een getal met [illegible] de komma.

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

[illegible]rgelijking van de regressielijn.

[illegible] plaatje in verticale stroken van bij[illegible]en halve centimeter en schat in een [illegible] het gemiddelde van de y-waarden. [illegible] strook het gemiddelde van de y-[illegible] een stip aan.

[illegible] twee lijnen in de puntenwolk is de [illegible]

[illegible]tingscoëfficiënt van de andere lijn?

[illegible]an de hand geweest zijn als alle [illegible]e andere lijn gelegen hadden?

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-gegevens'.

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

e en onafhankelijke variabelen

verband tussen gewicht en
bij Nederlanders van ongeveer
het meer zin te kijken naar hoe
ewicht verandert met de lengte
Vaarom is dat zo?

an gewicht y op lengte x geeft
van wat ongeveer het normale
bepaalde lengte.
hier de lengte, noemt men de
riabele; die kan niet beïnvloed
re variabele, hier het gewicht,
ankelijke variabele.

variabele geeft men meestal
r x, de afhankelijke variabele

tijd is er zo'n duidelijk verschil
variabelen als bij gewicht en
je bijvoorbeeld van de lengte x
de lengte y van de vrouw bij
rland?

ordt bijgehouden hoeveel reac-
tie oplevert. Een overzicht van
een advertentie en het aantal
de tabel op bladzijde 23.

ankelijke en wat is de onafhan-

rdt een product gemaakt. De
en de productiekosten worden

nkelijke en wat is de onafhan-

orbeeld waarbij het ook niet
geven welke de afhankelijke
ankelijke variabele is.

NISSAN tweedehands auto's

Hierboven zie je een advertentie uit de "Weekend-koerier" 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).

x	y
1	31,50
5	6,75

x	y

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.

15 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.

enschappelijke belangen

en verbouwt aardappelen en graan,
rse aardappelen en suikerbieten.
en van deze twee landbouwers komt
eer de helft uit de aardappelen. Een
engst en prijs van de aardappelen is
voor beiden.
k is er een duidelijke correlatie tus-
msten van de twee boeren.

---

jke samenhang kunnen we eenvoudig
met dobbelstenen.

n geven X_1, X_2 en X_3 ogen.

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

meenschappelijke term X_1 zijn X en Y
nkelijk.

ren van 2 tot (en met) 12, Y ook. Maar
ombinaties, bijvoorbeeld $X = 3$ en
niet mogelijk.

n het punt (3,9) niet optreden?

een paar combinaties die niet kunnen
Geef ook aan waarom ze niet kunnen

manieren kan het punt (4,5) optreden.
mogelijkheden op.

alle mogelijkheden op voor het punt

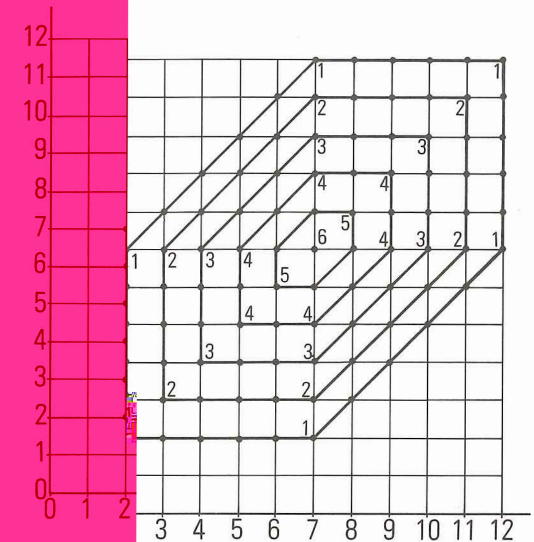
atje op de volgende bladzijde zie je
ten (x,y) kunnen optreden en op hoe-
ren.

et gemiddelde van de y-waarden bij
bij $x = 12$.

[illegible]elangen

Geef [illegible] gemiddelde y-waarde [illegible] 3 en de gemiddelde y-waarde [illegible] plaatje met [illegible] aan.

Bereken bij nog een paar waarden van X het gemiddelde van de y-waarden. Teken 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 ee… nemen, bijvoorbeeld door twintig k… dobbelstenen te gooien. De punter… dan krijgt zal een correlatiecoëfficiën… waarschijnlijk niet precies ½ is. Als … grotere steekproef neemt, dan zal … coëfficiënt daarvan steeds meer ga… de correlatiecoëfficiënt van de popu…

Opmerking 2.

We kunnen het gemeenschappelijk d… Y groter of kleiner maken:

Je gooit vier keer met een dobbelst…

$X = X_1 + X_2 + X_3$ en $Y = X_1 + X_2$ …

Het gemeenschappelijk deel is da… correlatiecoëfficiënt is dan ⅔.

Kun je verklaren waarom de regres… geven wordt door de formule: y = …

Je gooit vijf keer met een dobbelst…

$X = X_1 + X_2 + X_3$ en $Y = X_3 + X$ …

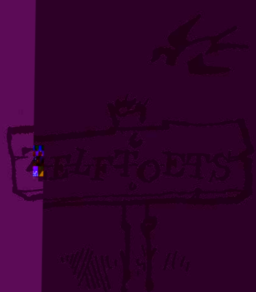
Het gemeenschappelijk deel is dan …

Hoe groot is, denk je, de correlatiec…

Wat is nu een vergelijking van de …

Opmerking 3.

De correlatiecoëfficiënt r = ½ bij d… vader en zoon (blz. 40) is in het licht va… wel aannemelijk. Waarom?



zelftoets

…elangen

…eekproef …met drie …lk die je …bben die …chter een …orrelatie- …lijken op …e.

…deel

…van X en …een

…en stelt: $X_2 + X_4$.

…roter, de …sielij… ⅔ • x … n nu ge- … + 3½ ?

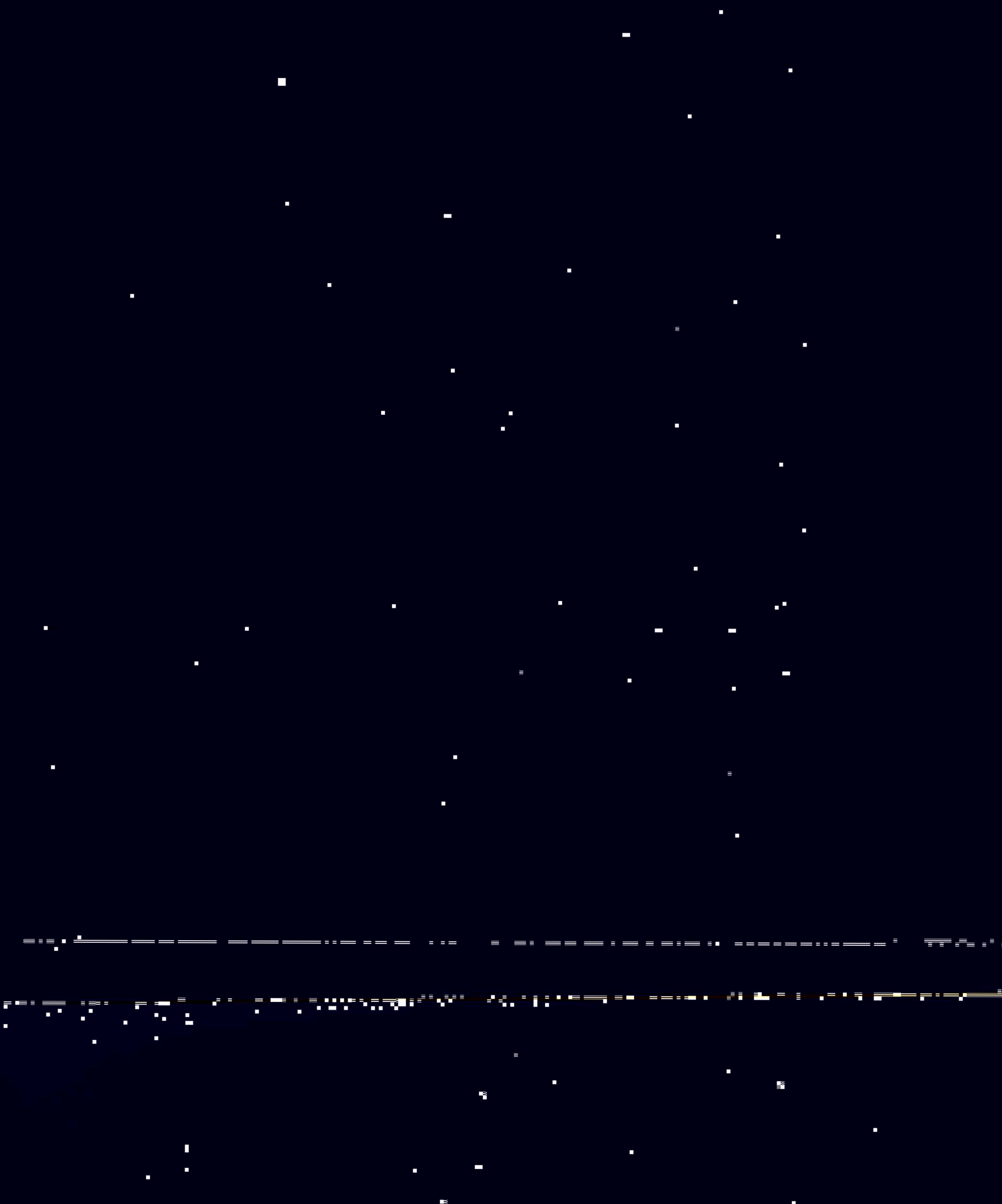
…en stelt: $X_4 + X_5$.

…klei… …coëffi… …ner.

…regre… …ciënt nu?

…essielijn?

…de le… …n dit …ngtes van …voorbeeld



aantekeningen/samenvatting

Nog meer formules:

Eerste formule voor de regressielijn (van y op x)

vorm $y = ax + b$

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

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

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.

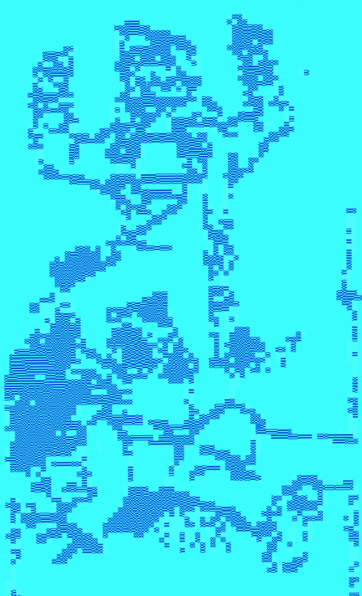
Formule voor de regressielijn van x op y

vorm $y = ax + b$

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

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

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



elopen jaar
ven.
ngenomen,

	man	vrouw	
aangenomen			
afgewezen			

citanten is
wen heeft

ken krijgt
nan koers

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

de aanvul-
banen en
vacatures

beter betaald	man	vrouw	
aangenomen			
afgewezen			

licitanten
l procent
k bij beter

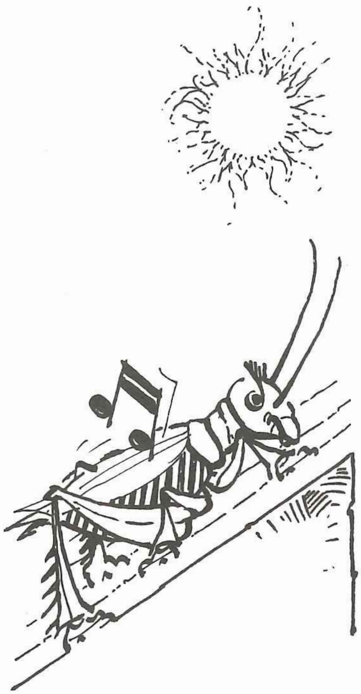
, relatief

die wer-
en hoger
wee groe-
es gelijk.

n daar de

extra werk

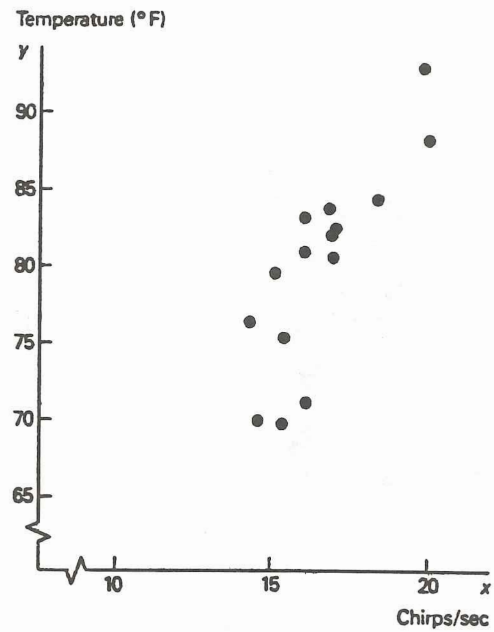
krekels



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.

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?

57

extra sterk

Ga na hoe uit de twee formules voor de variantie volgt:

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

Gevolg:

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

Stel: $z_i = x_i - t$.

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

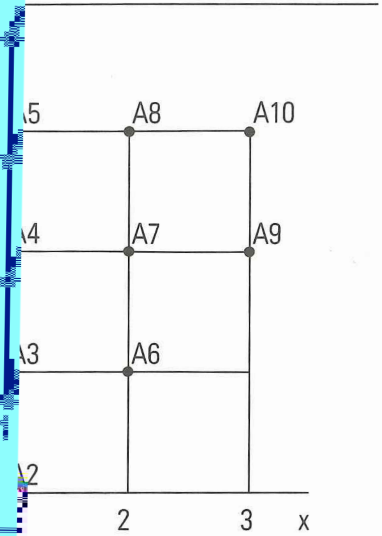
Toon nu aan:

$$Q(t) = (x_1 - t)^2 + \dots + (x_n - t)^2$$

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

Hieruit volgt direct: Q is minimaal als $t = \bar{x}$.

Hoe (precies) ?



A8
A10
A7
A9
A6
2
3
x

$y = ax + b$

x

$n(\bar{y} - a\bar{x} - b)^2$

$(x_1-\bar{x})(y_i-\bar{y}) + \sum_{i=1}^{n} (y_i-\bar{y})^2 + n(\bar{y}-a\bar{x}-b)^2$

$\mathrm{Var}(y) + n(\bar{y} - a\bar{x} - b)^2$

(a,b) gelijk is aan:

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

.

rste term 0 wordt?

erde term 0 wordt?

t het zwaartepunt $y = ax + b$ ligt.

dat

de correlatiecoëffi-

alle punten op een ?

s het omgekeerde een lijn liggen dan