

Wiskunde D Online VWO 6 Summary Blok 15-16 Updated 2022-04-06 Blok 16 Les 4 Edited

www.leennoordzij.me

Dr.l.noordzij@leennoordzij.nl

Samenvatting van de PowerPoint-Slides van de Universiteit van Groningen en de Wageningse Methode.

Deze samenvattingen zijn gemaakt voor mijn kleindochter bij de begeleiding van haar studie Wiskunde D Online, 2016-2019. De begeleiding betrof de theorie, het huiswerk en de toetsen.

De Open Universiteit(OU) heeft mij ondersteund bij de begeleiding.

Tags: Correlation, Chance, Education, Euclidean Geometry, Non-Euclidian Geometry

Regression, Mathematics, Statistics, Open University, Wageningse Methode, Wiskunde D Online, VWO 6, Prepcollege.

Inhoudsopgave

Blok 15 Correlatie en Regressie.....	2
Blok 15 Les 1 De associatiemaat	2
Oefenen en uitgewerkte Oefenopgaven.....	8
Blok 15 Les 2 Van $2 \cdot 2$ naar veel $\cdot$ veel.....	11
Oefenen en uitgewerkte Oefenopgaven.....	16
Blok 15 Les 3 Correlatie en toeval	22
Oefenen en uitgewerkte Oefenopgaven.....	26
Blok 15 Les 4 Rechthoekig verband	30
Oefenen en uitgewerkte Oefenopgaven.....	35
Appendix 1 De kleinste kwadraten methode: De Afleiding(paragrafen 56 t/m 60)	42
Appendix 2 Samenvatting formules	44
Blok 16 Niet-Euclidische meetkunde.....	46
Blok 16 Les 1 Wat Euclidische meetkunde?	46
Oefenen en uitgewerkte Oefenopgave.....	49
Blok 16 Les 2 Het parallellen postulaat.	50
Oefenen en uitgewerkte Oefenopgaven.....	52
Blok 16 Les 3 Meetkunde op de bol	55
Oefenen en uitgewerkte Oefenopgave.....	57
Blok 16 Les 4 Hyperbolische meetkunde	58
Epiloog Waarom de Samenvatting Online?	62

Blok 15 Correlatie en Regressie

Blok 15 Les 1 De associatiemaat

- Deze Les sluit aan bij de paragrafen 1 tot en met 11 van de tekst Correlatie en Regressie van de Wageningse Methode 1989
- **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.

De volgende vraag dringt zich op: is er verband tussen het kiezen van wiskunde en het bespelen van een muziek instrument?

Drie mogelijkheden:

1. Wiskunde(W) leerlingen kiezen even vaak een muziekinstrument(M) dan niet-wiskunde(NW) leerlingen. Nota Bene: NM→geen muziekinstrument.
2. Wiskunde leerlingen kiezen vaker een muziekinstrument dan niet-wiskunde leerlingen.
3. Wiskunde leerlingen kiezen minder vaak een muziekinstrument dan niet-wiskunde leerlingen.

ad 1. Wiskunde leerlingen kiezen even vaak een muziekinstrument dan niet-wiskunde leerlingen.

Stel dat een derde($1/3$) van de leerlingen, dus 34, een muziek instrument kiest:

In onderstaande tabel staat een voorbeeld van mogelijkheden. Er is geen verband tussen aanleg voor wiskunde en muziek. De factor $1/3$ vind je daarom in de rij M en $2/3$ vind je dus in de rij NM. Met deze verhoudingen kun je kolommen W en NW invullen.

	W	NW	
M	$\frac{1}{3} \times 75 = 25(a)$	$\frac{1}{3} \times 27 = 9(b)$	34
NM	$\frac{2}{3} \times 75 = 50(c)$	$\frac{2}{3} \times 27 = 18(d)$	68
	75	27	102

→ en twee keer zoveel NM als M . Dus $a : c = 1/2$, en $b : d = 1/2$.

Algemeen:

→ $a : c = b : d \rightarrow ad = bc \rightarrow ad - bc = 0$.

ad. 2. Wiskunde leerlingen kiezen vaker een muziekinstrument dan niet-wiskunde leerlingen. Er is blijkbaar een positief verband tussen aanleg voor wiskunde en muziek. Je verwacht dan verhoudingsgewijs veel muzikanten onder de leerlingen die wiskunde gekozen hebben. Als er een maximaal mogelijk aantal wiskunde leerlingen(34) voor M gekozen hebben, is dit aangegeven in onderstaande tabel.

(Onderstaande tabel: Extreem of maximaal verband $b = 0$).

Extreme situatie:

	W	NW	
M	$a = 34$	$b = 0$	34
NM	$c = 41$	$d = 27$	68
	75	27	102

ad.3. Wiskunde leerlingen kiezen minder vaak een muziekinstrument dan niet-wiskunde leerlingen. Er is blijkbaar een negatief verband tussen aanleg voor wiskunde en muziek (In onderstaande tabel is de extreme situatie weergegeven. Extreem of minimaal verband $d = 0$, wel wiskunde en niet muziek kan maximaal 68 zijn)

Extreme situatie:

	W	NW	
M	$a = 7$	$b = 27$	34
NM	$c = 68$	$d = 0$	68
	75	27	102

ad.2. Wiskunde leerlingen kiezen vaker een muziekinstrument (onderstaande tabel) dan niet-wiskunde leerlingen. Er is een zwak verband tussen aanleg voor wiskunde en muziek. Nota Bene: Dit geval is iets anders dan ad.2, *het extreme geval*.

Dus iets meer dan een derde van de wiskunde studenten kiest voor muziek. Stel: $a = 28 = 25 + 3 \rightarrow a > 25$.

Zwak positief verband:

	W	NW	
M	$a = 28$	$b = 6$	34
NM	$c = 47$	$d = 21$	68
	75	27	102

$$\frac{a}{c} > \frac{b}{d} \Leftrightarrow ad - bc > 0$$

Zwak positief verband, in procenten:

	W	NW	
M	37%	22%	34
NM	63%	78%	68
	100%	100%	102

Nu een zwak negatief verband $\rightarrow a < 25$. Stel $a = 14$ (of $a = 22$).

	W	NW	
M	$a = 14$	$b = 20$	34
NM	$c = 61$	$d = 7$	68
	75	27	102

$$\frac{a}{c} < \frac{b}{d} \Leftrightarrow ad - bc < 0$$

Opgave 3

De uitwerking in tabelvorm:

Zie volgende bladzijde.

Uitwerking

	W	NW	
M	$a = 14$	$b = 20$	34
NM	$c = 61$	$d = 7$	68
	75	27	102

$$\frac{a}{c} < \frac{b}{d} \Leftrightarrow ad - bc < 0$$

Samenvatting

1. Geen verband: $\frac{a}{c} = \frac{b}{d} \Leftrightarrow ad - bc = 0$
2. Positief verband: $\frac{a}{c} > \frac{b}{d} \Leftrightarrow ad - bc > 0$
3. Negatief verband: $\frac{a}{c} < \frac{b}{d} \Leftrightarrow ad - bc < 0$

Opgave 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 2×2 tabel weergegeven. Hierin is NA de groep die programma A niet gezien heeft. a is het aantal dat zowel A als B gezien heeft, b is het aantal dat Na en B gezien heeft, c is het aantal dat A en NB gezien heeft, d is het aantal dat NA en NB gezien heeft. In tabel:

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

Dus $b + d$ is het aantal dat programma A niet gezien heeft.

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 $\Rightarrow$ In formule: $\frac{a}{c} > \frac{b}{d} \therefore ad > bc \Rightarrow ad - bc > 0$.

Er is een negatief verband als tussen A en B $\Rightarrow \frac{a}{c} < \frac{b}{d} \therefore ad - bc < 0$.

Er is geen verband tussen A en B voor wat het kijkgedrag van de ondervraagden betreft als binnen de A-groep de verhouding B:NB, gelijk is aan de verhouding binnen de NA-groep $\Rightarrow$ In formule: $\frac{a}{c} = \frac{b}{d} \therefore ad = bc \Rightarrow ad - bc = 0$.

	A	NA
B	183	249
NB	654	453

	A	NA
B	345	678
NB	764	876

In nevenstaande twee tabellen zijn de letters door getallen vervangen.

Misschien zie je direct wat het verband tussen A en B is. Soms is het lastig, en dan zit er niets anders op om het verband uit te rekenen. Dan blijkt dat voor beide tabellen het verband tussen A en B negatief is.

- De associatiemaat R , definitie

De associatiemaat

	W	NW	
M	a	b	$a + b$
NM	c	d	$c + d$
	$a + c$	$b + c$	n

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

Gevolg: $-1 \leq R \leq 1$ (zonder bewijs).

Bewijs: De formule voor de associatiemaat R kwadrateren en uitschrijven, dan vind je het gevolg: $R^2 \leq 1 \rightarrow |R| \leq 1$.

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) in de 2×2 -tabel wordt aangegeven met de letter R (Zie bovenstaand kader).

Onderstaand een aantal getallenvoorbeelden:

Opgave 5 Associatie

	W	NW
M	25	9
NM	50	18

geen verband
R = 0

	W	NW
M	34	0
NM	41	27

positief verband
R = 0,429

	W	NW
M	7	27
NM	68	0

negatief verband
R = -0,85

	W	NW
M	28	6
NM	47	21

positief verband
R = 0,14

Opgave 6 Associatie:

1	2	3	4
2 3	3 2	20 30	21 14
3 2	2 3	30 20	14 21
R = -0,2	R = 0,2	R = -0,2	R = 0,2

5	6	7	8
1 4	5 0	0 5	2 1
4 1	0 5	3 0	3 0
R = -0,6	R = 1	R = -1	R = -0,45

In bovenstaande 8 tabellen, is de associatie maat berekend.

De getallen in tabel 3 zijn 10 keer zo groot als de getallen in tabel 1. De getallen in de vierde tabel zijn

7 keer zo groot als de getallen in tabel 2. Conclusie: R verandert niet als de verhoudingen tussen de getallen gelijk blijven.

a	b	
c	7	
20	10	

$R = 0$

Voor nevenstaande tabel is de associatie maat $R = 0$. Hoe a, b, c te vinden.

$$b + 7 = 10 \Rightarrow b = 3.$$

$$R = 0 \Rightarrow ad - bc = 0 \Rightarrow 7a - 3c = 0, \text{ en } a + c = 20 \Rightarrow a = 6 \text{ en } c = 14.$$

a	b		
c	d		
20	10	$R = 1$	

a	b		
c	d		
20	10	$R = -1$	

De associatiemaat van de linker tabel is 1, die van de

rechter tabel is -1.

$$\text{Voor } R = 1: ad - bc > 0, a + c = 20, b + d = 10$$

Met de formule voor $R = \frac{ad-bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} = 1, c + d = 30 - (a + b)$, vinden we:

$$1 = \frac{a(10-b)-b(20-a)}{\sqrt{200(a+b)[30-(a+b)]}} \Rightarrow a^2 - 20a - 20b + 2b^2 = 0. \text{ Op deze wijze hebben we een relatie tussen}$$

a en b gevonden. Dus: $a = 10 \pm \sqrt{150 - 2(b-5)^2}$. Dit resultaat is in overeenstemming met $0 \leq$

$a \leq 20$ en $0 \leq b \leq 10$. Gelet op $0 \leq a \leq 20$, moet gelden $0 \leq \sqrt{150 - 2(b-5)^2} \leq 10$.

De volgende aanname die we doen is dat, omdat a een geheel getal is, moet $\sqrt{150 - 2(b-5)^2}$, een geheel getal moet zijn. Daaruit volgt dat $150 - 2(b-5)^2$ een geheel getal is en tussen 0 en 100 moet liggen. Verder nemen we ook aan dat b een geheel getal is. Het resultaat is: $0 \leq 150 - 2(b-5)^2 \leq 100 \Rightarrow 0 \leq 75 - (b-5)^2 \leq 50$. Dit schrijven nog anders: $75 \geq (b-5)^2 \geq 25$

Met $0 \leq b \leq 10$ en b een geheel getal zijn er twee mogelijkheden: $b = 0$, of $b = 10$.

Als $b = 0$, hebben we twee waarden voor a : $a = 0$ of $a = 20$. Maar, als $a = 0$ en $b = 0 \Rightarrow R = 0$.

Dus met $b = 0$ geldt alleen $a = 20$. Dan hebben we met $a + c = 20, c = 0$, en $d = 10$.

Als $b = 10$, zijn er weer twee waarden voor a : $a = 0$ of $a = 20$. Voor $a = 0$, vinden we $c = 20$.

Omdat $b = 10$, geldt $d = 0$. Hieruit volgt $R < 0$. Voor $a = 20$, geldt $c = 0$. Omdat $b = 10$, geldt $d = 0$ met als totaal resultaat $R = 0$. De enige oplossing die voldoet is: $b = 0, a = 20, c = 0$, en $d = 10$.

Voor $R = -1$, geldt een vergelijkbare analyse. Ik geef alleen het resultaat: $a = 0, d = 0, c = 20$, en $b = 10$.

Opmerking:

In Oefenopgave 7 wordt aandacht besteed aan $R = \pm 1$. Door R uit te werken, zie je voor welke waarden van a, b, c of $d, R = \pm 1$ is.

• Links-of rechtshandig

Gegeven:

Een school met 1000 leerlingen

- 700 leerlingen in de onderbouw (Ob)
- 300 leerlingen in de bovenbouw (Bb)
- 100 linkshandige leerlingen (Lh)
- 900 rechtshandige leerlingen (Rh)

	Lh	Rh	
Ob	70	630	700
Bb	30	270	300
	100	900	1000

$$70 \times 270 = 30 \times 630 \Rightarrow R = 0 \rightarrow \text{Geen verband}$$

Met kansen:

	P(Lh)	P(Rh)	
P(Ob)	$\frac{70}{1000}$	$\frac{630}{1000}$	$\frac{700}{1000}$
P(Bb)	$\frac{30}{1000}$	$\frac{270}{1000}$	$\frac{300}{1000}$
	$\frac{100}{1000}$	$\frac{900}{1000}$	1

$$P(\text{Ob en Lh}) = \frac{70}{1000}$$

$$P(\text{Ob}) = \frac{700}{1000} \text{ en } P(\text{Lh}) = \frac{100}{1000} \rightarrow P(\text{Ob}) \times P(\text{Lh}) = \frac{700}{1000} \times \frac{100}{1000} = \frac{70}{1000}$$

→ De kansen P(Ob) en P(Lh) zijn onafhankelijk.

$$n = a + b + c + d = 70 + 630 + 30 + 270 = 1000,$$

en

$P(\text{Ob en Lh}) \rightarrow$ de kans op in de Onderbouw(Ob) en Linkshandig(Lh) zijn.

Onafhankelijke kansen en $R = 0$

	X	nX	
A	a	b	$a + b$
nA	c	d	$c + d$
	$a + c$	$b + d$	$n = a + b + c + d$

$$R = 0 \rightarrow ad = bc$$

Met kansen:

	P(X)	P(nX)	
P(A)	$\frac{a}{n}$	$\frac{b}{n}$	$\frac{a + b}{n}$
P(nA)	$\frac{c}{n}$	$\frac{d}{n}$	$\frac{c + d}{n}$
	$\frac{a + c}{n}$	$\frac{b + d}{n}$	1

Uitwerking

$R = 0 \Rightarrow$
onafhankelijke kansen

$$R = 0 \Rightarrow ad = bc$$

$$P(X) = \frac{a+c}{n}$$

$$P(A) = \frac{a+b}{n}$$

$$P(X \text{ en } A) = \frac{a}{n}$$

	P(X)	P(nX)	
P(A)	$\frac{a}{n}$	$\frac{b}{n}$	$\frac{a + b}{n}$
P(nA)	$\frac{c}{n}$	$\frac{d}{n}$	$\frac{c + d}{n}$
	$\frac{a + c}{n}$	$\frac{b + d}{n}$	1

$$P(X) \times P(A) = \frac{a+c}{n} \times \frac{a+b}{n} = \frac{(a+c) \cdot (a+b)}{n^2} = \frac{a^2 + ab + ac + bc}{n^2} =$$

$$= \frac{a^2 + ab + ac + ad}{n^2} = \frac{a(a+b+c+d)}{n^2} = \frac{a \cdot n}{n^2} = \frac{a}{n} = P(X \text{ en } A)$$

Lees bovenstaand bewijs van achter naar voren.

Opgave 9: Toevallige verbanden

H. trekt zes kaarten uit een volledig spel en let op het aantal harten(Ha) en het aantal plaatjes(Pl).

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

Vul de 2x2-tabel in en bereken R. Bij Opgave 6 staat de formule voor R.

Zie Tab.1 op de volgende pagina.

b) Hoe groot zou R geweest zijn als H. twee harten plaatjes, twee andere plaatjes en twee niet-harten plaatjes had getrokken? Zie onderstaande Tab. 2.

c) Geef een tabel waarbij $R=1$ en een tabel waarbij $R=-1$.

Zie de tabellen 3 en 4 hieronder.

	Ha	Nha	Tab.1
PI	0	2	
NPI	1	3	$R = -0.32$
	2	2	Tab.2
	0	2	$R = 0.5$
Tab.3	2	0	
	0	2	Tab.4
			$R = 1$
			$R = -1$
Tab.5	2	2	
$R=0$	1	1	Tab.6
			$R=0$
			R is onbepaald

Zoals je ziet kan R bij de trekking van zes kaarten allerlei waarden aannemen, die van het toeval afhangen. Zie Tab. 5 met $R=0$.

d) Soms kan R zelfs niet berekend worden. Zie de verdeling in Tab.6.

e) Hoe groot is R voor een volledig kaartspel?

Zie onderstaande Tabel.

	Ha	Nha	
PI	a=4	b=12	16
NPI	c=9	d=27	36
	13	39	52

Dus $R \Rightarrow ad - bc = 0 \Rightarrow R = 0$.

In dit vraagstuk is de nadruk gelegd op de mogelijkheid dat een steekproef toevallig de indruk kan wekken dat er een verband is tussen twee dingen, terwijl dat verband er in de populatie zelfs helemaal niet is. Vooral bij kleine steekproven spelen zulke toevalseffecten een rol

Hoe groter de som n van de getallen a , b , c , en d in de viervakjes, hoe kleiner de toevalseffecten.

In een kaartspel kan R alle mogelijke waarden aannemen.

Vuistregel

Is er geen verband, dan geldt meestal dat $-\frac{2}{\sqrt{n}} < R < \frac{2}{\sqrt{n}}$

en bijna nooit dat $R < -\frac{3}{\sqrt{n}}$ of $R > \frac{3}{\sqrt{n}}$.

Is $R < -\frac{3}{\sqrt{n}}$ of $R > \frac{3}{\sqrt{n}}$, dan mag je aannemen dat er een verband is.

Opgelet: maar als $|R| > \frac{2}{\sqrt{n}}$ dan nog $|R| \leq 1$.

Er is een situatie waar R onbepaald is. Voorbeeld: $a = 0$, $c = 0$ en $b = d$.

R verliest dan de betekenis.

Oefenen en uitgewerkte Oefenopgaven.

Opgave 7: Onafhankelijkheid en kansen.

De associatiemaat R :

$$R = \frac{ad-bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} \Rightarrow R^2 = \frac{(ad)^2 + (bc)^2 - 2adbc}{(ad+bc)^2 + a^2(bc+bd+cd) + b^2(ac+ad+cd) + c^2(ab+ad+bd) + d^2(ab+ac+bc)}$$

Gelet op de teller, $(ad - bc)^2$ en de eerste term in de noemer, $(ad + bc)^2$, is $R^2 < 1$. $R = 1$, treedt

op bij bijzondere combinaties: $b = 0$ en $c = 0 \Rightarrow R^2 = \frac{(ad)^2}{(ad)^2} = 1$.

a) 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;

PI en NPI: plaatjes en niet plaatjes.

Neem je twee van die indelingen, dan heb je ook een 2×2 – tabel. Van die tabel kun je weer de associatiemaat R berekenen. Er blijkt dat R in verband gebracht kan worden met het begrip onafhankelijkheid uit de kansrekening.

In onderstaande 4 tabellen zijn de getallen ingevuld voor de aangegeven groepen. Ook de associatiegraad is uitgerekend en ingevuld.

	Ro	NRo	
Ha	13	0	13
NHa	13	26	39
	26	26	52

$$R = 0.58$$

	PI	NPI	
Aa	4	0	4
NAa	12	36	48
	16	36	52

$$R = 0.44$$

	Ro	NRo	
PI	8	8	16
NPI	18	18	36
	26	26	52

$$R = 0$$

	Ha	NHa	
Aa	1	3	4
NAa	12	36	48
	13	39	52

$$R = 0.04$$

b) Bij de tabel met $R=0$, zie je dat $\frac{a}{d} = \frac{b}{c}$. De positie van a, b, c en d , is nog eens in de tabel van opgave 8 weergegeven.

c) In het kaartspel zijn rang en soort onafhankelijk. Als iemand uit een volledig spel (52 kaarten) een kaart trekt en vertelt dat het een aas is, dan weet je nog niets over de kleur: dus onafhankelijk. Als een kaart ontbreekt, krijg je bijvoorbeeld bij “rood” en “plaatje” al een positief of een negatief verband. Hoe is dat verband als harten 7 ontbreekt? Zie bovenstaande derde tabel links onder. Vervang 18 door 17. Het resultaat: $R=0.01$.

Opgave 8: Onafhankelijkheid en kansen, eerste helft van de opgave, de tweede helft staat in bovenstaande tekst van Blok 15 Les 1. Dit is een opgave waarbij the associatiemaat $R = 0$. Iemand trekt een kaart uit een volledig spel.

Hoe groot is $P(Ha)$, de kans op harten? Er zijn 52 kaarten in een volledig spel en er zijn 13 hartenkaarten. Dus, $P(Ha) = \frac{13}{52} = \frac{1}{4}$.

Hoe groot is de kans $P(PI)$ op een plaatje? Er zitten $4 \cdot 4$ plaatjes in een spel. Dus, $P(PI) = \frac{16}{52} = \frac{4}{13}$.

Bereken ook $P(Ha \text{ en } PI)$, de kans op een hartenplaatje. Er zitten 4 hartenplaatjes in een spel van 52 kaarten. Dus,

$$P(Ha \text{ en } PI) = \frac{4}{52} = \frac{1}{13}.$$

Wat is het verband tussen de drie zojuist uitgerekende kansen? De kansen zijn onafhankelijk.

Dat wil zeggen: $P(Ha) \cdot P(PI) = \frac{1}{4} \cdot \frac{4}{13} = \frac{1}{13} = P(Ha \text{ en } PI)$.

Maak een tabel van alle kansen. Er ontbreken nog 3 kansen: de kans $P(NPI)$, de kans op geen plaatje voor harten, de $P(NPIHa)$, de kans op geen harten plaatje en $P(NPINHa)$.

Tabel van de kansen:

	$P(Ha)$	$P(NHa)$	$P(Totaal)$
PI	$1/13, a/n$	$12/52, b/n$	$4/13$
NPI	$9/52, c/n$	$27/52, d/n$	$9/13$
$P(Totaal)$	$1/4$	$39/52$	1

In de kansrekening heten twee gebeurtenissen A en B onafhankelijk als $P(A \text{ en } B) = P(A) \cdot P(B)$.

In dit geval is de associatiemaat $R = 0$:

$$\frac{a}{n} \cdot \frac{d}{n} = \frac{b}{n} \cdot \frac{c}{n} \Rightarrow R = ad - bc = 4 \cdot 27 - 9 \cdot 12 = 0.$$

Opgave 10: Aspirine en hartinfarct

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 bloedklonteringen 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”.

a) Waarom werkt men bij een dergelijk onderzoek met foppillen? Er wordt met Placebo's gewerkt om een vergelijkingsgroep te hebben om conclusies te trekken over de werkzaamheid van aspirine. Nota Bene: aspirine $\Rightarrow$ ascorbinezuur, dus geen paracetamol.

b) Zet de resultaten uit in een 2×2 tabel en bereken de associatiemaat R

	Aspirine	Placebo	
Hartinf	104(a)	189(b)	293
Geen Inf	10896(c)	10811(d)	21707
	11000	11000	22000

De associatiemaat R:

$$R = \frac{ad-bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}} \Rightarrow \frac{104 \cdot 10811 - 189 \cdot 10896}{\sqrt{11000 \cdot 11000 \cdot 293 \cdot 21707}} = -0.03.$$

c) In de steekproef zijn de aantallen hartinfarcten nogal verschillend. Wat denk je, is dit verschil toevallig?

Om dit te beoordelen zijn er een aantal vuistregels:

Een steekproef kan toevallig de indruk 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 toevalseffecten een rol.

Hoe groter de som n van de getallen a, b, c en d in de viervakjes, hoe kleiner de toevalseffecten.

Vuistregel:

Als er geen verband is, krijg je maar zelden een R die meer dan $\frac{2}{\sqrt{n}}$ afwijkt van 0 en bijna nooit een R

die meer dan $\frac{3}{\sqrt{n}}$ van 0 afwijkt. Als je toch een dergelijke waarde vindt, mag je aannemen dat er een verband is. Tenzij de proef verkeerd is uitgevoerd: niet aselekt.

Bij bovenstaand onderzoek bedraagt $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{22000}} = 0.013 \Rightarrow$ geen verband.

Opgave 11: Vervolg opgave 10,

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

Als er geen verband zou zijn, dan kunnen we de groep van 293 hartinfarcten beschouwen als een aselekte steekproef uit een populatie van 22000 artsen.

Natuurlijk is dit een steekproef zonder terugleggen.

a) Waarom mogen we de steekproef toch beschouwen als een proef met terugleggen?

293 $\ll$ 22000. Dat wil zeggen $22293 \approx 22000$.

b) Ongeveer hoeveel procent van de steekproef zal, als er geen verband zou zijn, uit aspirine slikkers bestaan?

Bij benadering 50%. Dit is vrij grof. Want, 104 van 293 ligt dicht bij 30%. We doen het er mee.

Waarom? Dat wordt duidelijk in de volgende vraag.

c) Het probleem dat we hier onderzoeken kunnen we, met de benadering van 50%, opvatten als het 293 keer gooien van een munt. Het aantal keren kop, noem dat X , kan ook meer of minder afwijken van de helft. Vandaar de benadering met 50%.

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

Het is een binomiale verdeling met $p = 0.5$ en $n = 293$.

d) Bereken $E(X)$ en $Sd(X)$.

$$E(X) = X \cdot p = \frac{293}{2} = 146.5,$$

$$Sd(X) = \sqrt{X \cdot p \cdot (1 - p)} = 8.56.$$

e) Hoeveel keer de standaard $Sd(= \sigma)$ wijkt 104 af van $E(X)$ af, met andere woorden wat is de z -waarde van 104? Hieruit maak je op dat we de binomiale verdeling benaderen met de normale verdeling. Nota Bene: 104 → hartinfarcten bij aspirine.

$$\text{De } z\text{-waarde: } z = \frac{X - \mu}{\sigma} = \frac{X - E(X)}{Sd} = \frac{104 - 146.5}{8.56} \cong -5. \text{ Dus 5 keer } Sd.$$

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

Bij de normale verdeling weten we het percentage voor $-3 \leq z \leq 3 \Rightarrow 99,7\%$. Het resterende oppervlak is dan 0.3% . Dus voor $z = -5$ is het resterende oppervlak nog minder dan $0.3/2 = 0.15\%$.

g) Is het verschil tussen de aantallen hartinfarcten toevallig?

Gelet op wat we gevonden hebben is de kans dat het verschil tussen hartinfarcten, met en zonder aspirine, op toevalligheid berust nagenoeg uitgesloten.

h) Hoeveel keer $\frac{1}{\sqrt{22000}} = 0.0067$, is de associatiegraad R van nul verwijderd?

In Opgave 10 vonden we voor $R = -0.03$. Dus R is 4 á 5 keer van 0 verwijderd.

Caveat: Opgave 10 en 11 zijn een oefeningen om statistische gegevens te onderzoeken. Het onderzoek is van 1987. Inmiddels kunnen er nieuwe gegevens zijn. Dus gebruik de gegevens uit de opgave alleen om je kennis van statistiek te toetsen.

Blok 15 Les 2 Van 2 · 2 naar veel · veel

- Deze Les sluit aan bij de paragrafen 12 tot en met 23 van de tekst Correlatie en Regressie van de Wageningse Methode 1989
- Een intelligentie test

Opgave 12:

Test A: taalvaardigheid

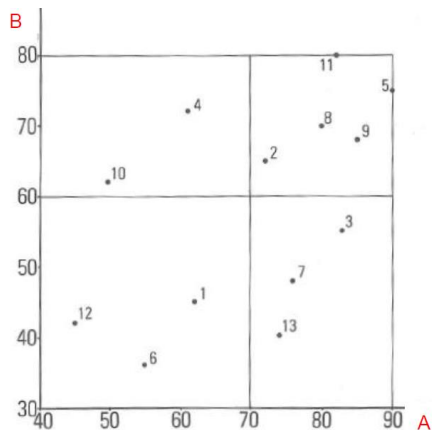
Test B: technisch inzicht

13 leerlingen,

De resultaten in tabel: x_i en y_i zijn de scores.

	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 resultaten in beeld, de scores zijn nog eens als punten weergegeven in onderstaande Figuur:



Vraag: geeft een hoge score op A ook een hoge score op B?

Je zou denken dat dit inderdaad het geval is, want er lijkt enig verband tussen de componenten van de test, A en B.

Een stap verder in de analyse. Eén manier om het verband te onderzoeken is er een 2×2 tabel van te maken.

Dit doe je door de testresultaten in 4 groepen te verdelen. Zie onderstaande tabel. We trekken ergens een verticale en een horizontale lijn, bijvoorbeeld bij $x = 70$, en $y = 60$ (Zie bovenstaande figuur).

	$x < 70$	$x \geq 70$	
$y \geq 60$	2, a	5, b	7
$y < 60$	3, c	3, d	6
	5	8	13

De 13 personen worden over de tabel verdeeld.

De associatiemaat $R = \frac{2 \cdot 3 - 5 \cdot 3}{\sqrt{5 \cdot 8 \cdot 7 \cdot 6}} = -0.22$ (zie Les 1 van dit blok). De grenzen voor de invulling van de tabel zijn willekeurig gekozen. Dien ten gevolge heeft R niet veel waarde.

Andere verdelingen zijn uiteraard mogelijk. Zelfs een verdeling met $R = 0$ en toch een positief verband. Het voorgaande maakt duidelijk dat er niet alleen gelet moet worden waar de punten liggen maar ook hoe de punten ten opzichte van elkaar liggen.

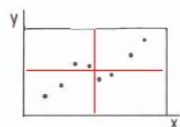
In onderstaande Figuur is nog een mogelijke verdeling weergegeven. Welk verband (positief, negatief of geen) is er tussen x en y bij onderstaande puntenwolk?

Het lijkt op een positief verband.

Breng met een horizontale lijn en een verticale lijn een verdeling in de puntenwolk aan, zodanig dat er in elk vakje 2 punten komen.

Wat is R ?

We weten bij deze verdeling dat: $ad - bc = 4 - 4 = 0 \Rightarrow R = 0$.



Een positief verband en toch is $R = 0$.

Met een andere verdeling kun je R ook aanpassen.

Maar uit bovenstaande figuur kun je niet concluderen, omdat $R = 0$, dat er geen verband zou zijn.

Einde opgave 12.

Uit bovenstaande opgave wordt duidelijk dat we er niet alleen moeten opletten in welke hoek(of vak) de punten liggen. We zullen er ook naar moeten kijken hoever de punten in de hoeken liggen, hoe ze ten opzichte van elkaar liggen. Spreiding speelt immers ook en

rol. Eerst wordt de spreiding in de x- en de y-waarden apart bepaald. Als maat voor de spreiding gebruiken we de standaardafwijking. De standaardafwijking Sd is de wortel uit de variantie.

- **Variantie**

We gaan gebruik maken van gemiddelden, variantie en spreiding. In het blok over de normale verdeling zijn deze begrippen besproken.

Zie, o.a., Blok 5(www.leennoordzij.me), Wiskunde D Online Summary 4 VWO-Blok1-7)

Een rij waarnemingen $x_1, x_2, \dots, x_n$

Het **gemiddelde** is: $\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$

De **variantie** is een maat voor de afwijkingen ten opzichte van het gemiddelde.

$$\begin{aligned} \text{Var}(x_1, x_2, \dots, x_n) &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n} = \\ &= \frac{(x_1)^2 + (x_2)^2 + \dots + (x_n)^2}{n} - \bar{x}^2 \end{aligned}$$

De **standaarddeviatie** $SD(x_1, x_2, \dots, x_n) = \sqrt{\text{Var}(x_1, x_2, \dots, x_n)}$

Dus het gemiddelde:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

variantie $\text{Var}(x_i)$:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2.$$

Opgave 14:

De resultaten van de intelligentie test 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

a) Bereken de gemiddelde en de standaarddeviatie

van de x-waarden. Dus: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \Rightarrow$ met de rekenmachine $\bar{x} = 70.6$. De standaarddeviatie:

$$Sd(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2} \Rightarrow Sd = 13.5.$$

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

Dus: $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \Rightarrow \bar{y} = 58,3$. De standaarddeviatie: $Sd(y) = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - (\bar{y})^2} = 14.2$.

Einde opgave 14.

Om de testresultaten verder te kunnen beoordelen hebben we ook de covariantie nodig. Zie het volgende onderwerp.

- **Covariantie**

Twee rijen waarnemingen $x_1, x_2, \dots, x_n$ en $y_1, y_2, \dots, y_n$

Met **gemiddelden** $\bar{x}$ en $\bar{y}$

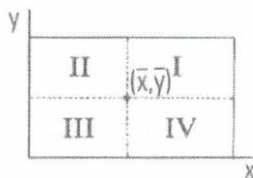
De **covariantie** is een maat voor de afwijkingen ten opzichte van beide gemiddelden.

$$\begin{aligned} \text{Covar}(x, y) &= \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y})}{n} = \\ &= \frac{x_1 \cdot y_1 + x_2 \cdot y_2 + \dots + x_n \cdot y_n}{n} - \bar{x} \cdot \bar{y} \end{aligned}$$

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

De tweede uitdrukking hierboven voor de covariantie is eenvoudiger dan de eerste. Bij de berekening kan volstaan worden met veel minder getallen, dus minder fouten.

Opgave 15:



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

a) In welke gebieden liggen de punten die een positieve bijdrage leveren aan de covariantie en aan welke formule kun je dat het beste zien? Voor de covariantie vind je hierboven twee formules. Uit de eerste formule, $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$, is direct te zien als (x_i, y_i) groter zijn dan $(\bar{x}, \bar{y})$, deze punten (x_i, y_i) een positieve bijdrage leveren aan de covariantie. Genoemde punten liggen in het gebied I. Dit geldt ook als (x_i, y_i) kleiner zijn dan $(\bar{x}, \bar{y})$. Deze punten liggen in het gebied III.

b) In welke gebieden liggen de punten die een negatieve bijdrage leveren aan de covariantie? Deze punten liggen in de gebieden II en IV. Bijvoorbeeld in gebied II: $x_i < \bar{x}$ en $y_i > \bar{y}$.

c) Welke bijdrage leveren de punten die op de stippellijn liggen aan de covariantie? Voor die punten geldt: $x_i = \bar{x}$ of $y_i = \bar{y}$. Deze punten leveren dus geen bijdrage.

d) Is de covariantie bij de resultaten van de intelligentie test, opgave 12 (zie boven), positief of negatief? Als je naar de ligging van de punten kijkt, lijkt het er op dat de covariantie positief is.

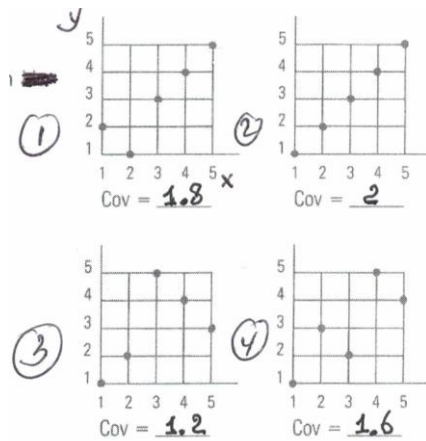
e) Bereken de covariantie bij de punten (x_i, y_i) van de intelligentie test, opgave 12. De gegevens staan in de tabel bij bovenstaande intelligentie test.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{13} x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^{13} y_i, \text{ en } \text{Covar} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Resteert een invuloefening met als resultaat: $\text{Covar} = 104.2$.

f) Bereken bij de 4 onderstaande figuren de Covar.

Ook dit is een invuloefening. De coördinaten zijn gegeven in de figuren. Wat op valt is, dat voor de 4 figuren geldt: $\bar{x} = \bar{y}$.



Einde opgave 15.

• De correlatiecoëfficiënt

De covariantie is een maat voor de samenhang tussen x - en y - waarden, maar is niet optimaal.

De covariantie is 'schaalgevoelig'.

Als je de x - waarden met 10 vermenigvuldigd, wordt de covariantie 10 keer zo groot (zie tweede formule).

Als je alle y - waarden door 12 deelt, wordt de covariantie 12 keer zo klein.

De correlatiecoëfficiënt is $r = \frac{\text{Cov}(x,y)}{\text{SD}(x) \cdot \text{SD}(y)}$

Laat zien dat r niet schaalgevoelig is.

Uitwerking

Als je de x - waarden met 10 vermenigvuldigd, wordt $\text{Cov}(x,y)$ 10 keer zo groot.

$\text{Var}(x)$ wordt 100 keer zo groot en dus $\text{SD}(x)$ weer 10 keer zo groot.

In de breuk valt dit tegen elkaar weg.

Opgave 16, vervolg op opgave 15(zie hierboven):

De covariantie is een maat voor de samenhang tussen de x - en y -waarden. Er kleven echter nadelen aan de covariantie. Zie de schaalgevoeligheid hierboven.

Dit volgt direct uit de formules:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{13} x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^{13} y_i, \text{ en } \text{Cov} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

en

$$\text{SD}(x) = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2}, \text{ en } \text{SD}(y) = \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - (\bar{y})^2}.$$

een maat voor de samenhang die niet schaalgevoelig is voor de x - en y - waarden is de correlatiecoëfficiënt.

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

Ook uit deze formule volgt direct dat de correlatiecoëfficiënt niet schaalgevoelig is.

a) bereken de correlatiecoëfficiënt voor de 13 punten van de intelligentietest.

Bij opgave 15(zie hierboven) is uitgerekend $\text{Cov} = 104.2$.

Bij opgave 14(zie hierboven) zijn $\text{SD}(x) = 13.5$ en $\text{SD}(y) = 14.2$ uitgerekend.

b) Bereken bij de 4 figuren van opgave 15(zie hierboven) de correlatiecoëfficiënt. De covariantie is voor elk van de 4 figuren al uitgerekend(opgave 15). Wat opvalt in de 4 figuren is de symmetrie: $\Rightarrow \bar{x} = \bar{y}$.

Ook zonder rekenmachine vind je: $\bar{x} = \bar{y} = 3$. Verder, $\text{SD}(x) = \text{SD}(y) = \sqrt{2}$.

Met deze gegevens vind je de vier correlatiecoëfficiënten: $r_1 = 0.9$, $r_2 = 1$ (zonder iets uit te rekenen), $r_3 = 0.6$ en $r_4 = 0.8$.

Einde opgave 16.

- Voor de correlatiecoëfficiënt geldt: $-1 \leq r \leq 1$
- $r = 1$ de punten liggen op één lijn met positieve richtingscoëfficiënt.
- $r = -1$ de punten liggen op één lijn met negatieve richtingscoëfficiënt

• Rangnummers

Nog eens de scores van de intelligentietest.

Aan de scores wordt een rangnummer toegekend.

Bijv.: persoon 12, $x_{12} = 45$, heeft rangnummer 1, etc.

Scores van de intelligentietest

	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

Rangnummers van de intelligentietest

	i	1	2	3	4	5	6	7	8	9	10	11	12	13
A	x_i	5	6	11	4	13	3	8	9	12	2	10	1	7
B	y_i	4	8	6	11	12	1	5	10	9	7	13	3	2

De $r(a,b)$ is de **rangcorrelatiecoëfficiënt** van Spearman als de data gegeven zijn als rangnummers (a,b) .

Als zowel scores (x,y) als rangnummers (a,b) gegeven zijn, dan geldt vaak dat $r(x,y) \approx r(a,b)$

Opmerking: Als blijkt dat $Sd(x) = Sd(y) \rightarrow Sd(x) \cdot Sd(y) \rightarrow (Sd(x))^2$.

Dit maakt de berekeningen nauwkeuriger, want je rekent de wortel niet uit.

Waarom rangnummers?

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. Spearman was een psycholoog. Vandaar.

Oefenen en uitgewerkte Oefenopgaven.

Opgaven 12, 14, 15 en 16 zijn in bovenstaande tekst opgenomen.

Opgave 17:

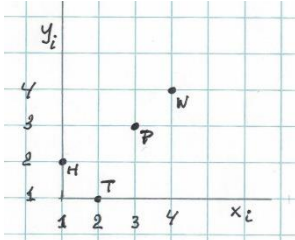
De correlatie coëfficiënt r is in bovenstaande les gedefinieerd. De representatie van r kan beknopter door gebruik te maken van de z-waarde van x_i : $z(x_i) = \frac{x_i - \bar{x}}{SD(x)}$, dus de afwijking van x_i van $\bar{x}$ gemeten in standaard afwijkingen. (Opgelet: met de z-waarde wordt geen nieuwe informatie gegenereerd. Zie ook Blok 5 Les 3: de z-waarde).

Op een sportdag blijken H, T, P en W 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 onderstaande tabel vind je de resultaten samengevat:

	x_i	y_i
H	1	2
T	2	1
P	3	3
W	4	4

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



Is het verband tussen x en y positief of negatief?

Het verband tussen x en y is positief.

Er zijn 4 punten: (1,2); (2,1); (3,3); (4,4) .

$$\bar{x} = \frac{1}{4} \sum_{i=1}^4 x_i \Rightarrow \bar{x} = 2.5, \text{ zo ook } \bar{y} = 2.5.$$

$$\text{Cov}(x, y) = \frac{1}{4} \sum_{i=1}^4 x_i y_i - \bar{x} \cdot \bar{y} = \frac{29}{4} - 6.25 = 1.$$

Nu de standaard deviaties. Gelet op de getallen x_i en y_i , zie je dat $\text{Sd}(x) = \text{Sd}(y)$. Daar ga je gebruik van maken, want:

$$r = \frac{\text{Cov}(x, y)}{\text{Sd}(x)\text{Sd}(y)} = \frac{\text{Cov}(x, y)}{(\text{Sd}(x))^2}.$$

$$(\text{Sd}(x))^2 = \text{Var}(x) = \frac{1}{4} \sum_{i=1}^4 (x_i)^2 - (\bar{x})^2 = \frac{1}{4} (1^2 + 2^2 + 3^2 + 4^2) - (2.5)^2 = 1.25.$$

Je rekent dus niet $\text{Sd}(x)$ uit. Je verricht dus twee handelingen minder: wortel trekken en daarna weer kwadrateren $\Rightarrow$ kans op fouten en onnauwkeurigheden.

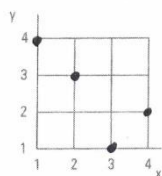
De correlatiecoëfficiënt: $r = \frac{\text{Cov}(x, y)}{(\text{Sd}(x))^2} = \frac{1}{1.25} = 0.8$. Hiermee is het positieve verband nog eens bevestigd.

Opgave 18, vervolg opgave 17:

Neem nu bij het hoogspringen een andere punten verdeling:

	x_i	y_i
H	1	4
T	2	3
P	3	1
W	4	2

a) Neem de punten weer op in een punten diagram:

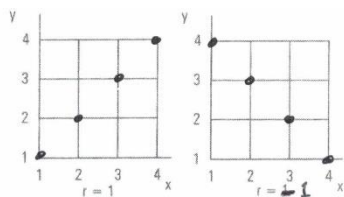


Het verband tussen x en y is nu negatief.

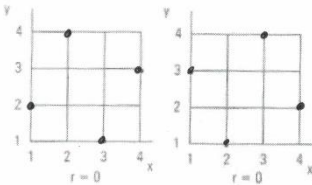
b) We houden steeds dezelfde uitslag bij het verspringen. Op hoeveel verschillende manieren kunnen er rangnummers voor het hoogspringen aan gekoppeld worden? Rangnummers: een gekozen volgorde met oplopende nummers.

Je kunt 1, 2, 3, 4 op 4! manieren rangschikken, permutaties dus zonder terugleggen.

c) Er is één manier waarbij $r = 1$, en er is één manier voor $r = -1$:



d) Er zijn twee manieren waarbij $r = 0$:



Er zijn nog twee manieren. Daarbij laat je ook toe dat er bijvoorbeeld twee 2^e plaatsen zijn en twee 3^e plaatsen zijn. Deze laat ik achterwege.

Nota Bene:

In de bovenstaande figuren bij vraag c) $r = 1$ en $r = -1$ is de richtingscoëfficiënt van de rechte lijn door de punten 1 en -1 respectievelijk.

$$r = 1 \Rightarrow x_i = y_i, \text{ en } \bar{x} = \bar{y} \Rightarrow \text{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \text{ en } \text{Sd}(x)\text{Sd}(y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \Rightarrow r = 1.$$

$$r = -1 \Rightarrow \bar{x} = \bar{y}, \text{ en } y_i = -x_i + 2\bar{x} \Rightarrow \text{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(-x_i + 2\bar{x} - \bar{x}) = -\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \text{ en } \text{Sd}(x)\text{Sd}(y) = \frac{1}{n} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2} = \frac{1}{n} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (-x_i + 2\bar{x} - \bar{x})^2} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \Rightarrow r = -1.$$

Vraag: is het relevant dat $x_i = y_i$, en $\bar{x} = \bar{y}$, voor $r = 1$? Of $\bar{x} = \bar{y}$, en $y_i = -x_i + 2\bar{x}$, voor $r = -1$?

Laten we eerst $r = 1$ onderzoeken. Merk op r is de correlatiecoëfficiënt en geen richtingscoëfficiënt. De punten (x_i, y_i) liggen op een rechte lijn, dus $y_i = ax_i \Rightarrow \bar{y} = \frac{1}{n} \sum_{i=1}^n ax_i = a\bar{x} \Rightarrow \text{Cov}(x, y) =$

$$= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(ax_i - a\bar{x}) = \text{Cov}(x, y) = \frac{a}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \text{ Verder, } \text{Sd}(x)\text{Sd}(y) = \frac{1}{n} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (ax_i - a\bar{x})^2} = \frac{a}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \therefore r = 1.$$

Nu, $r = -1$. De punten (x_i, y_i) , liggen op de rechte lijn $y = -ax + c$ en $y_i = -ax_i + c \Rightarrow$

$$\Rightarrow \text{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(-ax_i + c + a\bar{x} - c) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(-ax_i + a\bar{x}) = -\frac{a}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \text{ Wat betreft de standaard deviatie vind je met het gegeven } y_i = -ax_i + c, \text{ en } \bar{y} = -a\bar{x} + c : \text{Sd}(x)\text{Sd}(y) = \frac{a}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \therefore r = -1.$$

In **Opgave 19** wordt aandacht besteed aan de waarden van de correlatiecoëfficiënt voor verschillende puntenwolken. Je vindt altijd een uitkomst die tussen -1 en $+1$ ligt. De correlatiecoëfficiënt is alleen dan 1 als de punten op een rechte lijn liggen met positieve richtingscoëfficiënt. De correlatiecoëfficiënt is alleen -1 als de punten op een rechte lijn liggen met negatieve richtingscoëfficiënt. De correlatiecoëfficiënt kan onbepaald zijn: $\text{Sd}(x)\text{Sd}(y) = 0$ en $\text{Cov}(x, y) = 0$.

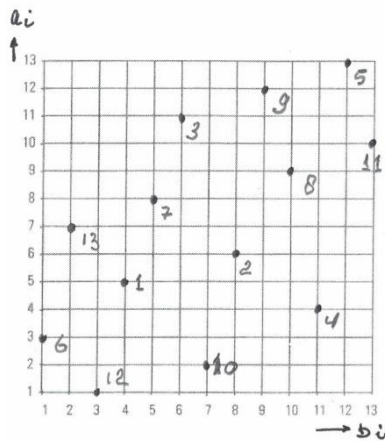
Opgave 20 Rangnummers:

Ook bij de intelligentietest (Opgave 12 Zie begin van Les 2) kunnen we de proefpersonen ordenen naar hun scores op onderdeel A en naar 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.

In onderstaande tabel is het overzicht gegeven van scores en rangnummers (a_i, b_i) :

	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

a) De puntenwolk (a_i, b_i) van de rangnummers is weergegeven in onderstaande figuur. Bij elk van deze punten (a_i, b_i) is het nummer i van de proefpersoon bijgeschreven.



b) Vergelijking van de punten wolk (x_i, y_i) uit opgave 12 (zie boven) met de puntenwolk (a_i, b_i) is aanleiding tot de conclusie dat er niet veel veranderd is.

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

$\bar{a} = \bar{b}$: rangnummers, dus een rekenkundige reeks $\rightarrow \bar{a} = \frac{1}{n} \sum_{i=1}^n a_i = \frac{1}{13} \cdot (13 + 1) \cdot \frac{13}{2} = 7$.

Opmerking: de rangnummers vormen een oplopende reeks. De rangnummers zijn allen ongelijk. Is dat bijzonder? Het is toch niet onmogelijk dat twee personen dezelfde score hebben bij de intelligentie test? Ga er van uit dat: als we n punten (x_i, y_i) hebben en alle x -waarden en 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 aan de y -waarden.

$Var(a) = Var(b)$: $Var(a) = \frac{1}{n} \sum_{i=1}^n (a_i - \bar{a})^2 = \frac{1}{n} \sum_{i=1}^n a_i^2 - \bar{a}^2$. Het doet er niet toe welke van de twee je kiest. Ik kies de eerste. Dus $Var(a) = \frac{1}{n} \sum_{i=1}^n (a_i - \bar{a})^2 = \frac{1}{13} (36 + 25 + 16 + 9 + 4 + 1) = 14$. Toch iets gemakkelijker dan de tweede. Je hebt geen rekenmachine nodig.

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

$$Cov(a, b) = \frac{1}{n} \sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{b}) = \frac{1}{n} \sum_{i=1}^n (a_i - \bar{a})(b_i - \bar{a}) = \frac{1}{n} \sum_{i=1}^n a_i b_i - \bar{a}^2 = \frac{1}{13} \cdot 736 - 49 = 7.6.$$

De correlatiecoëfficiënt $r = \frac{Cov(a, b)}{Sd(a)Sd(b)} = \frac{Cov(a, b)}{Var(a)}$, want $Sd(a) = Sd(b)$. We hebben alle ingrediënten:

$$r = \frac{Cov(a, b)}{Var(a)} = \frac{7.6}{14} = 0.54.$$

e) Het voordeel van het werken met rangnummers is dat het gehele getallen zijn.

f) Als je nu nog eens de correlatiecoëfficiënt uitrekent van de intelligentie test (een invuloefening) en deze vergelijkt met $r = 0.54$ van de rangnummers, dan vind je dat de correlatiecoëfficiënten dicht bij elkaar liggen. Opgelet: dit verschil kan in sommige situaties wel aanzienlijk zijn.

Opgave 21: Rangnummers(vervolg)

De opmerking in opgave 20 wordt gebruikt.

a) Samenvatting van opgave 20: bereken $\bar{a}$.

a_i zijn rangnummers, 1, 2, 3, 4, etc.: $\sum_{i=1}^n i = n(n+1)/2$.

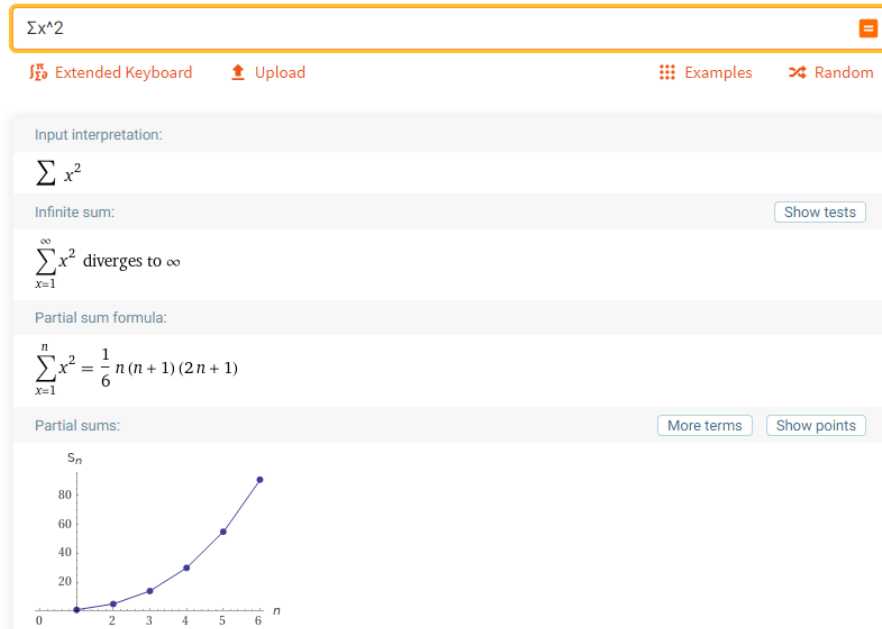
Dus $\bar{a} = (n+1)/2$. Als we een puntenwolk (a_i, b_i) hebben, geldt $\bar{a} = \bar{b}$.

b) We hebben in de vorige opgave $Var(a)$ berekend. In de formule van de variantie kom je de term $\sum_{i=1}^n a_i^2 = \sum_{i=1}^n i^2$, tegen. Dit lijkt een eenvoudige uitdrukking, maar de som is niet zo eenvoudig te vinden.

Je kunt die som opzoeken. Laten we dat doen: $\sum_{i=1}^n i^2 = \frac{1}{6} \cdot n \cdot (n+1) \cdot (2n+1)$. Ik heb de formule opgezocht in:

Beknopt Leerboek der Differentiaal- en Integraalberekening, Dr. N. G. De Bruijn, Tweede druk 1958. De Bruijn was toen Hoogleraar in Eindhoven. De Universiteit in Eindhoven (toen TH Eindhoven) bestond net 1 jaar.

Je kunt de formule voor $\sum_{i=1}^n i^2$, ook vinden met gebruik van de WolframAlpha-app.



Je ziet hoe eenvoudig dit is in dit screenshot. In het programmeer gedeelte heb ik slechts $\sum x^2$ ingevoerd.

Je wilt natuurlijk weten of de formule $\sum_{i=1}^n i^2 = \frac{1}{6} \cdot n \cdot (n+1) \cdot (2n+1)$ goed is. Het instrument daartoe is volledige inductie. Je neemt aan dat de formule goed is voor n , nadat je hebt vastgesteld dat de formule goed is voor $n = 1, 2$.

Geldt de formule ook voor $n+1$?

$$\begin{aligned} \sum_{i=1}^{n+1} i^2 &= \sum_{i=1}^n i^2 + (n+1)^2 = \frac{1}{6} \cdot n \cdot (n+1) \cdot (2n+1) + (n+1)^2 = \frac{1}{6} \cdot (n+1) \cdot [n \cdot (2n+1) + 6 \cdot (n+1)] = \\ &= \frac{1}{6} \cdot (n+1) \cdot [2n^2 + 7n + 6] = \frac{1}{6} \cdot (n+1) \cdot (n+2) \cdot (2n+3) = \frac{1}{6} \cdot (n+1) \cdot (n+2) \cdot [2(n+1) + 1], \text{ QED.} \end{aligned}$$

c) Wat vind je nu voor $Var(a) = \frac{1}{n} \sum_{i=1}^n i^2 - \bar{a}^2$?

De ingrediënten hebben we: $\frac{1}{n} \sum_{i=1}^n i^2 = \frac{1}{6} \cdot (n+1) \cdot (2n+1)$, en $\bar{a}^2 = (\frac{n+1}{2})^2 \Rightarrow$

$$\begin{aligned} \Rightarrow Var(a) &= \frac{1}{6} \cdot (n+1) \cdot (2n+1) - (\frac{n+1}{2})^2 = \frac{1}{12} \cdot (n+1) \cdot [2(2n+1) - 3(n+1)] = \frac{1}{12} \cdot (n+1)(n-1) \therefore \\ Var(a) &= \frac{1}{12} \cdot (n^2 - 1). \end{aligned}$$

d) Rest ons nog de rangorde correlatiecoëfficiënt $r(a, b) = \frac{Cov(a, b)}{Var(a)}$, zie opgave 20 hierboven.

We maken gebruik van: $(a_i - b_i)^2 = a_i^2 - 2a_i b_i + b_i^2 \Rightarrow a_i b_i = [a_i^2 + b_i^2 - (a_i - b_i)^2]/2$.

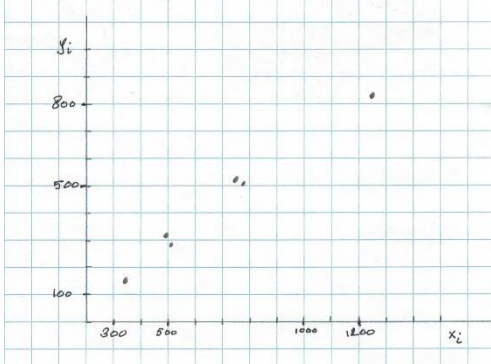
$$\begin{aligned} r(a, b) &= \frac{\frac{1}{n} \sum_{i=1}^n a_i b_i - \bar{a} \bar{b}}{Var(a)} = \frac{\frac{1}{n} \sum_{i=1}^n [a_i^2 + b_i^2 - (a_i - b_i)^2]/2 - \bar{a} \bar{b}}{Var(a)} = \frac{\frac{1}{n} \sum_{i=1}^n [a_i^2 + b_i^2]/2 - \bar{a} \bar{b}}{Var(a)} - \frac{1}{2n} \cdot \frac{1}{Var(a)} \cdot \sum_{i=1}^n (a_i - b_i)^2 = \\ &= \frac{\frac{1}{n} \sum_{i=1}^n [i^2 + i^2]/2 - \bar{a} \bar{b}}{Var(a)} - \frac{1}{2n} \cdot \frac{1}{Var(a)} \cdot \sum_{i=1}^n (a_i - b_i)^2 = \frac{Var(a)}{Var(a)} - \frac{1}{2n} \cdot \frac{1}{Var(a)} \cdot \sum_{i=1}^n (a_i - b_i)^2 \therefore \\ r(a, b) &= 1 - \frac{6}{n(n^2-1)} \cdot \sum_{i=1}^n (a_i - b_i)^2. \end{aligned}$$

Opgave 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 zijn opgenomen in onderstaande tabel. Ook zijn de rangnummers a en b weergegeven.

i	1	2	3	4	5	6
Kosten x_i	780	515	342	1257	496	768
Rangnummer a_i	5	3	1	6	2	4
Reacties y_i	509	298	176	819	312	520
Rangnummer b_i	4	2	1	6	3	5

In onderstaande grafiek is de punten wolk (x_i, y_i) weergegeven.



a) Zo, op het oog is er een positief verband tussen de advertentiekosten x en het aantal reacties y .

b) $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{6} \cdot 4158 = 693.$

$$Var(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2 = 87497.3.$$

$$Sd(x) = \sqrt{Var(x)} = 295.8.$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{6} \cdot 2634 = 439.$$

$$Var(y) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - (\bar{y})^2 = 43506.7.$$

$$Sd(y) = \sqrt{Var(y)} = 208.6.$$

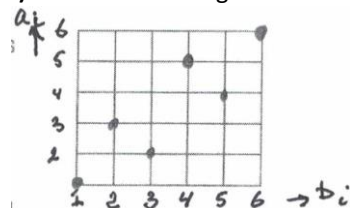
c) $Cov(x, y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}.$

$$\sum_{i=1}^n x_i y_i = 2193277.$$

$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y} = 61485.8.$$

d) De correlatiecoëfficiënt $r = \frac{Cov(x, y)}{Sd(x)Sd(y)} = \frac{61485.8}{295.8 \cdot 208.6} = 0.996.$

e) In onderstaande grafiek is de puntenwolk van de rangnummers weergegeven:



Ook de grafiek van rangnummers geeft een positief verband aan tussen de rangnummers van de advertentie kosten en de rangnummers van het aantal reacties.

f) De correlatiecoëfficiënt van de rangnummers. In opgave 21 is een elegante formule voor de correlatiecoëfficiënt afgeleid:

$$r(a, b) = 1 - \frac{6}{n(n^2-1)} \cdot \sum_{i=1}^n (a_i - b_i)^2 = 1 - \frac{6}{6 \cdot 35} \cdot 4 = 0.89.$$

g) Het is niet eenvoudig om de keuze tussen $r(x, y)$ en $r(a, b)$ te maken. Wel is het zo, dat de informatie om $r(x, y)$ te bepalen nauwkeuriger is dan de informatie voor $r(a, b)$.

Blok 15 Les 3 Correlatie en toeval

- Deze Les sluit aan bij de paragrafen 24 tot en met 33 van de tekst Correlatie en Regressie van de Wageningse Methode 1989
- Van de vierde naar de vijfde klas

Wat zegt het cijfer voor wiskunde in klas 4 VWO over het cijfer in klas 5 VWO?

De gegevens van 78 leerlingen.

Tabel, x-de horizontale as cijfers 4 t/m 9, y-de verticale as cijfers van 4 t/m 9

De gegevens (78 leerlingen)

		(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)						

2 leerlingen hadden een 4 in klas 4 en een 5 in klas 5.

7 leerlingen hadden een 7 in klas 4 en een 6 in klas 5.

Met behulp van de formules uit de vorige les, Blok 15 Les 2, kan de statistische analyse worden uitgevoerd.

De statistische analyse:

$$\bar{x} = 6,40, \bar{y} = 6,17$$

$$\text{Var}(x) = 0,98, \text{Var}(y) = 1,09$$

$$\text{Sd}(x) = 0,99, \text{Sd}(y) = 1,04$$

$$\text{Cov}(x, y) = 0,4722$$

$$r(x, y) = 0,4567$$

Conclusie:

- Gemiddelde gaat licht naar beneden
- Licht positief verband tussen cijfer in klas 4 en cijfer in klas 5.

Notatie: 6,40 → 6.40, etc.

Ook kan het verschil $d = y - x$ en het aantal malen dat het verschil d voorkomt, de frequentie f , in een tabel worden weergegeven.

$$\bar{d} = -0.23$$

$$\text{Var}(d) = 1.13$$

$$\text{Sd}(d) = 1.06$$

Ook kan bij elke leerling gekeken worden naar het aantal punten d dat de leerling vooruit is gegaan. Voorbeeld: 2 leerlingen zijn 2 punten vooruitgegaan: een 5 in klas 4 en een 7 in klas 5 $\Rightarrow$ 1; een 7 in klas 4 en een 9 in klas 5 $\Rightarrow$ 1. In de onderstaande tabel op de volgende bladzijde is het totale overzicht gegeven.

We weten dat $2 \cdot \text{Sd}(x) \cong 2$.

In bovenstaande tabel zie je dat bij 3 leerlingen de x-waarde meer dan twee keer van

$Sd(x)$ afwijkt, 2 keer een 4 en 1 keer een 9. Kiezen we de waarden uit de normale verdeling, zie Blok 5 Les 2, dan zie je dat $\geq 2\sigma = 0.05$. Nu: 3/78 is kleiner dan 5%.

verschil d	2	1	0	-1	-2	-3	-4
frequentie f	2	15	35	19	4	2	1
d·f	4	15	0	-19	-8	-6	-4
d ² ·f	8	15	0	19	16	18	16

- Kijk naar de verschillen van $y, x, \bar{x}$ en $\bar{y}$

$$d = y - x$$

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

$$\begin{aligned} \text{Var}(d) &= \sum \frac{(d - \bar{d})^2}{n} = \sum \frac{(y - x - (\bar{y} - \bar{x}))^2}{n} = \\ &= \sum \frac{(y - x - \bar{y} + \bar{x})^2}{n} = \sum \frac{((y - \bar{y}) - (x - \bar{x}))^2}{n} = \\ &= \sum \frac{(y - \bar{y} - (x - \bar{x}))^2}{n} = \sum \frac{(y - \bar{y})^2 - 2 \cdot (y - \bar{y}) \cdot (x - \bar{x}) + (x - \bar{x})^2}{n} = \\ &= \text{Var}(y) - 2 \cdot \text{Covar}(x, y) + \text{Var}(x) \end{aligned}$$

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

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

Hoe meer spreiding in de verschillen $(y - x)$ des te kleiner de correlatie.

Uit $\text{Var}(d)$ volgt ook dat, wanneer $\text{Var}(x) = \text{Var}(y)$, $\text{Var}(d) = 0$. Wat met $\text{Var}(x) = \text{Var}(y)$, vindt je $\text{Var}(x) + \text{Var}(y) = 2\text{Covar}(x, y)$.

Opgave 26, van de vierde naar de vijfde klas:

a) De formule voor $\text{Var}(d)$ is hierboven afgeleid.

b) Bij een verzameling punten (x, y) heeft men berekend:

$\text{Var}(x) = 169$, $\text{Var}(y) = 225$, en $\text{Var}(y - x) = \text{Var}(d) = 134$.

Bereken met deze gegevens $\text{Cov}(x, y)$ en $r(x, y)$.

Bij de afleiding van $\text{Var}(d = y - x)$ hebben we gevonden:

$\text{Var}(d) = \text{Var}(y) + \text{Var}(x) - 2 \cdot \text{Cov}(x, y) \Rightarrow \text{Cov}(x, y) = \frac{1}{2}[\text{Var}(x) + \text{Var}(y) - \text{Var}(d)]$. We hebben alle ingrediënten om

$\text{Cov}(x, y)$ uit te rekenen: $\text{Cov}(x, y) = \frac{1}{2}[169 + 225 - 134] = 130$.

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

$$\text{SD}(x) = \sqrt{\text{Var}(x)} = \sqrt{169} = 13, \text{ en } \text{SD}(y) = \sqrt{\text{Var}(y)} = \sqrt{225} = 15 \Rightarrow r = \frac{130}{13 \cdot 15} = \frac{2}{3}$$

c) Voor de correlatiecoëfficiënt kunnen we een derde formule afleiden.

De formule voor de correlatiecoëfficiënt is:

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

Omdat we twee formules hebben voor $\text{Cov}(x, y)$, zie bovenstaande les 2, hebben we hier dus twee formules voor de correlatiecoëfficiënt. Met de relatie voor $\text{Var}(d)$, vinden we een derde formule voor r :

$$r = \frac{\frac{1}{2}[\text{Var}(x) + \text{Var}(y) - \text{Var}(d)]}{\text{SD}(x) \text{SD}(y)}$$

d) Vijf leerlingen hebben voor twee opeenvolgende 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 $\text{Var}(x)$, $\text{Var}(y)$, $\text{Var}(d)$ en $r(x,y)$.

$$\text{Var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

$$\text{Var}(d) = \frac{1}{n} \sum_{i=1}^n [y_i - x_i - (\bar{y} - \bar{x})]^2.$$

$$r = \frac{\frac{1}{2}[\text{Var}(x) + \text{Var}(y) - \text{Var}(d)]}{\text{Sd}(x)\text{Sd}(y)}, \text{ met } \text{Sd}(x) = \sqrt{\text{Var}(x)}. \text{ Verder: } \text{Cov}(x, y) = \frac{1}{2}[\text{Var}(x) + \text{Var}(y) - \text{Var}(d)].$$

$$\bar{x} = \frac{1}{5} \sum_{i=1}^5 x_i = 6.64.$$

$$\bar{y} = \frac{1}{5} \sum_{i=1}^5 y_i = 6.72.$$

$$\text{Var}(x) = \frac{1}{5} \sum_{i=1}^5 (x_i - 6.64)^2 = 2.32.$$

Een invuloefening:

$$\text{Var}(y) = 1.49, \text{Var}(d) = 0.79, \text{Cov}(x, y) = 1.51 \text{ en } r = 0.81.$$

• Toevallige correlatie

Opgave 28

Aan drie toevallige voorbijgangers A, B, en C, wordt hun huisnummer x en leeftijd y gevraagd. Zie onderstaande tabel.

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

Eveneens: $\text{Var}(x)$, $\text{Var}(y)$, en $\text{Var}(d)$.

Bij bovenstaande opgave 26 vind je de formules.

Met de gegevens uit de tabel kun je hetgeen gevraagd wordt uitrekenen. Een invuloefening.

Leeftijd en huisnummers en de gevraagde grootheden.

Persoon	Huisnummer (x)	Leeftijd (y)
A	42	55
B	16	38
C	55	89

$$\bar{x} = 37,67, \bar{y} = 60,67, \bar{d} = 23$$

$$\text{Var}(x) = 262,89, \text{Var}(y) = 449,56, \text{Var}(d) = 74$$

$$r(x, y) = 0,928$$

Dus positief verband?

Je mag aannemen dat er in de populatie waaruit de steekproef is genomen geen enkel verband bestaat tussen leeftijd en huisnummer.

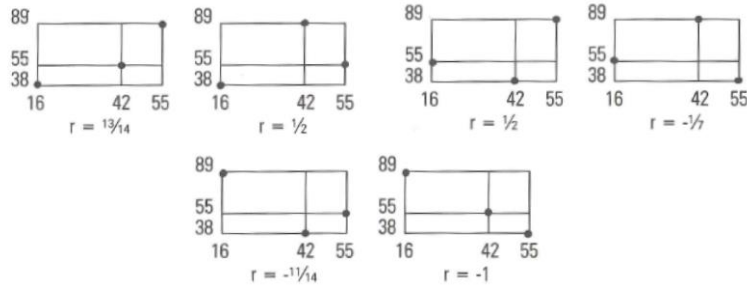
Je vindt toch een positieve correlatie terwijl er geen oorzakelijk verband is. De positieve correlatie moet op toeval berusten.

Verder onderzoek is nodig. Hou hier altijd rekening mee: als je geen oorzakelijk verband verwacht en je vindt het toch dan is inderdaad nader onderzoek nodig.

Je kent de uitdrukking: "*Correlation without causation*".

b) Voorbeeld:

Je kunt op 6 manieren de huisnummers bij de leeftijden plaatsen, bij elke mogelijkheid hoort een correlatie coëfficiënt. Dit wordt weergegeven door de punten in de volgende 6 (= 3!) verdelingen bij de drie leeftijden (zie volgende bladzijde):



Opgave 29, vervolg opgave 28.

a) Met de gegevens uit de bovenstaande 6 manieren, vind je:

$$\bar{r} = \frac{1}{6} \sum_{i=1}^6 r_i = 0, \text{ en } \text{Var}(r) = \frac{1}{2} (= \frac{1}{n-1}).$$

b) Hoeveel mogelijke punten (x, y) zijn er als we niet bij 3 maar bij n personen een x en een y genoteerd hebben? Hierbij neem je aan dat er geen verband is tussen x en y en dat alle x -waarden en alle y -waarden verschillend zijn.

Permutaties: er zijn dus $n!$ mogelijkheden. Dus zonder “terugleggen”. Je kan op $n!$ manieren de x -waarde bij de y -waarde plaatsen.

c) Hoeveel mogelijkheden zijn er als er twee x -waarden en 3 y -waarden, uit een totaal van n , x - en n , y -waarden, gelijk zijn? Uit de lessen van Blok 1-8 weet je dat het aantal mogelijkheden afneemt. En wel:

$$\frac{n!}{2!3!}. \text{ Denk bijvoorbeeld aan telefoonnummers met gelijke cijfers.}$$

Algemeen:

Gegeven:

- een aselechte steekproef van n personen met variabelen x en y .
- tussen x en y bestaat geen oorzakelijk verband.
- er zijn $n!$ mogelijkheden waarbij alle x en y waarden verschillend zijn.

Dan geldt:

- Het gemiddelde van de correlatiewaarden $r(x, y)$ is 0
- De standaardafwijking van deze r -waarden is $\frac{1}{\sqrt{n-1}}$.

Belangrijk voor het beoordelen van correlatiewaarden $r(x, y)$:

Voor $\bar{r} = 0$, is het bewijs eenvoudig(zie hieronder)

Voor $\sigma = \frac{1}{\sqrt{n-1}}$, wordt geen bewijs opgevoerd. Het bewijs is niet lastig maar kost veel werk.

Bijvoorbeeld voor $n = 3$ zijn er, zoals uit bovenstaand overzicht blijkt 6 mogelijkheden. Dus dit is nog te doen. Maar let op: verwar dubbele producten niet met de factor $n - 1$. Voor $n = 4$ kan de verwarring niet ontstaan, maar er zijn dan echter $24(4!)$ mogelijke waarden voor r .

Ook is het handig om bij de berekeningen: $\bar{r} = \frac{\text{Covar}(x, y)}{\text{Sd}(x)\text{Sd}(y)} = \frac{E(x, y) - E(x)E(y)}{\text{Sd}(x)\text{Sd}(y)}$, te gebruiken. Je

ziet ook gelijk als x en y onafhankelijk zijn dat $\bar{r} = 0$, want in dat geval geldt:

$$E(x, y) = E(x) \cdot E(y).$$

Een aanvullende conclusie is ook om je een onderzoek in omvang op te bouwen. Begin je gelijk met 10 huisnummers en leeftijden, dan heb je $10!$ (=faculteit) correlatiewaarden!(hier past een uitroepteken met een andere betekenis dan faculteit).

Oefenen en uitgewerkte Oefenopgaven.

Opgave 27: Associatie en correlatie

Op een middelbare school is er een leraar die als rapportcijfer altijd een 6 of een 7 geeft, $x_i(6,7)$. Een andere leraar geeft alleen maar zevens en achten, $y_i(7,8)$.

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

$\downarrow y_i; x_i \rightarrow$	$x_1 = 6$	$x_2 = 7$	
$y_1 = 7$	18 a	12 b	30
$y_2 = 8$	17 c	3 d	20
	35	15	Totaal aantal 50

a) Bereken de associatiemaat R .

$$R = \frac{ad-bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}. \text{ Na substitutie van de gegevens hou je over: } R = -\frac{1}{\sqrt{14}} = -0.2673.$$

b) Bereken:

$$\bar{x} = \frac{1}{a+b+c+d} [(a+c)x_1 + (b+d)x_2] = \frac{315}{50} = 6.3,$$

$$\bar{y} = \frac{1}{a+b+c+d} [(a+b)y_1 + (c+d)y_2] = \frac{370}{50} = 7.4,$$

$$\bar{d} = \frac{1}{a+b+c+d} [(a+b)y_1 - (a+c)x_1 + (c+d)y_2 - (b+d)x_2] = \frac{55}{50} = 1.1 \text{ of } 7.4 - 6.3.$$

c) Bereken:

$$Var(x) = \frac{1}{a+b+c+d} [(a+c)(x_1 - \bar{x})^2 + (b+d)(x_2 - \bar{x})^2].$$

$$\text{Substitueer de gegevens in bovenstaande uitdrukking: } Var(x) = \frac{2.1}{10} = 0.21.$$

$$Var(y) = \frac{1}{a+b+c+d} [(a+b)(y_1 - \bar{y})^2 + (c+d)(y_2 - \bar{y})^2] = 0.24.$$

$$Var(d) = ?$$

Ik realiseer mij dat het lastig is om met bovenstaande methode de formules toe te passen.

Daarom een aangepaste tabel zodat $\{x_i, y_i\}$ met betrekking tot de aantallen bij elkaar horen.

Cijfers	6 ($x_1 = x_2$)	7 ($x_3 = x_4$)	
7 ($y_1 = y_3$)	18 a (x_1, y_1)	12 b (x_3, y_3)	30 a + b
8 ($y_2 = y_4$)	17 c (x_2, y_2)	3 d (x_4, y_4)	20 c + d
	35 a + c	15 b + d	n = 50 = a + b + c + d

a)

$$\bar{x} = \frac{1}{a+b+c+d} [ax_1 + cx_2 + bx_3 + dx_4] = \frac{1}{a+b+c+d} [(a+c)x_1 + (b+d)x_4] = 6.3.$$

Op dezelfde wijze:

$$\bar{y} = \frac{1}{a+b+c+d} [ay_1 + cy_2 + by_3 + dy_4] = \frac{1}{a+b+c+d} [(a+b)y_1 + (c+d)y_2] = 7.4.$$

$$\bar{d} = \frac{1}{a+b+c+d} [a(y_1 - x_1) + c(y_2 - x_2) + b(y_3 - x_3) + d(y_4 - x_4)] = 1.1.$$

$$\text{b) } - Var(x) = \frac{1}{a+b+c+d} [a(x_1 - \bar{x})^2 + c(x_2 - \bar{x})^2 + b(x_3 - \bar{x})^2 + d(x_4 - \bar{x})^2] =$$

$$= \frac{1}{a+b+c+d} [(a+c)(x_1 - \bar{x})^2 + (b+d)(x_4 - \bar{x})^2] = 0.21.$$

$$- Var(y) = \frac{1}{a+b+c+d} [a(y_1 - \bar{y})^2 + c(y_2 - \bar{y})^2 + b(y_3 - \bar{y})^2 + d(y_4 - \bar{y})^2] =$$

$$\frac{1}{a+b+c+d} [(a+b)(y_1 - \bar{y})^2 + (c+d)(y_2 - \bar{y})^2] = 0.24.$$

Algemene formule voor $Var(d)$, zie hierboven op pagina 16.

$$- Var(d) = \frac{1}{a+b+c+d} \{a[(y_1 - x_1) - (\bar{y} - \bar{x})]^2 + c[(y_2 - x_2) - (\bar{y} - \bar{x})]^2 + b[(y_3 - x_3) - (\bar{y} - \bar{x})]^2 + d[(y_4 - x_4) - (\bar{y} - \bar{x})]^2\}.$$

Zoals je nu ziet, kan in de uitdrukking van $Var(d)$ geen termen bij elkaar worden genomen. Wel kan ik gebruik

maken van:

$(x_1 = x_2), (x_3 = x_4), (y_1 = y_3)$ en $(y_2 = y_4)$. Maar daar wordt de uitdrukking van $Var(d)$ niet eenvoudiger van.

c) Bereken de correlatiecoëfficiënt r .

$$-r = \frac{\frac{1}{2}[\text{Var}(x) + \text{Var}(y) - \text{Var}(d)]}{\sqrt{\text{Var}(x) \text{Var}(y)}} = -0.267.$$

Dus $r = R$.

Dit kan bewezen worden. Het kost nogal wat rekenwerk.

$$\text{Nota bene: } \text{Cov}(x, y) = \frac{1}{2} [\text{Var}(x) + \text{Var}(y) - \text{Var}(d)].$$

Appendix opgave 27, Bewijs $r = R$

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

Voor r kies ik:

$$r = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x) \text{Var}(y)}},$$

met

$$\text{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

Eerst wat ingrediënten uitzoeken:

$$x_1 - \bar{x} = x_1 - \frac{1}{a+b+c+d} [(a+c)x_1 + (b+d)x_4] = \frac{(a+c)x_1 + (b+d)x_1 - (a+c)x_1 - (b+d)x_4}{a+b+c+d} = \frac{(b+d)(x_1 - x_4)}{a+b+c+d}.$$

Op dezelfde wijze:

$$x_4 - \bar{x} = \frac{(a+c)(x_4 - x_1)}{a+b+c+d},$$

$$y_1 - \bar{y} = \frac{(c+d)(y_1 - y_2)}{a+b+c+d},$$

en

$$y_2 - \bar{y} = \frac{(a+b)(y_2 - y_1)}{a+b+c+d}.$$

Hou in gedachten: $x_1 = x_2, x_3 = x_4, y_1 = y_3$, en $y_2 = y_4$.

Wat hebben we nu voor de contravariantie?

$$\begin{aligned} \text{Cov}(x, y) &= \frac{(x_1 - x_4)(y_1 - y_2)[(c+d)\{a(b+d) - b(a+c)\} + (a+b)\{d(a+c) - c(b+d)\}]}{(a+b+c+d)^3} = \frac{(ad-bc)(a+b+c+d)(x_1 - x_4)(y_1 - y_2)}{(a+b+c+d)^3} = \\ &= \frac{(ad-bc)(x_1 - x_4)(y_1 - y_2)}{(a+b+c+d)^2}. \end{aligned}$$

$$\text{Nu de noemer van } r = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x) \text{Var}(y)}}.$$

$$- \text{Var}(x) = \frac{[(a+c)(x_1 - \bar{x})^2 + (b+d)(x_4 - \bar{x})^2]}{a+b+c+d},$$

en

$$- \text{Var}(y) = \frac{[(a+b)(y_1 - \bar{y})^2 + (c+d)(y_2 - \bar{y})^2]}{a+b+c+d}.$$

$$\text{Var}(x)\text{Var}(y) = \frac{(x_1 - x_4)^2 \{(a+c)(b+d)^2 + (b+d)(a+c)^2\} (y_1 - y_2)^2 \{(a+b)(c+d)^2 + (c+d)(a+b)^2\}}{(a+b+c+d)^6}.$$

Dus

$$\begin{aligned} \sqrt{\text{Var}(x)\text{Var}(y)} &= \frac{(x_1 - x_4)(y_1 - y_2)}{(a+b+c+d)^2} \sqrt{\{(a+c)(b+d)^2 + (b+d)(a+c)^2\} \{(a+b)(c+d)^2 + (c+d)(a+b)^2\}} = \\ &= \frac{(x_1 - x_4)(y_1 - y_2)}{(a+b+c+d)^3} \sqrt{(a+c)(b+d)(a+b)(c+d)(a+b+c+d)^2} = \\ &= \frac{(x_1 - x_4)(y_1 - y_2)}{(a+b+c+d)^2} \sqrt{(a+c)(b+d)(a+b)(c+d)}. \end{aligned}$$

Het resultaat:

$$r = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x) \text{Var}(y)}} = \frac{(ad-bc)(x_1 - x_4)(y_1 - y_2)}{(a+b+c+d)^2} \frac{1}{\frac{(x_1 - x_4)(y_1 - y_2)}{(a+b+c+d)^2} \sqrt{(a+c)(b+d)(a+b)(c+d)}} = \frac{(ad-bc)}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}.$$

Inderdaad, nogal wat rekenwerk.

Opgave 30: Toevallige correlatie.

Neem aan dat in een steekproef van 10 personen geen verband is tussen de x-waarden en de y-waarden.

a) Bereken de standaard afwijking in de bijbehorende r-waarden.

$$\text{De standaard afwijking } \sigma = \frac{1}{\sqrt{n-1}} = \frac{1}{\sqrt{10-1}} = \frac{1}{3}.$$

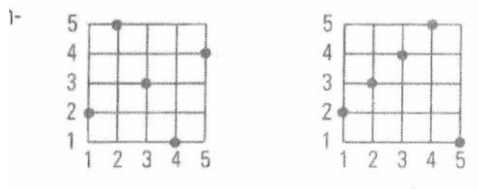
b) Kun je verklaren waarom een correlatiecoëfficiënt van 0.35 tussen de x- en y-waarden wel kan optreden maar dat $r=0.7$ onwaarschijnlijk is?

$r = 0.35$, is ongeveer gelijk aan de standaarddeviatie. $r = 0.7$, is ongeveer iets groter dan 2σ . Als je nu een normaalverdeling aanneemt dan weet je dat de kans op een r-waarde groter dan 2σ kleiner is dan 4.5%. Zie Blok 5 Les 2.

c) Bij $n = 100$ hoort $\sigma = \frac{1}{\sqrt{n-1}} \cong 0.1$.

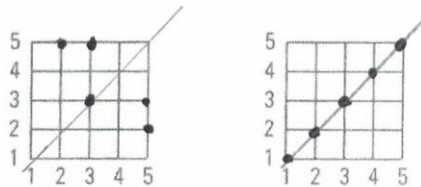
d) Voor $n = 2$ geldt $\sigma = \frac{1}{\sqrt{n-1}} = 1$. Dat hoort ook zo te zijn: $\frac{n}{2} = 1$.

e) Vervang de x-waarden $x_1, x_2, \dots, x_n$, door de rangnummers $1, 2, \dots, n$. Doe dit eveneens voor de y-waarden. Hieronder zijn bij $n=5$ twee mogelijke puntenwolken getekend:



Bij $n=5$ zijn $5!=120$, puntenwolken mogelijk.

f) Hieronder zijn twee puntenwolken getekend. De linker puntenwolk heeft een correlatiecoëfficiënt gelijk aan 0. De rechter puntenwolk heeft een correlatiecoëfficiënt van 1.



Let op: in de linker figuur zijn punten op 1 lijn getekend. Dat is niet de bedoeling, gelet op de permutaties! Ik heb alleen getoond dat voor $r=0$, links en rechts van de diagonaal de punten paarsgewijs op gelijke afstand van de diagonaal liggen. De punten paren (4,1), (1,4) in plaats van (5,3) en (3,5) zijn wel juiste keuzes.

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

Het aantal mogelijkheden is: $\frac{1}{n!} = \frac{1}{5!} = \frac{1}{120}$.

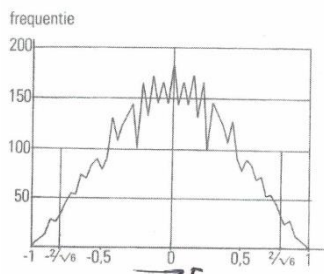
De kans is: $\frac{1}{120}$.

Opgave 31: Toevallige correlatie.

Hoeveel verschillende puntenwolken zijn er mogelijk bij $n = 7$? Zie bovenstaande les: $7!$, verschillende puntenwolken. Er zijn dus $7!$, verschillende r-waarden. Van deze r-waarden kun je een histogram maken.

Verbind de middens van de kolommen $\Rightarrow$ een frequentiepolygoon of lijndiagram.

Een dergelijk frequentiepolygoon is weergegeven in onderstaande figuur.



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

In de bovenstaande frequentiepolygoon zie je dat $r=0$ ongeveer 185 keer voorkomt. Het aantal mogelijkheden is $7!$.

De kans is $\frac{185}{7!} = 0.037$.

Hoe groot is de standaard afwijking bij $n=7$?

De standaard afwijking voor de r-waarden is: $\frac{1}{\sqrt{n-1}} = \frac{1}{\sqrt{6}}$.

Hoe groot is ongeveer de kans dat je een r -waarde vindt die meer dan twee keer de standaardafwijking van 0 af ligt?

Twee keer de standaard afwijking is aangegeven in de frequentiepolygoon: $\frac{2}{\sqrt{6}}$.

Bij $r = \frac{2}{\sqrt{6}}$, vind je een frequentie van ongeveer 35 in bovenstaande figuur. Let op: het gaat om groter dan $\frac{2}{\sqrt{6}}$.

Laten we het oppervlak onder de frequentie polygoon benaderen alsof de figuur een gelijkzijdige driehoek is met basis 2 en hoogte 185 $\Rightarrow$ oppervlak is: $185(\text{basis keer halve hoogte})$.

Het oppervlak van de twee kleine driehoeken voor $|r| > \frac{2}{\sqrt{6}} \Rightarrow 2 \cdot (1 - \frac{2}{\sqrt{6}}) \cdot \frac{35}{2}$.

de kans is: $\frac{(1 - \frac{2}{\sqrt{6}}) \cdot 35}{185} = 0.035$.

Ik heb de verdeling dus benaderd met een gelijkzijdige driehoek. Kies je voor de normale verdeling dan weet je uit de vuistregels dat de kans op $|r| > \frac{2}{\sqrt{6}}$ is 4.5%. Zie Blok 5, Les 2.

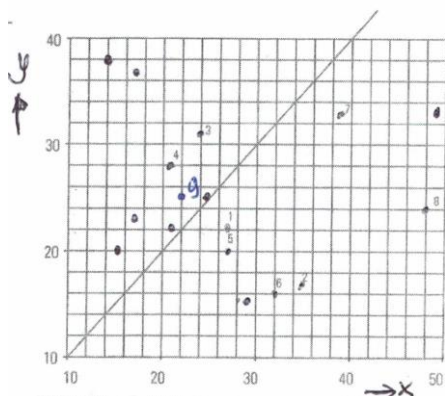
Opgave 32 Voetbal en toeval

Uit het onderstaande voetbal competitie overzicht volgt het aantal doelpunten dat elke club in de thuis- en uitwedstrijden heeft gescoord. De clubs zijn genummerd: 1 tot en met 17:

	doelpunten	
	thuis	uit
	x	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

Dit is een overzicht uit een geschiedenis boekje.

De puntenwolk is opgenomen in onderstaande grafiek.



De getrokken lijn suggereert een positief verband. Is dat ook zo?

Denk de lijn weg en het verband is dan niet duidelijk. We moeten dus gaan rekenen. Dan doen we in de volgende opgave.

Opgave 33 Voetbal en toeval.**a)**De som van de x-waarden $\sum_{i=1}^{17} x_i = 517$,de som van de y-waarden $\sum_{i=1}^{17} y_i = 380$,de som van de x^2 -waarden $\sum_{i=1}^{17} x_i^2 = 17123$,de som van de y^2 -waarden $\sum_{i=1}^{17} y_i^2 = 9084$,de som van de xy -waarden $\sum_{i=1}^{17} x_i y_i = 11698$.**b)** Bereken met bovenstaande gegevens:

$$Var(x) = \frac{1}{17} \sum_{i=1}^{17} (x_i - \bar{x})^2 = \frac{1}{17} \sum_{i=1}^{17} x_i^2 - \bar{x}^2, \text{ met } \bar{x} = \frac{517}{17} = 30.41 \Rightarrow Var(x) = 82.36.$$

Op vergelijkbare wijze met $\bar{y} = \frac{380}{17} = 22.35 \Rightarrow Var(y) = 34.7$.

$$Cov(x, y) = \frac{1}{17} \sum_{i=1}^{17} x_i y_i - \bar{x} \cdot \bar{y} \Rightarrow Cov(x, y) = 8.3.$$

We hebben alle ingrediënten voor de correlatiecoëfficiënt $r = \frac{Cov(x, y)}{\sqrt{Var(x)}\sqrt{Var(y)}} = 0.16$.**c)** De bovenstaande correlatiecoëfficiënt duidt er op dat er nauwelijks correlatie is tussen de uit- en thuiswedstrijden. r ligt te dicht bij 0.**d)** Laat nu de uitslag van nr 13 uit de tabel met doelpunten weg $\Rightarrow n = 16$.

Bovenstaande berekeningen doe je nog een keer, maar maak het je gemakkelijk, bijvoorbeeld:

$$\sum_{i=1}^{16} x_i = \sum_{i=1}^{17} x_i - 49 \Rightarrow \text{nieuwe } \bar{x}, \sum_{i=1}^{16} x_i^2 = \sum_{i=1}^{17} x_i^2 - 49^2, \dots$$

Je vindt dan een nieuwe correlatiecoëfficiënt $r = -0.10$. Er verandert dus weinig.

Blok 15 Les 4 Rechtlijnig verband

- Deze Les sluit aan bij de paragrafen 34 tot en met 46 van de tekst Correlatie en Regressie van de Wageningse Methode 1989
- A Warming-Up Rechtlijnig verband

Opgave 34 Een dagje uit:

Een dag naar Zandvoort aan de Zee met een gehuurde auto. Je huurt type B.

In onderstaande tabel staan de huurtarieven (niet actueel) van drie typen huurauto's, A, B en C:

Type	deuren	zitplaatsen	€	Per 24 uur	Per km
A	2	4	€	80	0.35
B	2	4	€	90	0.40
C	4	4/5	€	110	0.50

a) Wat moet je afrekenen als je die dag 154 km gereden hebt met type B?

$$154 \cdot €0.40 + €40 = €101.6.$$

b) Geef een formule voor de prijs p (in €), als je een dag x km gereden hebt met type B.

$$p = 0.4 \cdot x + 90.$$

c) Geef zo'n formule ook voor type C.

$$p = 0.5 \cdot x + 110.$$

Opgave 35, Vervolg opgave 34:

Behalve de huurprijs moet je ook de brandstof voor de auto betalen. Type B gebruikt 1 liter brandstof per 16 km, type C gebruikt 1 liter brandstof per 12 km. De brandstof kost €1.65 per liter.

a) Geef een formule voor de totale kosten y als je een dag x km rijdt met type B. Dus ook de kosten erbij die je in opgave 34 hebt berekend.

$$y = \left(0.4 + \frac{1}{16} \cdot 1.6\right) \cdot x + 90.$$

b) Geef een vergelijkbare formule voor type C.

$$y = \left(0.5 + \frac{1}{12} \cdot 1.6\right) \cdot x + 110.$$

c) Er is een aantal factoren dat er voor zorgt dat bovenstaande formules in de praktijk niet precies aangeven hoeveel de totale kosten zullen bedragen. Eén zo'n factor is de gereden snelheid.

Wanneer je een aantal keren zo'n auto huurt en steeds de gereden afstanden x en de totale kosten y noteert, zullen de punten (x, y) rond de bijbehorende rechte lijn liggen zoals hierboven is afgeleid. 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. In het vervolg van de les gaan we hier nader op in.

- Van klas 4 naar klas 5 (Zie ook vorige les)

Nog eens het overzicht van de cijfers, zie onderstaande tabel :

x is het cijfer voor wiskunde in klas 4.

y is het cijfer voor wiskunde A in klas 5.

2 leerlingen hadden een $x = 4$ in klas 4 en een $y = 5$ in klas 5.

7 leerlingen hadden een $x = 7$ in klas 4 en een $y = 6$ in klas 5.

Nota Bene: de omcirkelde getallen, in de onderstaande matrix, zijn de aantallen.

Dus:

- 2 "vijfen" voor y ,

- 2 "vieren" voor x .

Gemiddelden kunnen worden uitgerekend en grafieken getekend:

bijvoorbeeld $\bar{y}(x)$ als functie van x . De rood omcirkelde getallen zijn de aantallen leerlingen.

(vijfde klas)

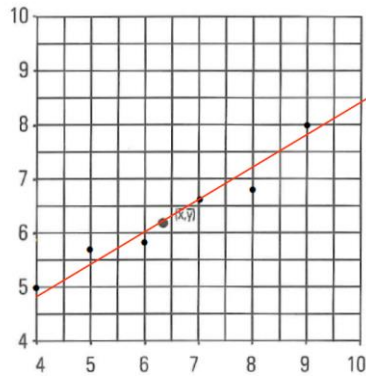
y	2	10	33	22	10	1	
9				1			1
8				3	3	1	7
7		1	5	8	5		19
6		5	20	7			32
5	2	4	6	2	1		15
4			2	1	1		4
	4	5	6	7	8	9	

x (vierde klas)

Bovenaan pag. 22, van de bovenstaande Samenvatting van Les 3, hebben we gezien dat er een duidelijke correlatie is: $r = 0.46$. Bij een hogere x -waarde vind je een hogere y -waarde. Deze trend kunnen we ook nog op een andere manier laten zien: Bereken bij elke x het gemiddelde $\bar{y}(x)$ van de bijbehorende y :

x	4	5	6	7	8	9
$\bar{y}(x)$	5,00	5,70	5,85	6,59	6,80	8,00

Bepaal de richtingscoëfficiënt van de lijn in de grafiek: De rode lijn wordt "op het oog" getrokken. Let op: het punt $(\bar{x}, \bar{y})$ is ook aangegeven(vaag) in de grafiek.



Richtingscoëfficiënt $\approx \frac{7.8-4.8}{5} = 0.60$.

Vergelijking van de lijn: $y = 0.60x + b$.

$\rightarrow b = 2.33 \rightarrow y = 0.60x + 2.33$.

De lijn gaat door het punt $(\bar{x}, \bar{y}) = (6.40; 6.17)$, het “zwaartepunt van de puntenwolk”. Dit volgt direct uit de formule van de lijn: vul in $y = \bar{y}$, en $x = \bar{x}$, en je ziet dat dit punt voldoet aan de vergelijking.

De richtingscoëfficiënt is ongeveer 0.6 .

- Regressie van y op x

x is de onafhankelijk variabele, y is de afhankelijk variabele.

De formule voor de regressie lijn is van de vorm $y = a \cdot x + b$.

De betekenis van de formulering y op x is: y is de afhankelijk variabele en x is de onafhankelijk variabele.

X is de onafhankelijke variabele, y is de afhankelijk variabele.
De formule van de regressielijn is van de vorm $y = a \cdot x + b$.

Eerste formule: $\frac{y - \bar{y}}{SD(y)} = r \cdot \frac{x - \bar{x}}{SD(x)}$ of: $z_y = r \cdot z_x$

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

De twee formules zijn aan elkaar gelijk.

De richtingscoëfficiënt hangt af van de correlatiecoëfficiënt r en de standaard afwijkingen $SD(x)$ en $SD(y)$. Bij positieve r is de regressielijn stijgend (bij hogere x is $\bar{y}$ groter), bij negatieve r is de regressielijn dalend. De formule voor de regressielijn in de vorm $y = a \cdot x + b$, de tweede formule, volgt direct uit de eerste formule.

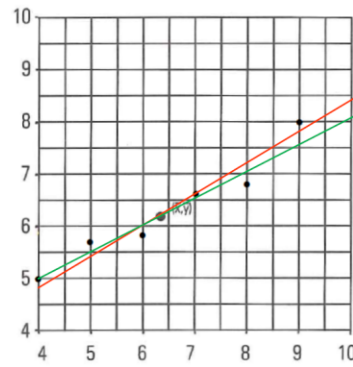
- Van klas 4 naar klas 5

x	4	5	6	7	8	9
$\bar{y}(x)$	5,00	5,70	5,85	6,59	6,80	8,00

$$\begin{aligned}\bar{x} &= 6,40 \text{ en } \bar{y} = 6,17 \\ SD(x) &= 0,99 \text{ en } SD(y) = 1,04 \\ r(x, y) &= 0,48 \\ \frac{SD(y)}{SD(x)} &= 1,05\end{aligned}$$

Vergelijking van de regressielijn:

$$y = 0,46 \cdot 1,05 \cdot x - 0,46 \cdot 1,05 \cdot 6,40 + 6,17 = 0,48 \cdot x + 3,11$$



De formule voor de regressielijn betreft de groene lijn. De rode lijn is voor: $y = 0.60x + 2.33$, de lijn die “op het oog” is getrokken. De berekende richtingscoëfficiënt 0.48 wijkt 25% af van de geschatte richtingscoëfficiënt 0.6 .

- De regressielijn (afleiding van de vergelijking: zie appendix 1)

Een formule voor de richtingscoëfficiënt:

$$\begin{aligned}a &= r \cdot \frac{SD(y)}{SD(x)} = \frac{\text{Cov}(x, y)}{SD(x) \cdot SD(y)} \cdot \frac{SD(y)}{SD(x)} = \frac{\text{Cov}(x, y)}{(SD(x))^2} = \\ &= \frac{\frac{x_1 \cdot y_1 + \dots + x_n \cdot y_n}{n} - \bar{x} \cdot \bar{y}}{\frac{x_1^2 + \dots + x_n^2}{n} - \bar{x}^2} = \\ &= \frac{x_1 \cdot y_1 + \dots + x_n \cdot y_n - n \cdot \bar{x} \cdot \bar{y}}{x_1^2 + \dots + x_n^2 - n \cdot \bar{x}^2}\end{aligned}$$

Hierboven is de formule voor de richtingscoëfficiënt a nog eens weergegeven. 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: $y = 0.48 \cdot x + 3.11$, een voorspelling doen voor het cijfer y dat deze leerling in de vijfde klas zal hebben.

- Wat is regressie?

De eerste formule: $\frac{y-\bar{y}}{SD(y)} = r \cdot \frac{x-\bar{x}}{SD(x)}$ of: $z_y = r \cdot z_x$

Gegeven: $x > \bar{x}$ en $r > 0$, dan is $y - \bar{y} > 0$ dus $y > \bar{y}$

Omdat: $0 \leq |r| \leq 1$ geldt: $|z_y| = |r| \cdot |z_x| \leq |z_x|$

→ De afwijking van y ten opzichte van het gemiddelde van y wordt kleiner ten opzichte van de afwijking van x ten opzichte van het gemiddelde van x .

→ Regressie: de waarde van y 'neigt' naar het gemiddelde.

Opmerking:

- regressie betekent terugvallen, teruggaan. Dus terug gaan naar het gemiddelde.
- de formules voor de regressielijn worden hier niet afgeleid.

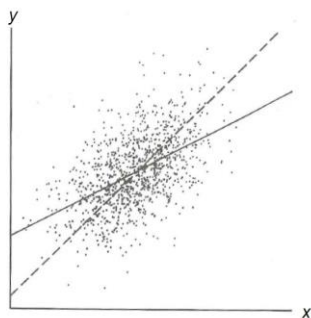
Voor de afleiding van de formule voor de regressielijn, zie appendix 1. De regressieformule volgt uit de kleinste kwadraten methode, appendix 1.

- Regressie van y op x of van x op y ?

In onderstaande figuur zie je een puntenwolk, 1078 punten (x, y) , waarbij x de lengte van een vader en y de lengte van diens volwassen zoon is. Het is het resultaat van een onderzoek dat in 1900 in Engeland werd gehouden. Men interesseerde zich toen in de overerving van eigenschappen. Francis Galton, (zie ook VWO 4 Het bord van Galton, www.leennoordzij.me) was een van de pioniers op dat gebied.

Verdere gegevens:

$\bar{x} = 172,72$ cm; $\bar{y} = 175,26$ cm
 $SD(x) = 6,86$ cm en $SD(y) = 6,86$ cm
 $r(x, y) = 0,5$



Met de formules voor de regressielijn,

$$y = 0.5 \cdot x + 88.90.$$

Nogmaals, x is de onafhankelijk variabele en y de afhankelijk variabele in deze formule voor de regressielijn met richtingscoëfficiënt 0.5.

Dit resultaat volgt uit:

$$\frac{y-\bar{y}}{SD(y)} = r \frac{x-\bar{x}}{SD(x)} \Rightarrow y = r(x - \bar{x}) \frac{SD(y)}{SD(x)} + \bar{y} \Rightarrow y = r \frac{SD(y)}{SD(x)} x - r \frac{SD(y)}{SD(x)} \bar{x} + \bar{y} = 0.5 \cdot x + 88.9,$$

waarbij bovenstaande gegevens zijn ingevuld

Je weet de lengte van de zoon, maar niet de lengte van de vader. Verwissel x en y in bovenstaande formule.

In dat geval pas je regressie van de zoon (y) op de vader (x) toe.

De regressievergelijking is dan:

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

Of:

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

Uitwerking

De vergelijking is in dit geval:

$$\frac{y - 175,26}{6,86} = 2 \cdot \frac{x - 172,72}{6,86}$$

Of:

$$y = 2 \cdot x - 170,18$$

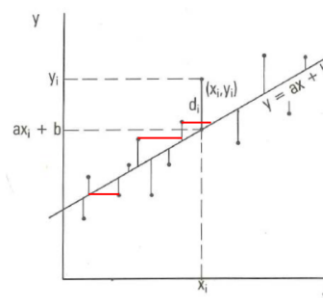
$$y = 180,00 \text{ cm} \rightarrow 175,09 \text{ cm}$$

Je ziet dus twee regressielijnen in bovenstaande grafiek (blz.34). De schaalverdeling ontbreekt. De gestippelde lijn is de regressie van de zoon (y) op de vader (x), met richtingscoëfficiënt 2.

Bij **regressie van x op y** is de lijn met vergelijking $x = a \cdot y + b$ is de beste lijn als de totale lengte van de horizontale lijntjes zo klein mogelijk is.

De lengte van zo'n horizontaal lijntje is de afwijking van x_i tot de lijn $x = a \cdot y + b$, dat is $|x_i - (a \cdot y_i + b)|$.

Dit leidt tot de vergelijking: $\frac{x - \bar{x}}{SD(x)} = r \cdot \frac{y - \bar{y}}{SD(y)}$



Oefenen en uitgewerkte Oefenopgaven.

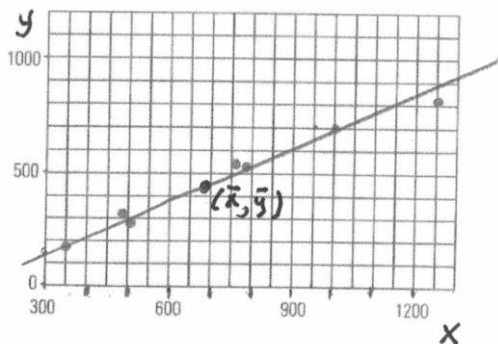
Opgave 39: Nog eens reclame, vervolg opgave 23 Blok 15 Les 2.

Van de punten wolk van (x,y) in opgave 23 hebben we de correlatiecoëfficiënt

$$r = \frac{\text{Cov}(x,y)}{\text{Sd}(x)\text{Sd}(y)} = \frac{61485.8}{295.8 \cdot 208.6} = 0.996, \text{ uitgerekend.}$$

a) de richtingscoëfficiënt a van de regressie lijn $y = ax + b$, is: $r \cdot \frac{\text{Sd}(y)}{\text{Sd}(x)} = \frac{\text{Cov}(x,y)}{[\text{Sd}(x)]^2} = \frac{\text{Cov}(x,y)}{\text{Var}(x)} = \frac{61485.8}{87497.3} = 0.7$.

b) Hieronder is de puntenwolk nog eens weergegeven. De regressielijn is in getekend en het zwaartepunt $(\bar{x}, \bar{y}) = (693, 439)$. De vergelijking van de regressie lijn wordt met deze gegevens: $y = 0.7 \cdot x - 46$.



c) Het aantal reacties bij een advertentie die €900 kost volgt uit substitutie in de vergelijking van de regressielijn: 584.

d) Het aantal reacties bij een advertentie die €2000 kost: 1354.

Deze laatste voorspelling is minder betrouwbaar dan de voorspelling onder c) omdat de voorspelling onder c) dichter bij de puntenwolk ligt.

Opgave 43, tweede hands auto's;

In onderstaande tabel is x de leeftijd van de auto in jaren y is de prijs in duizend euro's. Je ziet dat er 3 auto's zijn van 4 jaar oud.

x	y	x	y
1	31,50	3	14,95
5	6,75	0	26,0
1	15,25	3	15,5
3	14,95	4	16,75
6	7,75	4	11,5
6	16,95	4	14,5

Bij deze tabel hoort de volgende punten wolk:

Zie volgende pagina.

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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

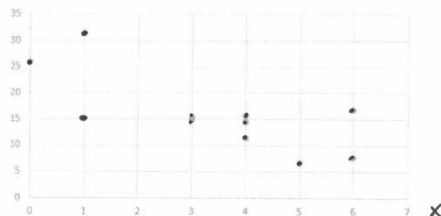
$$n = 12 \Rightarrow \bar{x} = \frac{40}{12} = 3.33.$$

$$\bar{y} = \frac{191.5}{12}, n = 12.$$

$$Sd(x) = \sqrt{Var(x)} = \sqrt{\frac{1}{12} \sum_{i=1}^{12} (x_i - \bar{x})^2} = 1.84.$$

$$Sd(y) = \sqrt{Var(y)} = \sqrt{\frac{1}{12} \sum_{i=1}^{12} (y_i - \bar{y})^2} = 6.58.$$

y



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

Voor de $Cov(x, y)$ zijn twee formules beschikbaar. Gebruik de beschikbare resultaten en kies de handigste formule.

$$Cov(x, y) = \frac{1}{n} \sum_{i=0}^n (x_i - \bar{x})(y_i - \bar{y}) = -8.82.$$

$$r = \frac{Cov(x, y)}{Sd(x) Sd(y)} = -0.73.$$

c) Geef een vergelijking van de regressielijn in de vorm $y = ax + b$.

De regressielijn:

$$y = r \cdot \frac{Sd(y)}{Sd(x)} \cdot x - r \cdot \frac{Sd(y)}{Sd(x)} \cdot \bar{x} + \bar{y} = r \cdot \frac{Sd(y)}{Sd(x)} [x - \bar{x}] + \bar{y} = -2.62x + 24.63.$$

Opgave 44, vervolg opgave 43:

a) Welke betekenis kun je aan a en b in $y = ax + b$ hechten?

Wiskundig: a is de richtingscoëfficiënt. Deze coëfficiënt geeft de richting aan van de prijsontwikkeling als functie van x , dus als functie van de tijd. Het getal van de gemiddelde waarde afname per jaar. b is de gemiddelde waarde bij het begin, dus $x = 0$.

b) Wat geeft de regressielijn ongeveer aan?

Hoe ouder, hoe goedkoper.

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

Het resultaat is afhankelijk van het type auto. Zet er maar een Ferrari tussen.

d) De populatie waaruit deze steekproef is getrokken, is uit de meeste tweedehands auto's. Hoe ziet de puntenwolk eruit als je de populatie beperkt tot één type?

De spreiding in de puntenwolk zal minder zijn.

e) Wat zal er dan gebeuren met de correlatiecoëfficiënt?

r wordt groter, in absolute zin, bij gelijkblijvende richtingscoëfficiënt.

f) De auto's zijn nog betrekkelijk 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?

De prijs van de auto zal niet negatief worden. De lijn zal afbuigen. Ook zal door allerlei oorzaken vaker oudere auto's uit de handel worden genomen. De conclusie is dat de regressielijn slecht over een bepaald interval van de tijd, betekenisvol kan worden toegepast.

Rechtlijnigheid heeft dus 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. Ook moet hier een kanttekening bij worden geplaatst. Gebruikelijk is om aan te nemen dat van een auto in de eerste twee jaren meer wordt afgeschreven dan in de volgende 5 jaren. Uitbreiding van de steekproef kan dit aantonen of verwerpen.

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

Een logaritmische presentatie komt in opgave 45 aan de orde.

Opgave 45: Groei

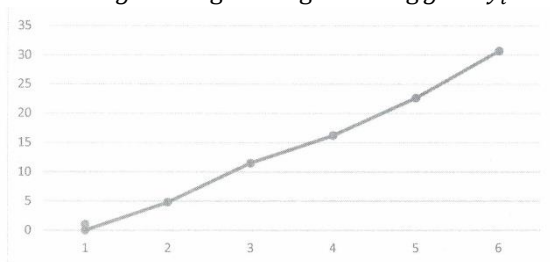
Een bioloog onderzoekt de groei van zonnebloemen. De bioloog denkt te kunnen bewijzen dat in een bepaalde fase van de groei de hoogte h van een zonnebloem exponentieel van de tijd t afhangt: $h = H \cdot g^t$.

In onderstaande tabel staan de gegevens die de bioloog verzameld heeft over genoemde fase:

t_i (in dagen)	0	4.8	11.5	16.2	22.6	30.7
h_i (in cm)	72	85	108	142	190	256

In onderstaande grafiek zijn de meetpunten, (t_i, y_i) , weergegeven, waarbij $y_i = \log(h_i)$.

Dus: $h = H \cdot g^t \Rightarrow \log h = \log H + t \log g \Rightarrow y_i = \log h_i = (\log g)t_i + \log H$.



Mag je, zo op het oog aannemen dat de bioloog een juiste veronderstelling maakte? Ja, een bijna rechte lijn van logwaarden.

Bereken de correlatiecoëfficiënt $r(t_i, y_i)$.

$$r(t, y) = \frac{\text{Cov}(t, y)}{Sd(t)Sd(y)},$$

$$\text{Cov}(t, y) = \frac{1}{6} \sum_{i=1}^6 (t_i - \bar{t})(y_i - \bar{y}),$$

$$Sd(t) = \sqrt{\frac{1}{6} \sum_{i=1}^6 (t_i - \bar{t})^2},$$

$$Sd(y) = \sqrt{\frac{1}{6} \sum_{i=1}^6 (y_i - \bar{y})^2}.$$

Zie tabel hierboven:

$$\bar{t} = \frac{1}{6} \sum_{i=1}^6 t_i = 14.3,$$

$$\bar{y} = \frac{1}{6} \sum_{i=1}^6 \log h_i = 2.11.$$

De correlatiecoëfficiënt:

$$r(t, y) = \frac{\text{Cov}(t, y)}{\text{sd}(t)\text{sd}(y)} = 1.00.$$

Een vergelijking voor de regressielijn:

$$y = r \cdot \frac{\text{sd}(y)}{\text{sd}(t)} \cdot t - r \cdot \frac{\text{sd}(y)}{\text{sd}(t)} \cdot \bar{t} + \bar{y}.$$

$$\text{Met } r \cdot \frac{\text{sd}(y)}{\text{sd}(t)} = 0.018, \text{ en } \bar{y} - r \cdot \frac{\text{sd}(y)}{\text{sd}(t)} \cdot \bar{t} = 1.86 \Rightarrow y = 0.018t + 1.86.$$

Geef met behulp van de voorgaande resultaten een formule voor de hoogte h als functie van de tijd t :

$$\log h = a \cdot t + b \Rightarrow \log h = \log A + \log B \Rightarrow h = A \cdot B.$$

$$\log A = 0.018t \Rightarrow A = 10^{0.018t},$$

$$\log B = 1.86 \Rightarrow B = 10^{1.86} = 72.44.$$

$$h = A \cdot B = 7.244 \cdot 10^{0.018 \cdot t}.$$

Opgave 56 kleinste kwadraten (zie ook appendix):

Uit de verzameling getallen $x_1, x_2, \dots, x_n$ 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 in deze voorspelling: $t - x_i$.

Neem aan dat je nu als 'straf' $(t - x_i)^2$ euro boete moet betalen.

Na de n trekkingen is je totale verlies:

$$Q(t) = (t - x_1)^2 + (t - x_2)^2 + \dots + (t - x_n)^2 = \sum_{i=1}^n (t - x_i)^2.$$

Het probleem is om $Q(t)$ zo klein mogelijk te houden.

a) Een 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 van deze getallen?

In woorden: de getallen optellen en delen door 4.

In formule:

$$\bar{x} = \frac{1}{4} \sum_{i=1}^4 x_i \Rightarrow \bar{x} = 1.5.$$

b) Schrijf $Q(t)$ als veelterm

$$Q(t) = \sum_{i=1}^n (t - x_i)^2 = \sum_{i=1}^n (t^2 - 2tx_i + x_i^2) = nt^2 - 2t \sum_{i=1}^n x_i + \sum_{i=1}^n x_i^2 = nt^2 - 2tn\bar{x} + n\bar{x}^2.$$

c) Laat met de afgeleide van $Q(t) \Rightarrow \frac{dQ}{dt} [\equiv Q'(t)]$ zien dat dat $Q(t)$ minimaal is als $t = \bar{x}$.

$$Q(t) = nt^2 - 2tn\bar{x} + n\bar{x}^2 \Rightarrow \frac{dQ}{dt} = 2nt - 2n\bar{x} = 0 \therefore t = \bar{x}.$$

Een minimum: de tweede afgeleide is gelijk aan $2n > 0$.

Anders geformuleerd: $Q(t)_{\min} = \sum_{i=1}^n (\bar{x} - x_i)^2 = n \cdot \text{Var}(x)$.

d) In het voorbeeld onder a) met $\bar{x} = 1.5$ is

$$Q(t) = 9 - 18 + 18 = 9.$$

Nota bene: dat in dit voorbeeld $-2tn\bar{x} + n\bar{x}^2 = 0$, komt door de gekozen getallen.

e) Een andere methode om de minimum waarde van $Q(t)$ te bepalen:

$$Q(t) = nt^2 - 2tn\bar{x} + n\bar{x}^2 = n(t^2 - 2t\bar{x} + \bar{x}^2) - n(\bar{x}^2 - \bar{x}^2) = n(t - \bar{x})^2 - n(\bar{x}^2 - \bar{x}^2).$$

Dus $Q(t) = n(t - \bar{x})^2 - n(\bar{x}^2 - \bar{x}^2)$, is minimaal voor $t = \bar{x}$.

Waarom? $(t - \bar{x})^2 \geq 0$ en $n(\bar{x}^2 - \bar{x}^2)$ is een getal dat niet afhankelijk is van t .

Verder: let op het verschil tussen $\bar{x}^2$ en $\bar{x}^2$.

Opgave 57, zie opgave 56

Variantie en $Q(t)$.

In Blok 15 Les 2 (zie onderstaande samenvatting samen vattting): "de variantie en de standaarddeviatie"

De **variantie** is een maat voor de afwijkingen ten opzichte van het gemiddelde.

$$\begin{aligned} \text{Var}(x_1, x_2, \dots, x_n) &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n} = \\ &= \frac{(x_1)^2 + (x_2)^2 + \dots + (x_n)^2}{n} - \bar{x}^2 \end{aligned}$$

$$\text{De standaarddeviatie SD}(x_1, x_2, \dots, x_n) = \sqrt{\text{Var}(x_1, x_2, \dots, x_n)}$$

is afgeleid dat:

$$\text{Var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n (x_i)^2 - (\bar{x})^2.$$

Dus:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i)^2 - n(\bar{x})^2.$$

Reminder:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n \cdot \bar{x}^2 = \sum_{i=1}^n x_i^2 - 2n \cdot \bar{x} \cdot \bar{x} + n \cdot \bar{x}^2.$$

In plaats van x_i in bovenstaande uitdrukking gebruik ik z_i .

We herschrijven bovenstaande uitdrukking:

$$\sum_{i=1}^n (z_i - \bar{z})^2 = \sum_{i=1}^n (z_i - \bar{z})^2 + n \cdot \bar{z}^2.$$

Nu weer terug naar x_i .

Stel nu $z_i = x_i - t$, waarbij t een getal is,

$$\frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n t = \bar{x} - \frac{n \cdot t}{n} = \bar{x} - t = \bar{z}.$$

Uit $z_i = x_i - t$, en $\bar{z} = \bar{x} - t \Rightarrow z_i - \bar{z} = x_i - \bar{x}$.

In opgave 56:

$$Q(t) = (t - x_1)^2 + (t - x_2)^2 + \dots + (t - x_n)^2 = \sum_{i=1}^n (t - x_i)^2 \Rightarrow Q(t) = \sum_{i=1}^n (x_i - t)^2.$$

De bovenstaande uitdrukking herschrijven:

$$\begin{aligned} Q(t) &= \sum_{i=1}^n (x_i - t)^2 = \sum_{i=1}^n x_i^2 - 2t \sum_{i=1}^n x_i + n \cdot t^2 = \sum_{i=1}^n x_i^2 - 2n \cdot t \cdot \bar{x} + n \cdot t^2 = \\ &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 + n\bar{x}^2 - 2nt\bar{x} + nt^2 = \sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2 + n(\bar{x} - t)^2 = \sum_{i=1}^n (x_i^2 - \bar{x}^2) + n(\bar{x} - t)^2. \end{aligned}$$

Dit resultaat volgt ook direct uit de formule voor de Variantie:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n (x_i)^2 - n(\bar{x})^2,$$

met deze uitdrukking vinden we

$$Q(t) = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - t)^2.$$

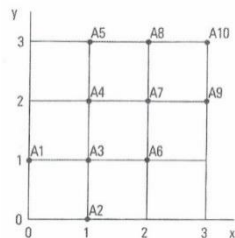
Nu zien we ook weer het resultaat voor minimale $Q(t)$ uit opgave 56, $\sum_{i=1}^n (x_i - \bar{x})^2$ is een getal, niet afhankelijk van t ,

$\therefore Q(t)$ is minimaal voor

$$\bar{x} = t, \text{ want } (\bar{x} - t)^2 \geq 0.$$

Opgave 58:

De beste lijn een bijzonder geval.



Hierboven zijn 10 punten, $A_1, A_2, \dots, A_{10}$ in een rooster aangegeven. Door het toeval wordt één van deze punten aangewezen. Van dat punt wordt de eerste coördinaat bekend gemaakt. Jij moet de tweede coördinaat raden. Bij gegeven eerste coördinaat x , schat je als tweede coördinaat y . Deze geschatte waarde noemen we $t(x)$.

Voorbeeld: als $x = 0$, dan kies je natuurlijk $t(0) = 1$, en $y = t(0)$. Er is maar 1 punt en de fout is ook nul.

Maar bij $x = 1$, heb je vier y -waarden.

Neem aan dat de boete voor verkeerd raden weer het kwadraat van de fout is: $[y_i - t(x_i)]^2$, zie opgave 56.

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

a) Welk getal kies je als $t(1)$?

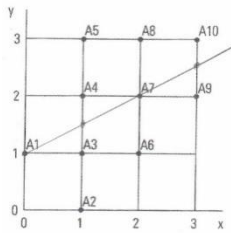
$x = 1$. Bij de vorige opgaven 56 en 57, heb je afgeleid dat de boete minimaal is bij $t = \bar{y}$.

Zoals je in de figuur ziet, is bij $x = 1$, de gemiddelde waarde $t = \bar{y} = 1.5 \left(= \frac{6}{4} \right)$.

Bij $x = 2 \Rightarrow t = \bar{y} = 2, A_7$,

en

bij $x = 3 \Rightarrow t = \bar{y} = 2.5$.



b) In bovenstaande figuur is een rechte lijn getekend door de geschatte waarden $t(x_i)$.

De lijn wordt weergegeven door de vergelijking $y = ax + b$.

$$x = 0, y = 1 \Rightarrow b = 1,$$

$$x = 1, y = 1.5 \Rightarrow a = 0.5.$$

De vergelijking is dus: $y = 0.5x + 1$.

Controle: $x = 2$, geeft met $y = 0.5x + 1 \Rightarrow y = 2$.

Merk op dat de lijn $y = 0.5x + 1$, de eigenschap heeft, dat de som van de kwadraten van de afstanden $d_i = y_i - t(x_i)$, zo klein mogelijk is.

$$Q(t) = \sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - t)^2.$$

De minimum waarde van $Q[t(x_i)]$ wordt gevonden bij $\bar{y} = t(x_i) \Rightarrow \bar{y}(x_i) = 0.5x + 1$, voor elke x_i .

c) Bereken de correlatiecoëfficiënt r van de 10 punten $A(x_i, y_i)$.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{10} (0 + 4 \cdot 1 + 3 \cdot 2 + 2 \cdot 3) = \frac{16}{10} = 1.6,$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{10} (1 + 0 + 1 + 2 + 3 + 1 + 2 + 3 + 2 + 3) = \frac{18}{10} = 1.8.$$

$$\text{Var}(x) = \frac{1}{10} \sum_{i=1}^{10} x_i^2 - \bar{x}^2 = \frac{1}{10} (0 + 4 \cdot 1^2 + 3 \cdot 2^2 + 2 \cdot 3^2) - \left(\frac{16}{10}\right)^2 = \frac{84}{100}$$

$$\text{Var}(y) = \frac{1}{10} \sum_{i=1}^{10} y_i^2 - \bar{y}^2 = \frac{1}{10} (1 + 0 + 1 + 4 + 9 + 1 + 4 + 9 + 4 + 9) - \left(\frac{18}{10}\right)^2 = \frac{96}{100}.$$

$$\text{Cov}(x, y) = \frac{1}{10} \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{10} \sum_{i=1}^{10} x_i y_i - \bar{x} \bar{y} = \frac{33}{10} - \frac{16}{10} \cdot \frac{18}{10} = \frac{42}{100}.$$

$$\text{dus } r = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x) \text{Var}(y)}} = \frac{42}{100} \cdot \frac{100}{\sqrt{84 \cdot 96}}. \text{ Het is handig om dit niet verder uit te rekenen.}$$

Want:

$$y = r \cdot \frac{\sqrt{\text{Var}(y)}}{\sqrt{\text{Var}(x)}} \cdot x - r \cdot \frac{\sqrt{\text{Var}(y)}}{\sqrt{\text{Var}(x)}} \cdot \bar{x} + \bar{y} \Rightarrow y = \frac{\text{Cov}(x, y)}{\text{Var}(x)} \cdot x - \frac{\text{Cov}(x, y)}{\text{Var}(x)} \cdot \bar{x} + \bar{y} \Rightarrow$$

$$\Rightarrow y = \frac{42}{100} \cdot \frac{100}{84} \cdot x - \frac{42}{100} \cdot \frac{100}{84} \cdot \frac{16}{10} + \frac{18}{10} \Rightarrow y = 0.5 + 1.$$

Conclusie: dus de lijn die gevonden is, is gelijk aan de regressielijn.

d) De verwachte boete voor $\bar{y} = t(x_i)$ en $\bar{y}(x_i) = t(x_i) = 0.5x_i + 1$.

De verwachte boete is de gemiddelde boete:

$$\frac{1}{n} Q(t) = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 + (\bar{y} - t)^2 = \frac{1}{10} \sum_{i=1}^n (y_i - \frac{1}{2} x_i - 1)^2$$

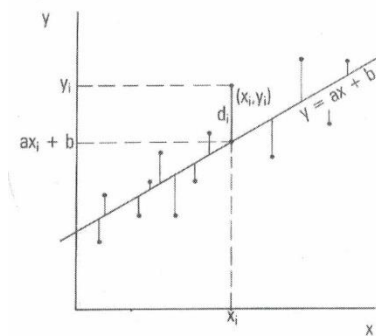
en

$$(\bar{y} - t)^2 = 0.$$

Opgave 59:

De beste lijn, het algemene geval. Vervolg vraagstuk 58.

Bij een puntenwolk (x_i, y_i) , zoeken we naar de lijn $y = ax + b$, zo dat de som $Q(a, b)$, van de kwadraten van de afstanden $d_i = y_i - ax_i - b$, zo klein mogelijk is. Zie onderstaande figuur.

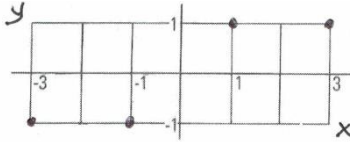


We kiezen aselekt een punt uit deze wolk en maken de eerste coördinaat x_i bekend. We proberen weer de tweede coördinaat te raden, te voorspellen met behulp van $ax_i + b$ (een

eerstegraads functie).

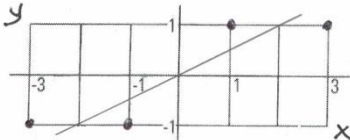
Bij deze voorspelling is de verwachte, gemiddelde fout, $\frac{1}{n} Q(a, b)$. Waarbij n het aantal punten in de puntenwolk is. In opgave 58, de verwachte boete (gemiddelde fout) is de gemiddelde boete $\frac{1}{n} Q(t)$, was de oplossing simpel omdat de gemiddelde y -waarden al op een rechte lijn lagen.

In het algemeen is dit niet te verwachten. In onderstaande figuur zijn 4 punten gegeven.



a) Bepaal de lijn $y = ax + b$, waarbij $Q(a, b)$, minimaal is.

Bij dit voorbeeld zie je direct dat $\bar{y} = 0$, en $\bar{x} = 0$.



In bovenstaande figuur is de lijn $y = ax + b$, getekend. De lijn gaat door het punt $x = 0, y = 0$. Gelet op de punten die op deze lijn liggen, is de vergelijking: $y = \frac{1}{2}x$. Is $Q(a, b)$ minimaal?

b) Is de lijn $y = \frac{1}{2}x$ correct?

We weten dat $a = r \cdot \frac{Sd(y)}{Sd(x)}$, en $r = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} \Rightarrow a = \frac{Cov(x,y)}{Var(x)}$.

Welke formule kiezen we voor $Cov(x, y)$? Het bovenstaande voorbeeld is eenvoudig. Het doet er dus niet zoveel toe.

$$Cov(x, y) = \frac{1}{4} \sum_{i=1}^4 x_i \cdot y_i - \bar{x} \cdot \bar{y}.$$

We weten al dat $\bar{y} = 0$, en $\bar{x} = 0 \Rightarrow Cov(x, y) = \frac{1}{4} \sum_{i=1}^4 x_i \cdot y_i$, en $Var(x) = \frac{1}{4} \sum_{i=1}^4 x_i^2$.

Vul de gegevens in: $Cov(x, y) = \frac{1}{4} \sum_{i=1}^4 x_i \cdot y_i = 2$, en $Var(x) = \frac{1}{4} \sum_{i=1}^4 x_i^2 = 5$.

Dus $a = \frac{Cov(x,y)}{Var(x)} = \frac{2}{5}$. De vergelijking is: $y = \frac{2}{5}x$, want $\bar{y} = 0$, en $\bar{x} = 0$. Het punt $x = 0, y = 0$, wordt het zwaartepunt $(\bar{x}, \bar{y})$ genoemd.

Conclusie: de schatting $y = \frac{1}{2}x$ is fout!

En dan heb je ook gelijk de betekenis van zwaartepunt: de punten die verder weg liggen van het zwaartepunt tellen zwaarder mee.

c) $Q(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2$.

Substitueer in bovenstaande uitdrukking: $z_i = y_i - ax_i \Rightarrow Q(a, b) = \sum_{i=1}^n (z_i - b)^2$.

Uit $z_i = y_i - ax_i$ volgt:

$$\frac{1}{n} \sum_{i=1}^n z_i = \bar{z} = \frac{1}{n} \sum_{i=1}^n (y_i - ax_i) = \frac{1}{n} \sum_{i=1}^n y_i - a \frac{1}{n} \sum_{i=1}^n x_i = \bar{y} - a\bar{x}.$$

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

Verder:.

$$Q(a, b) = \sum_{i=1}^n (z_i - b)^2.$$

We weten $\bar{z} = \bar{y} - a\bar{x}$ en $\bar{y} = a\bar{x} + b \Rightarrow b = \bar{y} - a\bar{x} = \bar{z} \Rightarrow Q(a, b) = \sum_{i=1}^n (z_i - b)^2 = \sum_{i=1}^n (z_i - \bar{z})^2$.

Opmerking: $b = \bar{z}$. In bovenstaand voorbeeld. In het algemeen zal niet gelden: $\bar{y} = 0$, en $\bar{x} = 0$.

Opgave 60:

In opgave 59 hebben we gevonden dat: $Q(a, b) = \sum_{i=1}^n (z_i - b)^2 = \sum_{i=1}^n (z_i - \bar{z})^2$.

Hierbij is de term $n(\bar{z} - b)^2$, weggelaten, want $b = \bar{z}$, in het voorbeeld van opgave 59.

In het algemeen zal $n(\bar{z} - b)^2 \neq 0$, zijn. Dus:

$$Q(a, b) = \sum_{i=1}^n (z_i - b)^2 = \sum_{i=1}^n (z_i - \bar{z})^2 + n(\bar{z} - b)^2.$$

We maken gebruik van:

$z_i = y_i - ax_i$, en $\bar{z} = \bar{y} - a\bar{x}$. Deze uitdrukkingen zijn in de vorige opgave(n) afgeleid.

Dan:

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

Uitwerken geeft:

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

In bovenstaande uitdrukking gaan we gebruik maken van:

$$r = \frac{\text{Cov}(x,y)}{\text{Sd}(x)\text{Sd}(y)} = \frac{1}{n} \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\text{Sd}(x)\text{Sd}(y)} \Rightarrow \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) = n \cdot r \cdot \text{Sd}(x)\text{Sd}(y).$$

Verder gebruiken we:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = n \cdot \text{Var}(x),$$

en

$$\sum_{i=1}^n (y_i - \bar{y})^2 = n \cdot \text{Var}(y).$$

Met bovenstaande ingrediënten vinden we:

$$Q(a,b) = n \cdot a^2 \text{Var}(x) - 2a \cdot n \cdot r \cdot \text{Sd}(x)\text{Sd}(y) + n \text{Var}(y) + n(\bar{y} - a\bar{x} - b)^2.$$

Maken we gebruik van $[\text{Sd}(x)]^2 = \text{Var}(x)$ en $[\text{Sd}(y)]^2 = \text{Var}(y)$, dan kunnen we $Q(a,b)$ herschrijven tot:

$$Q(a,b) = n(a \cdot \text{Sd}(x) - r \cdot \text{Sd}(y))^2 + n(1 - r^2)\text{Var}(y) + n(\bar{y} - a\bar{x} - b)^2,$$

waarbij termen zijn toegevoegd die tegen elkaar wegvallen.

Of:

$$\bar{Q}(a,b) = (a \cdot \text{Sd}(x) - r \cdot \text{Sd}(y))^2 + (1 - r^2)\text{Var}(y) + (\bar{y} - a\bar{x} - b)^2.$$

We gaan deze vergelijking nader onderzoeken.

Als:

$$a = r \cdot \frac{\text{Sd}(y)}{\text{Sd}(x)} \Rightarrow (a \cdot \text{Sd}(x) - r \cdot \text{Sd}(y))^2 = 0.$$

Als:

$$b = \bar{y} - a\bar{x} \Rightarrow (\bar{y} - a\bar{x} - b)^2 = 0.$$

Dus met $y = ax + b$, ligt het zwaartepunt $(\bar{x}, \bar{y})$ op deze lijn $\Rightarrow \bar{y} = a\bar{x} + b$.

Het gevolg van deze twee keuzes is:

$$\bar{Q}(a,b) = (1 - r^2)\text{Var}(y).$$

Uit deze relatie volgt de conclusie

$-1 \leq r \leq 1$, want $Q(a,b)$ zijn afstanden die niet negatief zijn.

Als $Q(a,b) = 0$, liggen alle punten op één lijn: $r = \pm 1$.

Als $r = \pm 1$ en $\text{Sd}(x) \neq 0 \neq \text{Sd}(y)$ dan liggen alle punten op een lijn, want $Q(a,b) = 0$. Hierbij is aangenomen dat het zwaartepunt in ieder geval op de lijn $y = ax + b$ ligt.

- **Oefenen**

Lezen Paragraaf 42

Appendix 1 De kleinste kwadraten methode: De Afleiding(paragrafen 56 t/m 60)

- **Afleiding van de formule voor de regressielijn**

In tegenstelling tot de horizontale afstanden d_i ga je nu het kwadraat van de som van de verticale afstanden tot de regressie lijn minimaliseren:

$$d_i = |y_i - (a \cdot x_i + b)|.$$

X is de onafhankelijke variabele, y is de afhankelijke variabele.
De formule van de regressielijn is van de vorm $y = a \cdot x + b$.

Eerste formule: $\frac{y - \bar{y}}{SD(y)} = r \cdot \frac{x - \bar{x}}{SD(x)}$ of: $z_y = r \cdot z_x$

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

De vergelijking waar het omgaat luidt dan:

$$\sum_{i=1}^n d_i^2 = \sum_{i=1}^n [y_i - (a \cdot x_i + b)]^2.$$

De punten (x_i, y_i) zijn gegeven. Het gaat er nu om de functie $\sum_{i=1}^n d_i^2(a, b)$ te minimaliseren. Dat wil zeggen wanneer is de eerste afgeleide gelijk aan 0?

Zoals je ziet is de functie afhankelijk van twee variabelen (a, b) . Het minimum bepalen van een dergelijke functie is nieuw voor jou. Dat is niet in de voorgaande lessen behandeld.

Je mag aannemen dat je de functie eerst naar a differentieert, terwijl b constant beschouwd wordt, en het resultaat gelijk aan 0 stelt:

$$\sum_{i=1}^n 2x_i[y_i - (a \cdot x_i + b)] = 0 \Rightarrow \sum_{i=1}^n x_i y_i = b \cdot \sum_{i=1}^n x_i + a \cdot \sum_{i=1}^n x_i^2.$$

Daarna wordt de functie $\sum_{i=1}^n d_i^2(a, b)$ naar b gedifferentieerd, terwijl a constant beschouwd wordt, en het resultaat weer gelijk aan 0 wordt gesteld:

$$\sum_{i=1}^n 2[y_i - (a \cdot x_i + b)] = 0 \Rightarrow \sum_{i=1}^n y_i = n \cdot b + a \cdot \sum_{i=1}^n x_i$$

Uit deze twee bovenstaande vergelijkingen (met twee onbekenden a en b) vind je a en b :

$$a = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2},$$

en

$$b = \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)(\sum_{i=1}^n x_i y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}.$$

Nu ga je de formules voor de Covariantie, Variantie en de Correlatiecoëfficiënt gebruiken om bovenstaande formules voor a en b in termen van Covariantie, Variantie en Correlatiecoëfficiënt te formuleren:

$$Covar(x, y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y},$$

$$Var(x) = \frac{1}{n} \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2:$$

$$a = \frac{n^2 [Covar(x, y) + \bar{x} \bar{y}] - n^2 \bar{x} \bar{y}}{n^2 Var(x)} = r \frac{Sd(y)}{Sd(x)},$$

$$b = \frac{n \bar{y} Var(x) - n \bar{x} Var(y)}{n Var(x)} = \bar{y} - \bar{x} \frac{Covar(x, y)}{Var(x)} = \bar{y} - \bar{x} r \frac{Sd(y)}{Sd(x)}.$$

Dit resulteert in:

$$y = ax + b = r \cdot \frac{Sd(y)}{Sd(x)} \cdot x - r \cdot \frac{Sd(y)}{Sd(x)} \cdot \bar{x} + \bar{y}.$$

Hiermee hebben we de tweede formule in bovenstaand blok gevonden.
In de paragrafen 56 t/m 60 van de Wageningse methode wordt door middel van oefeningen de kleinste kwadraten methode bestudeerd.
In Blok 15 Les 4, zijn de meeste oefenopgaven uitgewerkt.

Appendix 2 Samenvatting formules

Samenvatting formules

- Covariantie
- Correlatiecoëfficiënt
- Correlatiecoëfficiënt van de rangnummers
- De regressielijn

Eerste formule voor de covariantie

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

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

Tweede formule voor de covariantie

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

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

Eerste formule voor de correlatiecoëfficiënt

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

Tweede formule voor de correlatiecoëfficiënt

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

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

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

Derde formule voor de correlatiecoëfficiënt

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

Correlatiecoëfficiënt van de rangnummers:

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

Eerste formule voor de regressielijn (van y op x)

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

$$\text{vorm } y = ax + b$$

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

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

Tweede formule voor de regressielijn

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

ofwel

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

Formule voor de regressielijn van x op y

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

$$\text{vorm } y = ax + b$$

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

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

Blok 16 Niet-Euclidische meetkunde

Blok 16 Les 1 Wat Euclidische meetkunde?

Deze Les sluit aan bij Hoofdstuk 1, De elementen van Euclides, van de tekst Niet-Euclidische Meetkunde van de Wageningse Methode.(2013)

- **Euclides**

- Leefde 300 jaar voor Christus.
- Schreef: **De Elementen**.
- De eerste systematische opbouw van de meetkunde.
- De originele tekst is verloren gegaan.
- Via de islamitische cultuur tot ons gekomen.
- Eerste druk in 1482 (in het latijn).



Euclides, ook Euclides van Alexandrië genoemd, was een Hellenistisch wiskundige werkzaam in de bibliotheek van Alexandrië.

- **De Elementen**

- Ga uit van een **aantal basisbegrippen**
 - Definities (23)
 - Algemene inzichten (5)
 - Postulaten (5)
- Propositions: wiskundige beweringen die bewezen worden vanuit de basisbegrippen met de strikte wetten van de logica.
- Gebruik deze basisbegrippen en geen andere.
- Wat al bewezen is, mag je weer gebruiken.



De elementen van Euclides

De Elementen. Dit werk geeft in 13 boeken een systematische verhandeling over de meetkunde. Het is het meest succesvolle handboek en een van de invloedrijkste werken in de geschiedenis van de wiskunde.

Resultaat:

- Meetkundige figuren die je alleen met '**passer en liniaal**' kunt tekenen.
- Een liniaal zonder maatstrepen.
- Geen gradenboog of tekendriehoek zodat je lengtes en hoeken zou kunnen meten.

Trisectrice: Verdeel een hoek in drie gelijke hoeken met passer en een liniaal zonder merktekens (maatstrepen). De Grieken loste dit probleem op met een liniaal met merktekens. In *ARCHIMEDE's LEVER, Trisectrice of an angle by Geometrical Iteration*, www.leennoordzij.me, wordt een oplossing gepresenteerd met een liniaal zonder merktekens en passer in een vaste stand.

- **Definities**

Hier wordt vastgelegd over welke objecten het gaat en wat de basiskenmerken zijn van deze objecten:

1. Een **punt** is een niet deelbaar object.
2. Een **lijn(stuk)** heeft geen breedte.
3. Een **lijnstuk** heeft punten als uiteinde.
4. Een **rechte lijn** is een lijn die gelijk is met de punten erop.
5.

- **Wat is een lijn?**

Volgens Euclides:

Definitie 4. *Een rechte lijn is een lijn die gelijk is met de punten erop.*

Dit is een geheimzinnige omschrijving die tot veel discussie heeft geleid.

Volgens Proclus (412 – 485 na Christus) betekent dit dat op een lijn de afstand tussen twee willekeurige punten gelijk is aan de lengte van het lijnstuk tussen die punten.

- **Algemene inzichten**

Dit zijn regels die je kunt gebruiken bij het bewijzen.

Samen vormen ze een soort van ‘meetkundige bewijslogica’.

Ze zijn de regels van het ‘bewijs-spel’.

- 1 *Dingen, gelijk aan hetzelfde, zijn ook aan elkaar gelijk.*
Als $a = c$ en $b = c$ dan is $a = b$.
- 2 *Als men bij gelijke dingen gelijke voegt, dan zijn de totalen gelijk.*
Als $a = b$ dan is $a + c = b + c$
- 3 *Als men van gelijke dingen gelijke afneemt, dan zijn de totalen gelijk.*
Als $a = b$ dan is $a - c = b - c$
- 4 *Dingen die op elkaar passen zijn gelijk.*
Hier kun je geen algebraïsche formulering bij geven.
Deze regel gebruik je om de congruentie van figuren aan te tonen door ze op elkaar te leggen.
- 5 *Het geheel is groter dan het deel.*
 $a + b > a$

Euclides heeft de bekende term QED: “quod erat demonstrandum” voor het eerst geïntroduceerd.

- **Postulaten**

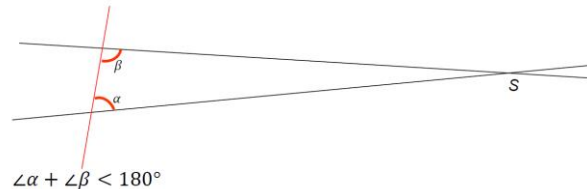
Dit zijn grondwaarheden, waar niemand aan twijfelt en die je bij het bewijzen als basis kunt gebruiken.

Je kunt:

- 1 van elk punt naar elk (ander) punt precies één rechte lijn (lijnstuk) trekken; dus een liniaal gebruiken,
- 2 een lijnstuk verlengen tot een lijn; met een liniaal,
- 3 je kunt met elk punt en elk lijnstuk een cirkel tekenen met het punt als middelpunt en het lijnstuk als straal; je kunt een passer gebruiken,
- 4 alle rechte hoeken zijn aan elkaar gelijk;
(dit betekent niet dat je een rechte hoek kunt tekenen, je mag dus geen geodriehoek gebruiken, een rechte hoek moet je construeren, maar elke constructie leidt tot dezelfde rechte hoek).

- **Het vijfde postulaat**

5 Gegeven twee rechte lijnen en een derde lijn die deze twee lijnen snijdt.
 Gegeven is ook dat de binnenhoeken van deze derde lijn met de twee snijdende lijnen aan dezelfde kant, samen kleiner zijn dan een gestrekte hoek.
 Dan snijden de twee rechten elkaar aan dezelfde kant, tot in het oneindige verlengd.



- **Proposities**

Stellingen (beweringen) over meetkundige objecten en eigenschappen die je moet bewijzen.

Propositie 1

Op een gegeven lijnstuk kun je een gelijkzijdige driehoek construeren.

Gegeven een lijnstuk AB . Construeer een gelijkzijdige driehoek.

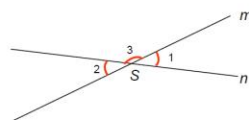
Constructie: volgens het derde postulaat is het mogelijk een cirkel te tekenen met A als middelpunt en AB als straal. Ook is het mogelijk een cirkel te tekenen met B als middelpunt en AB als straal. Deze twee cirkels snijden elkaar in een punt C . Volgens definitie 15 is $AC = AB$, maar ook $BC = AB$. Dus is (algemeen inzicht 1) : $AC = BC = AB$ en figuur ABC (definitie 20) een gelijkzijdige driehoek.

“quod erat faciendum”, wat te construeren was, is een andere term die Euclides heeft geïntroduceerd.

- **Bewijzen en proposities**

Propositie 15 (Overstaande hoeken)

Gegeven twee lijnen die elkaar snijden. Dan zijn de overstaande hoeken aan elkaar gelijk (X-hoeken).



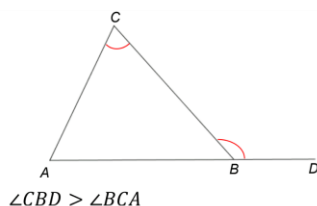
Gegeven: lijn m en lijn n snijden elkaar in S .

Te bewijzen: $\angle S_1 = \angle S_2$

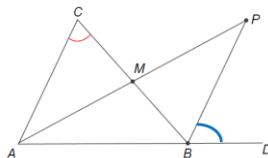
Bewijs: $\angle S_1 + \angle S_3$ is een gestrekte hoek (prop 13).
 $\angle S_2 + \angle S_3$ is een gestrekte hoek (prop 13).
 $\rightarrow \angle S_1 + \angle S_3 = \angle S_2 + \angle S_3$ (aanname 1)
 $\rightarrow \angle S_1 + \angle S_3 - \angle S_3 = \angle S_2 + \angle S_3 - \angle S_3$ (aanname 3)
 $\angle S_1 = \angle S_2$

Propositie 16: De stelling van de buitenhoek

In een driehoek is de buitenhoek (de hoek tussen een zijde en het verlengde van een andere zijde) groter dan elke niet aanliggende hoek.



Bewijs (van Euclides)
 M is het midden van BC .
 $|AM| = |MP|$
 $\rightarrow \triangle AMC \cong \triangle BMP$
 $\rightarrow \angle ACM = \angle PBM$
 $\rightarrow \angle PBM < \angle PBM + \angle DBP$
 $\rightarrow \angle C < \angle DBM$

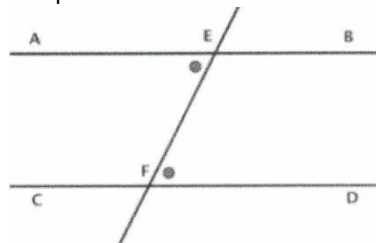


Oefenen en uitgewerkte Oefenopgave.

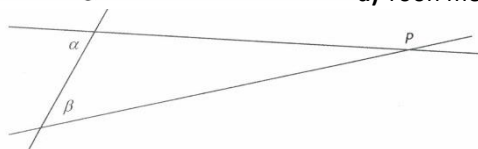
Opgave 9:

In deze opgave bewijzen we propositie 1.27 en 1.28 (als er sprake is van Z-hoeken of F-hoeken dan zijn lijnen evenwijdig).

In de Figuur hieronder: stel dat de hoeken (met stip) gelijk zijn, en dat de lijnen AB en CD elkaar toch ergens in een punt P aan de rechterkant in de Figuur snijden. Dan is er een $\triangle PEF$.

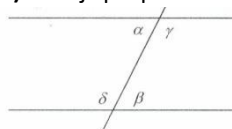


a) Toon met behulp van propositie 1.16 aan dat dit onmogelijk is.



Er is gegeven dat $\alpha = \beta$. Maar stel dat de twee lijnen elkaar snijden in P, dan is er sprake van een driehoek. Volgens propositie 16 is een buitenhoek groter dan ieder van de niet-aanliggende binnenhoeken: $\alpha > \beta$. Dit is in tegenspraak met het gegeven, dus het kan niet.

b) Bewijs propositie 1.27.



In onderdeel a) hebben we bewezen dat er geen snijpunt is aan de rechterzijde (zie bovenstaande Figuur). Als de α en β gelijk zijn, dan zijn ook de hoeken γ en δ gelijk. Immers: $\delta = \pi - \beta$, en $\gamma = \pi - \alpha$, en als je van π gelijke delen afhaalt ($\alpha = \beta$), dan resteren gelijke delen (algemeen inzicht 3). We kunnen nu op soortgelijke wijze als in onderdeel a) bewijzen dat er links geen snijpunt is van de lijnen. De lijnen hebben dus nergens een snijpunt $\therefore$ de lijnen zijn evenwijdig.

Opgave 10

We gebruiken de figuren uit opgave 9.

In deze opgave bewijzen we propositie 1.29 (als lijnen evenwijdig zijn dan is er sprake van gelijke Z-hoeken en F-hoeken). Dit is de omgekeerde stelling van propositie 1.27 en 1.28 en tevens de eerste propositie waarvoor Euclides het parallellenpostulaat nodig had.

Parallellenpostulaat

"En laat geëist zijn dat, als een rechte, die twee rechten treft, de binnenhoeken aan dezelfde kant kleiner dan twee rechte hoeken maakt, de twee rechten, tot in het oneindige verlengd, elkaar ontmoeten aan de kant, waar de hoeken kleiner zijn dan twee rechte hoeken."

a) Stel dat de twee hoeken in de figuur hierboven niet aan elkaar gelijk zijn, bijvoorbeeld $\angle AEF > \angle DFE$.

Bij beide hoeken tellen we $\angle FEB$ op:

$$\angle AEF + \angle FEB > \angle DFE + \angle FEB,$$

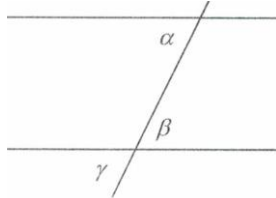
dus

$$180^\circ > \angle DFE + \angle FEB.$$

Nu volgt uit postulaat 5 dat er een snijpunt is.

b) Als we er vanuit gaan dat er geen snijpunt is, dan volgt uit a) dat niet geldt $\angle AEF > \angle DFE$. Maar er mag ook niet gelden $\angle AEF < \angle DFE$, want dan is er aan de andere kant een snijpunt. De enige mogelijkheid die overblijft is: $\angle AEF = \angle DFE$.

c) We maken gebruik van de volgende figuur:



Uit onderdeel a) volgt: $\alpha = \beta$.

Uit propositie 15 volgt: $\beta = \gamma$.

Met algemeen inzicht volgt dan: $\alpha = \gamma$.

- **Oefenen**

Lezen: bladzijde 7

Blok 16 Les 2 Het parallellen postulaat.

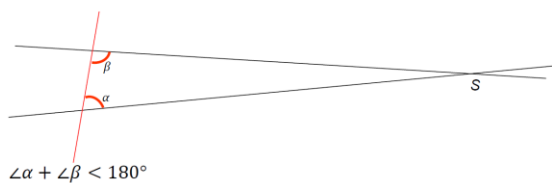
Deze les sluit aan bij paragraaf 1.3 van de tekst Niet-Euclidische meetkunde van de Wageningse Methode 2013

- **Het vijfde postulaat**

5 Gegeven twee rechte lijnen en een derde lijn die deze twee lijnen snijdt.

Gegeven is ook dat de binnenhoeken van deze derde lijn met de twee snijdende lijnen aan dezelfde kant, samen kleiner zijn dan een gestrekte hoek.

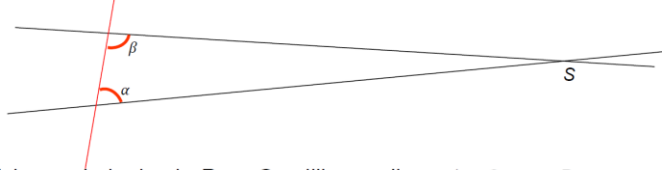
Dan snijden de twee rechten elkaar aan dezelfde kant, tot in het oneindige verlengd.



“tot in het oneindige verlengd”: hoe dichter $\angle \alpha + \angle \beta$ bij π komen hoe verder S naar oneindig verschuift.

De logische structuur van het vijfde postulaat.

$\angle\alpha + \angle\beta < 180^\circ \Rightarrow$ de lijnen snijden elkaar



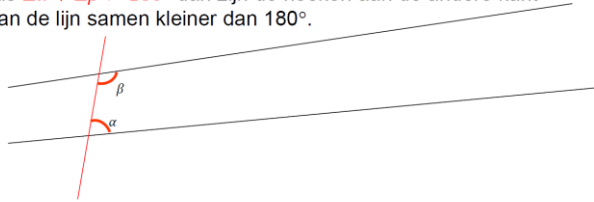
Volgens de logica is $P \Rightarrow Q$ gelijkwaardig met $\neg Q \Rightarrow \neg P$

Waarom is de volgende bewering onjuist:

de lijnen zijn evenwijdig $\Rightarrow \angle\alpha + \angle\beta > 180^\circ$

Uitwerking

Als $\angle\alpha + \angle\beta > 180^\circ$ dan zijn de hoeken aan de andere kant van de lijn samen kleiner dan 180° .



Aan deze kant zijn de hoeken samen kleiner dan 180° .

Volgens het vijfde postulaat snijden de lijnen aan de linkerkant.

De bewering:

$\angle\alpha + \angle\beta < 180^\circ \Rightarrow$ de lijnen snijden elkaar

is dus volgens de logica gelijkwaardig met:

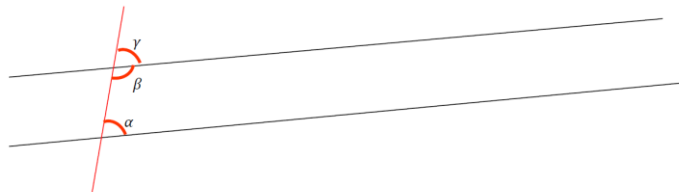
de lijnen zijn evenwijdig $\Rightarrow \angle\alpha + \angle\beta = 180^\circ$

Bewijs hiermee de volgende stelling (opgave 10):

de lijnen zijn evenwijdig $\Rightarrow$ F- en Z-hoeken zijn gelijk

Uitwerking

$\angle\alpha$ en $\angle\gamma$ zijn F-hoeken.



Gegeven: $\angle\alpha + \angle\beta = 180^\circ$

Ook geldt: $\angle\gamma + \angle\beta = 180^\circ$

$\rightarrow \angle\alpha = \angle\gamma$

Op dezelfde manier bewijs je de gelijkheid van de Z-hoeken.

Als $P \Rightarrow Q$ waar is geldt niet automatisch dat ook $Q \Rightarrow P$ waar is.

Dus de volgende bewering is niet vanzelfsprekend waar:

de lijnen snijden elkaar $\Rightarrow \angle\alpha + \angle\beta < 180^\circ$

Dit is volgens de logica gelijkwaardig met:

$\angle\alpha + \angle\beta = 180^\circ \Rightarrow$ de lijnen zijn evenwijdig

Of met:

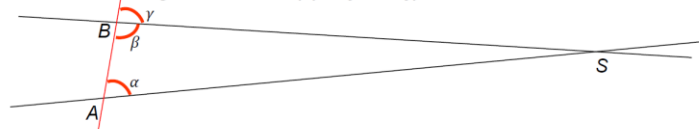
F- en Z-hoeken zijn gelijk $\Rightarrow$ de lijnen zijn evenwijdig

Deze laatste stelling wordt meestal gebruikt in bewijzen van evenwijdigheid.

• Het omgekeerde vijfde postulaat Les 1

Gegeven (zie figuur)(opgave 9):

Twee lijnen m en n worden gesneden door een derde lijn.
 m en n hebben een gemeenschappelijk snijpunt.



Te bewijzen: $\angle\alpha + \angle\beta < 180^\circ$

Bewijs

In $\triangle ASB$ is hoek $\angle\gamma$ een buitenhoek.

Dus $\angle\alpha < \angle\gamma$ volgens propositie 16.

Dan ook: $\angle\alpha + \angle\beta < \angle\gamma + \angle\beta$ (aanname 2)

$\rightarrow \angle\alpha + \angle\beta < 180^\circ$

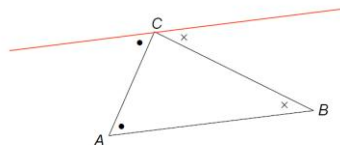
• De som van de hoeken in een driehoek

Het vijfde postulaat $\Rightarrow$ De som van de hoeken in een driehoek is 180°

Gegeven: Het vijfde postulaat.

Te bewijzen:

De som van de hoeken in $\triangle ABC$ is 180° .



Bewijs

Volgens het vijfde postulaat kun je een lijn door C evenwijdig aan AB tekenen.

De som van de hoeken van een driehoek is $180^\circ \Rightarrow$ het vijfde postulaat.

Conclusie:

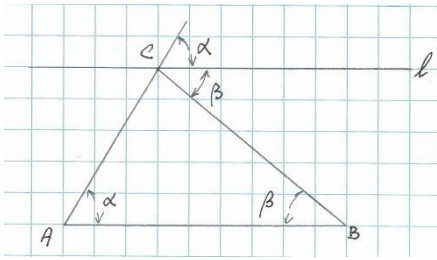
De propositie "De som van de hoeken in een driehoek is 180° " is gelijkwaardig aan het vijfde postulaat.

Oefenen en uitgewerkte Oefenopgaven.

Opgave 14:

Het postulaat over de hoekensom van een driehoek is gelijkwaardig aan het parallellenpostulaat. In bovenstaande les, is al bewezen dat uit de postulaten (inclusief het parallellenpostulaat) volgt dat de som van de hoeken van een driehoek gelijk is aan π . Om de gelijkwaardigheid aan te tonen moeten we bewijzen dat als we de vier postulaten aannemen en aannemen dat de hoekensom in een driehoek gelijk is aan π , dat hieruit het parallellen postulaat volgt. De eerste 28 proposities zijn bewezen zonder gebruik te maken van het parallellen postulaat. Daarom mogen we de 28 proposities voor waar aannemen.

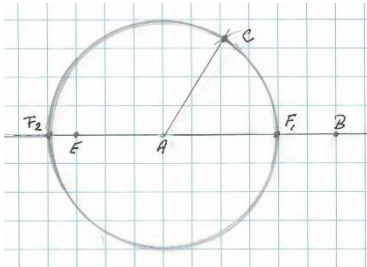
Bewijs dat in een driehoek de buitenhoek gelijk is aan de som van de twee niet aanliggende binnenhoeken.



Volgens het vijfde postulaat kun je een lijn l door C evenwijdig aan AB tekenen. De tekening maakt in een oogopslag duidelijk dat de buitenhoek bij C , $\alpha + \beta$, gelijk is aan de som van de twee, α en β , niet aanliggende binnenhoeken $\alpha + \beta$.

Opgave 15:

Gegeven een lijn AB en een C niet op deze lijn gelegen. Verder is punt F_1 op lijn AB geconstrueerd zo dat $AC = CF_1$. Zie onderstaande tekening.



a) Bewijs dat $\angle AF_1C = \frac{1}{2} \angle CAE$. Het resultaat van de voorgaande opgave 14 kan worden gebruikt. Door de cirkel met straal AF_1 te tekenen is het resultaat duidelijk en wel op twee manieren:

- met opgave 14. $\triangle AF_1C$ is een gelijkzijdige driehoek. Hieruit volgt dat $\angle AF_1C = \angle CAF_1 = \angle F_1CA$.

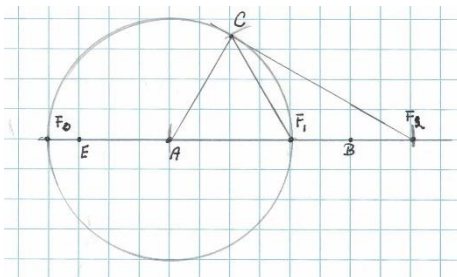
$$\angle CAE = \angle F_1CA + \angle AF_1C = 2 \cdot \angle AF_1C \therefore \angle AF_1C = \frac{1}{2} \angle CAE$$

- uit een vorige les. $\angle CAE = \sphericalangle CF_2$ en $\angle AF_1C = \frac{1}{2} \cdot \sphericalangle CF_2$.

b) Zie onderstaande figuur:

Construeer punt F_2 op lijn AB , rechts van F_1 , zodat $F_1C = F_1F_2$.

Bewijs dat $\angle AF_2C = \frac{1}{2} \angle AF_1C = \left(\frac{1}{2}\right)^2 \angle CAE$.



We kunnen weer verschillende methoden gebruiken. Ik gebruik nu ook de buitenhoek. $\angle AF_1C$ is de buitenhoek van de gelijkbenige driehoek $\triangle CF_1F_2 \therefore \angle AF_1C = 2 \cdot \angle AF_2C \therefore$

$$\angle AF_2C = \frac{1}{2} \angle AF_1C.$$

Onder a) is afgeleid $\angle AF_1C = \frac{1}{2} \angle CAE \therefore \angle AF_2C = \frac{1}{2} \angle AF_1C = \left(\frac{1}{2}\right)^2 \angle CAE$.

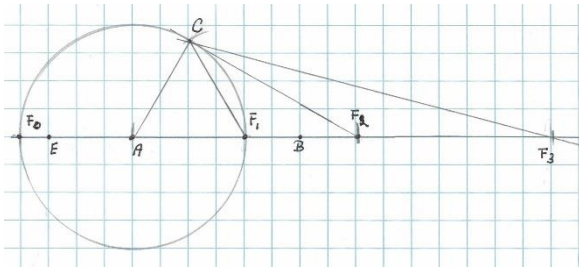
c) Zie onderstaande figuur:

Construeer punt F_3 op lijn AB , rechts van F_2 , zodat $F_2C = F_2F_3$.

Bewijs dat $\angle AF_3C = \left(\frac{1}{2}\right)^3 \angle CAE$.

Construeer punt F_2 op lijn AB , rechts van F_1 , zodat $F_1C = F_1F_2$.

Bewijs dat $\angle AF_3C = \left(\frac{1}{2}\right)^3 \angle CAE$.



De buitenhoek methode wordt weer gebruikt. $\angle AF_2C$ is de buitenhoek van de gelijkbenige driehoek $\triangle CF_2F_3 \therefore \angle AF_2C = 2 \cdot \angle AF_3C \therefore \angle AF_3C = \frac{1}{2} \angle AF_2C$. Onder b) is afgeleid

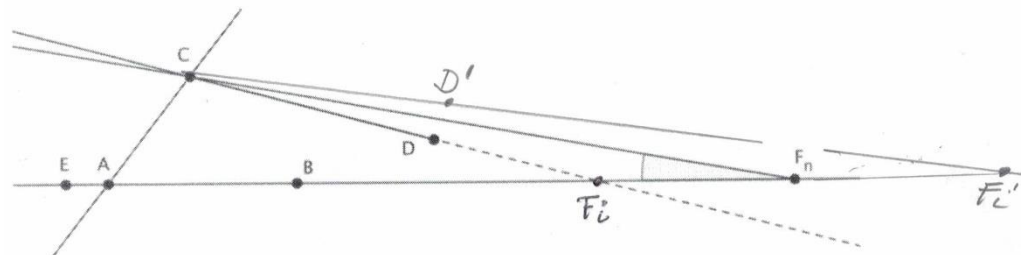
$\angle AF_1C = \frac{1}{2} \angle CAE \therefore \angle AF_2C = \frac{1}{2} \angle AF_1C = \left(\frac{1}{2}\right)^2 \angle CAE$. We weten $\angle AF_3C = \frac{1}{2} \angle AF_2C \therefore \angle AF_3C = \left(\frac{1}{2}\right)^3 \angle CAE$.
Opmerking: $\angle CAE = \frac{2}{3} \pi \Rightarrow \angle AF_3C = \frac{1}{12} \pi \Rightarrow 15^\circ$.

d) Bewijs dat het mogelijk is een driehoek AF_nC te construeren waarbij $\angle AF_nC < 10^\circ$.

Zoals je uit het resultaat onder c) kunt opmaken, gebeurt dit al bij de volgende stap: Construeer punt F_4 op lijn AB , rechts van F_3 , zodat $F_3C = F_3F_4$. Ik laat dit niet in een volgende figuur zien. De procedure spreekt voor zich zelf: $\angle AF_3C$ is de buitenhoek van de gelijkbenige driehoek $\triangle CF_3F_4 \therefore \angle AF_3C = 2 \cdot \angle AF_4C \therefore \angle AF_4C = \frac{1}{2} \angle AF_3C$. Onder c) is afgeleid: $\angle AF_3C = \left(\frac{1}{2}\right)^3 \cdot \angle CAE \Rightarrow \angle AF_4C = \left(\frac{1}{2}\right)^4 \cdot \angle CAE = \frac{1}{24} \pi \Rightarrow 7.5^\circ < 10^\circ$.

e) Gelet op het resultaat onder d), mogen we concluderen dat in elke volgende stap van $\angle AF_iC$ naar $\angle AF_{i+1}C$, $\angle AF_{i+1}C$ de helft is van $\angle AF_iC$. Conclusie: $\lim_{n \rightarrow \infty} \angle AF_nC = \lim_{n \rightarrow \infty} \frac{1}{2^n} \cdot \angle CAE = 0$. Dit duidt erop dat we met dit resultaat, het parallellenpostulaat kunnen bewijzen.

Opgave 16:



In bovenstaande tekening zie je twee lijnen AB en CD waarbij lijn AC lijnen AB en CD zodanig snijdt dat $\angle BAC + \angle DCA < 170^\circ$. Verder is een driehoek geconstrueerd waarbij $\angle AF_nC < 10^\circ$. Opgave 15 heeft laten zien dat dit mogelijk is.

a) Toon aan $\angle ACF_n > \angle ACD$.

Om duidelijk te maken dat je geen conclusies kunt/mag trekken uit de tekening, heb ik ook de lijn CF'_i toegevoegd.

$\angle BAC + \angle DCA < 170^\circ$, en $\angle BAC + \angle DCA + \angle CF_iA = 180^\circ \therefore \angle CF_iA > 10^\circ$.

$\angle BAC + \angle DCA + \angle CF_iA = 180^\circ$ en $\angle BAC + \angle ACF_n + \angle CF_nA = 180^\circ$. Trek beide vergelijkingen van elkaar af:

$\angle DCA + \angle CF_iA = \angle ACF_n + \angle CF_nA$. Substitueer in deze vergelijking $\angle CF_iA > 10^\circ$ en het gegeven $\angle AF_nC < 10^\circ$:

$\angle ACF_n = \angle DCA + \angle(> 10^\circ) - \angle(< 10^\circ) \therefore \angle ACF_n > \angle ACD (\equiv \angle DCA)$.

Opmerking: Als ik de lijn CF'_i , had gebruikt had ik hetzelfde resultaat gevonden.

b) Waarom volgt hieruit dat de lijn CD , de lijn AB ergens tussen A en F_n snijdt?

Deze vraag heeft mij er toe gebracht om ook de lijn CF'_i in de bovenstaande figuur op te nemen. Dit heb ik gedaan om te voorkomen dat je het bewijs onder a) zou ontleenen aan de figuur. Onder a) is het bewijs volledig gebaseerd op de gegevens:

$\angle BAC + \angle DCA < 170^\circ$ en $\angle AF_nC < 10^\circ$.

Onder a) hebben we gevonden dat: $\angle CF_iA > 10^\circ$. Verder weten we dat $\angle AF_nC < 10^\circ$.

Kan F_i rechts van F_n liggen (F'_i)? Met gebruikmaking van de buitenhoek, zou dit betekenen dat

$\angle CF_iA + \angle F_nCF_i = \angle AF_nC < 10^\circ$. Maar, $\angle CF_iA > 10^\circ$, dus $\angle CF_iA + \angle F_nCF_i < 10^\circ$, is onmogelijk en dus

een tegenspraak. Hieruit volgt dat de lijn CD, de lijn AB ergens tussen A en F_n snijdt.

c) Stel nu $\angle BAC + \angle DCA < 179^\circ$. Toon aan dat de lijn CD, de lijn AB ergens snijdt.

In opgave 15^e hebben we laten zien dat $\angle AF_n C < 1^\circ$, kan zijn. Hier uit volgt dat de lijn CD, de lijn AB ergens snijdt.

d) Toon het parallellenpostulaat aan. Het parallellen postulaat ofwel het vijfde postulaat wordt in bovenstaande Les besproken.

“En laat geëist zijn dat, als een rechte, die twee rechten treft, de binnenhoeken aan dezelfde kant kleiner dan twee rechte hoeken maakt, de twee rechten, tot in het oneindige verlengd, elkaar ontmoeten aan de kant, waar de hoeken kleiner zijn dan twee rechte hoeken.” .” Of met de moderne formulering: Gegeven een lijn en een punt buiten deze lijn: In de Euclidische meetkunde gaat er door dit punt precies één lijn parallel aan de gegeven lijn.

Het vijfde postulaat is dus geldig. Dit hebben we in opgave 15 aangetoond: $\lim_{n \rightarrow \infty} \angle AF_n C = 0$, met F_n tot in het oneindige naar rechts verschoven

- Oefenen

Lezen: Bladzijde 9 en 10

Blok 16 Les 3 Meetkunde op de bol

Deze les sluit aan bij paragrafen 2.1 en 2.2 van de tekst Niet-Euclidische meetkunde van de Wagoningse Methode 2013

- Kun je het vijfde postulaat afleiden uit de voorgaande vier?

Vanwege de bijzondere formulering is lange tijd gezocht naar een bewijs dat je het vijfde postulaat met de voorgaande vier postulaten kunt bewijzen.

Dus ook dat je de stelling over de som van de hoeken in een driehoek zou kunnen bewijzen met de eerste vier postulaten.

- Niet-Euclidische meetkunde

Euclidische meetkunde:

- het vijfde postulaat doet mee: door een punt buiten een lijn gaat **precies één** lijn die parallel is met die lijn.

Niet- Euclidische meetkunde:

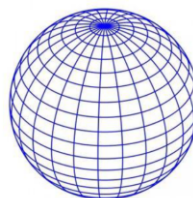
- het vijfde postulaat doet niet mee
 - **Elliptische meetkunde:** door een punt buiten een lijn gaat **geen** lijn die parallel is met die lijn.
 - **Hyperbolische meetkunde:** door een punt buiten een lijn gaan **oneindig veel** lijnen die parallel zijn met die lijn.

- Meetkunde op de bol

Op de aarde wordt een plaats aangegeven door middel van coördinaten in het netwerk van **meridianen** en **breedtecirkels (parallellellen)**.

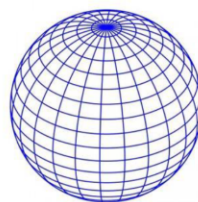
De verticale cirkels zijn de meridianen.
Ze geven de lengtegraden aan.

De horizontale cirkels zijn de breedtecirkels,
Ze geven de breedtegraden aan.



Op de aarde wordt een plaats aangegeven door middel van coördinaten in het netwerk van **meridianen** en **breedtecirkels (parallelle)**.

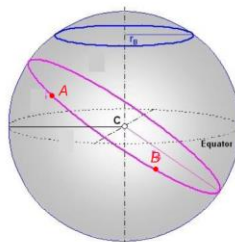
De meridianen zijn allemaal even groot en elke meridiaan is de doorsnede van de bol met een verticaal vlak door het middelpunt van de aarde.



Breedtecirkels verschillen in grootte. Ze zijn de doorsnede van een horizontaal vlak met de bol. De evenaar is de enige breedtecirkel die een doorsnede is van de bol met een vlak door het middelpunt van de aarde.

De evenaar is de doorsnede vlak door het middelpunt van de aarde en de aardbol.

Gegeven twee willekeurige punten A en B op de aardbol.
Neem het vlak door A , B en het middelpunt C .
De doorsnede van dit vlak met de bol is een cirkel door A en B .
Een dergelijke cirkel heet een **grootcirkel**.

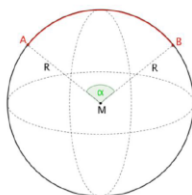


De kortste verbinding tussen A en B op de bol is het cirkelsegment op de grootcirkel door A en B .

In het platte Euclidische vlak is de rechte lijn de kortste verbinding tussen twee punten.

Op dezelfde manier wordt een grootcirkel op de bol opgevat als een (rechte) lijn op de bol.

Voor een lijnstuk tussen A en B kun je kiezen tussen de rode cirkelboog (boven) of de zwarte (onder).



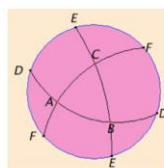
Voor de kortste boog geldt dat $\alpha < \pi$ (radialen)
Het rode cirkelsegment is nu per definitie in deze meetkunde de kortste afstand tussen de punten A en B .
Dit noemen we het lijnstuk AB op de bol.

• Diametrale punten

De Noordpool en de Zuidpool liggen op één lijn met het middelpunt M . Dat betekent dat er oneindig veel vlakken door deze drie punten gaan. En dus dat beide punten door oneindig veel grootcirkels (de meridianen) worden verbonden.

Het zelfde geldt voor andere **diametrale** punten op de bol, dat zijn punten die op één lijn liggen met het middelpunt.

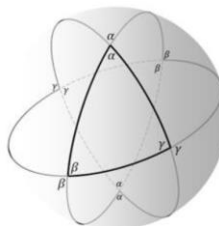
Om problemen met het eerste postulaat te voorkomen, vatten we een dergelijk diametraal puntenpaar op als één punt (met twee locaties).



• Het vijfde postulaat

Volgens Euclides zijn twee lijnen evenwijdig (parallel) als ze geen snijpunt hebben.

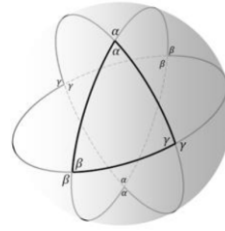
In de meetkunde op de bol geldt dezelfde afspraak:
twee lijnen (= grootcirkels) zijn evenwijdig (parallel) als ze geen snijpunt hebben.



Verklaar waarom twee verschillende lijnen op de bol altijd één snijpunt hebben.

Uitwerking

De twee grootcirkels zijn de doorsnijdingen van verschillende vlakken door M .
De twee vlakken hebben een snijlijn.
De snijlijn van die vlakken doorsnijden de bol in twee diametrale punten.
Dat geldt als één punt.

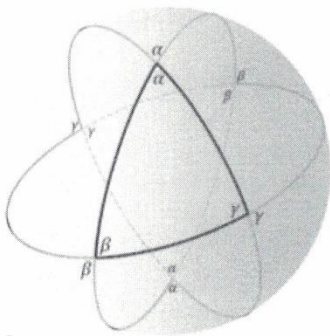


Conclusie

Voor de meetkunde op de bol geldt het vijfde postulaat niet.

Oefenen en uitgewerkte Oefenopgave.

Opgave 21:



a) Is tussen ieder tweetal punten op een bol precies één lijn te trekken?

Ja. Dus het eerste postulaat is geldig.

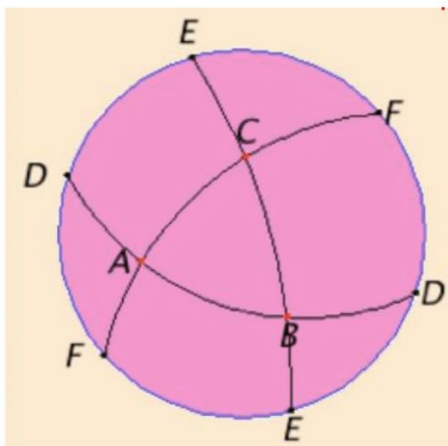
b) Kan ieder lijnstuk verlengd worden tot een rechte lijn? Ja. Dus het tweede postulaat is geldig.

c) In de Euclidische meetkunde verdeelt een lijn het platte vlak in twee delen. Als we een punt in het ene deel verbinden met een punt in het andere deel dan snijdt deze verbindende lijn de gegeven lijn. Geldt dit ook op de bol? Ja.

d) Zijn het derde en het vierde postulaat geldig? Ja.

Opgave 22:

Construeer willekeurige lijnen (grootcirkels) op een bol en onderzoek of ze elkaar snijden.



a) Zie nevenstaande tekening. Verwijder alle lijnen op één na (**FACF**) en kies een punt dat niet op die lijn ligt: B. Hoeveel lijnen gaan er door dit punt die de gegeven lijn niet snijdt?

Denk de lijnen **DABD** [zo wordt een (groot) cirkel aangegeven] en **ECBE** weg en behoudt het punt B.

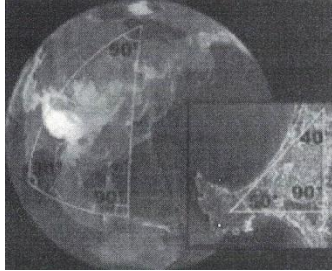
Hoeveel lijnen gaan er dan door het punt B die de lijn **FACF** niet snijden? Er gaan nul lijnen door het punt B die de cirkel **FACF** niet snijden.

b) Is het vijfde postulaat geldig?

Het vijfde postulaat: *“En laat geëist zijn dat, als een rechte, die twee rechten treft, de binnenhoeken aan dezelfde kant kleiner dan twee rechte hoeken maakt, de twee rechten, tot in het oneindige verlengd, elkaar ontmoeten aan de kant, waar de hoeken kleiner zijn dan twee rechte hoeken.”* Of met de moderne formulering: Gegeven een lijn en een punt buiten deze lijn: In de Euclidische meetkunde gaat er door dit punt precies één lijn parallel aan de gegeven lijn.

Het vijfde postulaat is dus niet geldig.

Opgave 23:



De hoekensom is niet altijd 180° .

De hoekensom is minimaal 180° en maximaal 360° .

- Oefenen

Lezen: Opgave: 18.

Blok 16 Les 4 Hyperbolische meetkunde

Deze les sluit aan bij paragraaf 3 van de tekst Niet-Euclidische meetkunde van de Wageningse Methode 2013

- Euclidische meetkunde

Het vijfde postulaat doet mee:

Door een punt buiten een lijn gaat **precies één** lijn die parallel is met die lijn.

Dat is de vlakke meetkunde zoals je die al kent: met lijnen, driehoeken, veelhoeken, cirkels, parabolen,.....

NB **parallel** betekent: geen gemeenschappelijk snijpunt.

Het vijfde postulaat doet niet mee

- **Elliptische meetkunde**: door een punt buiten een lijn gaat **géén** lijn die parallel is met die lijn.
- **Hyperbolische meetkunde**: door een punt buiten een lijn gaan **oneindig veel** lijnen die parallel zijn met die lijn.

- Elliptische meetkunde

Meetkunde op de bol: door een punt buiten een lijn gaat **géén** lijn die parallel is met die lijn.

- Grootcirkels zijn lijnen.
- Som van de hoeken in een driehoek groter dan 180°
- Toepassingen bij navigatie op de aardbol
- Concreet voorstelbaar

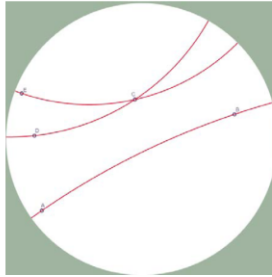
- Hyperbolische meetkunde

Meetkunde op de Poincaré-schijf

Dit is een bijzondere meetkunde binnen een cirkelschijf.

Is dit nog wel meetkunde?

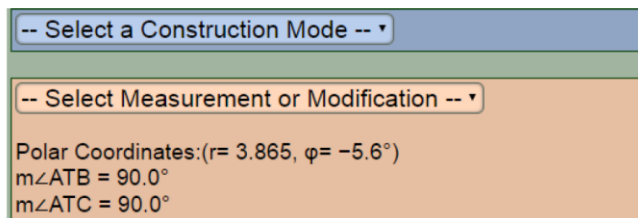
Lees vooraf paragraaf 2.3



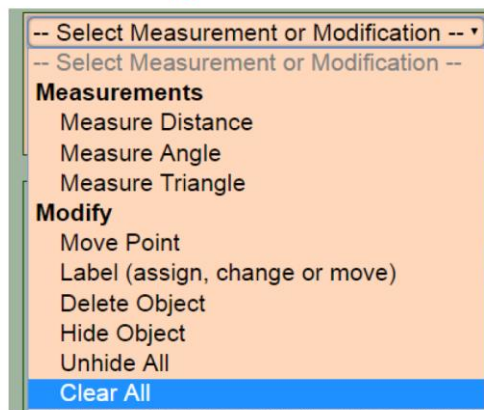
- De Poincaré schijf

Een computerpracticum

- Ga naar <https://www.cs.unm.edu/~joel/NonEuclid/NonEuclid.html>
- Je ziet eerst een animatie.
- Met de menu's kun je experimenteren



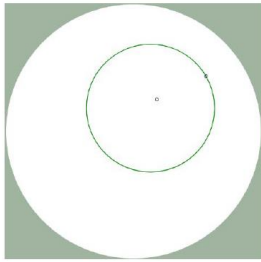
Maak eerst je scherm leeg(Clear all)



- Klik op het pijltje van -- Select a Construction Mode --
- Teken twee punten
- Teken een lijnstuk door beide punten
- Teken een lijn
- Teken een punt buiten die lijn
- Teken een lijn door het punt die parallel is met de eerste lijn.

Ook in deze meetkunde is een cirkel de verzameling van alle punten met een vaste afstand tot het middelpunt.

Teken een cirkel:



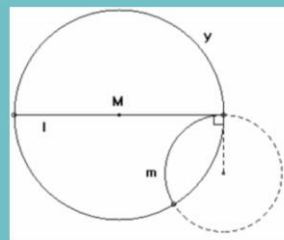
Je ziet dat het middelpunt niet in het midden van de cirkel zit. Dat komt omdat de afstanden worden berekend met een formule, waarbij je de afstand tot de rand als oneindig ver weg ervaart.

- Met Move Point kun je het middelpunt verplaatsen. Verschuif het middelpunt zo dicht mogelijk naar de rand en zie wat er gebeurt.

Een **lijn** op de Poincaré schijf is

- een 'open' cirkelboog die de rand γ loodrecht snijdt.
- een 'open' middellijn door het midden M van de schijf. ('open' betekent de rand doet niet mee.)

Een **hoek** tussen twee snijdende lijnen is de hoek tussen de raaklijnen aan de cirkelbogen.



Met opgave 27 krijg je een idee hoe de afstand gemeten wordt op de Poincaré-schijf. Wees bewust van de "fallacy", dat je vanuit jouw Euclidische ruimte naar de Poincaré-schijf kijkt.

Computerpracticum: Is dit nog wel meetkunde?

Net als bij de Euclidische meetkunde kun je het midden van een lijnstuk construeren met passer en liniaal.

1. Teken een lijnstuk op de Poincaré schijf en construeer het midden van dat lijnstuk.
2. Teken een driehoek en construeer de zwaartelijnen in de driehoek en laat zien dat de zwaartelijnen door één punt gaan.

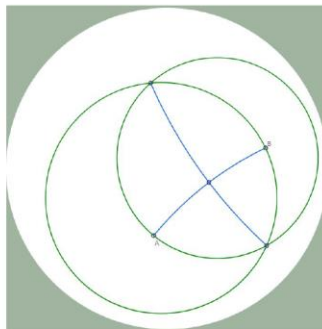
Op NonEuclid.jar kun je je constructies controleren via het menu:

-- Select from Gallery / Activities or Save --

Laat met het programma NonEuclid.jar zien dat de eerste vier postulaten op de Poincaré-schijf geldig zijn en dat het vijfde niet geldig is.

Uitwerking (postulaat 4)

In de figuur zie je hoe je een rechte hoek kunt construeren. Door punt A of B te verslepen kun je controleren dat alle rechte hoeken gelijk zijn.



Uitwerking (postulaat 5)

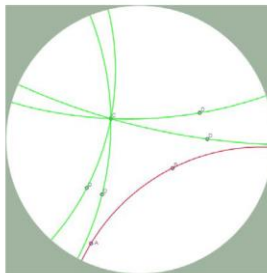
Gegeven een lijn door A en B.

C is een punt buiten de lijn.

De lijnen door C en D zijn parallel met de lijn door A en B.

Door D te verslepen krijg je oneindig veel lijnen die parallel zijn met de lijn AB.

Postulaat 5 is niet geldig.



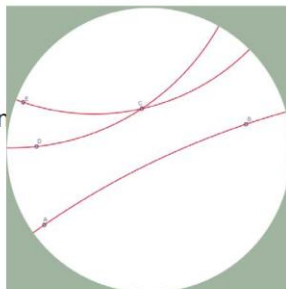
Propositie 32 stelt dat de som van de hoeken in een driehoek 180° .
Bij het bewijs hiervan heb je het vijfde postulaat nodig.

Onderzoek de geldigheid van propositie 32 op de Poincaré cirkel.

• Hyperbolische meetkunde

Meetkunde op de Poincaré-schijf: door een punt buiten een lijn gaan oneindig veel lijnen die parallel zijn met die lijn.

- Meetkunde op een cirkelschijf
- De rand doet niet mee
- Hoe dichter één punt bij de rand, hoe groter de 'afstand' tussen twee punten
- Cirkelbogen die loodrecht staan op de rand zijn lijnen.
- Som van de hoeken in een driehoek kleiner dan 180°
- Abstract model
- In beeld gebracht door Maurits Escher



• Oefenen

Met Geogebra-HM.ggb kun je hyperbolische constructies uitvoeren.

Voor de Poincaré-schijf maak je gebruik van NonEuclid.jar.

Maken: De opgaven 26 tot en met 30.

Lezen : Bladzijde 27.

Verder:

- Laat zien hoe je een rechthoekige driehoek kunt construeren op de Poincaré schijf.
- Onderzoek hiermee de geldigheid van de stelling van Pythagoras op de Poincaré schijf.

Epiloog Waarom de Samenvatting Online?

Deze samenvatting van Wiskunde D-Online VWO 4 heb ik gemaakt voor mijn kleindochter.

Waarom deze samenvatting online geplaatst?

Mijn kleindochter wilde voor haar VWO opleiding Wiskunde D als keuzevak. Helaas, haar school had Wiskunde D niet meer opgenomen in het lesprogramma.

Wat te doen?

Het bleek dat er meer scholen waren/zijn die Wiskunde D niet meer opgenomen hebben in het lesprogramma. Dit is aanleiding geweest om de Stichting Wiskunde D-Online op te zetten.

Om gebruik te kunnen maken van het Wiskunde D programma door leerlingen die toch Wiskunde D willen kiezen, dient een school zich aan te melden. (Het zal duidelijk zijn dat deze scholen door zich aan te melden hun leerlingen willen steunen.) Op deze wijze krijgen de scholen toegang tot het online-lesprogramma. Het lesmateriaal is samengesteld uit: PowerPoint presentaties, video's, boeken (De Wageningse Methode) met opgaven, antwoorden en toetsen. Daarnaast hebben de docenten van participerende scholen toegang tot het docentenmateriaal.

Zoals gezegd, de scholen die toch buiten het reguliere lesprogramma hun leerlingen willen faciliteren bij de Wiskunde D keuze moeten zich aanmelden. Tevens een docent ter beschikking stellen om de leerlingen te begeleiden bij de studie en het maken van opgaven. Toetsen kunnen ter beoordeling worden opgestuurd.

De school van mijn kleindochter wilde niet participeren en evenmin een docent ter beschikking stellen. Mijn kleindochter was de enige leerling met Wiskunde D keuze. Met als gevolg dat ik samen met mijn kleindochter Wiskunde D-Online heb gedaan.

Nu de reden dat ik de samenvatting online (www.leennoordzij.me) heb geplaatst: het kan nog steeds zo zijn dat er leerlingen Wiskunde D kiezen zonder dat een school voor begeleiding zorg wil dragen. Door deze samenvattingen te publiceren hoop ik de belangstelling voor Wiskunde D te stimuleren.

Remark: To whom it concerns.

When reading the above text in your mother's tongue, the translation can sometimes be hilarious. For example, in Dutch: "geluk". The meaning of this word in my text is usual: by chance, by coincidence or by good fortune. However, various translation programs come up with happiness. In my text, I am not involved with "happiness".

Obviously, it made me happy to support my granddaughter!

