Defensie Ondersteuningscommando
Ministerie van Defensie

Statistiek: college 13

Correlatie en lineaire regressie

DOSCO
Dr. ir. Danny Blom
Nederlandse Defensie Academie
Faculteit Militaire Wetenschappen

Juli 2025

Recap: Statistiek deel 2

- Schatten en betrouwbaarheidsintervallen

- Algemeen stappenplan voor hypothesetoetsen

- Chikwadraattoetsen:
  - onafhankelijkheid van twee categorische variabelen
  - Aanpassingstoets voor een specifieke kansverdeling (“goodness-of-fit test”)

- Verschiltoetsen:
  - Toetsen voor gelijke verwachtingswaardes van twee onafhankelijke populaties
  - F-toets voor gelijke varianties.

Leerdoelen

Aan het eind van dit college kunnen studenten:

- het verschil uitleggen tussen correlatie en regressie, inclusief het doel en toepassingsgebied.

- correlatiecoëfficiënten van Pearson en Spearman uitrekenen en aan de hand hiervan de sterkte en richting van een correlatie duiden.

- eenvoudige regressieanalyses uitvoeren en interpreteren binnen een praktische context.

- de uitkomsten toepassen op bedrijfskundige en militaire vraagstukken, met aandacht voor beperkingen en aannames van deze methoden

Verbanden tussen twee variabelen

In de vorige twee colleges hebben we verbanden tussen variabelen behandeld:

Chikwadraattoetsen	Verschiltoetsen
• Verband tussen twee categorische variabelen (nominaal / ordinaal)	• Verbanden tussen twee onafhankelijke populaties op een enkele continue variabele (interval / ratio)

Vandaag: (lineaire) samenhang van twee ratio variabelen X en Y

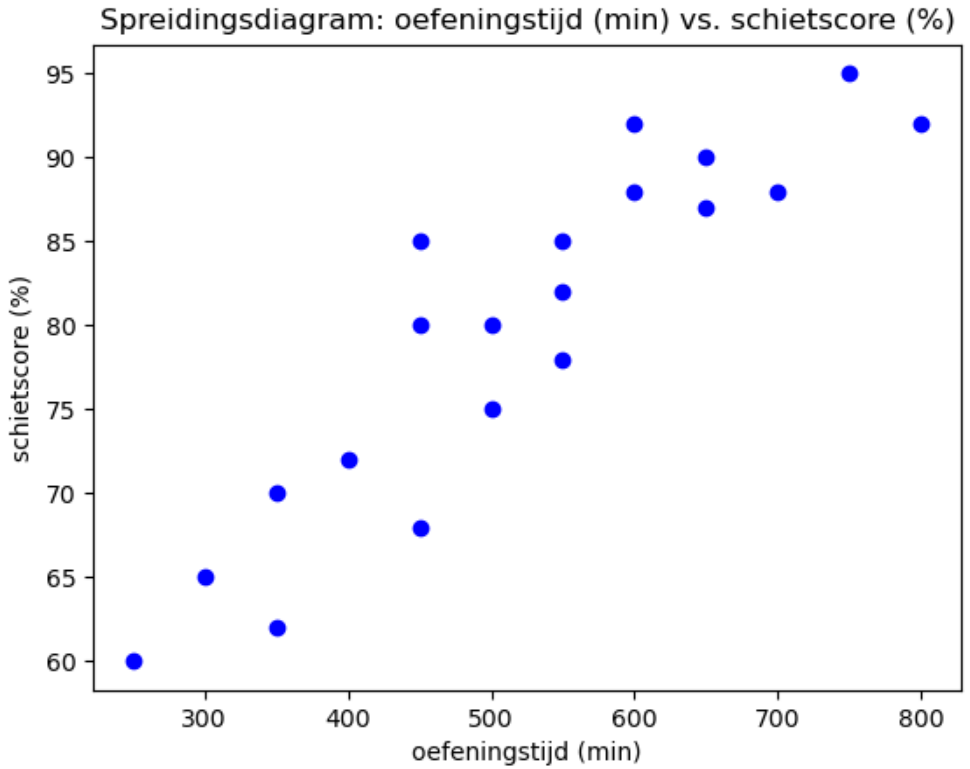
Voorbeeld: jaarlijkse keuring van actieve militairen

ID	Rang	Eenheid	Oefeningstijd (min)	Schietscore (%)	Aantal missies	Geslacht
1	Sergeant	Bravo	450	85	12	Man
2	Korporaal	Alpha	600	92	8	Vrouw
3	Luitenant	Charlie	550	78	15	Man
4	Majoor	Bravo	700	88	20	Man
5	Soldaat	Alpha	300	65	5	Vrouw
6	Kapitein	Charlie	500	80	10	Vrouw
7	Sergeant	Bravo	650	90	18	Man
8	Soldaat	Alpha	400	72	6	Man
9	Soldaat	Charlie	350	70	4	Man
10	Kapitein	Bravo	750	95	25	Vrouw
11	Luitenant	Alpha	550	82	12	Man
12	Korporaal	Charlie	450	68	7	Vrouw
13	Sergeant	Bravo	600	88	14	Man
14	Majoor	Alpha	800	92	30	Vrouw
15	Soldaat	Charlie	250	60	3	Vrouw
16	Kapitein	Bravo	650	87	18	Man
17	Luitenant	Alpha	500	75	9	Man
18	Sergeant	Charlie	550	85	13	Vrouw
19	Korporaal	Bravo	450	80	10	Man
20	Soldaat	Alpha	350	62	5	Vrouw

Welke variabelen zijn geschikt om te testen op lineaire samenhang?

Spreidingsdiagram (scatter plot)

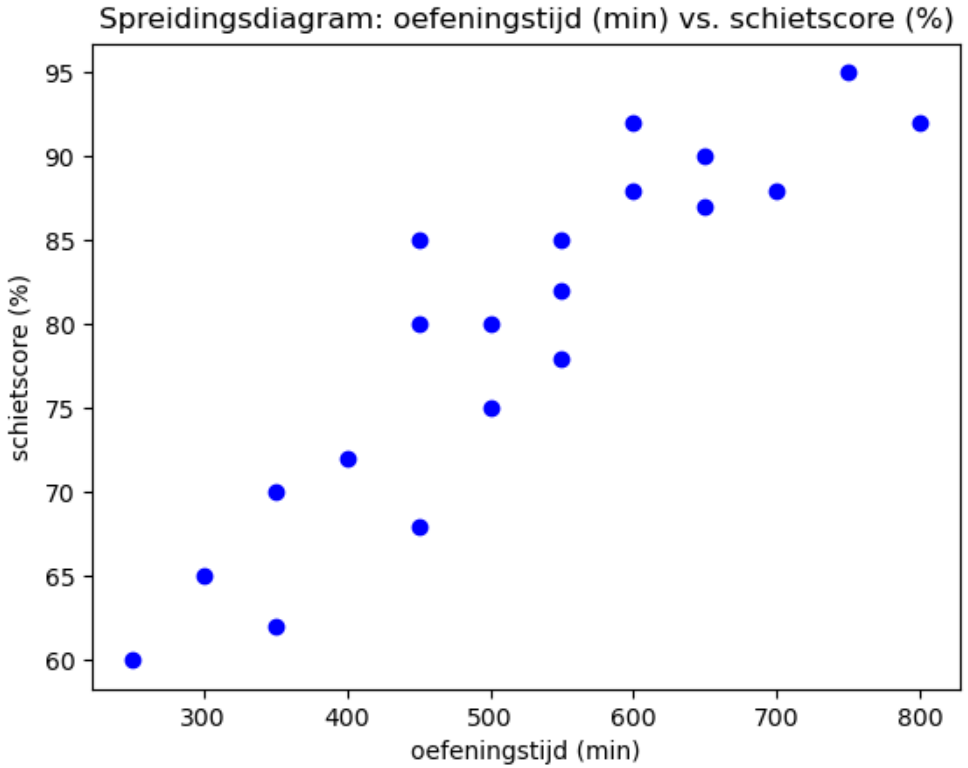
Oefeningstijd (min)	Schietscore (%)
450	85
600	92
550	78
700	88
300	65
500	80
650	90
400	72
350	70
750	95
550	82
450	68
600	88
800	92
250	60
650	87
500	75
550	85
450	80
350	62



Hoe zouden we statistisch verantwoord de samenhang tussen oefeningstijd en schietscore kunnen analyseren?

Correlatie

In het spreidingsdiagram is een duidelijke lineaire trend te zien.



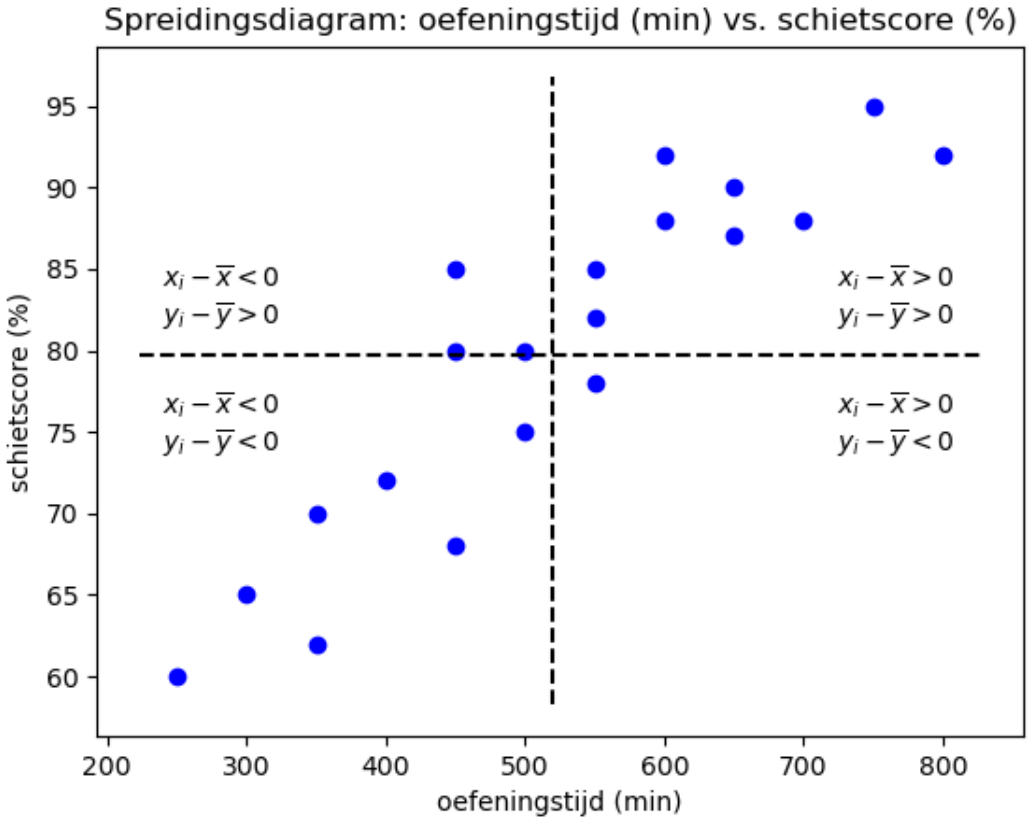
Correlatie: kwantitatieve maatstaf voor de mate van samenhang van twee variabelen.

Covariantie

Gegeven: geobserveerde steekproef $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

- Verdeel de steekproef in kwadranten
- Merk op:
  - Meeste datapunten linksonder en rechtsboven
  - In deze twee kwadranten is het product

    $(x_i - \overline{x}) \cdot (y_i - \overline{y}) > 0$



De eerste bouwsteen is de **covariantie:**

$$\mathrm{Cov}(x, y) = \frac{\sum (x_i - \overline{x}) \cdot (y_i - \overline{y})}{n - 1}$$

Pearson's correlatiecoëfficiënt

De interpretatie van de covariantiewaarde is erg lastig en hangt af van meeteenheden (als X in seconden zou worden gemeten, dan wordt de covariantie $60\times$ zo groot)

Wat betekent een covariantie van 8? Is dat hoog of is dat laag?

Om die reden wordt vaak naar een gestandaardiseerde waarde gekeken tussen -1 en 1, de **correlatiecoëfficiënt** (gedeeld door beide steekproefstandaardafwijkingen s_x en s_y):

$$r(x,y) = \frac{\mathrm{Cov}(x,y)}{s_x \cdot s_y} = \frac{\frac{\sum(x_i-\overline{x})\cdot(y_i-\overline{y})}{n-1}}{\sqrt{\frac{\sum(x_i-\overline{x})^2}{n-1}}\cdot\sqrt{\frac{\sum(y_i-\overline{y})^2}{n-1}}}$$

Dit wordt ook wel **Pearson's correlatiecoëfficiënt** genoemd.

Pearson's correlatiecoëfficiënt

Aangezien de formule op de vorige slide moeilijk dan wel onwerkbaar is, gebruiken we een andere vorm.

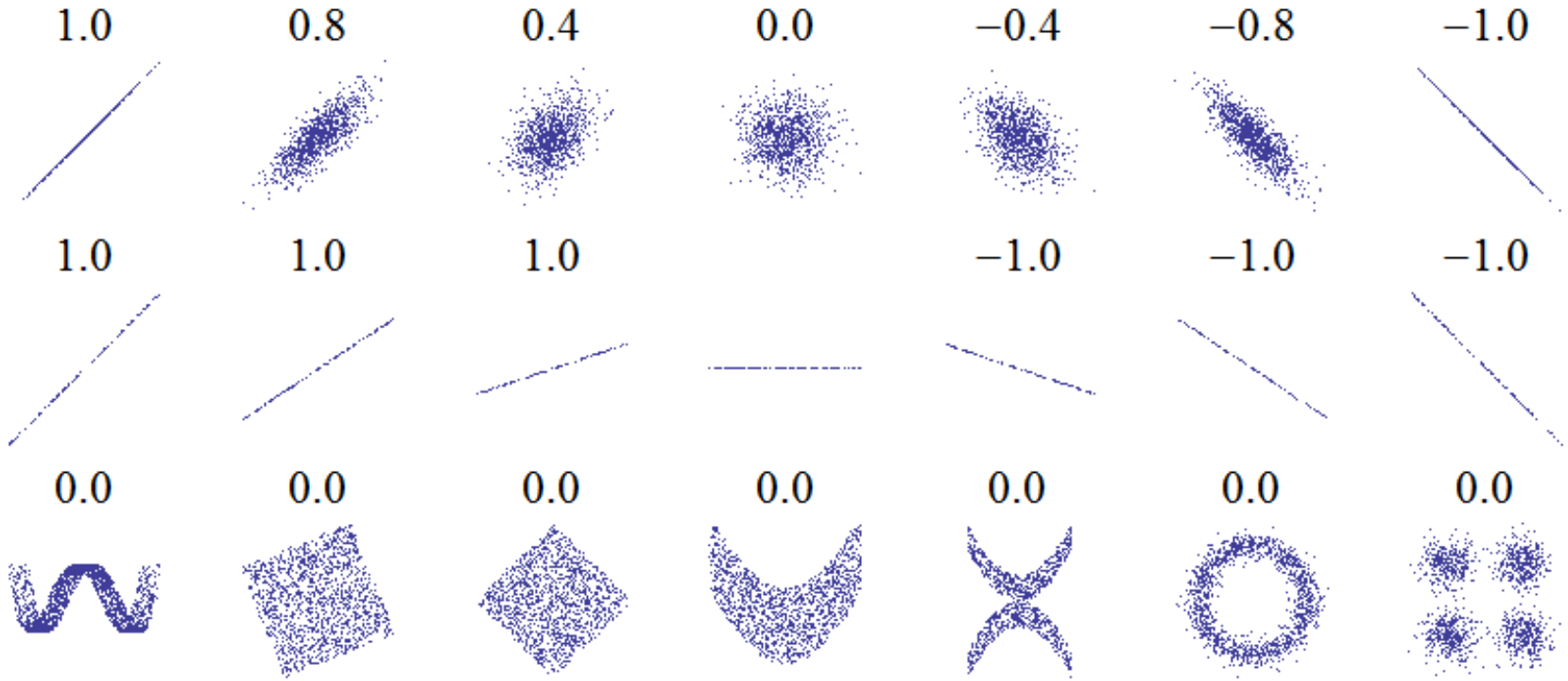
$$r(x,y) = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2) \cdot (\overline{y^2} - \overline{y}^2)}}$$

Dit wordt ook wel **Pearson's correlatiecoëfficiënt** genoemd.

- $\boldsymbol{r(x,y)}$ **ligt altijd tussen** $\boldsymbol{-1}$ **en** $\boldsymbol{1}$
  - $r(x,y) < 0$: **negatieve** correlatie (dalende trend)
  - $r(x,y) > 0$: **positieve** correlatie (stijgende trend)

Pearson's correlatiecoëfficiënt

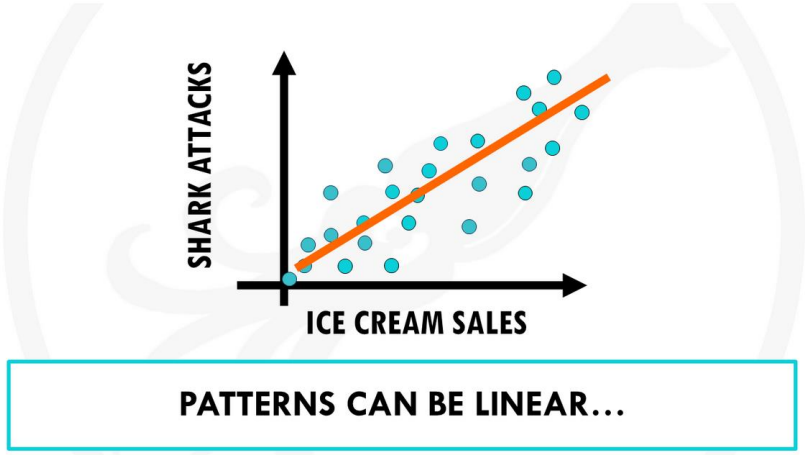
https://en.wikipedia.org/wiki/Correlation#/media/File:Correlation_examples2.svg

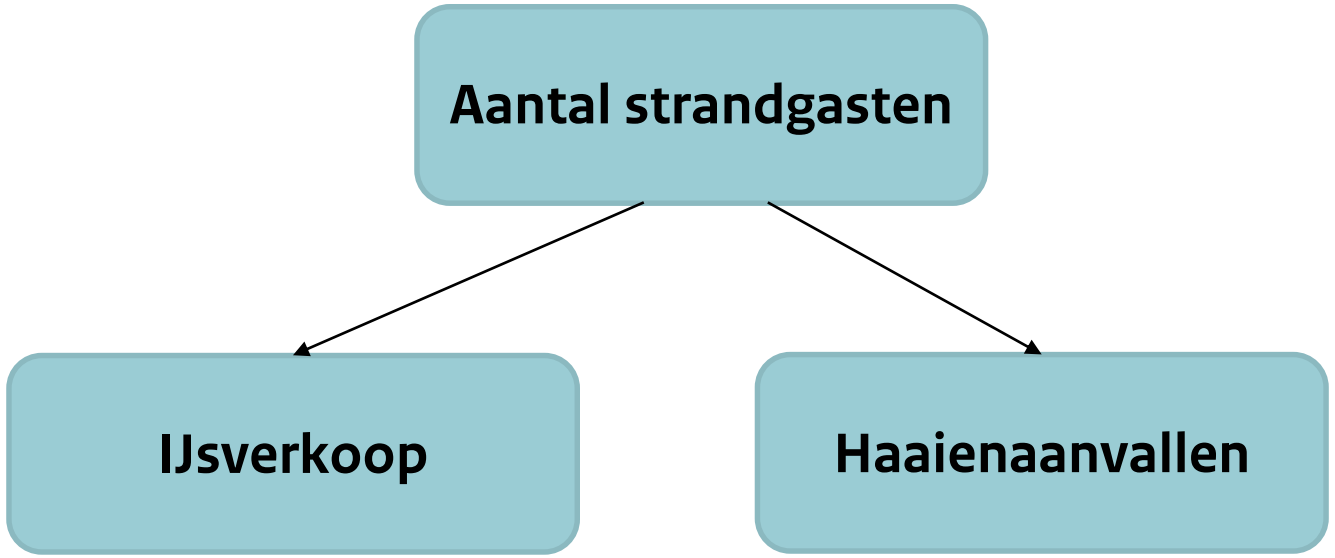


1. Hoe dichter de punten op een lijn liggen, hoe verder van o de correlatiecoëfficiënt ligt.
2. Sterke correlatie zegt niets over de richtingscoëfficiënt van de betreffende lijn.
3. Bij een correlatiecoëfficiënt $r = 0$ is de samenhang niet lineair, maar kan het nog steeds niet-lineair zijn.

Correlatie ≠ causaliteit

https://biostatsquid.com/correlation-does-not-imply-causation/





IJsverkoop en haaienaanvallen zijn duidelijk gecorreleerd, maar ijsverkoop is niet de **oorzaak** van een haaienaanvallen.
We hebben hier te maken met een derde variabele (“aantal strandgasten”) die invloed heeft op beiden.

Voorbeeld

Defensie onderzoekt het benzineverbruik van de PNOD voertuigen. Hiertoe wordt op een speciaal circuit een tijdlang met constante snelheid gereden, waarna het benzineverbruik wordt vastgesteld in liters per 100 kilometer.

Snelheid (km/h)	90	100	110	120	130
Verbruik (ltr / 100km)	6,8	7,5	7,9	8,3	9,1

Wat kunnen we op statistische verantwoorde wijze zeggen over de correlatie tussen snelheid en verbruik?

Voorbeeld: Pearson's correlatiecoëfficiënt

We hebben een steekproef $(x_1, y_1), (x_2, y_2), \ldots, (x_5, y_5)$ van datapunten met x_i: snelheid (in km / uur) en y_i: brandstofverbruik (in liters per 100 km)

x	y	xy	x^2	y^2
90	6,8	$90 \cdot 6{,}8 = 612$	$90 \cdot 90 = 8100$	$6{,}8 \cdot 6{,}8 = 46{,}24$
100	7,5	$100 \cdot 7{,}5 = 750$	$100 \cdot 100 = 10000$	$7{,}5 \cdot 7{,}5 = 56{,}25$
110	7,9	$110 \cdot 7{,}9 = 869$	$110 \cdot 110 = 12100$	$7{,}9 \cdot 7{,}9 = 62{,}41$
120	8,3	$120 \cdot 8{,}3 = 996$	$120 \cdot 120 = 14400$	$8{,}3 \cdot 8{,}3 = 68{,}89$
130	9,1	$130 \cdot 9{,}1 = 1183$	$130 \cdot 130 = 16900$	$9{,}1 \cdot 9{,}1 = 82{,}81$
$\overline{x} = \mathbf{110}$	$\overline{y} = \mathbf{7,92}$	$\overline{xy} = \mathbf{882}$	$\overline{x^2} = \mathbf{12300}$	$\overline{y^2} = \mathbf{63,32}$

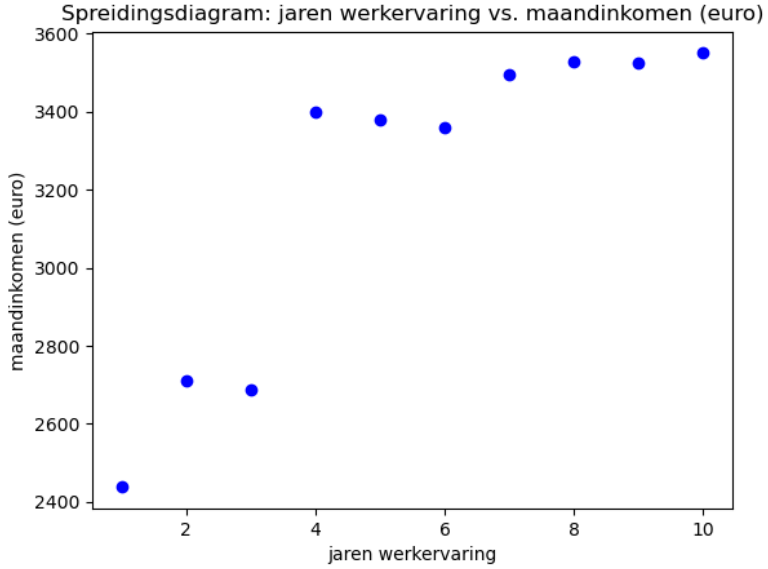
Correlatiecoëfficiënt: $r(x,y) = \dfrac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2) \cdot (\overline{y^2} - \overline{y}^2)}} = \dfrac{882 - 110 \cdot 7{,}92}{\sqrt{(12300 - 110^2) \cdot (63{,}32 - 7{,}92^2)}} \approx 0{,}9912$

Beperkingen van Pearson's correlatiecoëfficiënt

- Het vereist dat de variabelen op interval / rationiveau gemeten worden.
- Erg gevoelig voor outliers
- Het vereist dat er een lineaire samenhang is tussen de twee variabelen.
- Het vereist de aanname van normaal verdeelde kansvariabelen.

Spearman's correlatiecoëfficiënt

De correlatiecoëfficiënt van Pearson zegt alleen iets over lineaire samenhang.
We kunnen ook correlaties berekenen als de samenhang niet lineair is, maar **monotoon.**



Jaren werkervaring	1	2	3	4	5	6	7	8	9	10
Maandinkomen	2438	2712	2687	3398	3380	3358	3496	3527	3524	3550

Spearman's correlatiecoëfficiënt

Bij de **correlatiecoëfficiënt van Spearman** bekijken we niet de absolute waardes van de variabelen, maar maken we een ranking (van laag naar hoog) en berekenen we de verschillen d_i tussen de rankings.

Jaren werkervaring	1	2	3	4	5	6	7	8	9	10
Maandinkomen	2438	2712	2687	3398	3380	3358	3496	3527	3524	3550
Maandinkomen (ranking)	1	3	2	6	5	4	7	9	8	10
Verschil d_i in rankings	0	-1	1	-2	0	2	0	-1	1	0
d_i^2	0	1	1	4	0	4	0	1	1	0

Correlatiecoëfficiënt van Spearman:

$$r_s(x,y) = 1 - \frac{6 \cdot \sum_i d_i^2}{n^3 - n}$$

In dit geval: $r_s(x,y) = 1 - \frac{6 \cdot 12}{10^3 - 10} \approx 0{,}9273$

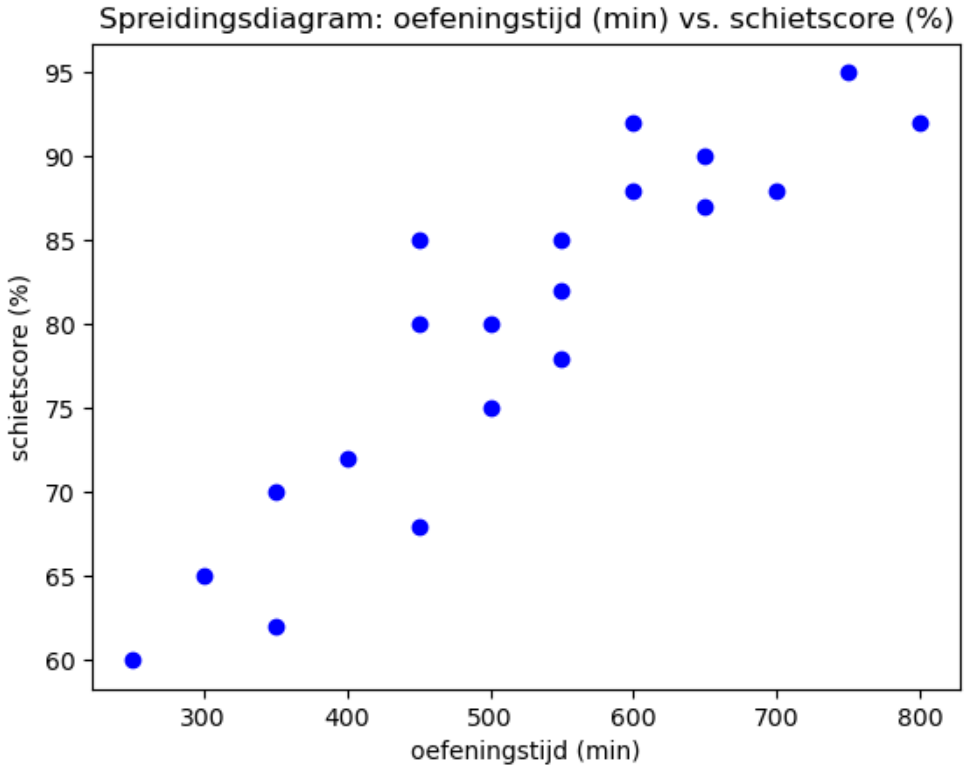
Voorbeeld: jaarlijkse keuring van actieve militairen

ID	Rang	Eenheid	Oefeningstijd (min)	Schietscore (%)	Aantal missies	Geslacht
1	Sergeant	Bravo	450	85	12	Man
2	Korporaal	Alpha	600	92	8	Vrouw
3	Luitenant	Charlie	550	78	15	Man
4	Majoor	Bravo	700	88	20	Man
5	Soldaat	Alpha	300	65	5	Vrouw
6	Kapitein	Charlie	500	80	10	Vrouw
7	Sergeant	Bravo	650	90	18	Man
8	Soldaat	Alpha	400	72	6	Man
9	Soldaat	Charlie	350	70	4	Man
10	Kapitein	Bravo	750	95	25	Vrouw
11	Luitenant	Alpha	550	82	12	Man
12	Korporaal	Charlie	450	68	7	Vrouw
13	Sergeant	Bravo	600	88	14	Man
14	Majoor	Alpha	800	92	30	Vrouw
15	Soldaat	Charlie	250	60	3	Vrouw
16	Kapitein	Bravo	650	87	18	Man
17	Luitenant	Alpha	500	75	9	Man
18	Sergeant	Charlie	550	85	13	Vrouw
19	Korporaal	Bravo	450	80	10	Man
20	Soldaat	Alpha	350	62	5	Vrouw

Bereken de correlatiecoëfficiënten van Pearson en Spearman voor oefeningstijd en schietscore.

Lineaire regressie

In het spreidingsdiagram is een duidelijke lineaire trend te zien.



Vraag: welke formule van een lijn beschrijft het “best” deze lineaire trend?

Lineaire regressie

Bij een regressieanalyse willen we een **afhankelijke variabele** Y beschrijven als functie van één of meerdere **onafhankelijke variabelen** $X_1, X_2, \ldots$

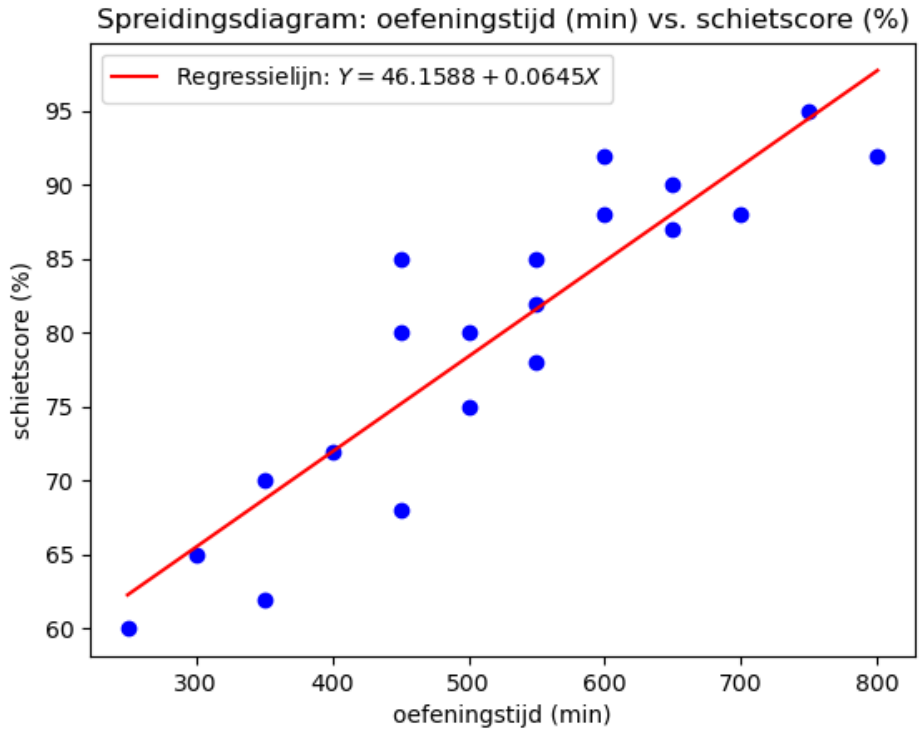
Focus: lineaire functie van één onafhankelijke variabele (enkelvoudig lineaire regressie)

$$Y = \alpha + \beta \cdot X + \varepsilon$$

- $\alpha + \beta \cdot X$: **lineaire term** met X
- $\varepsilon \sim N(0, \sigma)$: **storingsterm** die externe factoren van onzekerheid meeneemt
  - Een verband is (bijna) nooit perfect lineair, dus hiermee moet rekening worden gehouden.

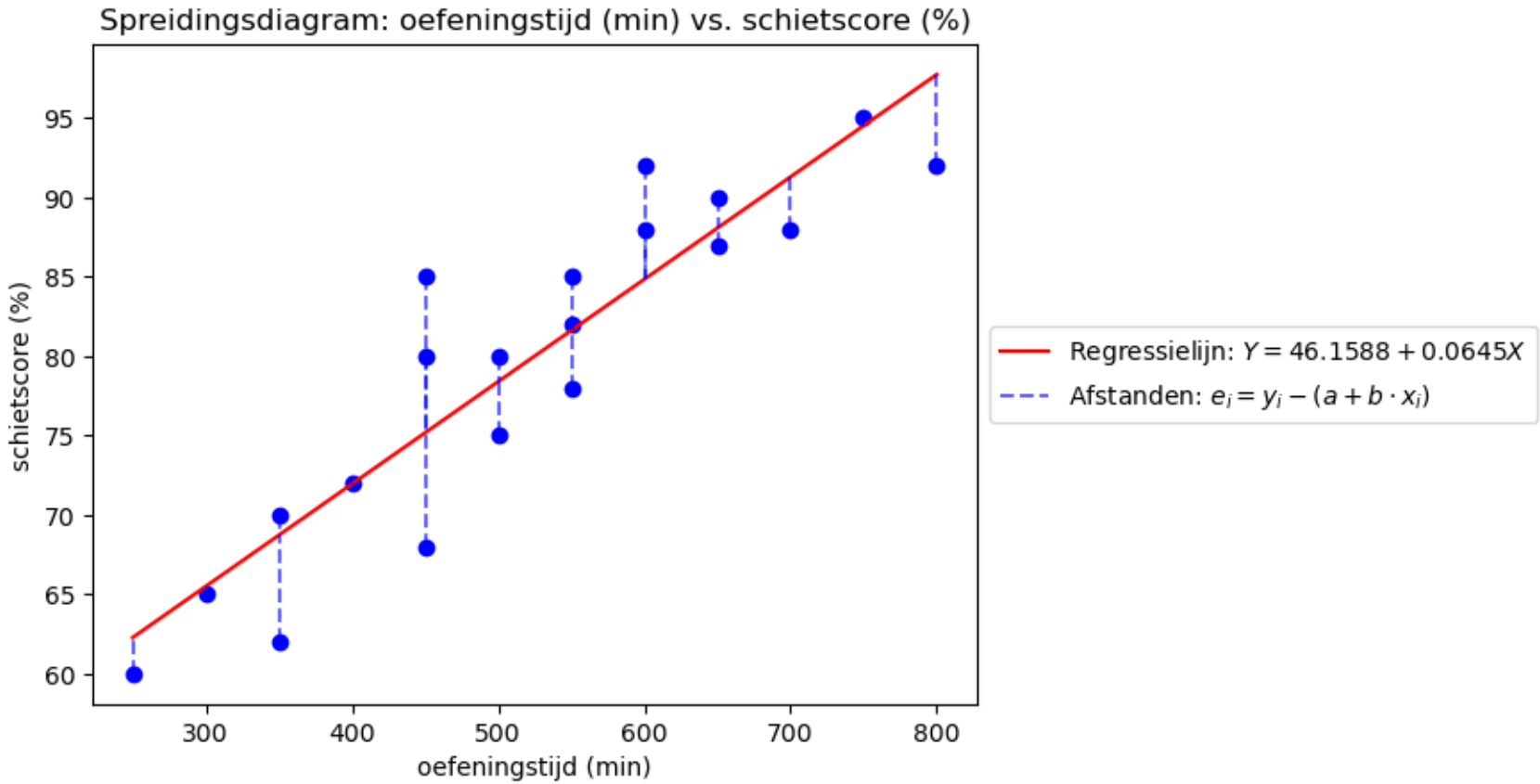
Lineaire regressie

Populatiemodel $Y = \alpha + \beta \cdot X + \varepsilon$



Doel: met een **steekproef** de (onbekende) parameters α en β schatten ($Y = a + b \cdot X$)

Kleinstekwadratenmethode



Kleinstekwadratenmethode

De datapunten liggen niet perfect op een lijn, dus moeten we een “best passende” lijn bepalen:

$$Y = a + b \cdot X$$

1. **Voor ieder datapunt (x_i, y_i) bekijken we de verticale afstand tussen de “daadwerkelijke” y-waarde (y_i) en de “voorspelde” y-waarde $(a + b \cdot x_i)$**

$$e_i = y_i - (a + b \cdot x_i)$$

2. **Kwadrateer deze afstanden (zodat we met alleen positieve getallen werken):**

$$e_i^2 = \big(y_i - (a + b \cdot x_i)\big)^2$$

3. **Minimaliseer de som van deze afstanden:**

$$\min \sum e_i^2 = \min \sum \big(y_i - (a + b \cdot x_i)\big)^2$$

Kleinstekwadratenmethode

Omdat de steekproef$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ bekend is, is $\sum\left(y_i - (a + b \cdot x_i)\right)^2$ een functie van a en b.

Deze functie wordt geminimaliseerd zodra a en b de volgende waardes aannemen:

$$b = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}$$

$$a = \overline{y} - b \cdot \overline{x}$$

Voorbeeld

Defensie onderzoekt het benzineverbruik van de PNOD voertuigen. Hiertoe wordt op een speciaal circuit een tijdlang met constante snelheid gereden, waarna het benzineverbruik wordt vastgesteld in liters per 100 kilometer.

Snelheid (km/h)	90	100	110	120	130
Verbruik (ltr / 100km)	6,8	7,5	7,9	8,3	9,1

- **Als we regressie toepassen, welke variabele is dan de verklarende variabele X?**
- **Stel de vergelijking $Y = a + b \cdot X$ op van de regressielijn.**

Voorbeeld: lineaire regressie

We kunnen de tabel van eerder dit college hergebruiken:

x	y	xy	x^2
90	6,8	$90 \cdot 6{,}8 = 612$	$90 \cdot 90 = 8100$
100	7,5	$100 \cdot 7{,}5 = 750$	$100 \cdot 100 = 10000$
110	7,9	$110 \cdot 7{,}9 = 869$	$110 \cdot 110 = 12100$
120	8,3	$120 \cdot 8{,}3 = 996$	$120 \cdot 120 = 14400$
130	9,1	$130 \cdot 9{,}1 = 1183$	$130 \cdot 130 = 16900$
$\overline{x} = \mathbf{110}$	$\overline{y} = \mathbf{7,92}$	$\overline{xy} = \mathbf{882}$	$\overline{x^2} = \mathbf{12300}$

$$b = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{882 - 110 \cdot 7{,}92}{12300 - 110^2} = 0{,}054$$

$$a = \overline{y} - b \cdot \overline{x} = 7{,}92 - 0{,}054 \cdot 110 = 1{,}98$$

Regressielijn:

$$Y = 1{,}98 + 0{,}054 \cdot X$$

Belangrijke opmerkingen: lineaire regressie

- De lineaire regressielijn $Y = 1{,}98 + 0{,}054 \cdot X$ zou voor $X = 0$ (stilstand) nog steeds een positief verbruik voorspellen: $Y = 1{,}98$
  - Je mag niet zomaar extrapoleren buiten de dataset!

- Hetzelfde principe is mogelijk voor meerdere verklarende variabelen $X_1, X_2, \ldots$
  - Dit maakt de formules nog complexer

- Een regressielijn op basis van één steekproef geeft een **puntschatting** voor y gegeven een waarde voor $x = x_0$

Schatten en voorspellen met de regressielijn

Recap: we starten vanuit het **populatiemodel**

$$Y = \alpha + \beta \cdot X + \varepsilon$$

Om verantwoorde voorspellingen voor Y gegeven $X = X_0$ te kunnen doen, moeten we ook de storingsterm $\varepsilon \sim N(0, \sigma)$ bestuderen. Dit doen we door allereerst de variantie σ^2 te schatten:

$$s_\varepsilon^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum \left(y_i - (a + b \cdot x_i)\right)^2}{n-2} = \frac{n}{n-2} \cdot \left(\overline{y^2} - a \cdot \overline{y} - b \cdot \overline{xy}\right)$$

We hebben te maken met $n - 2$ vrijheidsgraden, omdat we a en b al hebben geschat met behulp van de steekproef $(x_1, y_1), \ldots, (x_n, y_n)$.

Betrouwbaarheidsinterval voor de gemiddelde Y bij een gegeven $x = x_0$

Omdat het verband niet perfect lineair is door de storingsterm ε, is Y ook een kansvariabele bij een gegeven $x = x_0$.

Stel dat we de verwachtingswaarde willen bepalen, $E[\,Y \mid X = x_0\,]$.
De standaardafwijking van $E[Y \mid X = x_0]$ kunnen we schatten met behulp van

$$s_\mu = s_\varepsilon \cdot \sqrt{\frac{1}{n} \cdot \left(1 + \frac{(x_0 - \overline{x})^2}{\overline{x^2} - \overline{x}^2}\right)}$$

Een **betrouwbaarheidsinterval voor $E[\,Y \mid X = x_0\,]$** vinden we met de t-verdeling ($df = n - 2$):

$$t = \text{InvT}(\text{opp} = 1 - \frac{\alpha}{2}; df = n - 2)$$

$$[(a + b \cdot x_0) - t \cdot s_\mu; (a + b \cdot x_0) + t \cdot s_\mu]$$

Voorspellingsinterval voor Y bij een gegeven $x = x_0$

Omdat het verband niet perfect lineair is door de storingsterm ε, is Y ook een kansvariabele bij een gegeven $x = x_0$.

We kunnen ook een losse toekomstige uitkomst $Y \mid X = x_0$ voorspellen ("forecast").
De standaardafwijking van $Y \mid X = x_0$ kunnen we schatten met behulp van

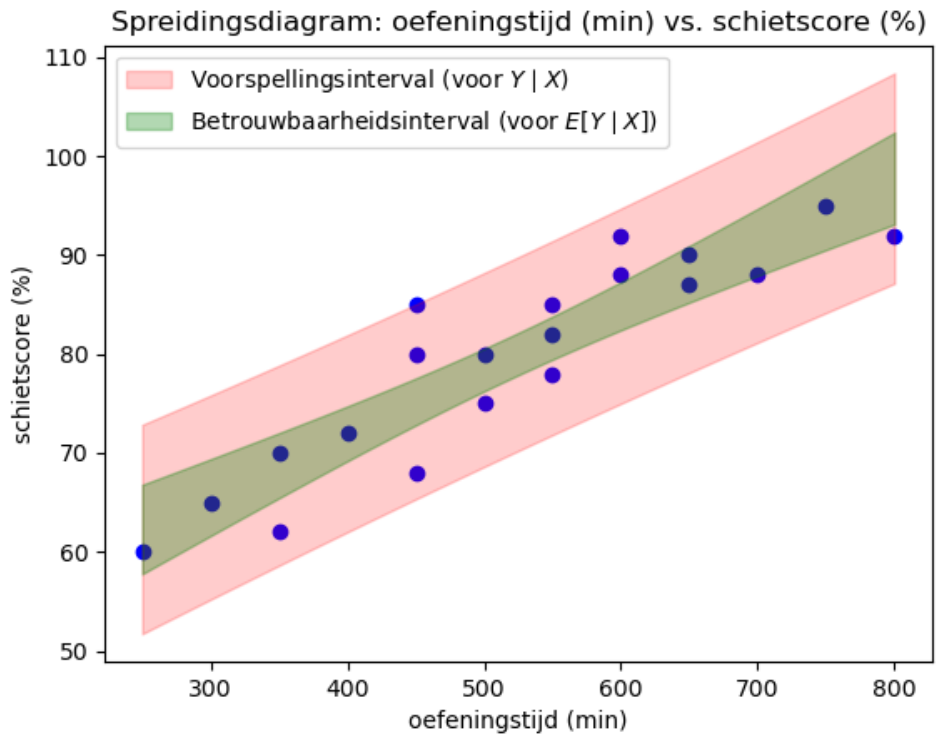
$$s_f = s_\varepsilon \cdot \sqrt{1 + \frac{1}{n} \cdot \left(1 + \frac{(x_0 - \overline{x})^2}{\overline{x^2} - \overline{x}^2}\right)}$$

Een **voorspellingsinterval voor $Y \mid X = x_0$** vinden we met de t-verdeling ($df = n - 2$):

$$t = \text{InvT}(\text{opp} = 1 - \frac{\alpha}{2}; df = n - 2)$$

$$[(a + b \cdot x_0) - t \cdot s_f; (a + b \cdot x_0) + t \cdot s_f]$$

Schatten en voorspellen met de regressielijn



- Voorspellingsintervallen zijn breder dan betrouwbaarheidsintervallen!
  - Bij gemiddelden vallen storingstermen tegen elkaar weg, bij een losse uitkomst niet.
- Intervallen zijn het smalst voor $x = \overline{x}$ (steekproefgemiddelde)
- Intervallen zijn heel breed naarmate x ver weg ligt van $\overline{x}$.
  - Extrapoleren leidt tot onnauwkeurige voorspellingen!

Samenvatting

- Correlatie en regressie
  - Correlatiecoëfficiënten van Pearson en Spearman
  - Lineaire regressie
  - Betrouwbaarheidsintervallen / voorspellingsintervallen met een regressielijn.

Huiswerk:

- Lezen van A. Buijs: hoofdstuk 13.1 (blz. 401-408), 13.2.1-13.2.3 (blz. 408-417), 13.3 (blz. 354-357)

- Opdrachten:
  - Hoofdstuk 13: m1-m4, m6, 13.1, 13.4, 13.5, 13.9, 13.11, 13.15