



Relaties tussen continue variabelen

Zie ook:
Cursusboek hoofdstuk 28

Inhoud

1. Lineaire relaties in een steekproef
 - Wat is een lineaire relatie?
 - Pearsons correlatiecoëfficiënt r .
 - Covariantie
 - Formule voor de correlatiecoëfficiënt
2. Hypothesetoets van correlatiecoëfficiënt
3. De regressielijn
4. Betrouwbaarheidsintervallen en toetsen voor de regressielijn
5. Interpolatie en extrapolatie
6. Niet-lineaire relaties fitten

Inhoud

1. Lineaire relaties in een steekproef

- Wat is een lineaire relatie?
- Pearsons correlatiecoëfficiënt r .
- Covariantie
- Formule voor de correlatiecoëfficiënt

2. Hypothesetoets van correlatiecoëfficiënt

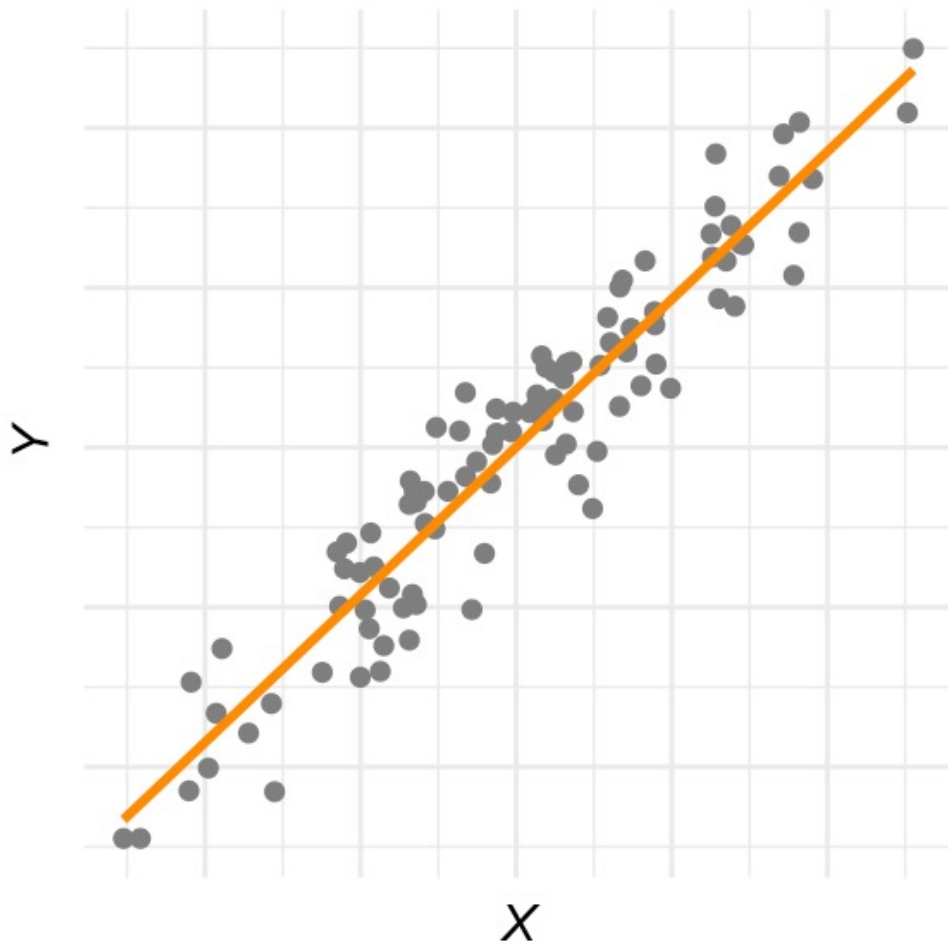
3. De regressielijn

4. Betrouwbaarheidsintervallen en toetsen voor de regressielijn

5. Interpolatie en extrapolatie

6. Niet-lineaire relaties fitten

Lineaire relaties tussen variabelen in een steekproef

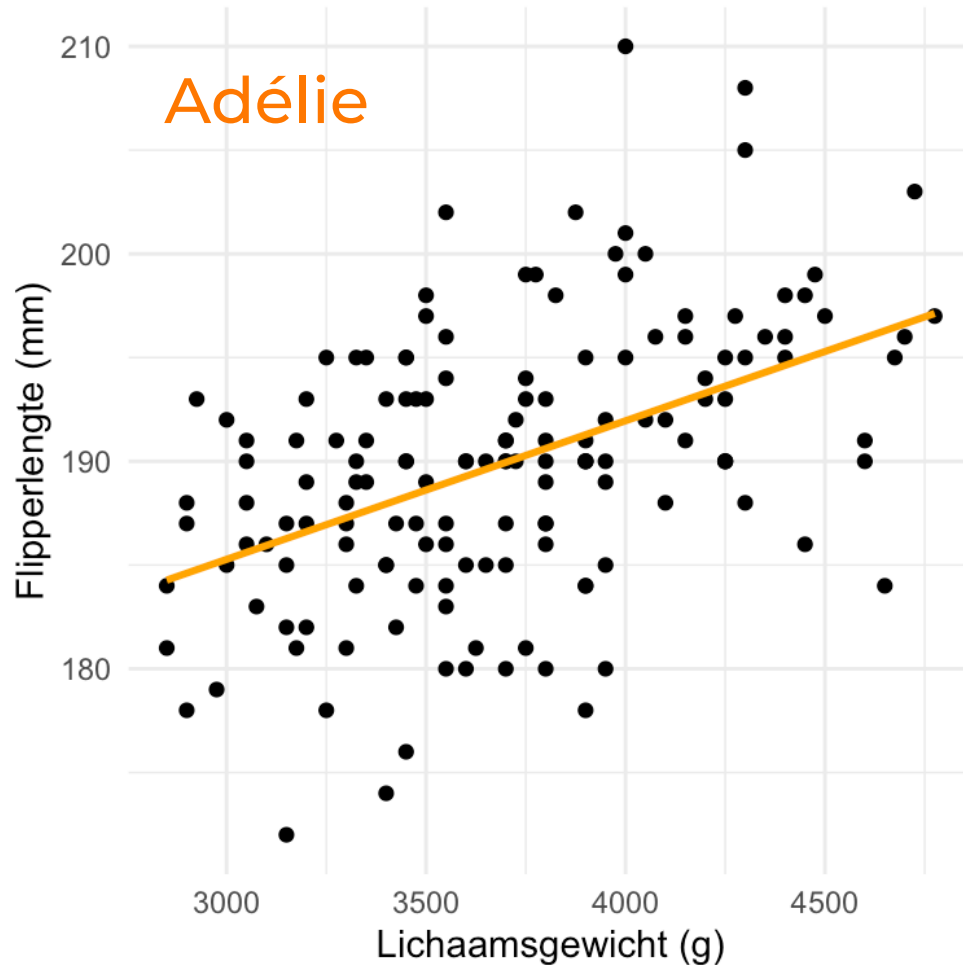


Lineaire relatie:

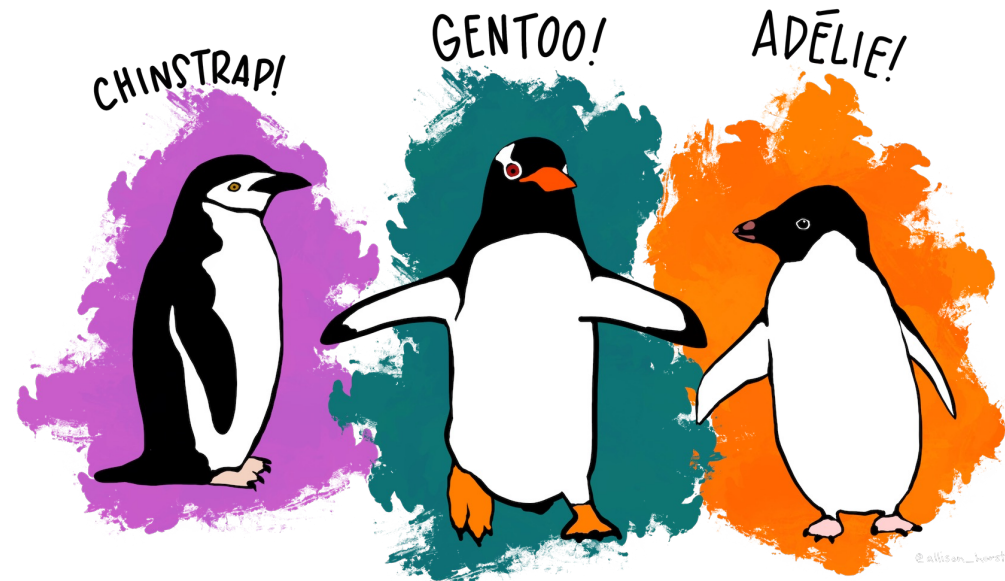
de datapunten bevinden zich op of in de buurt van een rechte lijn die stijgt of daalt.

- Hoe sterk is de relatie? (Hoe precies volgen de punten de lijn?)
- Welke lijn beschrijft de relatie het beste?
- Kan ik Y voorspellen uit X ?

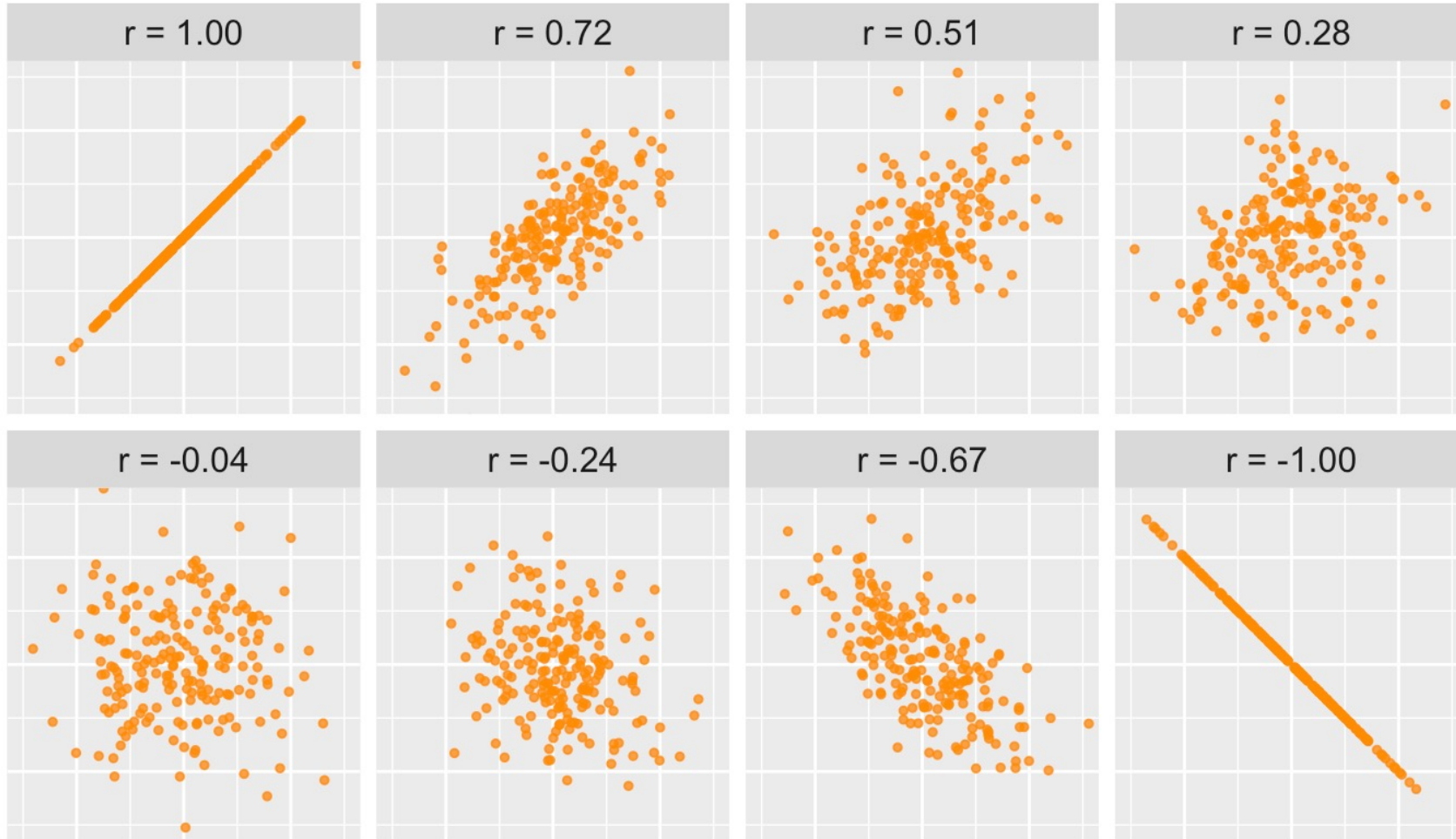
Flipperlengte vs lichaamsgewicht



Flipperlengte en lichaamsgewicht lijken bij Adélie lineair samen te hangen.

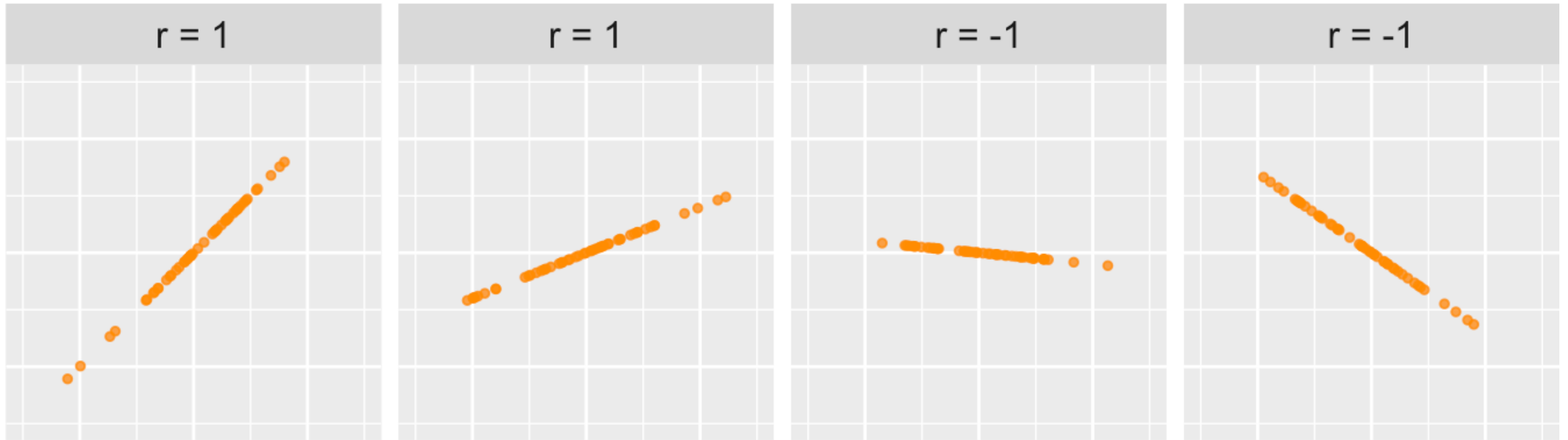


Pearson's correlatiecoëfficiënt



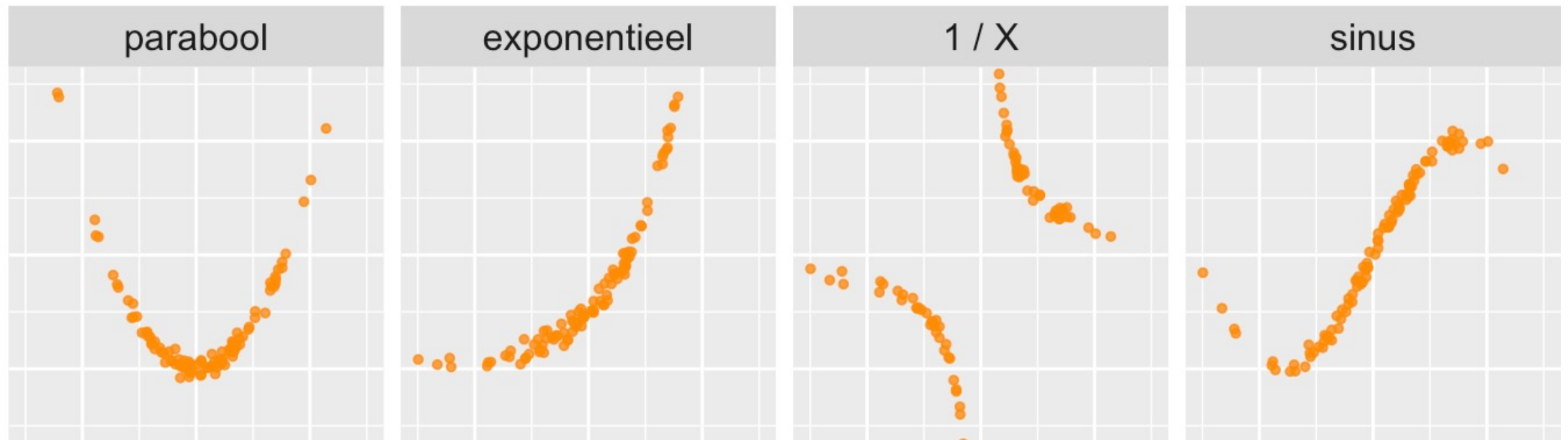
- **Dimensieloos** getal.
- $-1 \leq r \leq 1$
- **Teken** van r geeft richting van verband aan.
- **Absolute waarde** geeft sterkte van verband aan.

Correlatiecoëfficiënt is geen richtingscoëfficiënt

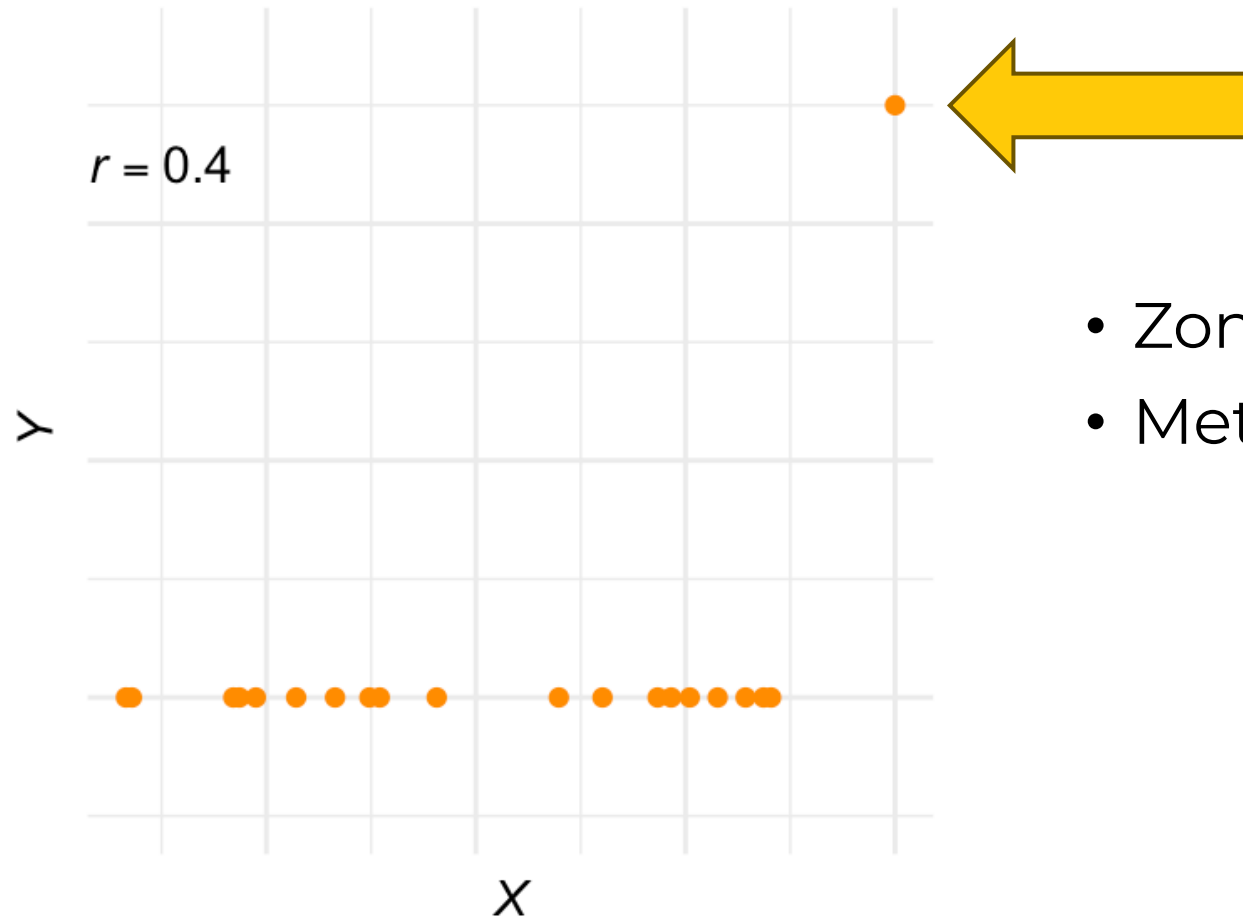


Alleen betekenisvol door lineaire verbanden

Voorbeelden van niet-lineaire verbanden



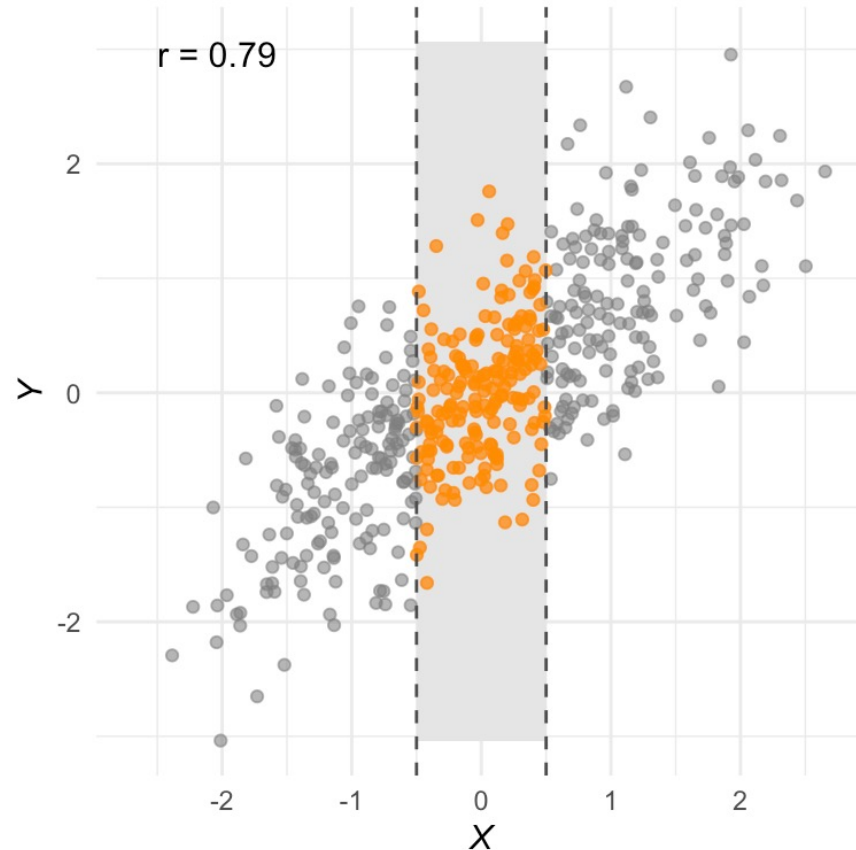
Correlatiecoëfficiënt gevoelig voor uitbijters



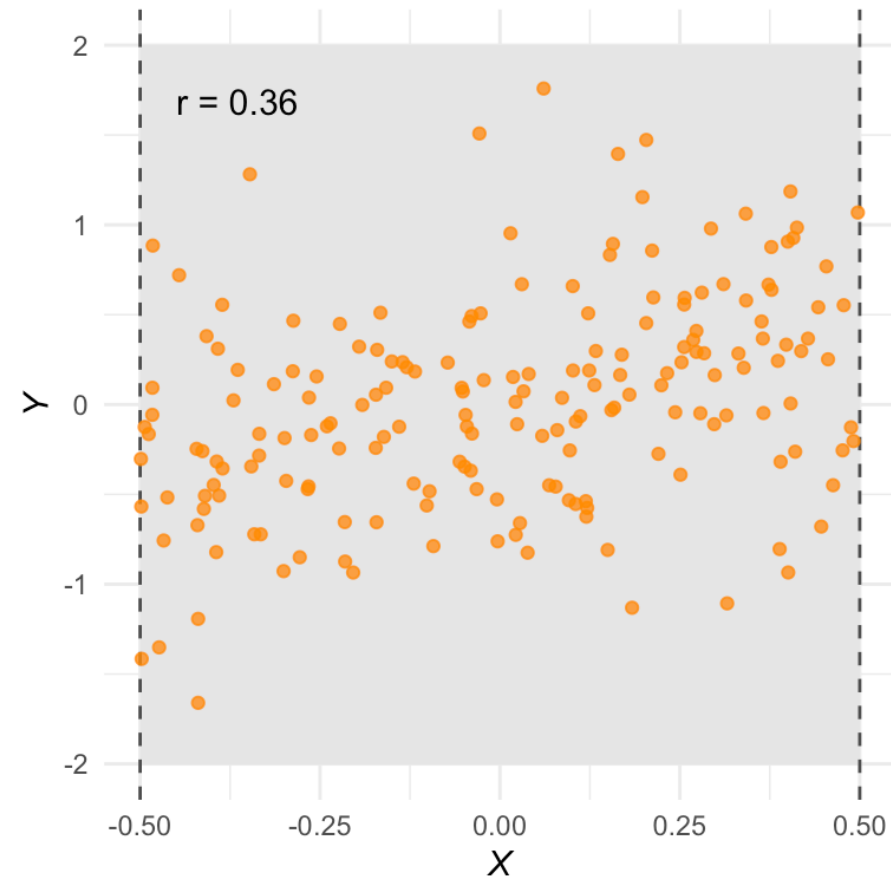
- Zonder de uitbijter: $r = 0$.
- Met de uitbijter: $r = 0.4$.

Correlatiecoëfficiënt gevoelig voor X -bereik

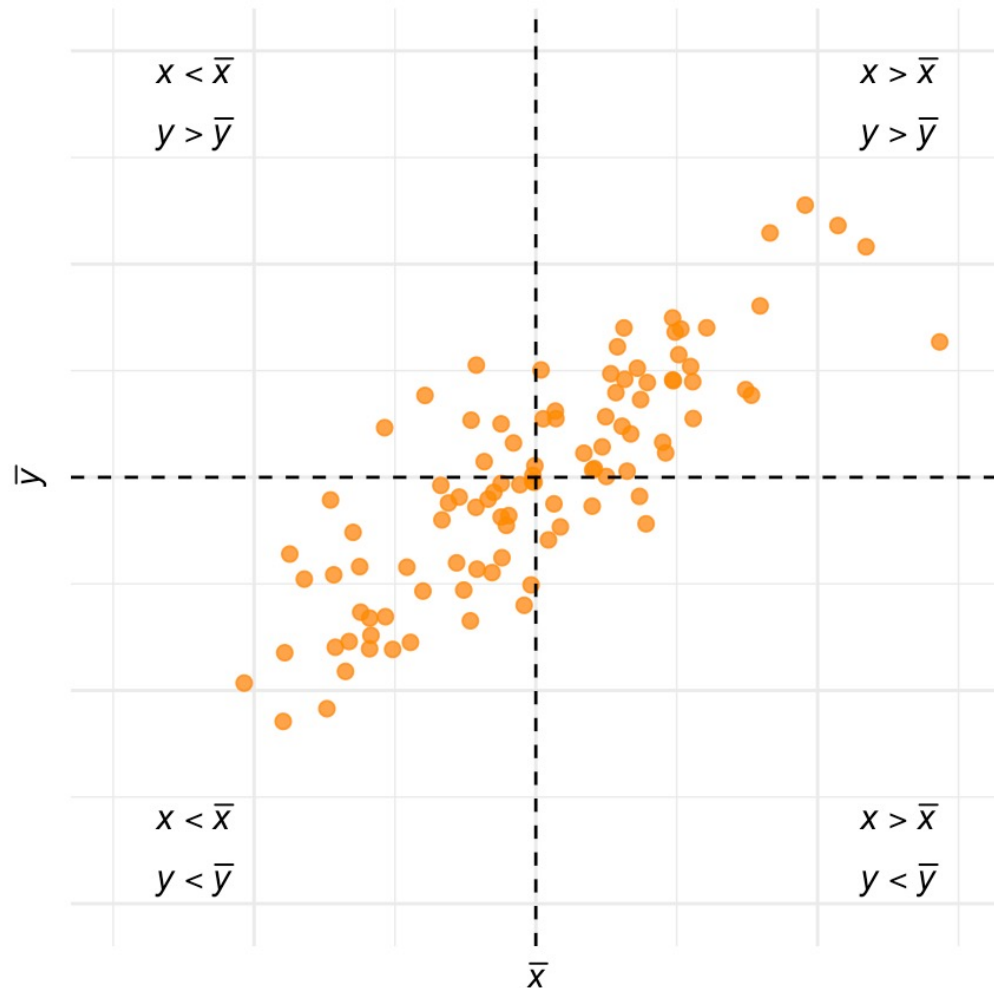
Volledige dataset



Middelste deel



De covariantie van X en Y in een steekproef

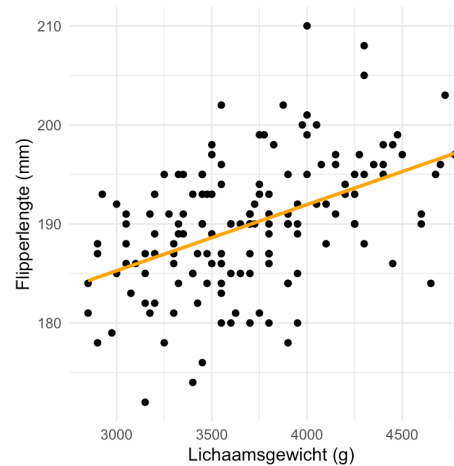


- Meet de lineaire samenhang tussen X en Y .

- Formule:

$$C_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- Heeft de dimensies van X en Y .



Voorbeeld:

- Flipperlengte Y is een lengte.
- Lichaamsgewicht X een gewicht.

Dus:

Dimensie C_{XY} is lengte x gewicht.

Formule voor Pearsons correlatiecoëfficiënt

Formule voor de correlatiecoëfficiënt:

$$r = \frac{c_{XY}}{s_X s_Y}.$$

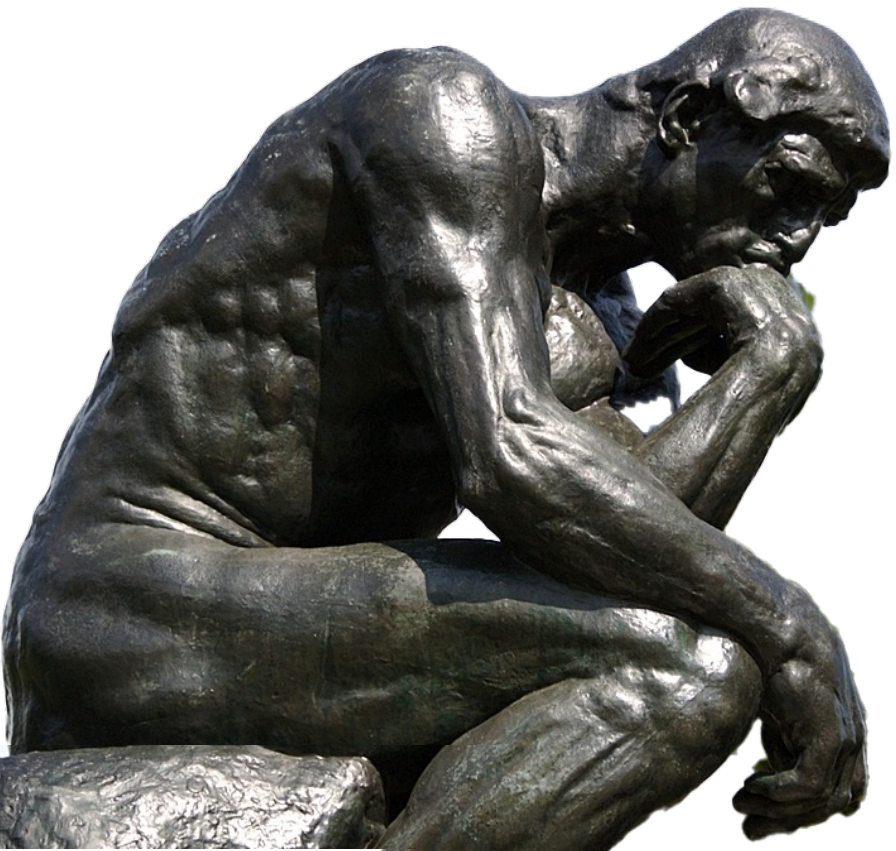
(dimensie van X)(dimensie van Y)

dimensie van X dimensie van Y

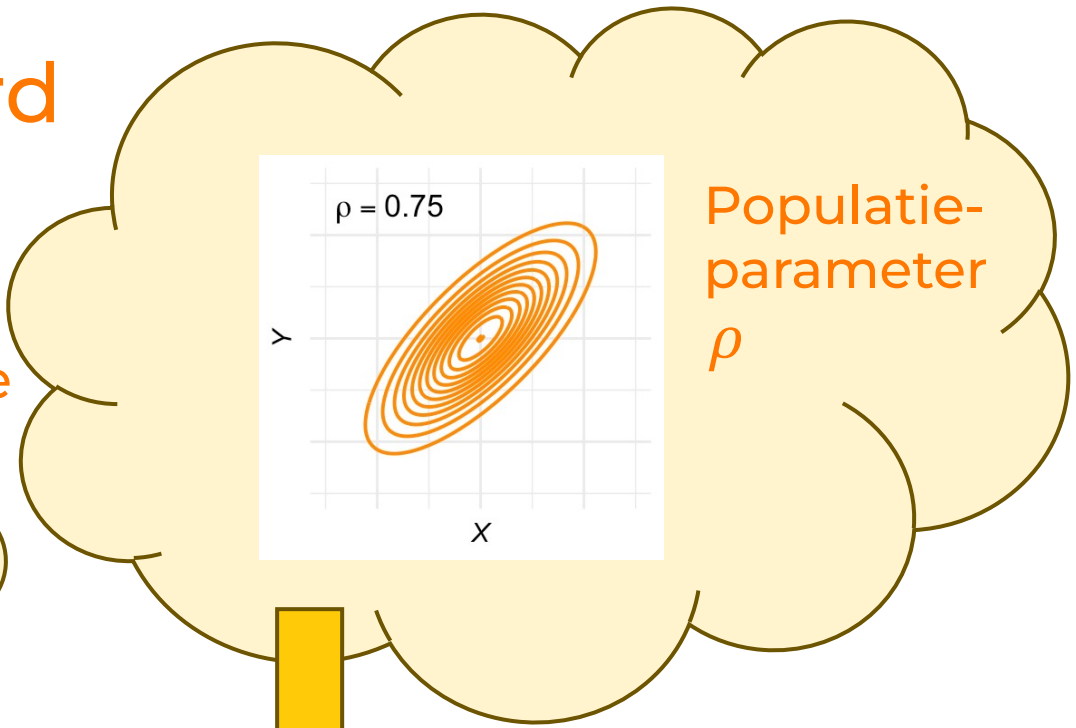
Inhoud

1. Lineaire relaties in een steekproef
 - Wat is een lineaire relatie?
 - Pearsons correlatiecoëfficiënt r .
 - Covariantie
 - Formule voor de correlatiecoëfficiënt
2. Hypothesetoets van correlatiecoëfficiënt
3. De regressielijn
4. Betrouwbaarheidsintervallen en toetsen voor de regressielijn
5. Interpolatie en extrapolatie
6. Niet-lineaire relaties fitten

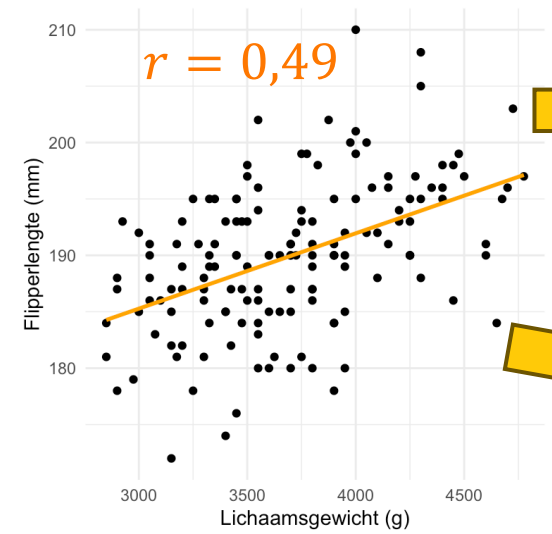
Zijn de variabelen gecorreleerd in de populatie?



Populatie



Steekproef

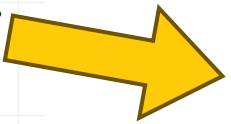


Toets

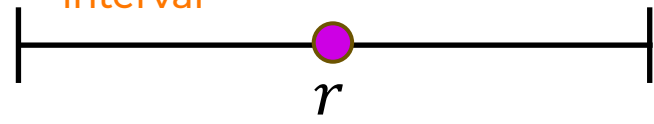


$H_0: \rho = 0,$
 $H_A: \rho \neq 0.$
 $P = ?$

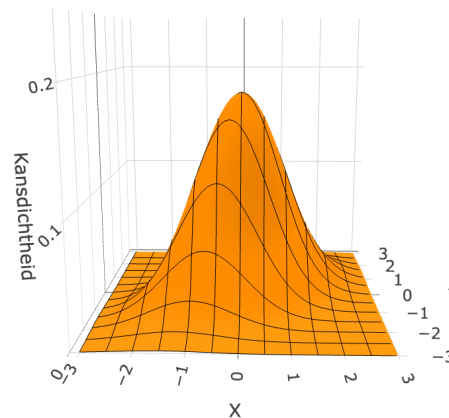
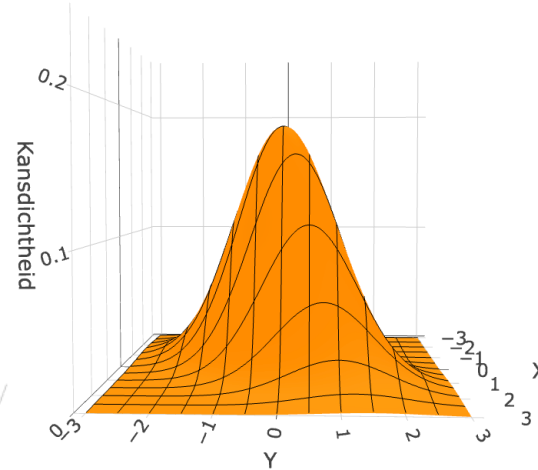
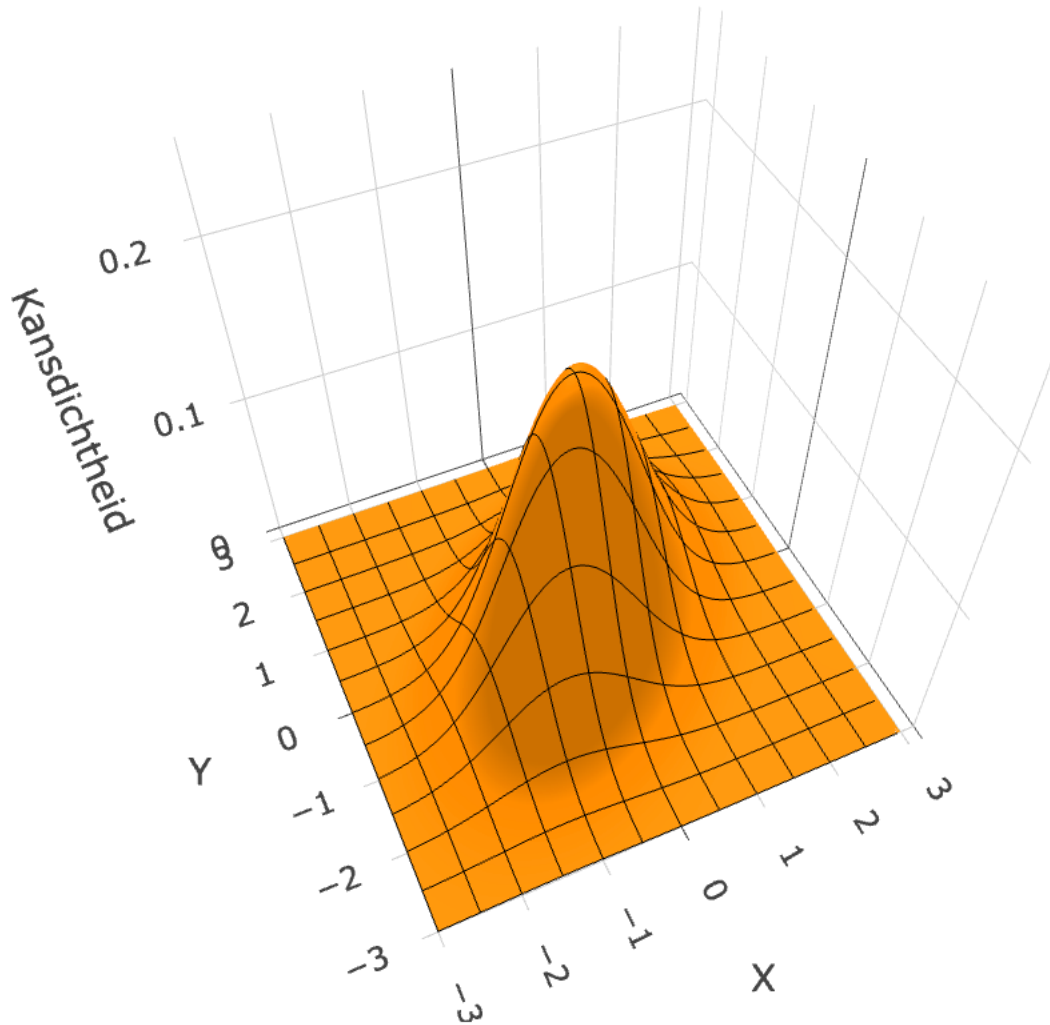
Schatting



Betrouwbaarheids-interval



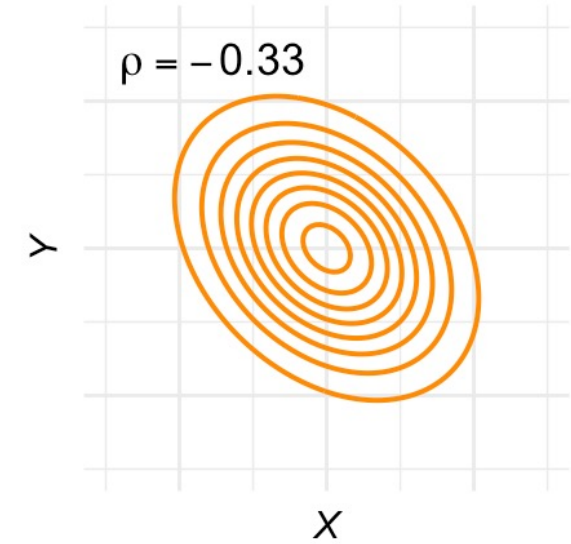
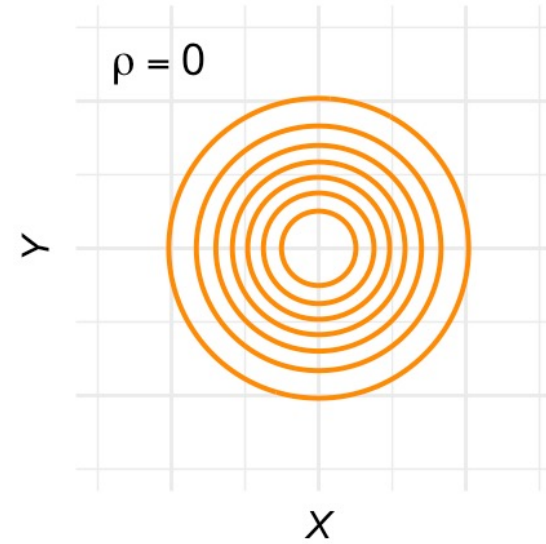
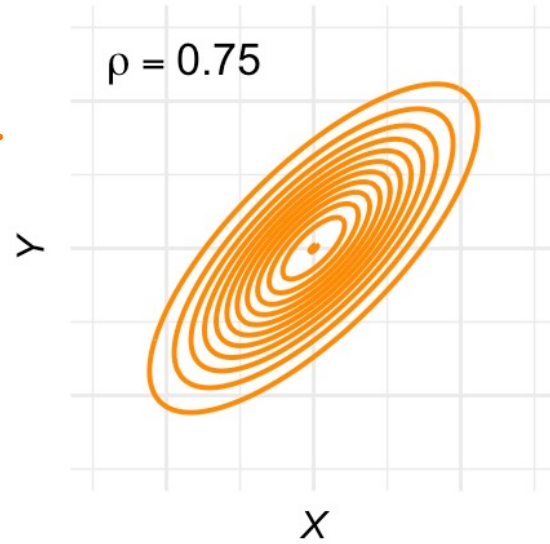
Stap 1, statistisch model: de bivariate normale verdeling



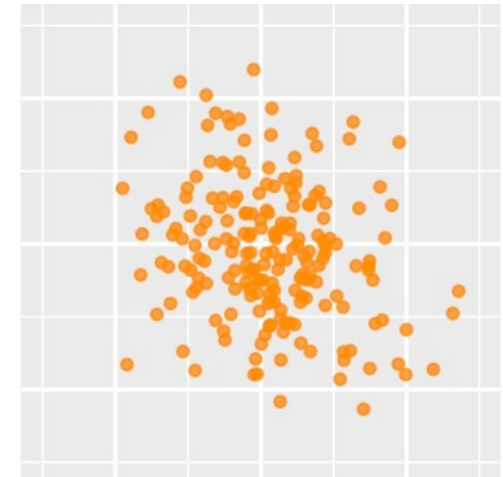
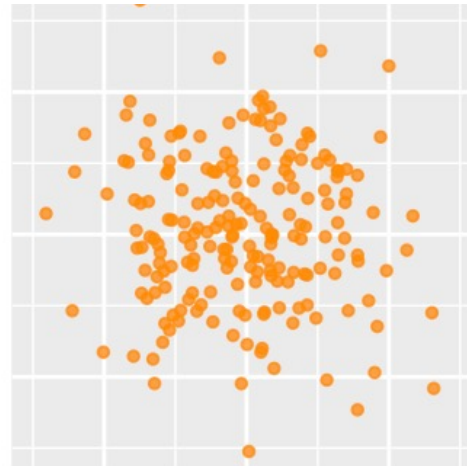
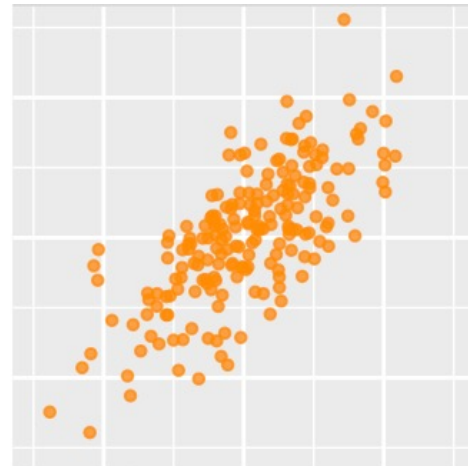
- Beide variabelen normaal verdeeld.
- Extra parameter: correlatiecoëfficiënt van de populatie, ρ .

Hoogtelijnen van de bivariate normale verdeling

Hoogtelijnen van de kansverdeling.



Steekproef uit de kansverdeling.



Toets van de correlatiecoëfficiënt: overige stappen

Stap 2: $H_0: \rho = 0, H_A: \rho \neq 0$.

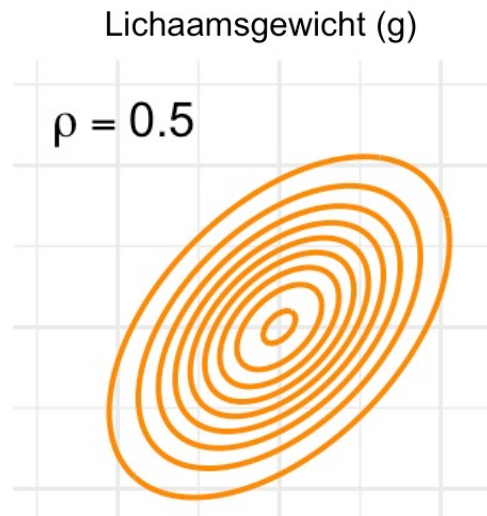
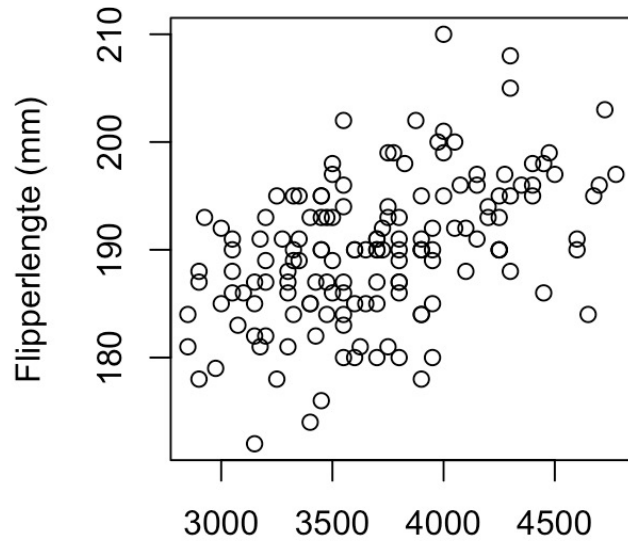
Stap 3: Toetsingsgrootheid: $t = \frac{r}{SE_r}$, waarin $SE_r = \sqrt{\frac{1-r^2}{n-2}}$.
Kansverdeling: t -verdeling, met $df = n - 2$

Stap 4: t uitrekenen.

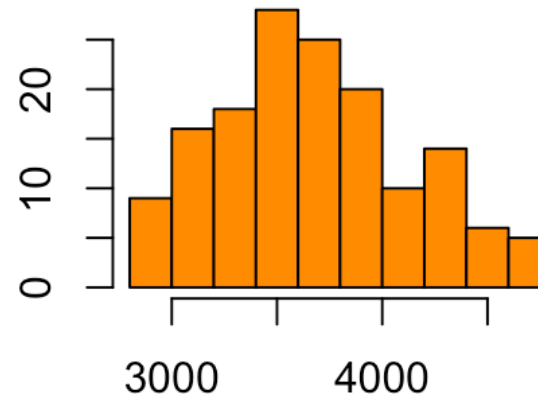
Stap 5: P uitrekenen (met R): `(P <- 2 * pt(-abs(t), df = n - 2))`

Stap 6: Conclusie.

Is het model van toepassing?

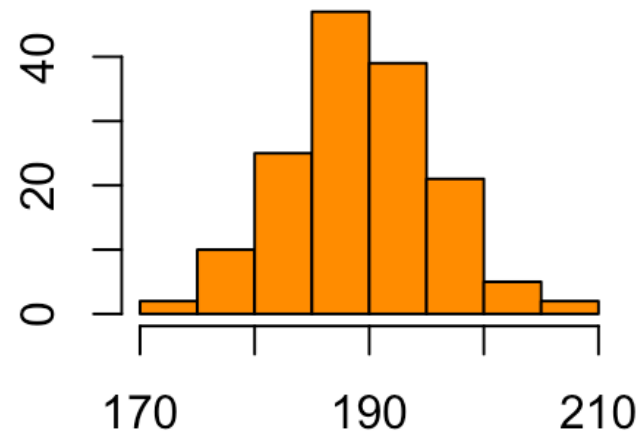


Frequency



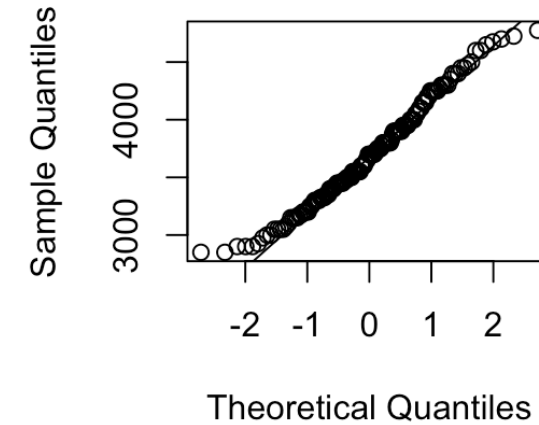
Lichaamsgewicht (g)

Frequency

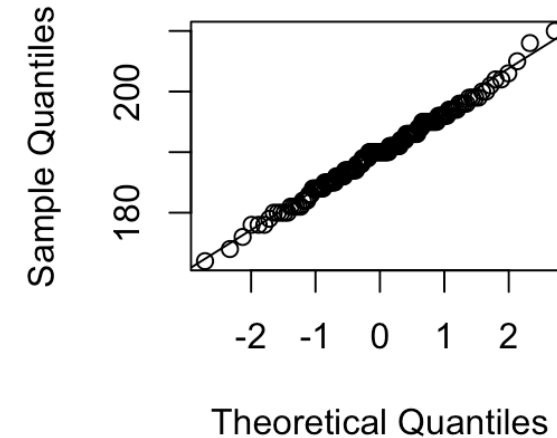


Flipperlengte (mm)

QQ-plot lichaamsgewicht



QQ-plot flipperlengte



Toets van correlatie met R uitvoeren



```
cor.test(adelie$body_mass_g, adelie$flipper_length_mm)
```

Pearson's product-moment correlation

data: adelie\$body_mass_g and adelie\$flipper_length_mm

t = 6.4678, df = 149, p-value = 1.343e-09

alternative hypothesis: true correlation is not equal to 0

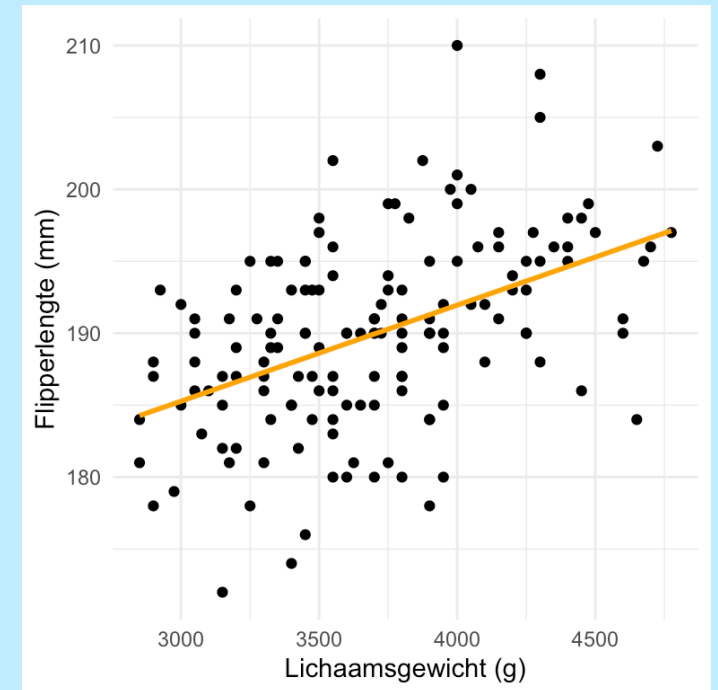
95 percent confidence interval:

0.3334071 0.5842379

sample estimates:

cor

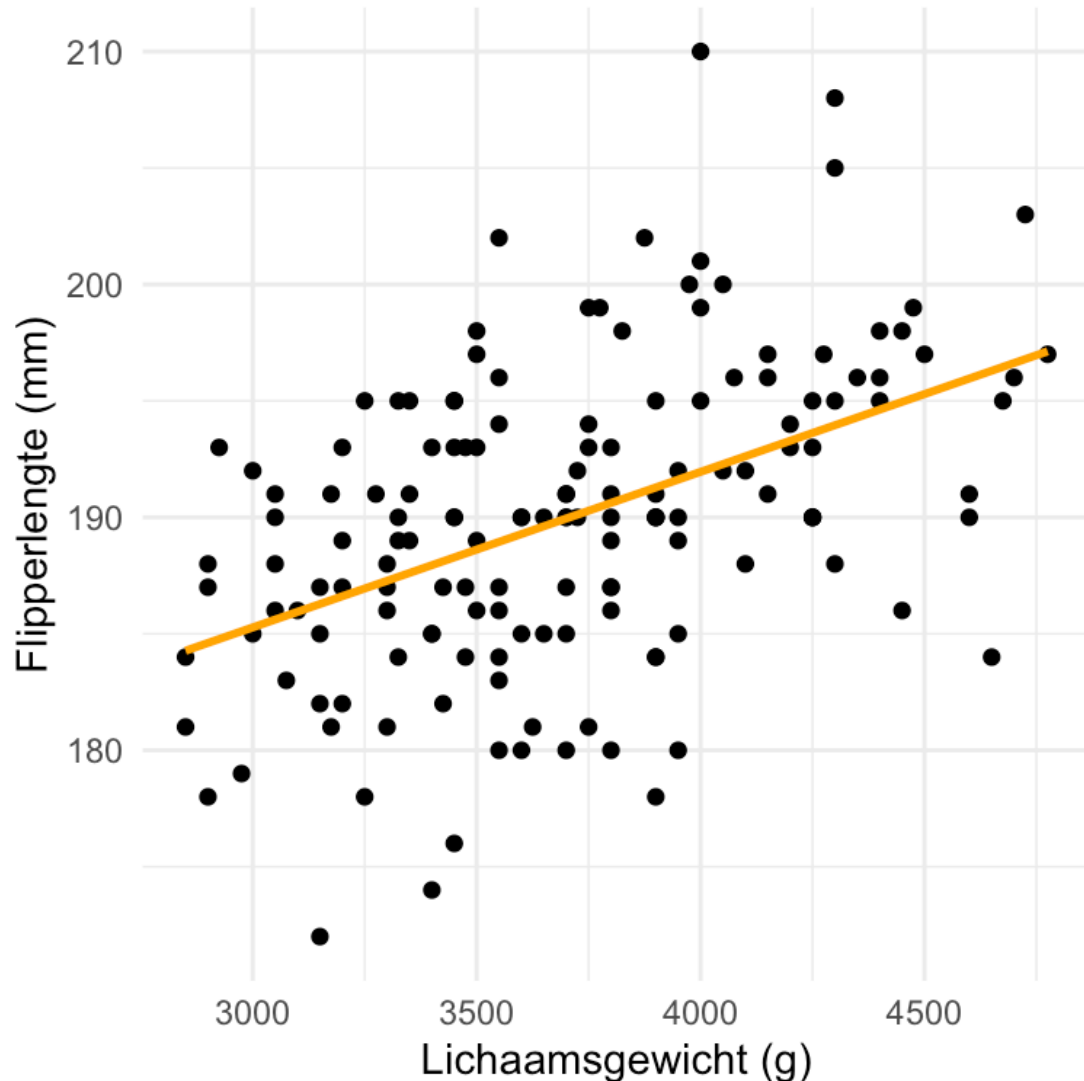
0.4682017



Inhoud

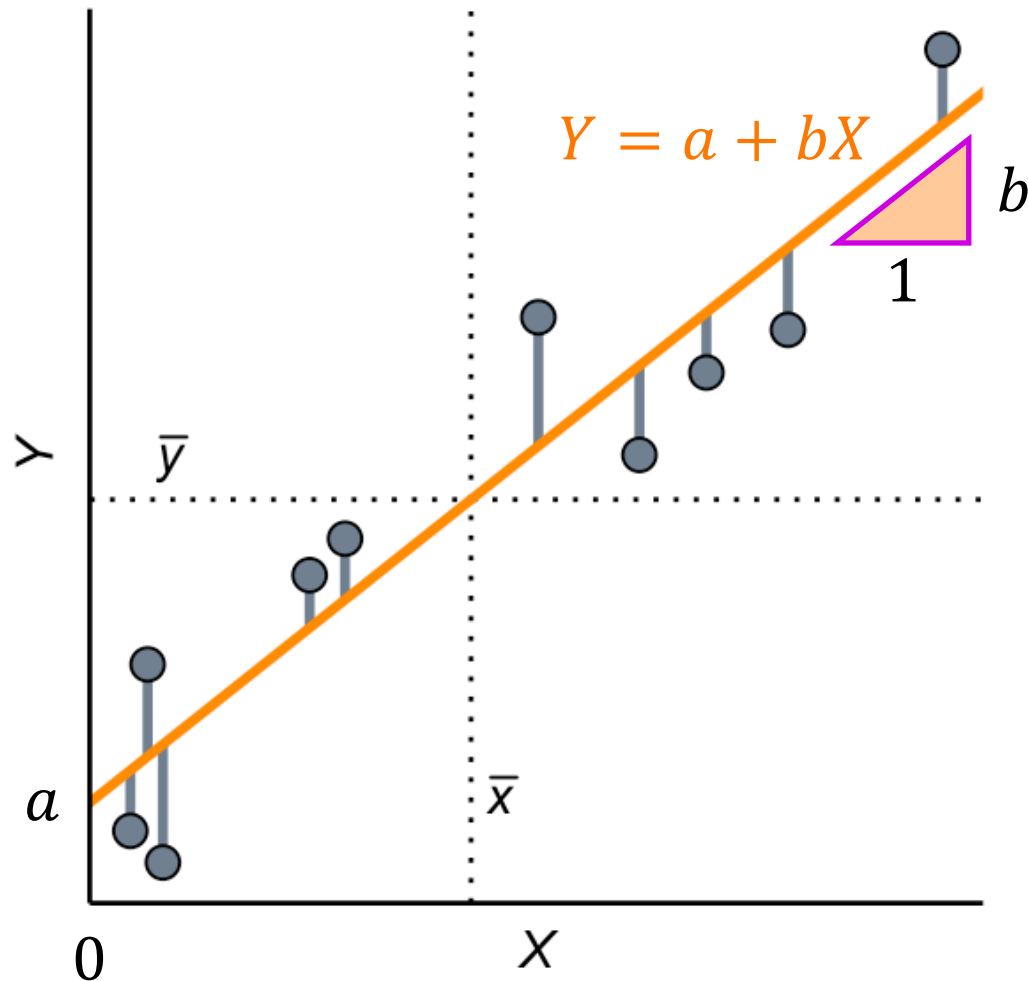
1. Lineaire relaties in een steekproef
 - Wat is een lineaire relatie?
 - Pearsons correlatiecoëfficiënt r .
 - Covariantie
 - Formule voor de correlatiecoëfficiënt
2. Hypothesetoets van correlatiecoëfficiënt
3. De regressielijn
4. Betrouwbaarheidsintervallen en toetsen voor de regressielijn
5. Interpolatie en extrapolatie
6. Niet-lineaire relaties fitten

De regressielijn: beste lineaire “fit” van de data



- Kan gebruikt worden om Y te voorspellen uit X .
- Hoe groter de correlatiecoëfficiënt, hoe preciezer de voorspelling.

De lijn die de kwadratensom minimaliseert



Definitie van de regressielijn:

De lijn waarbij a en b zo gekozen zijn dat de kwadratensom zo klein mogelijk is.

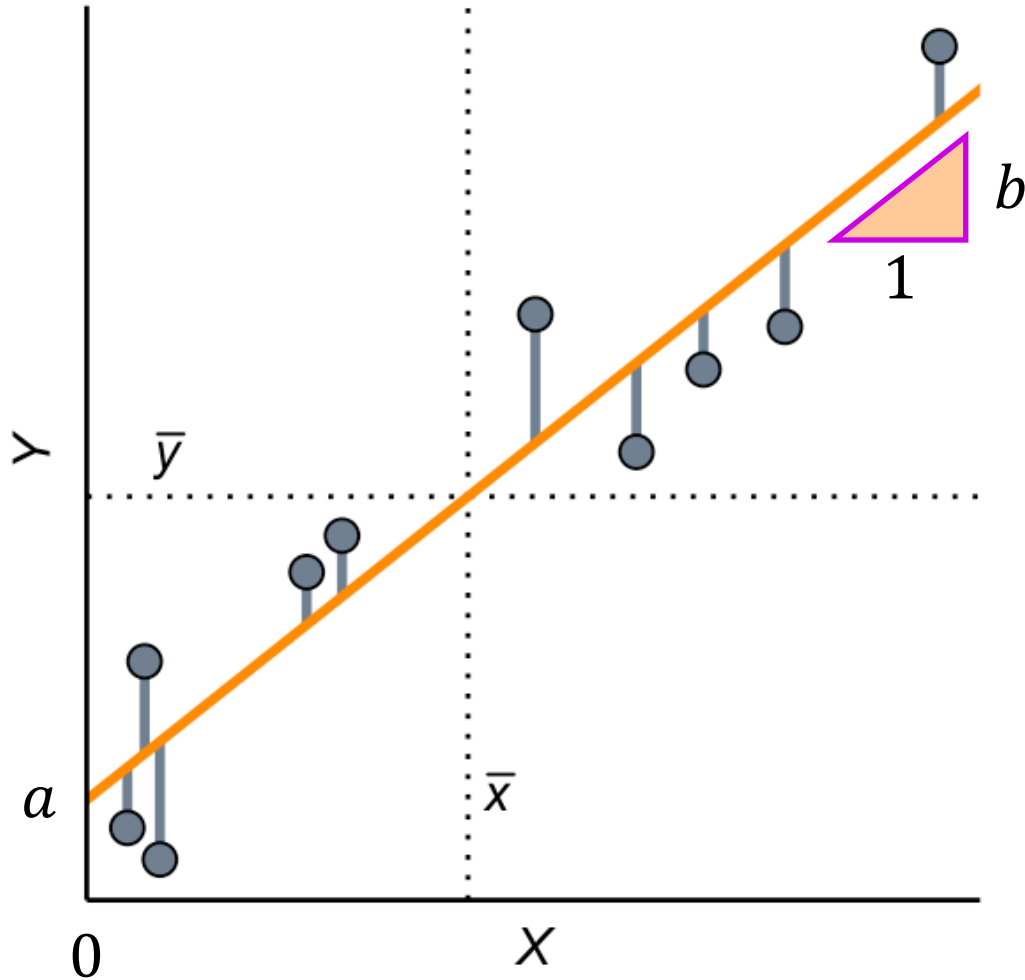
(**Kwadratensom**: som van de gekwadrateerde grijze lijntjes.)

Metafoor: grijze lijntjes zijn elastiekjes.

Consequentie:

Gaat altijd door het zwaartepunt van de puntenwolk, $(\bar{x}, \bar{y})$.

Formules voor b .



Een gesloten formule voor b :

$$b = \frac{C_{XY}}{V_X}.$$

Eerder zagen we de definitie van r :

$$r = \frac{C_{XY}}{s_X s_Y}.$$

Daaruit volgt:

$$b = \left(\frac{s_Y}{s_X} \right) r.$$

Dus: b and r hangen samen, maar zijn niet hetzelfde!

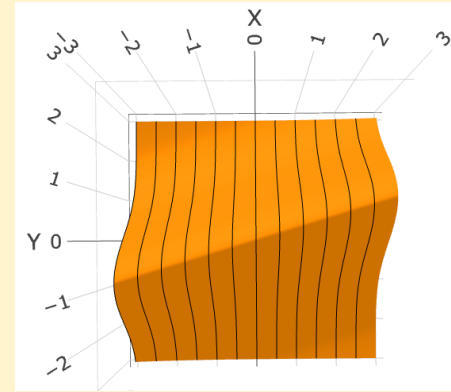
Inhoud

1. Lineaire relaties in een steekproef
 - Wat is een lineaire relatie?
 - Pearsons correlatiecoëfficiënt r .
 - Covariantie
 - Formule voor de correlatiecoëfficiënt
2. Hypothesetoets van correlatiecoëfficiënt
3. De regressielijn
4. Betrouwbaarheidsintervallen en toetsen voor de regressielijn
5. Interpolatie en extrapolatie
6. Niet-lineaire relaties fitten

Is de richtingscoëfficiënt
van de regressielijn
in de populatie
ongelijk aan 0?



Populatie

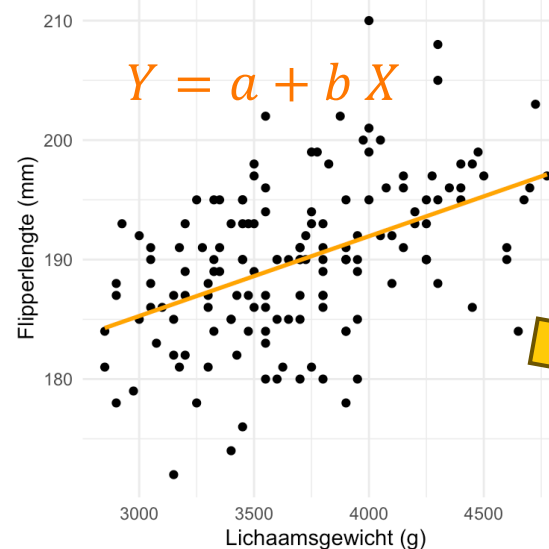


Populatie-
parameter
 β

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



Steekproef

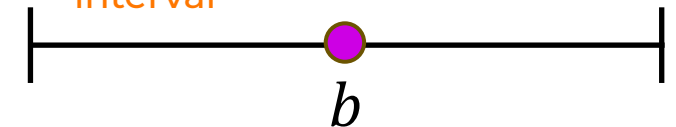


Toets

$$H_0: \beta = 0, \\ H_A: \beta \neq 0. \\ P = ?$$

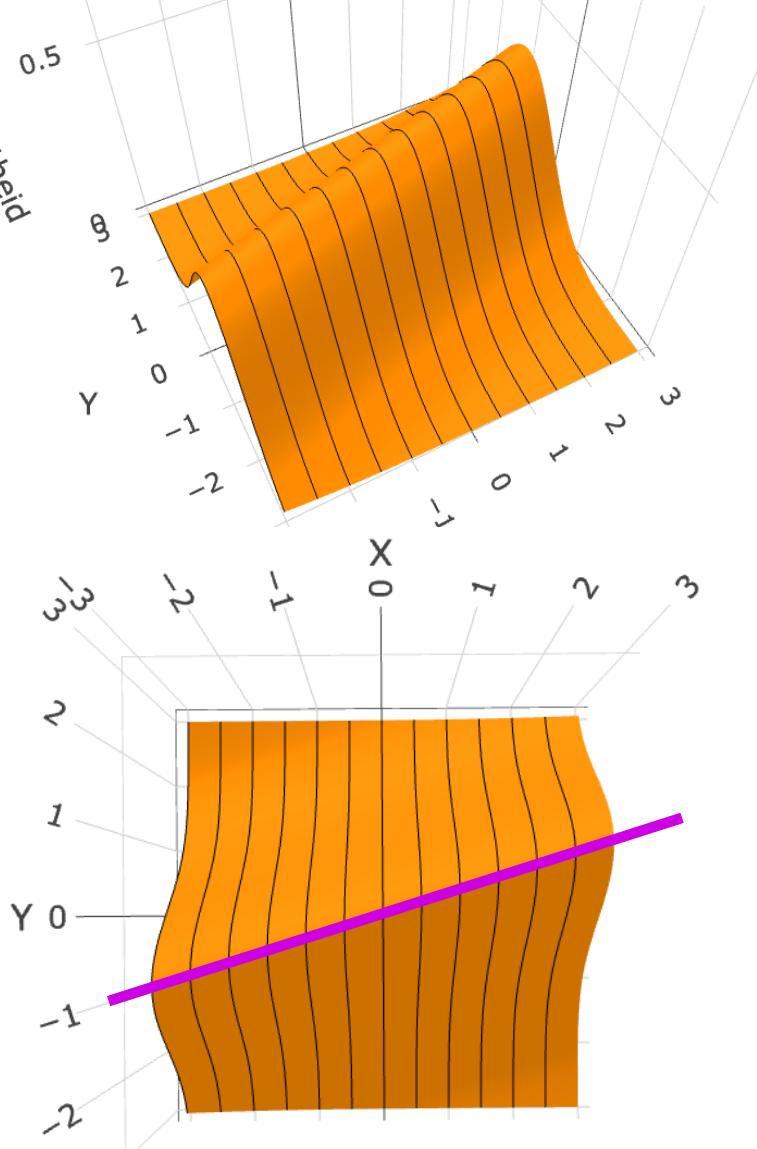
Schatting

Betrouwbaarheids-
interval



Stap 1: Het statistische model

Kansdichtheid

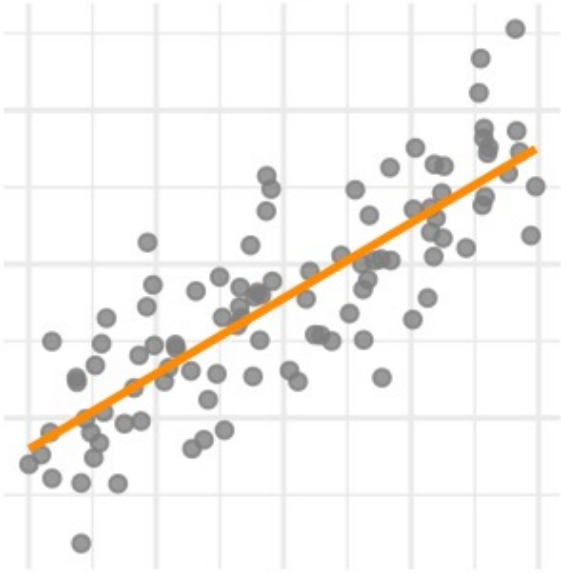


Voor de populatie geldt:

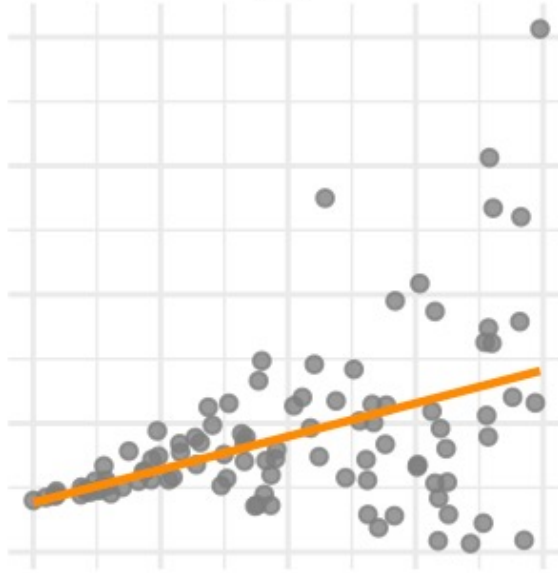
1. De verwachtingswaarde van Y hangt linear af van de waarde van X :
$$\mathbb{E}[Y|X] = \alpha + \beta X.$$
2. Voor iedere waarde van X is Y normaal verdeeld.
3. De standaarddeviatie van Y is voor alle waarden van X hetzelfde.
4. Iedere Y -waarde is onafhankelijk uit deze verdeling getrokken.

Hoe moeten samples eruit zien?

(a)



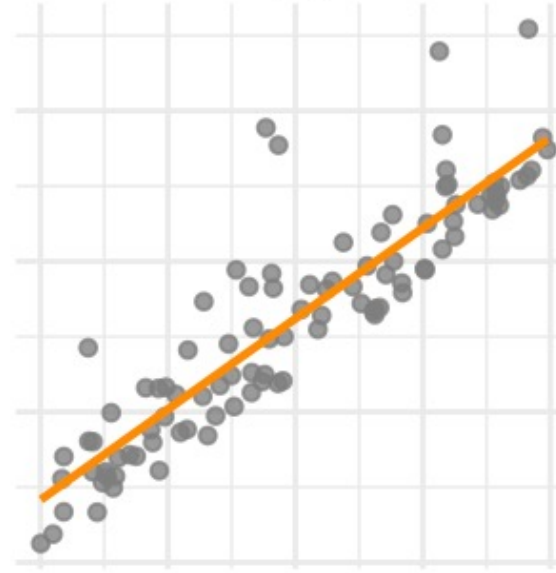
(b)



Variantie van Y niet
gelijk voor al alle X .



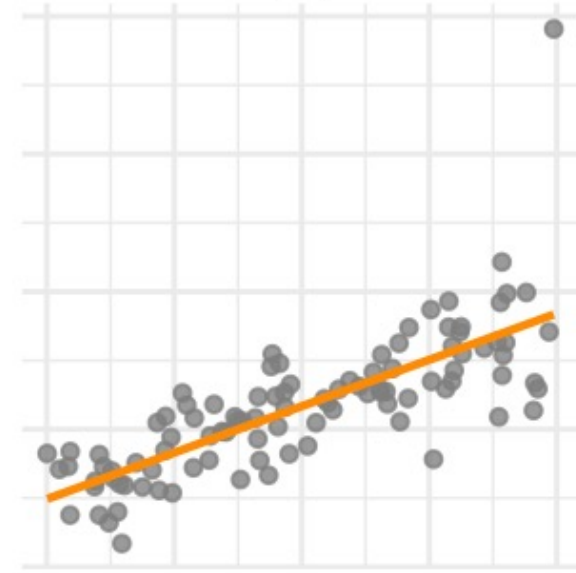
(c)



Verdeling van Y scheve
verdeling.



(d)

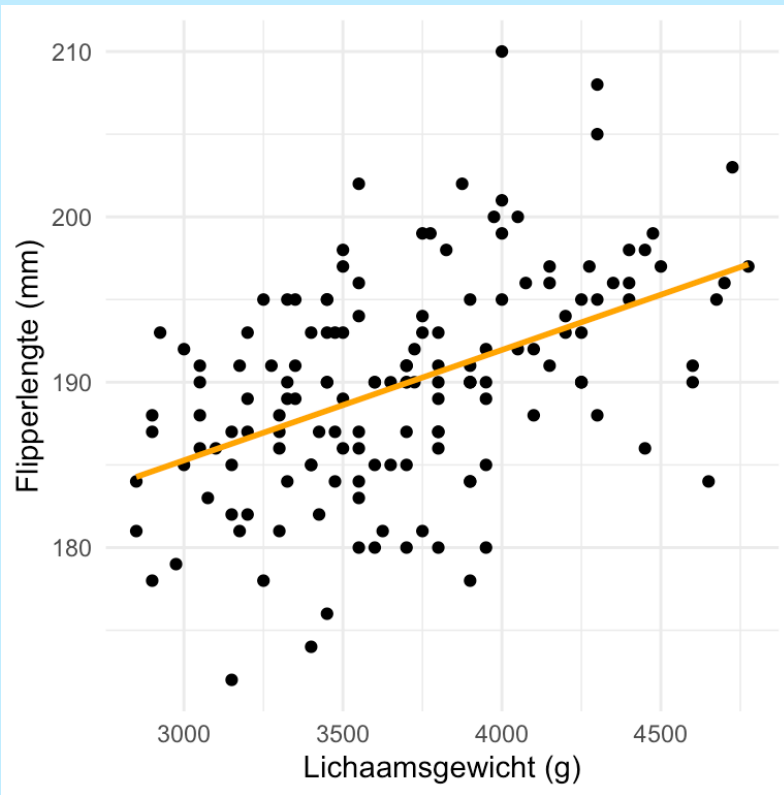


Extreme uitbijter.



BMS

Regressiecoëfficiënten uitrekenen met R



```
resultaatLM <- lm(flipper_length_mm ~ body_mass_g, data = adelie)
summary(resultaatLM)$coefficients
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.652448e+02	3.849280576	42.928752	8.681123e-86
body_mass_g	6.676867e-03	0.001032317	6.467848	1.343265e-09

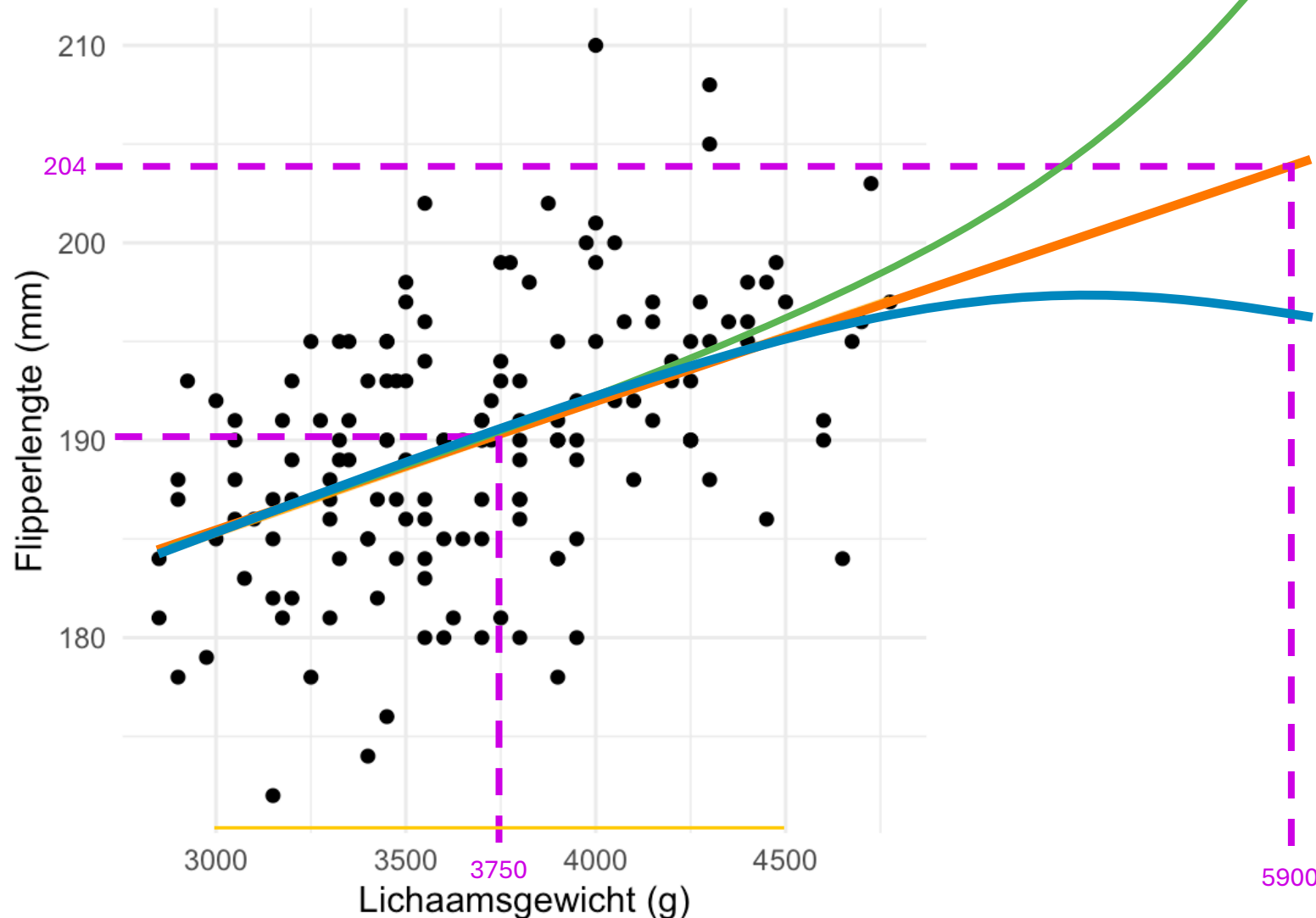
```
confint(resultaatLM)
```

	2.5 %	97.5 %
(Intercept)	1.576386e+02	1.728510e+02
body_mass_g	4.636996e-03	8.716738e-03

Inhoud

1. Lineaire relaties in een steekproef
 - Wat is een lineaire relatie?
 - Pearsons correlatiecoëfficiënt r .
 - Covariantie
 - Formule voor de correlatiecoëfficiënt
2. Hypothesetoets van correlatiecoëfficiënt
3. De regressielijn
4. Betrouwbaarheidsintervallen en toetsen voor de regressielijn
5. Interpolatie en extrapolatie
6. Niet-lineaire relaties fitten

Interpoleren en extrapoleren



Interpoleren:

Een model toepassen binnen het domein waarbinnen het model gefit was.

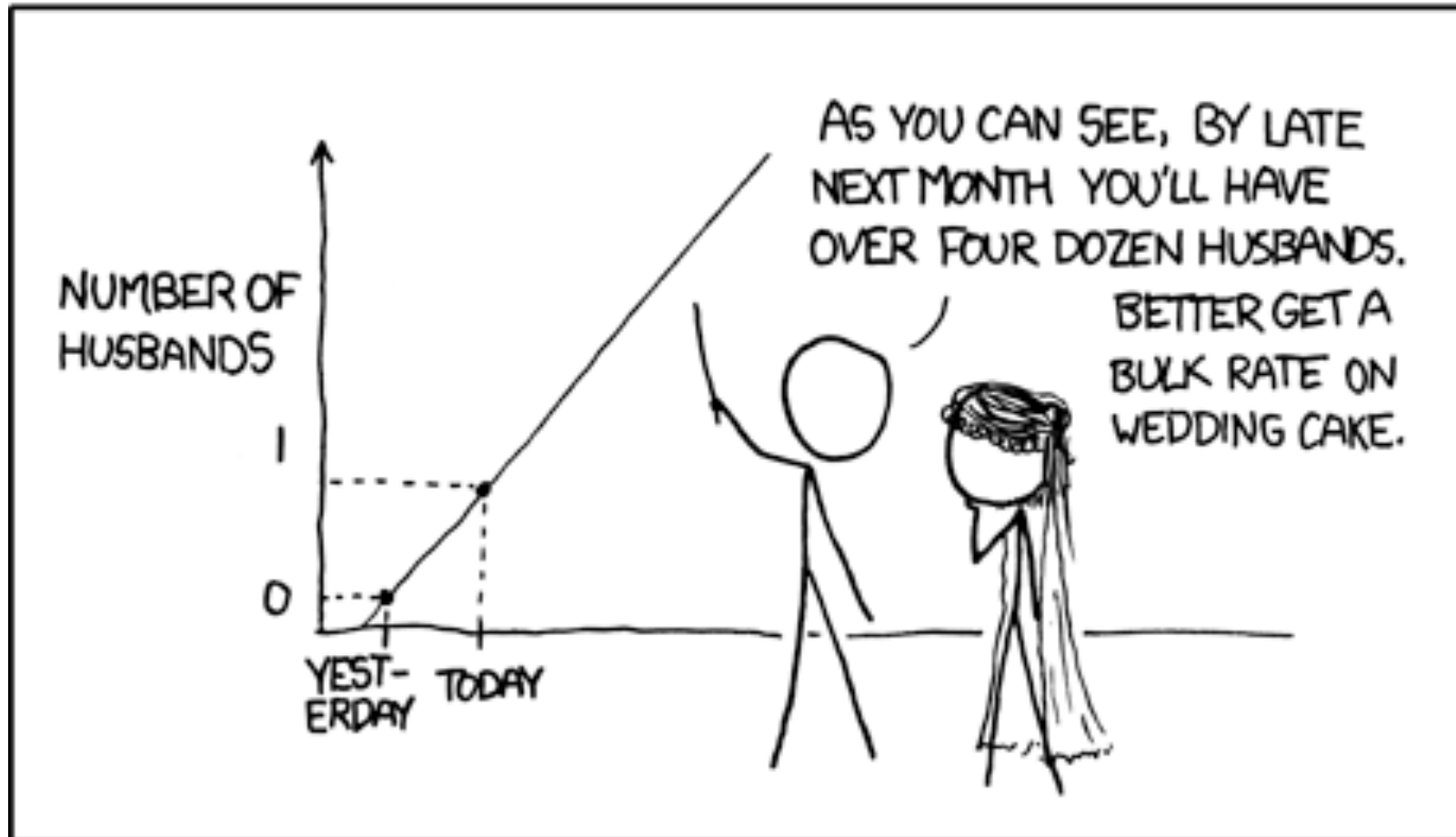
Extrapoleren:

Een model toepassen buiten het domein waarbinnen het model gefit was.

Gevaarlijk!

Interpoleren en extrapoleren

MY HOBBY: EXTRAPOLATING



Inhoud

1. Lineaire relaties in een steekproef
 - Wat is een lineaire relatie?
 - Pearsons correlatiecoëfficiënt r .
 - Covariantie
 - Formule voor de correlatiecoëfficiënt
2. Hypothesetoets van correlatiecoëfficiënt
3. De regressielijn
4. Betrouwbaarheidsintervallen en toetsen voor de regressielijn
5. Interpolatie en extrapolatie
6. Niet-lineaire relaties fitten

Niet-lineaire verbanden fitten

