

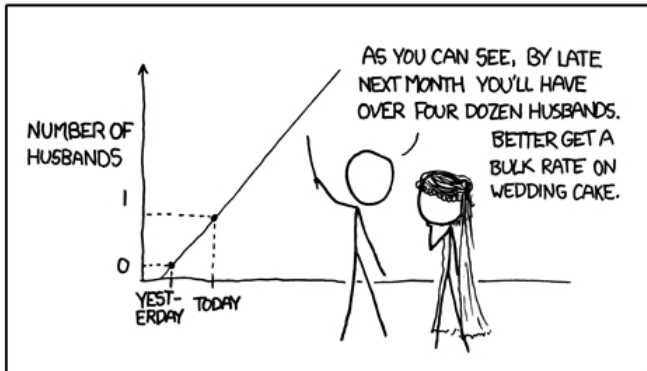
Statistiek voor A.I.

College 2

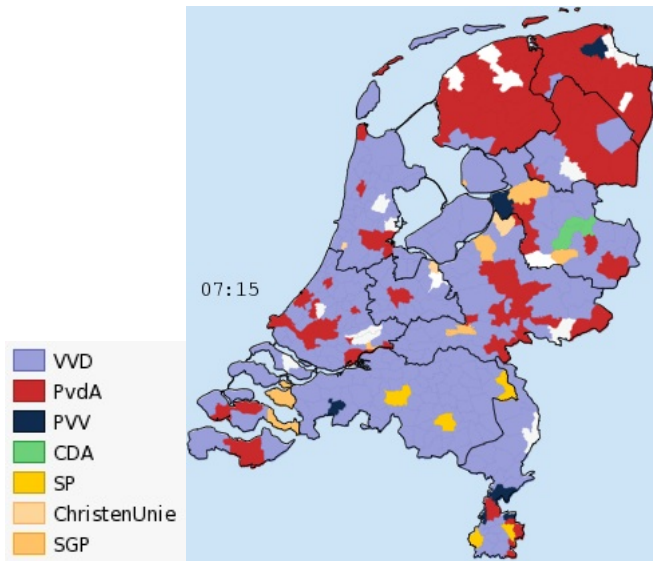
Donderdag 13 September 2012

1 Beschrijvende statistiek

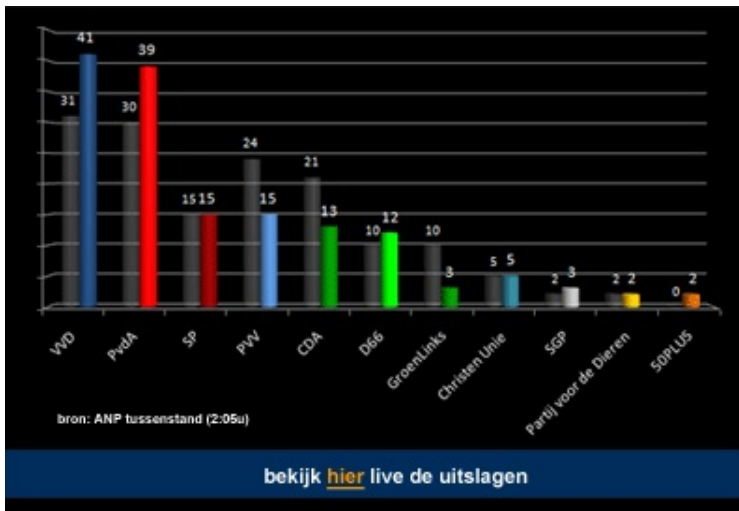
MY HOBBY: EXTRAPOLATING



Verkiezingen 2012



Verkiezingen 2012



1 Beschrijvende statistiek

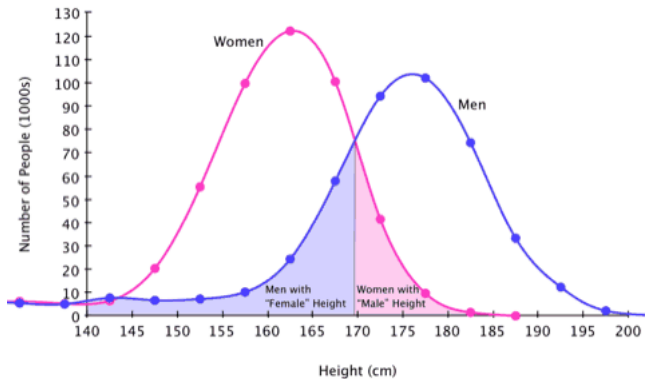
Vandaag:

- Normale verdeling
- Correlatie
- Regressie

Normale verdeling

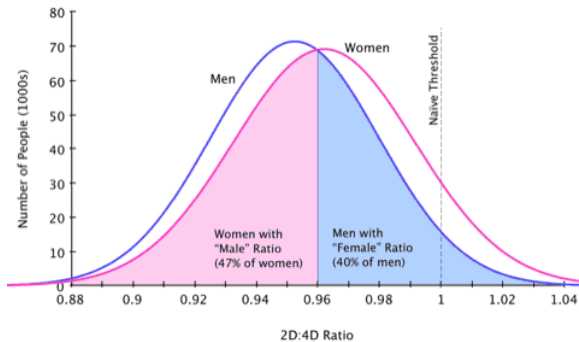
Normale verdeling

Lengte van mannen en vrouwen:



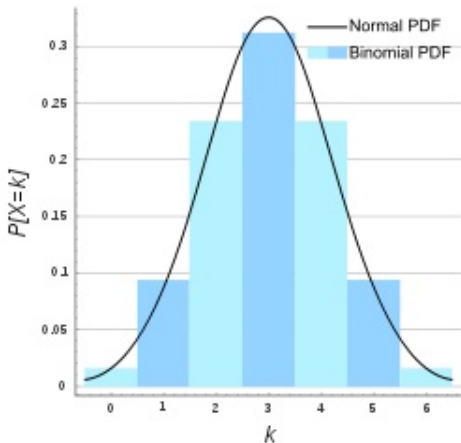
Normale verdeling

De verhouding tussen wijs- en ringvinger bij mannen en vrouwen:

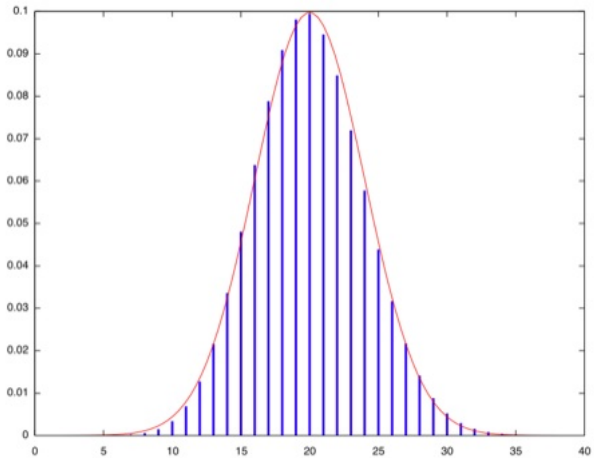


Normale verdeling en binomiale verdeling

De binomiale verdeling is discreet, de normale verdeling continu.



Normale verdeling en binomiale verdeling



Normale verdeling

Definitie De curve van de *normale verdeling* met *gemiddelde* μ en *standaardafwijking* σ wordt gegeven door de functie:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

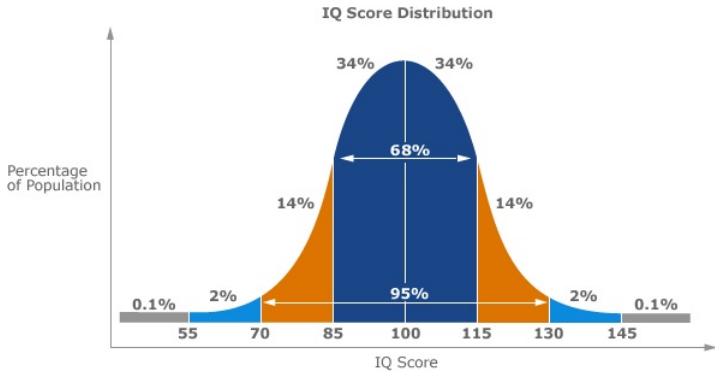
De *standaard normale verdeling* is de normale verdeling met gemiddelde $\mu = 0$ en standaardafwijking $\sigma = 1$. De bijbehorende functie is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

De verdeling wordt vaak aangeduid met P_s .

Normale verdeling

Het IQ:



Normale verdeling

Stelling Als de waarden van een variabele X normaal verdeeld zijn, met gemiddelde μ en standaardafwijking σ , dan geldt:

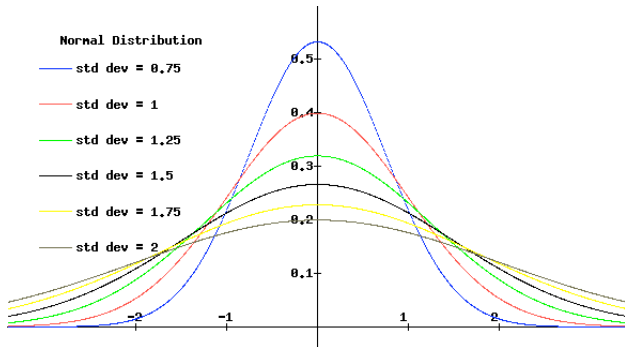
68% van de waarden liggen in het interval $[\mu - \sigma, \mu + \sigma]$,

95% van de waarden liggen in het interval $[\mu - 2\sigma, \mu + 2\sigma]$ en

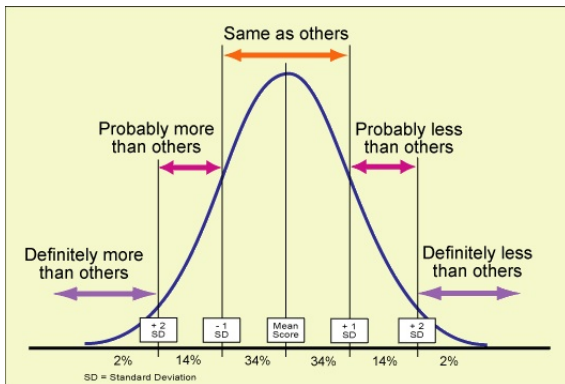
99% van de waarden liggen in het interval $[\mu - 3\sigma, \mu + 3\sigma]$.

Normale verdeling

Voorbeeld Hoe groter de standaardafwijking hoe breder de klok:



Normale verdeling



Correlatie

Correlatie

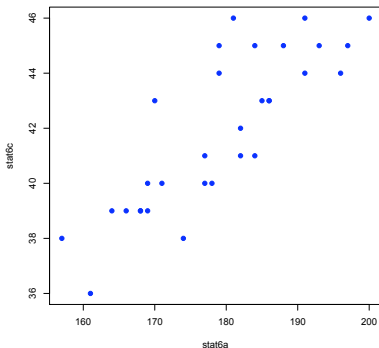
Definitie Een *scatterplot* op grond van een steekproef waarbij twee scores gemeten worden, stochasten X en Y , is een grafiek waarin de paren (X_i, Y_i) voor elke element i in de steekproef weergegeven worden.

Voor scatterplots moeten de variabelen interval- of ratioschaal zijn.

Correlatie: lengte en schoenmaat KI studenten 2012

Voorbeeld Lengte (x-as) tegen schoenmaat (y-as):

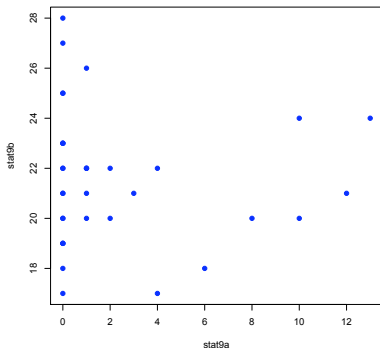
lengte	157	161	...	170	170	...
schoenmaat	38	36	...	39	40	...



Correlatie: carnaval en BMI KI studenten 2012

Voorbeeld Aantal maal carnaval gevierd (x-as) tegen BMI (y-as):

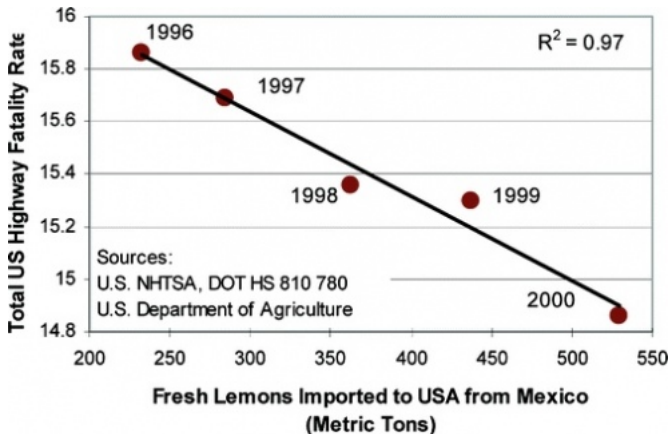
carnaval	0	0	0	...	13
BMI	17	18	19	...	24



Correlatie

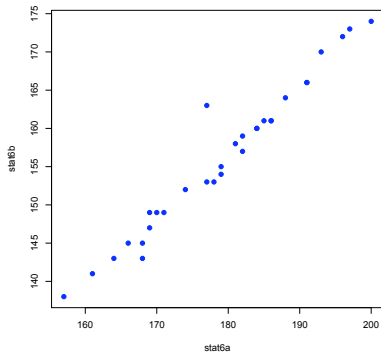
Definitie Een *correlatie* tussen de variabelen betekent dat er op grond van de steekproef een verband lijkt te zijn.

Een correlatie impliceert niet noodzakelijk een causaal verband (hidden variables).



Correlatie: lengte en hoogte kin KI studenten 2012

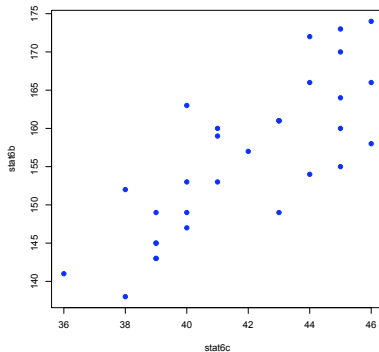
Voorbeeld Lengte (x-as) tegen hoogte kin (y-as):



Er bestaat een sterke correlatie tussen lengte en de hoogte van de kin.

Correlatie: lengte en hoogte kin KI studenten 2012

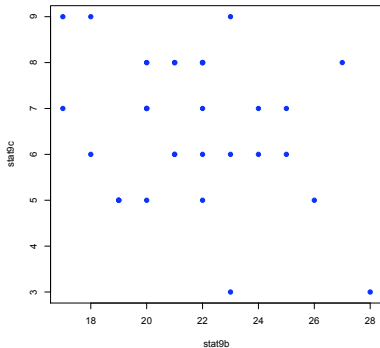
Voorbeeld Schoenmaat (x-as) tegen hoogte kin (y-as):



Er bestaat een correlatie tussen schoenmaat en de hoogte van de kin.

Correlatie: lengte en hoogte kin KI studenten 2012

Voorbeeld BMI (x-as) tegen aantal vakken behaald in het eerste jaar (y-as):



Er bestaat geen correlatie tussen BMI en het aantal vakken behaald in het eerste jaar.

Correlatie: correlatiecoëfficiënt

Definitie Gegeven twee stochasten X en Y , waarbij (X_i, Y_i) de score van element i is in een steekproef ter grootte n , is de *Pearson correlatiecoëfficiënt*:

$$r_{XY} = \frac{\sum z_X z_Y}{n-1} = \frac{\sum_{i=1}^n z_{X_i} z_{Y_i}}{n-1}.$$

Waarbij z_{X_i}, z_{Y_i} de standaarscores van X_i, Y_i t.o.v. de steekproef zijn:

$$z_{X_i} = \frac{X_i - \bar{X}}{s_X} \quad z_{Y_i} = \frac{Y_i - \bar{Y}}{s_Y}.$$

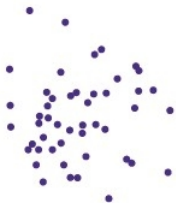
Gebruikmakend van de notatie $x = X - \bar{X}$:

$$r_{XY} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}.$$

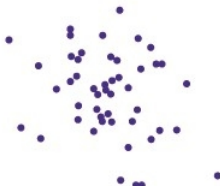
Ook geldt

$$r_{XY} = \frac{n \sum XY - \sum X \sum Y}{\sqrt{(n \sum X^2 - (\sum X)^2)(n \sum Y^2 - (\sum Y)^2)}}.$$

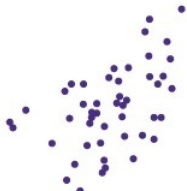
Correlatie



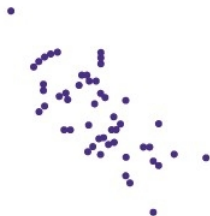
Correlation $r = 0$



Correlation $r = -0.3$



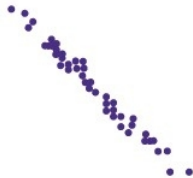
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

Merk op: $-1 \leq r_{XY} \leq 1$.

De Pearson correlatiecoëfficiënt is een maat voor het *lineaire* verband tussen twee variabelen.

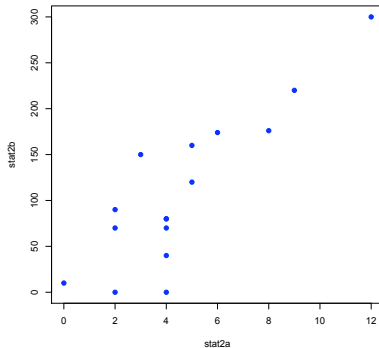
Hoe homogener een van de variabelen over de populatie verdeeld is, hoe kleiner de absolute waarde van de Pearson correlatiecoëfficiënt wordt.

De Pearson correlatiecoëfficiënt is ordinaal.

Correlatie: (facebook) vrienden KI studenten 2012

Voorbeeld Goede vrienden (x-as) tegen aantal facebook vrienden (y-as).

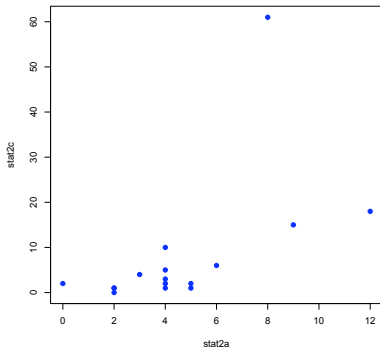
Gemiddel aantal goede vrienden: 4.6. Gemiddeld aantal facebook vrienden: 108.8.



De Pearson correlatiecoëfficiënt: 0.82.

Correlatie: vrienden en zoenen KI studenten 2012

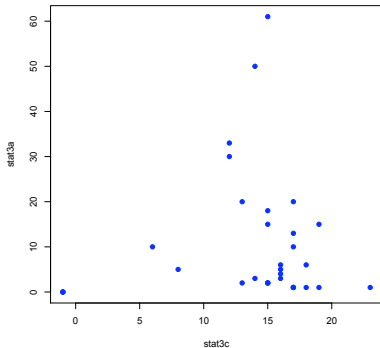
Voorbeeld Goede vrienden (x-as) tegen aantal mensen waarmee gezoend (y-as).
Gemiddeld aantal mensen waarmee gezoend: 8.3.



De Pearson correlatiecoëfficiënt: 0.56.

Correlatie: (eerste) zonen KI studenten 2012

Voorbeeld Leeftijd eerste zoon (x-as) tegen aantal mensen waarmee gezoend (y-as).

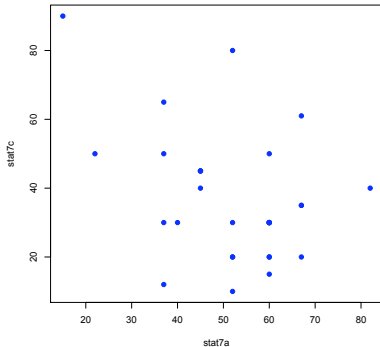


De Pearson correlatiecoëfficiënt: 0.08.

Correlatie: Studie versus vrij tijd KI studenten 2012

Voorbeeld Aantal ECTS vorig jaar behaald (x-as) tegen aantal uren per week besteed aan hobbies/uitgaan/werk (y-as).

Gemiddeld aantal ECTS: 52.1. Gemiddeld aantal uren hobbies/uitgaan/werk: 36.71.



De Pearson correlatiecoëfficiënt: -0.40.

Regressie

Definitie Op grond van twee variabelen X en Y worden constanten a en b bepaald zodat de lijn $\hat{Y} = bX + a$ zo goed mogelijk het lineaire verband tussen X en Y weergeeft.

De *regressiecoëfficiënt* b is gedefiniëerd als:

$$b = r \frac{s_Y}{s_X}.$$

waarbij s_X en s_Y de standaardafwijking van respectievelijk X en Y zijn.

De *regressieconstante* a is gedefiniëerd als:

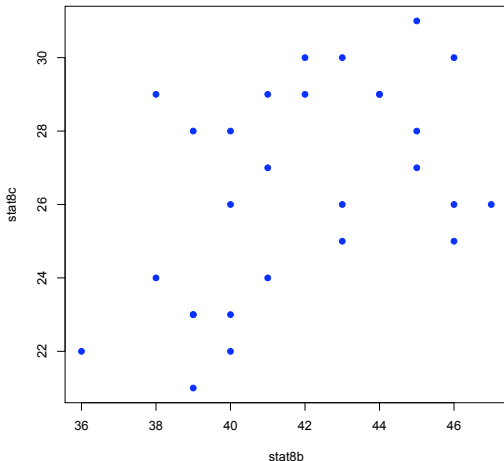
$$a = \bar{Y} - b\bar{X}.$$

Stelling Voor de regressielijn is de waarde van $\sum(Y - \hat{Y})^2$ minimaal.

Regressie: schoenmaat en elleboog KI studenten 2012

Voorbeeld Schoenmaat (x-as) tegen de afstand van de binnenkant van de elleboog tot de handpalm (y-as).

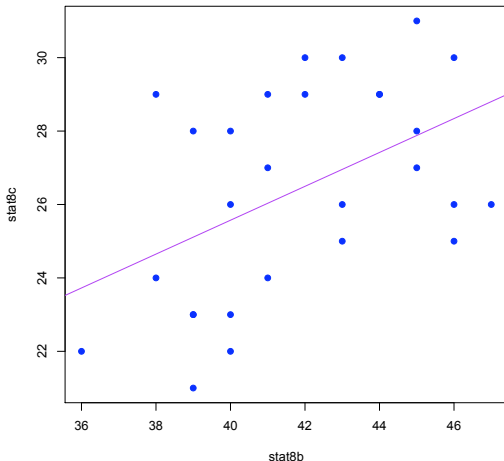
Pearson correlatiecoëfficiënt $r = 0.47$, regressieconstante $a = 7.11$ en regressiecoëfficiënt $b = r \frac{s_Y}{s_X} = 0.46$.



Regressie: schoenmaat en elleboog KI studenten 2012

Voorbeeld Schoenmaat (x-as) tegen de afstand van de binnenkant van de elleboog tot de handpalm (y-as).

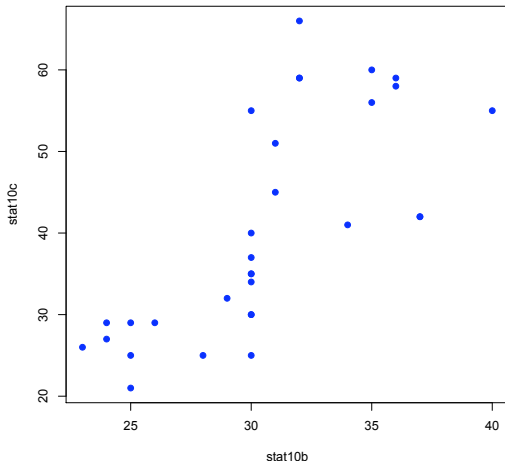
Pearson correlatiecoëfficiënt $r = 0.47$, regressieconstante $a = 7.11$ en regressiecoëfficiënt $b = r \frac{s_Y}{s_X} = 0.46$.



Regressie: vingerkootjes KI studenten 2012

Voorbeeld Lengte middelste vingerkootje (x-as) tegen lengte onderste vingerkootje (y-as).

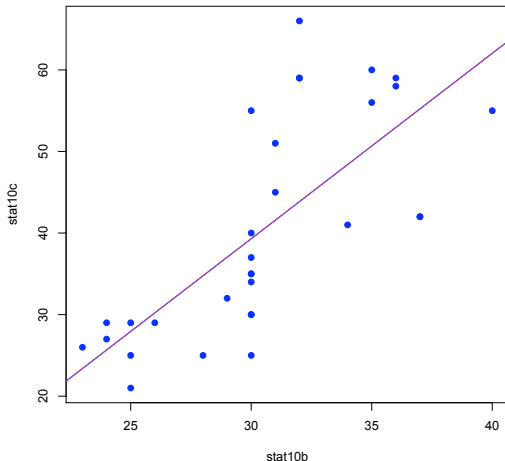
Pearson correlatiecoëfficiënt $r = 0.73$, regressieconstante $a = -28.89$ en regressiecoëfficiënt $b = r \frac{s_Y}{s_X} = 2.27$.



Regressie: vingerkootjes KI studenten 2012

Voorbeeld Lengte middelste vingerkootje (x-as) tegen lengte onderste vingerkootje (y-as).

Pearson correlatiecoëfficiënt $r = 0.73$, regressieconstante $a = -28.89$ en regressiecoëfficiënt $b = r \frac{s_Y}{s_X} = 2.27$.



Stelling Laat $\hat{Y} = bX + a$ de regressielijn van (X, Y) zijn en r de Pearson correlatiecoëfficiënt. $z_{\hat{Y}}$ zijn de standaardscores van $\hat{Y}$ en z_X van X . Dan geldt:

$$z_{\hat{Y}} = rz_X.$$

Laat $\hat{Y} = bX + a$ de regressielijn van (X, Y) zijn.

Met behulp van de *regressielijn* kunnen de scores/waarden voor elementen uit de populatie die niet in de steekproef bevat zijn voorspeld worden.

Definitie Wanneer er n scores zijn en de Pearson correlatiecoëfficiënt is r , dan is de *standaardfout van de benadering*, $s_{Y \cdot X}$, gedefiniëerd als de standaardafwijking van de stochast $(Y - \hat{Y})$ bij $n - 1$ scores. D.w.z., voor $Z = (Y - \hat{Y})$:

$$s_{Y \cdot X} = \sqrt{\frac{\sum (Z - \bar{Z})^2}{n - 2}}.$$

Stelling Wanneer er n scores zijn en de Pearson correlatiecoëfficiënt is r , dan geldt

$$s_{Y \cdot X} = s_Y \sqrt{1 - r^2} \sqrt{\frac{n - 1}{n - 2}}.$$

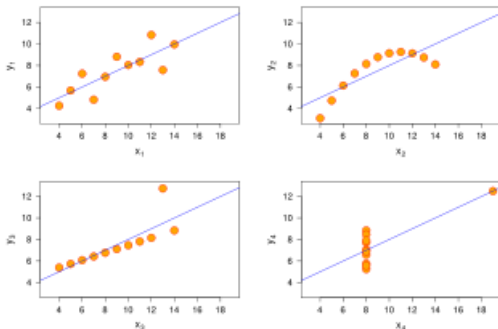
Voor grote n geldt:

$$s_{Y \cdot X} \approx s_Y \sqrt{1 - r^2}.$$

Hoe groter de correlatie, hoe kleiner de standaardfout van de benadering.

Regressie: het quartet van Anscombe

Het quartet van Anscombe bestaat uit vier verzamelingen data die dezelfde statistische eigenschappen hebben, maar die verschillend zijn wanneer ze grafisch worden weergegeven.



In alle vier de verzamelingen data zijn voor X (x -as) gemiddelde en variantie gelijk. Evenzo voor Y . Voor alle vier is de correlatiecoëfficiënt en regressielijn gelijk.

Finis