Extra uitleg bij de begrippen steekproefcovariantie en correlatiecoëfficiënt.

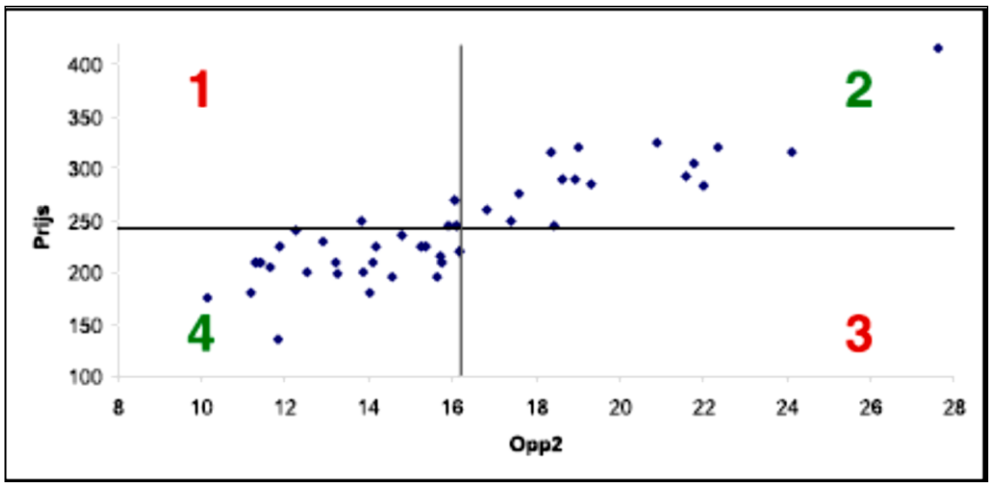
Soms willen we 2 grootheden tegelijkertijd bestuderen. We hebben dan een steekproef waarbij we 2 sets van waarden krijgen x_i en y_i, die we dikwijls als koppels noteren (x_i, y_i). We willen nagaan of er een verband is tussen beide grootheden. De steekproefcovariantie en de correlatiecoëfficiënt zijn 2 getallen waarmee we dit kunnen doen.

Definities:

De **steekproefcovariantie** $cov(x,y) = s_{xy}$ is:	De **correlatiecoëfficiënt** $r(x,y)$ is:
$cov(x,y) = s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	$r(x,y) = \frac{cov(x,y)}{\sqrt{var(x)var(y)}} = \frac{s_{xy}}{s_x s_y}$

Intuïtieve interpretatie

De waarde en het teken van de steekproefcovariantie en de correlatiecoëfficiënt geven het verband, de associatie, aan tussen de 2 gegeven sets waarden x_i en y_i.
Zijn de x_i en de y_i waarnemingen van 2 grootheden die een sterk gelijklopend gedrag hebben, een zogenaamde positieve associatie hebben, dan zal de covariantie positief en groot zijn. Wat is groot? Dit hangt af van de spreiding van de waarden voor x_i en y_i rond hun gemiddelde (= de standaardafwijking). Heb je bvb voor x_i en y_i een spreiding van de grootteorde 10 en bekom je voor de covariantie 70 dan is dit groot. Is de spreiding op x_i en y_i van de grootteorde 1000, dan is een covariantie van 70 klein. De covariantie is dus in feite geen goede waardemeter voor het verband tussen x_i en y_i. Vandaar dat we de covariantie herschalen zodanig dat we steeds een getal krijgen tussen -1 en 1. Dit is de correlatiecoëfficiënt. Als $r(x,y) = 0.9$ dan betekent dat inderdaad dat x_i en y_i heel gelijklopend zijn, positief geassocieerd zijn. De correlatiecoëfficiënt geeft dus aan in welke mate de 2 grootheden een invloed hebben op mekaar, in welke mate er een verband is tussen beide.
Is er een perfect lineair verband tussen *x* en *y*, bvb $y = ax + b$ dan is de correlatiecoëfficiënt exact gelijk aan 1 als $a > 0$ (als *x* stijgt, stijgt *y* ook), en exact gelijk aan -1 als $a < 0$ (als *x* stijgt, daalt *y*) . Vandaar dat de correlatie enkel het lineaire verband meet tussen *x* en *y*. Op de grafiek p 26 zie je dit goed.



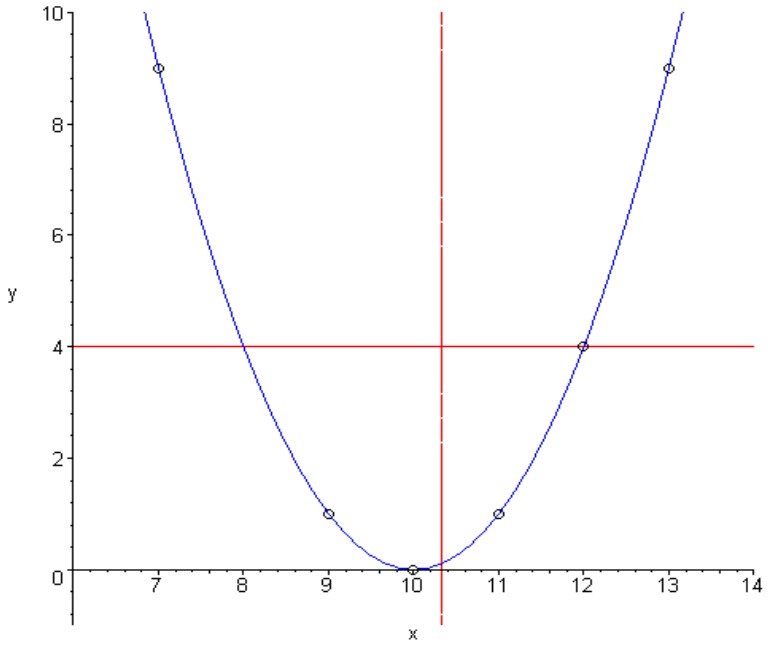
Deze 2 grootheden hebben een sterke correlatie ($r(x,y) = 0.886$ zie p27) en je ziet dit op de grafiek want de punten liggen allemaal in een strook die van linksonder naar rechtsboven

loopt, je kunt er als het ware bijna een rechte door trekken. (Dit laatste zal je trouwens volgend jaar doen in de cursus Toegepaste Statistiek II bij lineaire regressie).

Liggen de punten wild verspreid in het vlak dan is de correlatie heel klein.
Desondanks kan er wel een niet lineair verband bestaan tussen *x* en *y*, zie het voorbeeld op pag. 28 dat voorgesteld is op onderstaande figuur.

x_i	11	12	10	9	13	7
y_i	1	4	0	1	9	9

De punten liggen helemaal verspreid over het vlak. Na berekening blijkt de correlatie inderdaad 0 te zijn. Desondanks liggen de punten allemaal op dezelfde parabool en is er dus een exact kwadratisch verband tussen de *x* en de *y* nl. $y_i = (x_i - 10)^2$



<u>Praktische berekening van de steekproefcovariantie en de correlatiecoëfficiënt?</u>

In de praktijk gebruiken we voor de covariantie meestal de volgende uitdrukking:

$$cov(x,y) = s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)$$

of kortweg:

$$cov(x,y) = s_{xy} = \frac{1}{n-1}\left(\sum xy - \frac{1}{n}\sum x \sum y\right)$$

Die onmiddellijk volgt uit de definitie door de haakjes uit te werken:

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y})$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y}\right)$$

Door gebruik te maken van de definitie van het gemiddelde, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$en analoog voor *y*, vind je hieruit onmiddellijk de gewenste formule.

Voorbeelden:

- Laat ons eens met het RM de covariantie en correlatie berekenen van het voorbeeld hierboven (vb 2.32 p 28).

x_i	11	12	10	9	13	7
y_i	1	4	0	1	9	9

Ga als volgt te werk:
Kies STAT, EDIT, 1:Edit
Stop de waarden voor *x* in de kolom L1, de waarden voor *y* in L2
Kies STAT, CALC, 2:2-Var Stats ENTER
We gebruiken de kolommen L1 en L2 dus vul het commando 2-Var Stats nu aan:
2-Var Stats L1,L2 en ENTER
Op het scherm lees je o.a. af:

$$\sum x = 62, \sum x^2 = 664, S_x = 2.160247$$
$$\sum y = 24, \sum y^2 = 180, S_y = 4.09878$$
$$\sum xy = 248$$
$$n = 6$$

Hiermee is

$$cov(x,y) = \frac{1}{5}\left(\sum xy - \frac{1}{6}\sum x \sum y\right) = \frac{1}{5}\left(248 - \frac{62.24}{6}\right) = 0$$

en dus ook $r(x,y) = \frac{cov(x,y)}{\sqrt{var(x)var(y)}} = 0$

(Je kan natuurlijk ook de waarden opvragen via VARS, typ dan:

$1 \div 5 \times (VARS\ 5{:}\sum 5{:}\ ENTER - 1 \div 6 \times VARS\ 5{:}\sum 1{:}\ ENTER \times VARS\ 5{:}\ \sum 3{:}\ ENTER)$

- Verband prijs/verkochte hoeveelheid

Volgende metingen werden gedaan van de eenheidsverkoopprijs en het aantal verkochte eenheden van een bepaald goed. Bereken de correlatie tussen beide grootheden.

p	3.00	5.20	6.60	4.00	3.58	4.23	5.00	6.00
q	710	468	332	593	648	561	503	394

Oplossing:
n = 8
De formule voor de covariantie is dus: $cov(x,y) = \frac{1}{7}\left(\sum xy - \frac{1}{8}\sum x \sum y\right)$

Stop de waarden voor p en q elk in een kolom, bvb L3 en L4. (via STAT, Edit)

Dan: STAT, CALC, 2:2-Var Stats L3,L4

Voor de covariantie typ dan:

$1 \div 7 \times (VARS\ 5{:}\,Statistics\ \sum 5{:}\,\sum xy - 1 \div 8\ \times VARS\ 5{:}\,Statistics\ \sum 1{:}\,\sum x\ \times VARS\ 5{:}\,Statistics\ \sum 3{:}\,\sum y)$

resultaat: cov(p,q) = -155.5558929

Sla dit op: STO> C

De correlatie vind je dan via :

$$C \div VARS\ 5{:}\,Statistics\ XY\ 3{:}\,S_x \div VARS\ 5{:}\,Statistics\ XY\ 6{:}\,S_y$$

Resultaat: - 0.998

De correlatie is dus heel sterk en tegengesteld. Als je de gegevens voorstelt op een pq grafiek zullen de punten dus ongeveer op een dalende rechte moeten liggen. Dit blijkt uit volgende screenshot van het RM, dat jullie zelf ook eens kunnen maken thuis. Wat de vergelijking is van deze best passende rechte zullen jullie volgend jaar leren berekenen. (Lineaire regressie)