# Chapitre 7:

*Corrélation*

## 7.1. Corrélation entre deux variables

Jusqu'à présent, nous nous sommes intéressés à des questions du type:

- quelle est la taille moyenne des garçons belges âgés d'une vingtaine d'années ?
- quelle est la probabilité pour qu'un médicament soit efficace ?
- quel pourcentage de voix un parti politique recueillera-t-il aux prochaines élections ?
- quelle fraction des barres métalliques produites par une usine sera-t-elle rejetée par le client ?
- le poids moyen des pains produits dans une boulangerie est-il supérieur à 800 grammes ?

Dans toutes ces questions, nous étudions le comportement statistique d'une seule variable: taille, efficacité du médicament, pourcentage de voix, longueur des barres, poids des pains.

Il existe cependant toute une gamme de problèmes statistiques où l'on s'intéresse à la *relation entre plusieurs variables*.

*Exemples:*

- les individus les plus grands sont-ils les plus lourds ?
- le revenu d'une famille a-t-il une influence sur les résultats scolaires des enfants ?
- y a-t-il une relation entre le tabagisme et les cancers du poumon ?
- le rendement en céréales dépend-il de la quantité d'engrais utilisée ?
- la productivité d'une entreprise est-elle liée au salaire des ouvriers ou employés ?

Dans ces questions, nous désirons savoir si le comportement d'une variable est influencé par la valeur d'une autre variable:

taille ⟷ poids　　revenu ⟷ résultats

tabagisme ⟷ cancer　　rendement ⟷ engrais

La relation peut être *causale* ou non

Pour étudier les *relations* ou *corrélations* entre deux variables statistiques, on peut les porter sur un graphique.

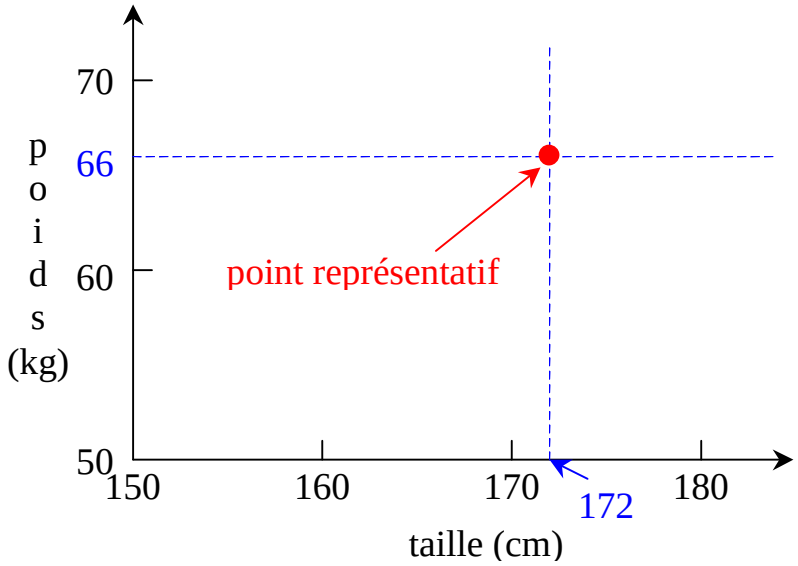
*Exemple:* relation entre la taille et le poids des individus

pour chaque individu de l'échantillon, on porte sur un graphique:

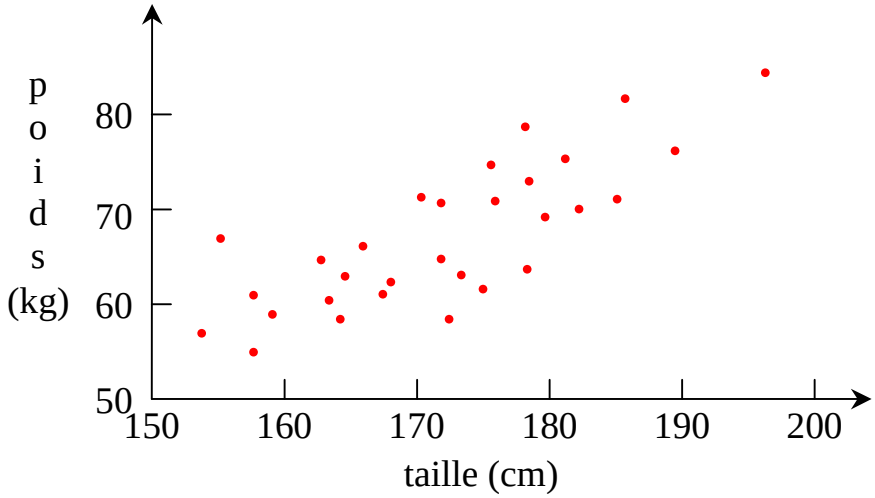
- sa taille en abscisse (*l'abscisse d'un point correspond à sa projection sur l'axe horizontal*)
- son poids en ordonnée (*l'ordonnée d'un point correspond à sa projection sur l'axe vertical*)

chaque individu est donc, dans ce graphique, *représenté* par un point *(point représentatif)*

soit un individu mesurant 172 cm et pesant 66 kg:




Dans le graphe, il y aura donc autant de points qu'il y a d'individus dans l'échantillon.




*Relation entre le poids et la taille dans un échantillon de 30 individus.*

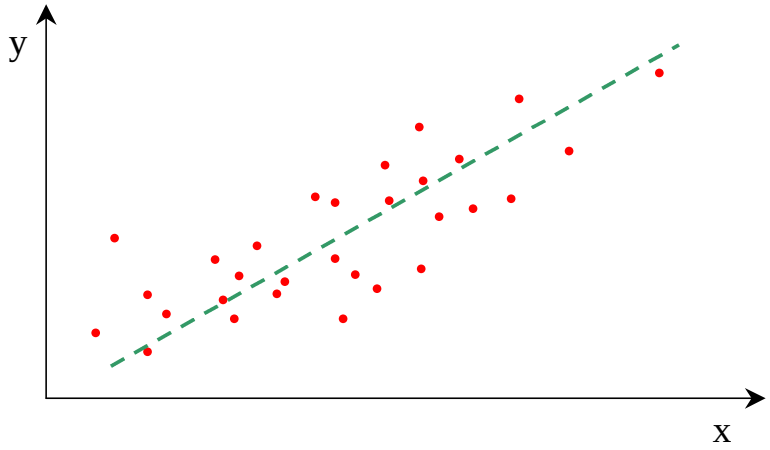
On peut (par la pensée ou réellement) tracer une droite qui passe au mieux par ces points (au milieu du "nuage" de points).

Si cette droite "monte", on dira qu'il y a *corrélation positive* entre les deux variables.

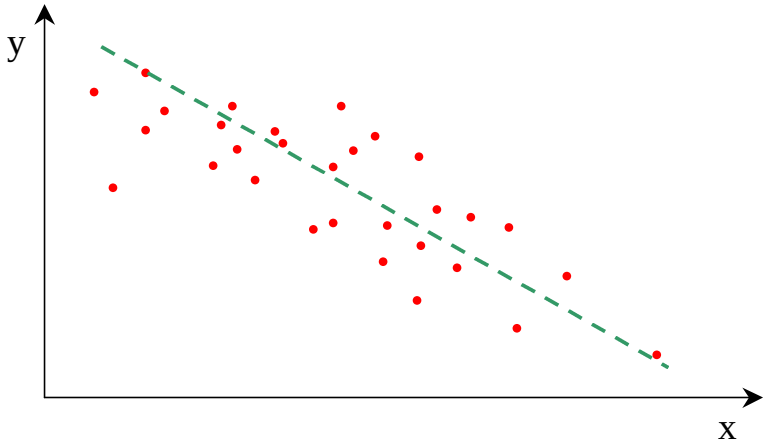
Si elle "descend", c'est une *corrélation négative*.

Si elle est "horizontale", ou si on ne peut pas décider, c'est qu'il y a *absence de corrélation*.

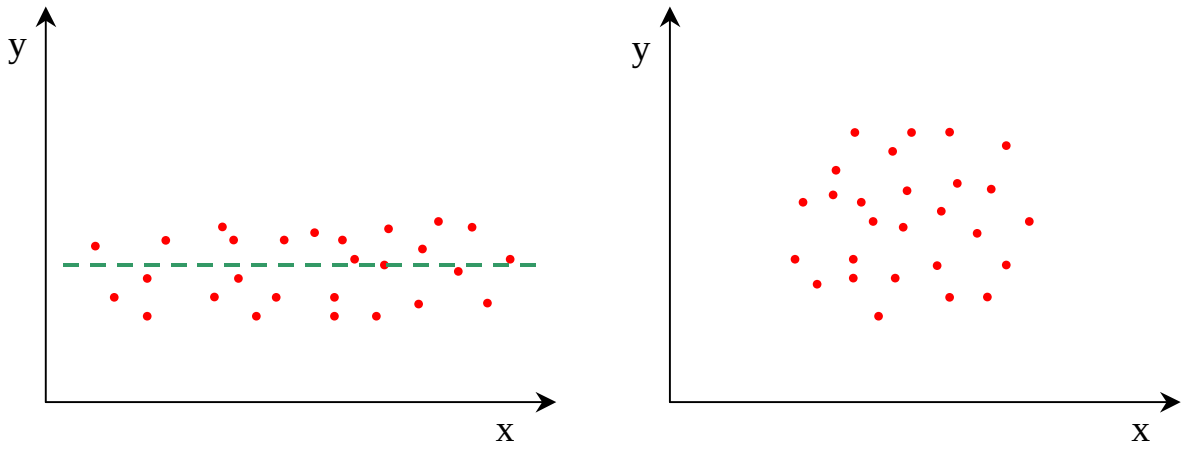
Corrélation positive:




Corrélation négative:

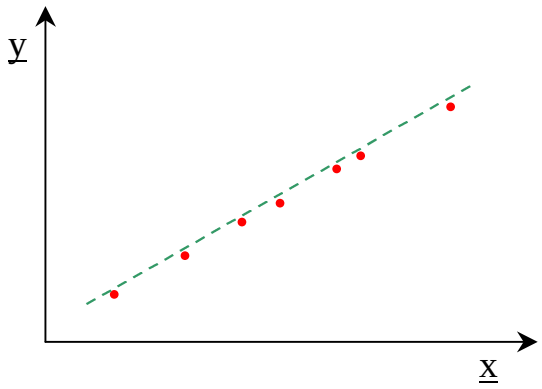



Absence de corrélation:

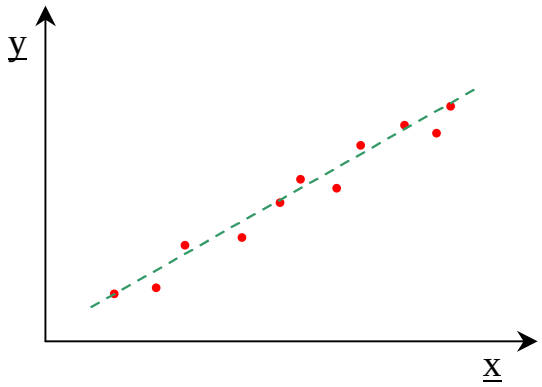


La *qualité* de la corrélation entre deux variables peut se mesure par la *dispersion* des points autour de la relation moyenne.

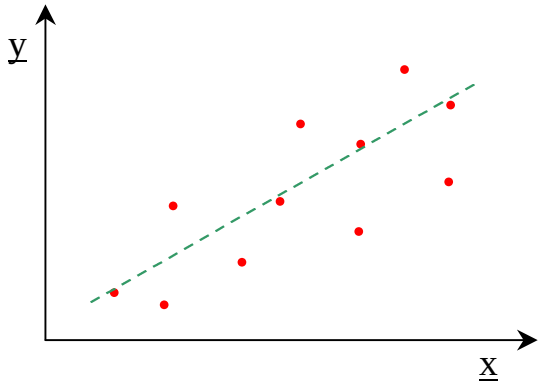
Corrélation parfaite:




Bonne corrélation (corrélation forte):

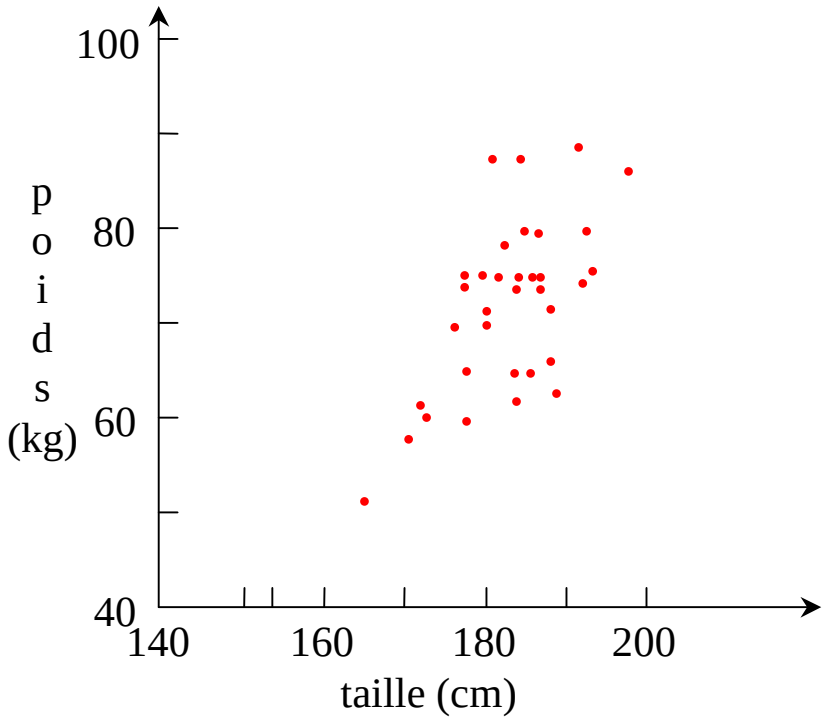



Mauvaise corrélation (corrélation faible):



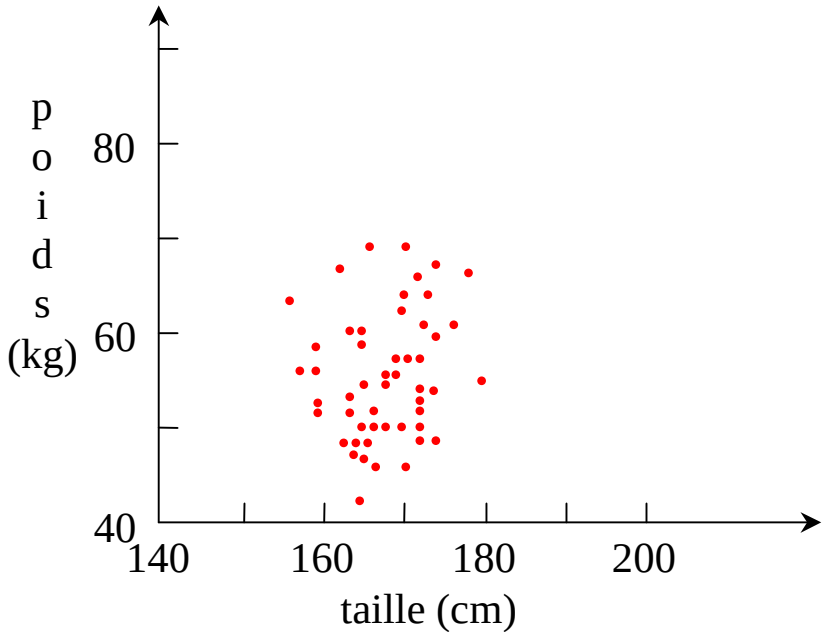
*Exemple:*

1. Corrélation entre le poids et la taille pour les garçons de 2ème candidature communication (1998).




On constate une augmentation du poids avec la taille (*corrélation positive*): les garçons les plus grands sont généralement les plus lourds.
Mais la dispersion des points est assez grande: *la corrélation est assez faible*.

2. Corrélation entre le poids et la taille pour les filles de 2ème candi. commu.



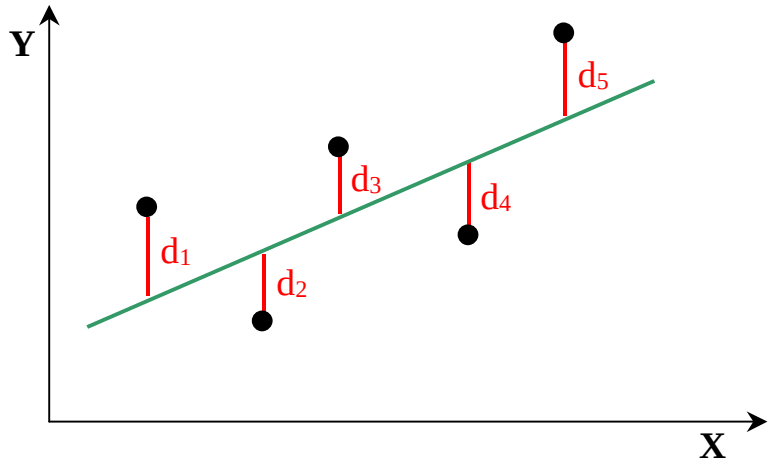

On ne constate pas de relation entre le poids et la taille (*absence de corrélation*): le poids des filles est indépendant de leur taille.
(Les filles les plus grandes sont donc les plus minces)

## 7.2. Méthode des moindres carrés

Si on se contente de tracer à main levée la droite qui "passe au mieux" par les points représentatifs, différentes personnes vont obtenir des résultats différents.

Il existe une méthode mathématique pour déterminer la "meilleure" droite: c'est la *méthode des moindres carrés*.

Elle consiste, dans sa version la plus simple, à trouver la droite qui minimise les carrés des écarts des points représentatifs à cette droite.




Trouver la droite telle que la somme des carrés des écarts $d_1$, $d_2$,… soit minimale:

$$\sum d^2 = \text{minimum}$$

Soit

$$Y = aX + b$$

l'équation de la droite cherchée (*droite de régression*)

Les coefficients *a* et *b* peuvent être calculés à partir des formules suivantes:

Pente:

$$a = \frac{(X_1 - \overline{X})(Y_1 - \overline{Y}) + (X_2 - \overline{X})(Y_2 - \overline{Y}) + \cdots + (X_n - \overline{X})(Y_n - \overline{Y})}{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}$$

ou:

$$a = \frac{\sum (X - \overline{X}).(Y - \overline{Y})}{\sum (X - \overline{X})^2}$$

Ordonnée à l'origine:

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

Rappels:

$$\overline{X} = \frac{1}{n}\sum X$$

$$\overline{Y} = \frac{1}{n}\sum Y$$

## 7.3. Coefficient de corrélation

Le signe de la pente *a* donne le *sens* de corrélation, mais pas sa *qualité*.

| | |
|---|---|
| $a > 0$ | corrélation positive |
| $a < 0$ | corrélation négative |
| $a = 0$ | pas de corrélation |

La qualité de la corrélation peut être mesurée par un *coefficient de corrélation r*

$$r = \frac{\sum (X - \overline{X}).(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2} \times \sqrt{\sum (Y - \overline{Y})^2}}$$

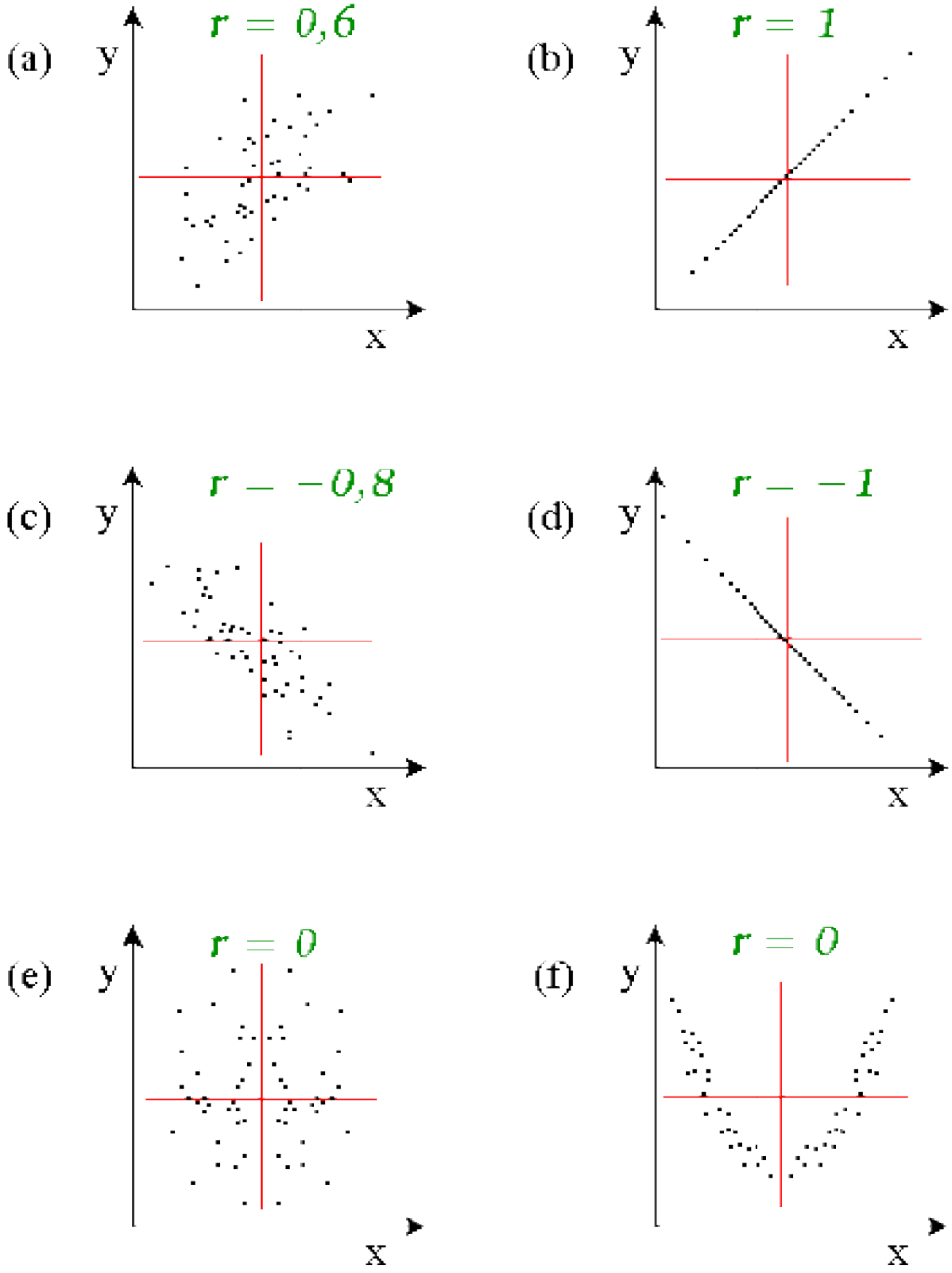
Le coefficient de corrélation est compris entre –1 et +1.

Plus il s'éloigne de zéro, meilleure est la corrélation

| | |
|---|---|
| $r = +1$ | corrélation positive parfaite |
| $r = -1$ | corrélation négative parfaite |
| $r = 0$ | absence totale de corrélation |

Quelques exemples de corrélation

(le coefficient de corrélation *r* est indiqué dans chaque cas)



*Exemples:*

1. Supposons un échantillon aléatoire de 4 firmes pharmaceutiques présentant les dépenses de recherche X et les profits Y suivants (en milliers de dollars):

| X | Y |
|---|---|
| 40 | 50 |
| 40 | 60 |
| 30 | 40 |
| 50 | 50 |

Trouvez la droite de régression et le coefficient de corrélation.

Calculons tout d'abord $\overline{X}$ et $\overline{Y}$:

$$\overline{X} = \frac{1}{n}\sum X = \frac{1}{4}(40+40+30+50) = \frac{160}{4} = 40$$
$$\overline{Y} = \frac{1}{n}\sum Y = \frac{1}{4}(50+60+40+50) = \frac{200}{4} = 50$$

Complétons le tableau suivant:

| $X$ | $Y$ | $X-\overline{X}$ | $Y-\overline{Y}$ | $(X-\overline{X})^2$ | $(Y-\overline{Y})^2$ | $(X-\overline{X}).(Y-\overline{Y})$ |
|---|---|---|---|---|---|---|
| 40 | 50 | 0 | 0 | 0 | 0 | 0 |
| 40 | 60 | 0 | +10 | 0 | +100 | 0 |
| 30 | 40 | −10 | −10 | +100 | +100 | +100 |
| 50 | 50 | +10 | 0 | +100 | 0 | 0 |

On a donc:

$$\sum (X-\overline{X})^2 = 200$$
$$\sum (Y-\overline{Y})^2 = 200$$
$$\sum (X-\overline{X}).(Y-\overline{Y}) = 100$$

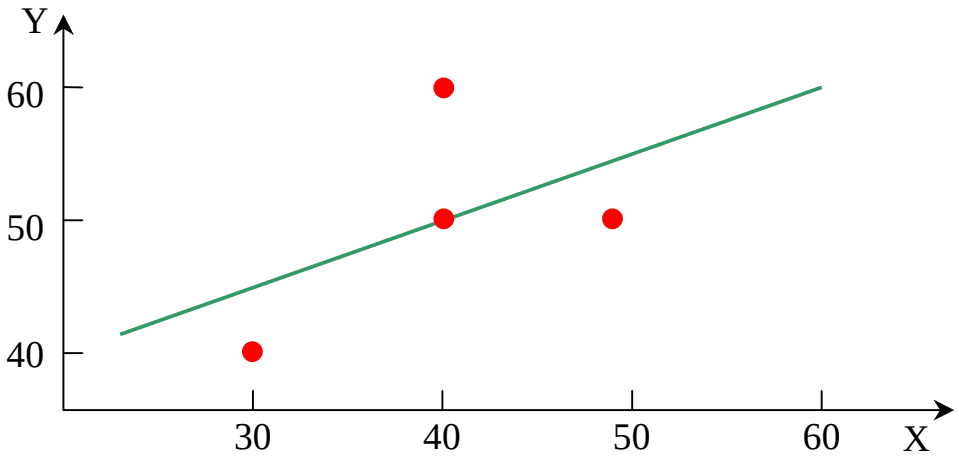
Les coefficients de la droite de régression sont:

$$a = \frac{\sum (X-\overline{X}).(Y-\overline{Y})}{\sum (X-\overline{X})^2} = \frac{100}{200} = 0{,}5$$
$$b = \overline{Y} - a.\overline{X} = 50 - 0{,}5 \times 40 = 50 - 20 = 30$$

Et le coefficient de corrélation:

$$r = \frac{\sum (X-\overline{X}).(Y-\overline{Y})}{\sqrt{\sum (X-\overline{X})^2} \times \sqrt{\sum (Y-\overline{Y})^2}} = \frac{100}{\sqrt{200} \times \sqrt{200}} = \frac{100}{200} = 0{,}5$$

La corrélation est *positive* et de *qualité moyenne*




2. La corrélation entre la taille (X) et le poids (Y) pour les garçons de 2ème candi. commu. donne les résultats suivants:

   (a) droite de régression $Y = aX + b$
   $a = 0{,}816$ $b = -77{,}0$

   (b) coefficient de corrélation
   $r = 0{,}61$

   la corrélation est donc positive, de qualité moyenne

3. De la même manière, pour les filles, on obtient:

   (a) droite de régression
   $a = 0{,}239$ $b = 16{,}6$

   (b) coefficient de corrélation
   $r = 0{,}20$

   la corrélation est positive (les filles les plus grandes tendent à être les plus lourdes), mais de très mauvaise qualité ($r$ proche de zéro).

*Remarques:*

1. Le coefficient de corrélation nous donne des informations sur l'existence d'une relation *linéaire* (sous forme d'une droite) entre les deux grandeurs considérées.

   Un coefficient de corrélation nul ne signifie pas l'absence de toute relation entre les deux grandeurs. Il peut exister une relation *non linéaire* entre elles.

   (cf. exemple (f) ci-dessus: la connaissance de X nous donne des informations sur la valeur de Y).

2. Il ne faut pas confondre *corrélation* et *relation causale*.

   Une bonne corrélation entre deux grandeurs peut révéler une relation de cause à effet entre elles, mais pas nécessairement.

   *Exemples:*

   1. Si on compare la durée de vie des individus à la quantité de médicaments pour le cœur qu'ils ont absorbée, on observera probablement une corrélation négative. Il serait imprudent de conclure que la prise de médicaments pour le cœur abrège la vie des individus…

      (en fait, dans ce cas, la corrélation est l'indice d'une cause commune: la maladie de cœur).

   2. Le soleil tire son énergie de réactions nucléaires transformant l'hydrogène en hélium. Notre société tire une bonne part de son énergie de la combustion du pétrole. Si on compare, année après année, la quantité d'hélium contenue dans le soleil au prix moyen du pétrole, on obtiendra une bonne corrélation positive, sans qu'il y ait la moindre relation de cause à effet, ni aucune cause commune.

   3. Depuis une dizaine d'années, la taille de mon fils cadet, né en 1989, est très bien corrélée avec la puissance de calcul des ordinateurs personnels. Cette excellente corrélation ne révèle bien évidemment aucune relation de cause à effet, ni cause commune.

   *L'existence d'une corrélation, aussi bonne soit elle, n'est jamais la preuve d'une relation de cause à effet.*