Calculer un coefficient de corrélation

Etudier la corrélation entre deux ou plusieurs variables, c'est étudier l'intensité de la liaison qui peut exister entre ces variables. La liaison recherchée est une relation dont la représentation graphique est une droite.
Une mesure de cette corrélation est obtenue par le calcul du **coefficient de corrélation linéaire de Bravais-Pearson**.

1. CALCUL DU COEFFICIENT DE CORRELATION

Par exemple, pour calculer le coefficient de corrélation entre deux séries de même longueur (cas typique : une régression), on suppose qu'on a les tableaux de valeurs suivants : $X(x_1,\ldots x_n)$ et $Y(y_1,\ldots,y_n)$ pour chacune des deux séries. Alors, pour connaître le coefficient de corrélation liant ces deux séries, on applique la formule suivante :

$$r_p = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$, ce qui donne : $$\boxed{r_p = \frac{\sum_{i=1}^{N}(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{N}(y_i-\bar{y})^2}}}$$

où $\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})\cdot(y_i-\bar{y})$ est la covariance entre X et Y et

où $\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}$ est l'écart-type de X et $\sigma_y = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y})^2}$ est l'écart-type de Y.

2. INTERPRETATION DU COEFFICIENT DE CORRELATION

Le coefficient de corrélation est compris entre -1 et +1. Plus le coefficient est proche des valeurs extrêmes -1 et 1, plus la corrélation entre les variables est forte ; on emploie simplement l'expression « fortement corrélées » pour qualifier les deux variables. Une corrélation égale à 0 signifie que les variables sont linéairement indépendantes.

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

Le coefficient de corrélation nous donne des informations sur **<u>l'existence d'une relation linéaire</u>** (sous forme d'une droite) entre 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.

Le coefficient de corrélation n'est **<u>pas sensible aux unités</u>** de chacune des variables. Ainsi, par exemple, le coefficient de corrélation linéaire entre l'âge et le poids d'un individu sera identique que l'âge soit mesuré en semaine, en mois ou en année(s).

En revanche, ce coefficient de corrélation est extrêmement **<u>sensible à la présence de valeurs aberrantes</u>** ou extrêmes dans notre ensemble de données (valeurs très éloignées de la majorité des autres, pouvant être considérées comme des exceptions).

Attention ! Il ne faut pas confondre corrélation et relation causale !

Une erreur courante est de croire qu'un coefficient de corrélation élevé induit une relation de causalité entre les deux phénomènes mesurés. En réalité, les deux phénomènes peuvent être corrélés à un même phénomène-source : une troisième variable non mesurée, et dont dépendent les deux autres. Une bonne corrélation entre deux grandeurs peut révéler une relation de cause à effet entre elles, mais pas nécessairement !

Exemples :

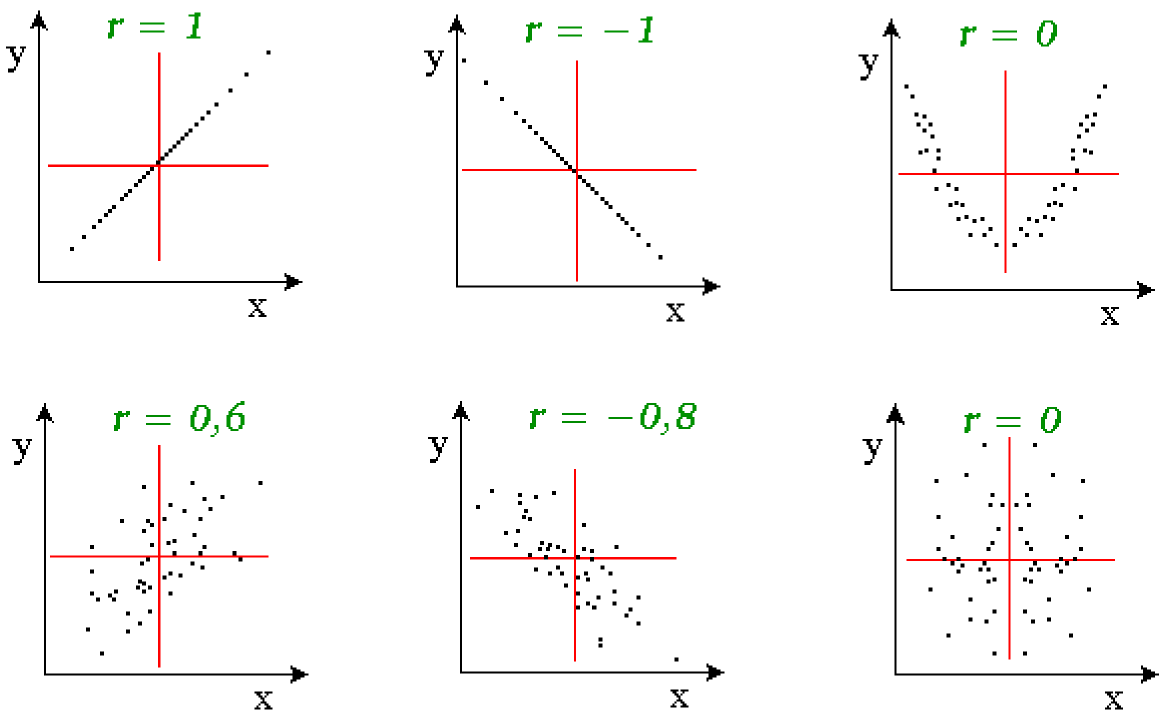
1. Si on compare la durée de vie des individus à la quantité de médicaments pour le coeur qu'ils ont absorbée, on observera probablement une corrélation négative. Il serait imprudent de conclure que la prise de médicaments pour le coeur abrège la vie des individus (en fait, dans ce cas, la corrélation est l'indice d'une cause commune: la maladie de coeur).

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 1999, 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 !

3. REPRESENTATION GRAPHIQUE DU COEFFICIENT DE CORRELATION



Sources : BEGUIN H., *Méthodes d'analyse géographique quantitative*, Litec, Paris, 1979
http://www.astro.ulg.ac.be/cours/magain/stat/
http://fr.wikipedia.org