Semaine 6 : Covariance et corrélation

1 La covariance entre X et Y

1.1 La covariance théorique

La covariance théorique
La covariance théorique entre deux variables aléatoires X et Y est une mesure d'association linéaire définie comme suit :

$$Cov(X,Y) \;=\; E\big[(X-\mu_X)\,(Y-\mu_Y)\big]$$

Remarques :

- Si $Cov(X,Y) > 0$, alors la relation est ...
- Si $Cov(X,Y) < 0$, alors la relation est ...
- Si $Cov(X,Y) = 0$, alors ...
- Si X et Y sont indépendantes, alors $Cov(X,Y) =$
- Unités :
- Valeurs possibles :

Propriétés de la covariance théorique

1) $Cov(X,X) =$

2) $Cov(X,Y) =$

3) $Cov(aX,Y) =$

4) $Cov(X+b,Y) =$

5) $Cov(X_1+X_2,Y) =$

6) $Var(X+Y) =$

7) $Var(X-Y) =$

1.2 La covariance échantillonnale

La covariance échantillonnale
On estime en général la covariance théorique par la covariance échantillonnale :

$$\begin{aligned}\widehat{Cov}(X,Y) &= \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1}\\ &= \frac{S_{XY}}{n-1}\\ &=\end{aligned}$$

2 La corrélation entre X et Y

2.1 La corrélation théorique

Coefficient de corrélation théorique : ρ
Pour régler le problème des unités de la covariance, on utilise plus souvent le coefficient de corrélation entre X et Y.

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$$

Ce coefficient est compris entre -1 et 1.

Unités :

2.2 La corrélation échantillonnale

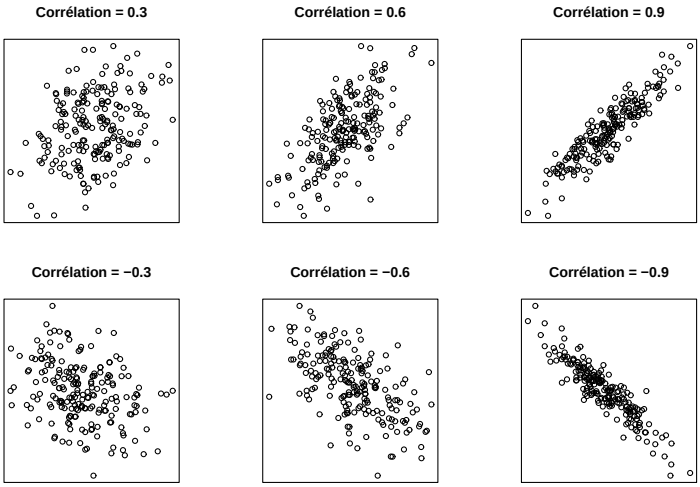
Coefficient de corrélation échantillonnal : r
On estime la corrélation par le coefficient de corrélation échantillonnal r :

$$r = \frac{\widehat{Cov}(X,Y)}{S_X\ S_Y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\ \cdot\ \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2}}$$

$$= \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} =$$

Propriétés de r

-
-
-
-



Questions

- Si $r = 1$, que vaut $\hat{\sigma}^2 = MSE$?

- Si $r = 0$, qu'est-ce que cela signifie ?

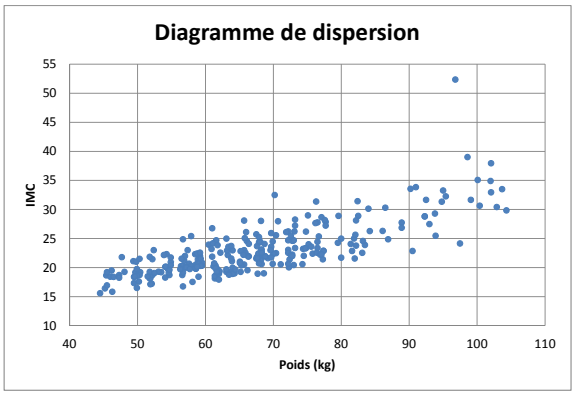
- Que vaut r si tous les points sont sur une droite de pente nulle ?

Causalité

Peut-on dire que plus une corrélation est forte entre 2 variables, plus cela indique la présence d'un lien causal entre elles ?

Exemple sur l'IMC

Que valent la covariance et le coefficient de corrélation échantillonnaux entre l'IMC et le poids selon l'étude chez les jeunes de 16-17 ans ?

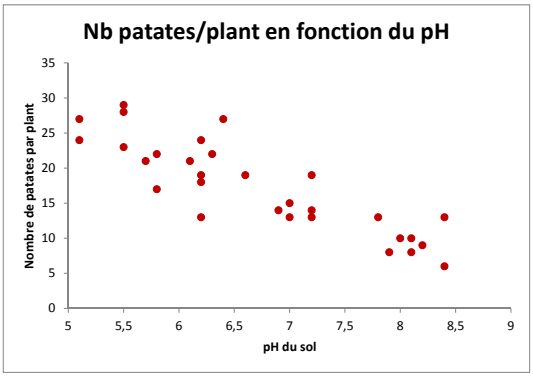


Rappels :

$n = 278, \quad S_{XX} = 49\,442, \quad S_{YY} = 5\,411, \quad S_{XY} = 12\,841$

2.3 La méthode du rectangle

Estimation de ρ par le rectangle



Source : *Visions*, Sciences naturelles, manuel de l'élève, volume 1, p. 17.

Calcul par la méthode du rectangle

Dans les livres du secondaire, on estime le coefficient de corrélation par la formule suivante :

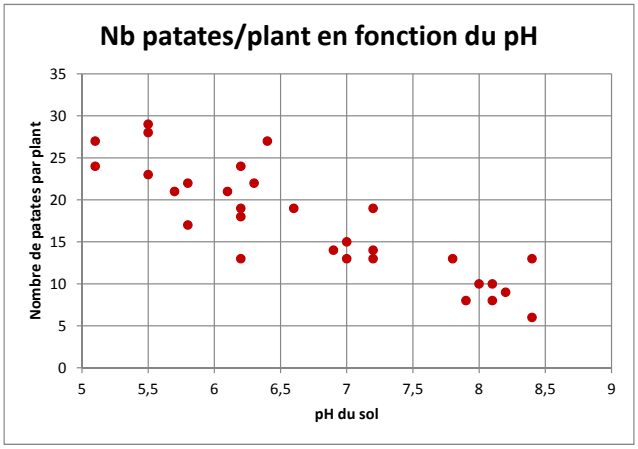
$r \approx$

2.4 Discrétisation de variables continues

Discrétisation de variables continues
On peut aussi décrire de façon qualitative la corrélation entre X et Y à partir du tableau de fréquences (à double entrée), où les valeurs sont regroupées en classes.

On dira qu'il y a corrélation si on observe ...

Exemple



Complétez le tableau de fréquences ci-dessous en vous basant sur le nuage de points.

Nb patates \ pH	[5, 6[	[6, 7[	[7, 8[	[8, 9[	Total
[5, 10[					
[10, 15[					
[15, 20[					
[10, 25[					
[25, 30[					
Total					

Que remarquez-vous ?

Sauriez-vous calculer une valeur approximative du coefficient de corrélation entre le pH et le nombre de patates par plant à partir de ce tableau ?