

Module 4 - Covariance et corrélation

Département de mathématiques et de statistique



UNIVERSITÉ
LAVAL

Distributions conjointes

Une expérience, plusieurs variables d'intérêt

On peut être intéressé **simultanément** par la valeur de plusieurs variables aléatoires.

Exemple : Le réservoir d'un barrage est alimenté par 2 rivières. Soit X_i = le volume d'eau déversé dans le réservoir par la rivière i en une journée ($i = 1, 2$).

Le niveau d'eau dans le réservoir dépend du volume d'eau apporté par les 2 rivières, et ces volumes ne sont **pas indépendants** les uns des autres.

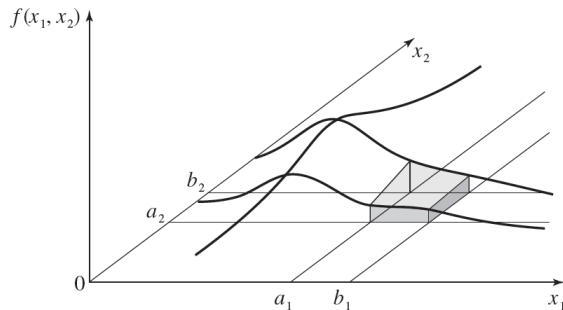
On voudra déterminer la **loi conjointe du vecteur aléatoire** $\mathbf{X} = (X_1, X_2)$.

Besoin de modèles conjoints

Quand les variables formant un vecteur aléatoire évoluent de manière indépendante, on peut utiliser la théorie précédente pour calculer des probabilités, moyennes, variances, etc.

Quand la valeur d'une variable contient de l'information sur la distribution des autres variables, on doit **étudier les probabilités de façon conjointe**.

Densité conjointe à 2 dimensions



Une fonction de densité conjointe, où $P(a_1 \leq X_1 \leq b_1 \text{ et } a_2 \leq X_2 \leq b_2)$ correspond au volume ombré

Densité conjointe à 2 dimensions

Dans ce cours, nous ne ferons aucun calcul à partir de densités conjointes.

Nous nous intéresserons seulement au lien linéaire entre deux variables aléatoires.

Nous introduirons la covariance et la corrélation.

Covariance et corrélation

Covariance entre deux variables

Soit X et Y deux variables aléatoires. On peut mesurer leur association linéaire.

Covariance

Soit deux variables X et Y d'espérances respectives μ_X et μ_Y .

La **covariance** entre X et Y est définie par

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Covariance entre deux variables

Soit X et Y deux variables aléatoires. On peut mesurer leur association linéaire.

Covariance

Soit deux variables X et Y d'espérances respectives μ_X et μ_Y .

La **covariance** entre X et Y est définie par

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Covariance = indicateur de la **force du lien linéaire** qui unit deux variables.

Le signe de la covariance (+ ou -) nous dit si les variables ont tendance à prendre de fortes valeurs par rapport à leur moyenne en même temps ou non.

Covariance entre deux variables

Soit X et Y deux variables aléatoires. On peut mesurer leur association linéaire.

Covariance

Soit deux variables X et Y d'espérances respectives μ_X et μ_Y .

La **covariance** entre X et Y est définie par

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Covariance = indicateur de la **force du lien linéaire** qui unit deux variables.

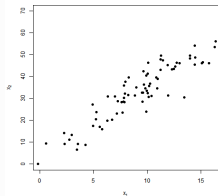
Le signe de la covariance (+ ou -) nous dit si les variables ont tendance à prendre de fortes valeurs par rapport à leur moyenne en même temps ou non.

Attention : $X \perp Y \rightarrow \text{Cov}(X, Y) = 0$, mais $\text{Cov}(X, Y) = 0 \nrightarrow X \perp Y$

Interprétation de la covariance

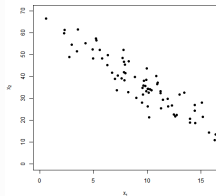
Cov. positive

X et Y prennent de fortes ou faibles valeurs par rapport à leur moyenne en même temps



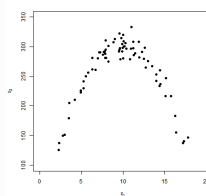
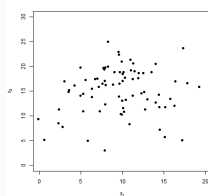
Cov. négative

X prend de fortes valeurs quand Y en prend de faibles par rapport à leur moyenne et vice versa



Covariance nulle

Pas de tendance linéaire dominante.



Propriétés de la covariance

- $Cov(X, X) = V(X)$
- $Cov(aX, bY) = abCov(X, Y)$
- $Cov(X + c, Y + d) = Cov(X, Y)$
- $V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$
- $V(X - Y) = V(X) + V(Y) - 2Cov(X, Y)$
- $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X, Y)$
- $V(a_0 + a_1X_1 + \dots + a_kX_k) = a_1^2\sigma_1^2 + \dots + a_k^2\sigma_k^2 + 2 \sum_{i < i'} a_i a_{i'} Cov(X_i, X_{i'})$

QUIZ!

- Que peut-on dire de $Cov(5, X)$?
- Vrai ou Faux : $Cov(X, 3Y) = Cov(Y, 3X)$?
- Si X est mesuré en $^{\circ}\text{C}$ et Y en secondes, quelles sont les unités de la covariance?

Corrélation

Soit les variables aléatoires W, X, Y et Z . Nous ne pouvons pas directement comparer $Cov(W, X)$ et $Cov(Y, Z)$ pour savoir quelle paire de variables a la plus forte association linéaire, parce que la covariance dépend de l'échelle des variables.

Coefficient de corrélation

Soit deux variables X et Y de variances respectives σ_X^2 et σ_Y^2 .

Le **coefficient de corrélation** entre X et Y est défini par

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{Cov(X, Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}}.$$

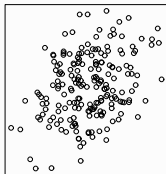
Plus les valeurs de X et Y sont fortement corrélées, plus la valeur de $\rho_{X,Y}$ sera proche de -1 ou 1 .

Pour alléger l'écriture, si seulement deux variables sont présentes, on écrit ρ .

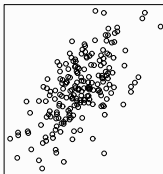
Quelques valeurs de corrélations

Voici quelques nuages de points échantillonnés à partir de distributions conjointes ayant les corrélations théoriques précisées.

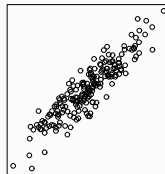
Corrélation = 0.3



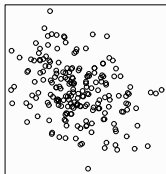
Corrélation = 0.6



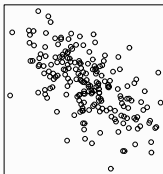
Corrélation = 0.9



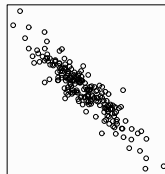
Corrélation = -0.3



Corrélation = -0.6



Corrélation = -0.9



QUIZ!

Sauriez-vous tracer un nuage de points pour lequel...

- $\rho_{X,Y} = 0$?
- $\rho_{X,Y} = -1$?

Exemple 1

Soit X , Y et W trois variables avec

- $E(X) = 1$, $E(Y) = 0$, $E(W) = 0$
- $V(X) = 16$, $V(Y) = 4$, $V(W) = 25$
- $Cov(X, Y) = -2$, et W est indépendante de X et de Y .

Calculez les quantités suivantes.

(i) $E(3 + 4X - Y + W)$

Exemple 1 (suite)

(ii) $V(3 + 4X - Y + W)$

(iii) $\text{corr}(X, Y)$.

Réponses aux exemples

1. $E(3 + 4X - Y + W) = 7$
 $V(3 + 4X - Y + W) = 301$
 $corr(X, Y) = -0.25.$