# Séance 10 Séries statistiques à deux variables

**Notions abordées**
Analyse bivariée. Tableaux croisés.
Covariance et coefficient de corrélation

## I. Variables indépendantes

En statistiques comme en probabilités, on est souvent amené à considérer plusieurs variables et à mesurer la dépendance ou l'indépendance de deux d'entre elles. Ainsi, deux variables X et Y sont **indépendantes** si pour tous les intervalles A et B de $R$, l'égalité suivante est vérifiée :

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B)$$

Dans le cas discret, ceci est équivalent à dire que pour toutes les valeurs $x$ et $y$, l'égalité suivante est vérifiée :

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$$

Dans le cas continu, si l'on note $f_X$ et $f_Y$ les densités respectives de X et de Y, on obtient :

$$P(X \in A, Y \in B) = \int_A f_X(x)dx \bullet \int_B f_Y(y)dy$$

### Exemple 1

On dispose de 2 dés à six faces équilibrés : un bleu et un rouge. On désigne par X la variable qui donne le résultat du dé bleu et Y le résultat du dé rouge. Quels que soient les nombres x et y compris entre 1 et 6, on obtient :

$$P(X = x, Y = y) = P(X = x) \cdot P(Y = y) = \frac{1}{36}$$

Les variables X et Y sont indépendantes : le résultat du premier dé ne dépend pas du résultat du second.

### Exemple 2

On dispose d'une urne contenant 1 boule numérotée 1 et 4 boules numérotées 2. On effectue deux tirages sans remise. On désigne par X la variable qui donne le résultat premier tirage et Y le résultat du deuxième tirage.

$$P(X = 1, Y = 1) = 0$$

$$\text{Pourtant} \quad P(X = 1) \cdot P(Y = 1) = \frac{1}{5} \bullet \frac{4}{5} \times \frac{1}{4} = \frac{1}{25}$$

Les variables X et Y sont dépendantes : le résultat du premier tirage influence le résultat du second.

Dans le cas de 2 variables dépendantes, on peut alors se demander comment les traiter et comment évaluer dans un deuxième temps leur lien de dépendance.

## II. Variables dépendantes

On ne traitera dans la suite du cours que du cas discret.
Soient X et Y deux variables discrètes. La loi du couple (X, Y) est définie par l'ensemble des probabilités

$$P(X = x,\ Y = y)$$

Pour toutes les valeurs possibles $x$ et $y$.
Notons $D_X$ et $D_Y$ l'ensemble des valeurs possibles respectivement pour X et pour Y. On peut retrouver les lois de chacune des variables à partir de la loi du couple.
Soit $x \in D_X$

$$P(X = x) = P\left(X = x \text{ et } Y \in D_Y\right) = \sum_{y \in D_Y} P(X = x, Y = y)$$

De la même façon, pour $y \in D_Y$

$$P(Y = y) = P\left(X \in D_X \text{ et } Y = y\right) = \sum_{x \in D_X} P(X = x, Y = y)$$

Ainsi, on peut retrouver les lois marginales à partir de la loi du couple mais l'inverse n'est pas possible.

Dans le cas où les variables discrètes prennent un petit nombre de valeurs, on utilise souvent un tableau pour écrire la loi du couple. Dans le tableau ci-dessous, on suppose que X peut prendre $n$ valeurs et Y $p$ valeurs.

| $Y \backslash X$ | $x_1$ | $x_2$ | … | $x_n$ | *Total* |
|---|---|---|---|---|---|
| $y_1$ | $P(X = x_1, Y = y_1)$ | $P(X = x_2, Y = y_1)$ | … | $P(X = x_n, Y = y_1)$ | $P(Y = y_1)$ |
| $y_2$ | $P(X = x_1, Y = y_2)$ | $P(X = x_2, Y = y_2)$ | … | $P(X = x_n, Y = y_2)$ | $P(Y = y_2)$ |
| … | … | … | … | … | … |
| $y_p$ | $P(X = x_1, Y = y_p)$ | $P(X = x_2, Y = y_p)$ | … | $P(X = x_n, Y = y_p)$ | $P(Y = y_p)$ |
| *Total* | $P(X = x_1)$ | $P(X = x_2)$ | … | $P(X = x_n)$ | *1* |

**Exemple**

On lance trois fois une pièce truquée. La probabilité de tomber sur « pile » est de 2/3. Soit X le nombre de « face » obtenu dans les deux premiers jets et Y le nombre de « face » obtenu dans les deux derniers jets. La loi de (X, Y) est donnée par

| $Y \backslash X$ | 0 | 1 | 2 | $P(Y = y)$ |
|---|---|---|---|---|
| 0 | $\left(\frac{2}{3}\right)^3 = \frac{8}{27}$ | $\frac{1}{3} \bullet \left(\frac{2}{3}\right)^2 = \frac{4}{27}$ | *0* | $\frac{12}{27} = \frac{4}{9}$ |
| 1 | $\frac{1}{3} \bullet \left(\frac{2}{3}\right)^2 = \frac{4}{27}$ | $\frac{1}{3} \bullet \left(\frac{2}{3}\right)^2 + \frac{2}{3} \bullet \left(\frac{1}{3}\right)^2 = \frac{6}{27}$ | $\frac{2}{3} \bullet \left(\frac{1}{3}\right)^2 = \frac{2}{27}$ | $\frac{12}{27} = \frac{4}{9}$ |
| 2 | 0 | $\frac{2}{3} \bullet \left(\frac{1}{3}\right)^2 = \frac{2}{27}$ | $\left(\frac{1}{3}\right)^3 = \frac{1}{27}$ | $\frac{3}{27} = \frac{1}{9}$ |
| $P(X = x)$ | $\frac{12}{27} = \frac{4}{9}$ | $\frac{12}{27} = \frac{4}{9}$ | $\frac{3}{27} = \frac{1}{9}$ | *1* |

## III. Covariance de deux variables

Dans ce paragraphe, on va se donner deux nouveaux outils pour mesurer la dépendance de 2 variables : la covariance et la corrélation.

La **covariance** de deux variables X et Y est donnée par la formule

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X) \cdot E(Y)$$

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

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

L'espérance E(XY) est calculée à partir de la loi (X, Y). Dans le cas discret, lorsque la somme a un sens, on a

$$E(XY) = \sum_{x, y} xy \cdot P(X = x, Y = y)$$

**Remarque**

Si les deux variables X et Y sont **indépendantes**, on peut en déduire la formule suivante

$$E(XY) = E(X) \cdot E(Y)$$

**Propriété 1**

Soient X et Y deux variables aléatoires

$$V(X + Y) = V(X) + V(Y) + 2 \cdot Cov(X, Y)$$

**Preuve**

$V(X + Y) = E\left[(X + Y - E(X + Y))^2\right] = E\left[(X - E(X) + Y - E(Y))^2\right]$

$V(X + Y) = E\left[(X - E(X))^2\right] + E\left[(Y - E(Y))^2\right] + 2 \bullet E[(X - E(X))(Y - E(Y))]$

$V(X + Y) = Var(X) + Var(Y) + 2 \bullet E[XY - X \cdot E(Y) - E(X) \cdot Y + E(X) \cdot E(Y)]$

$V(X + Y) = Var(X) + Var(Y) + 2 \cdot (E(XY) - E(X \cdot E(Y)) - E(E(X) \cdot Y) + E(X) \cdot E(Y))$

Or

$E(X \cdot E(Y)) = E(E(X) \cdot Y) = E(X) \cdot E(Y)$
Donc
$Var(X + Y) = Var(X) + Var(Y) + 2 \cdot (E(XY) - E(X) \cdot E(Y))$
$Var(X + Y) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$

**Propriété 2**
Si X et Y sont deux variables aléatoires **indépendantes**, alors

$$Cov(X, Y) = 0$$

Et donc

$$V(X + Y) = V(X) + V(Y)$$

Attention, la réciproque est fausse, si la covariance est nulle, on ne peut pas savoir si les variables sont dépendantes ou non. Par contre, si la covariance n'est pas nulle, les variables sont dépendantes.

Voici un contre-exemple. On lance un dé équilibré à 6 faces et on jette une pièce de monnaie équilibrée dont les faces sont marquées -1 et 1. On note X la variable aléatoire qui donne le résultat du dé, Y la variable aléatoire qui désigne le nombre inscrit sur la face visible de la pièce. On définit la variable aléatoire $Z = XY$. X et Y sont clairement indépendantes donc $X^2$ et Y aussi mais Z et X sont clairement dépendantes. Ainsi,

$$cov(X, Z) = E(XZ) - E(X) \bullet E(Z)_{\backsim=0} = E(XZ) = E\left(X^2 Y\right) = E\left(X^2\right) \bullet E(Y)_{\backsim=0} = 0$$

**Exemple**
Une étude médicale sur l'effet du tabac est menée dans un hôpital. Les 2278 patients sont divisés en deux groupes : ceux atteints d'un cancer pulmonaire ($X = 1$) et les autres ($X = 0$). Les membres de chaque groupe sont ensuite répartis selon le nombre Y de paquets de cigarettes fumées par jour.

| *Cancer pulmonaire* | *Nombre de paquets de cigarettes* | | | | | *Total* |
|---|---|---|---|---|---|---|
| | *0* | *1* | *2* | *3* | *4* | |
| *0* | *1247* | *492* | *319* | *58* | *9* | *2125* |
| *1* | *66* | *50* | *28* | *6* | *3* | *153* |
| *Total* | *1313* | *542* | *347* | *64* | *12* | *2278* |

On souhaite étudier l'association entre cancer pulmonaire et la consommation de cigarette en calculant la covariance. La proportion de personnes atteintes d'un cancer pulmonaire est **6,72%**. Le nombre moyen de paquets de cigarettes consommés est **0,65**. On obtient

$$Cov(X, Y) = \frac{1\times1\times50+1\times2\times28+1\times3\times6+1\times4\times3}{2278} - 0,0672\times0,65 = 0,02$$

La covariance est positive, le résultat indique qu'il y a un lien positif entre la déclaration du cancer et la consommation de cigarettes (plus on consomme de cigarettes, plus le risque de cancer est grand).

Contrairement à la covariance, la corrélation ne dépend pas des unités de mesure des variables.

**Propriété 3**
**$\rho(X, Y)$ est comprise entre $-1$ et 1.**
**Plus $|\rho(X, Y)|$ est proche de 1, plus les variables sont dépendantes.**
**Si $|\rho(X, Y)| = 1$, alors il existe des réels a et b tels que $aX + bY = 0$. Les variables sont entièrement liées.**

# Séance 11 Régression

**Notions abordées**
Nuage de points
Régression affine par la méthode des moindres carrés ordinaires
Coefficient de détermination $R^2$

## I. Notion d'ajustement

Dans toute la suite de ce cours, X et Y sont des variables statistiques quantitatives pouvant prendre $n$ valeurs respectivement $x_1, x_2, \ldots, x_n$ et $y_1, y_2, \ldots, y_n$ avec la probabilité $1/n$.

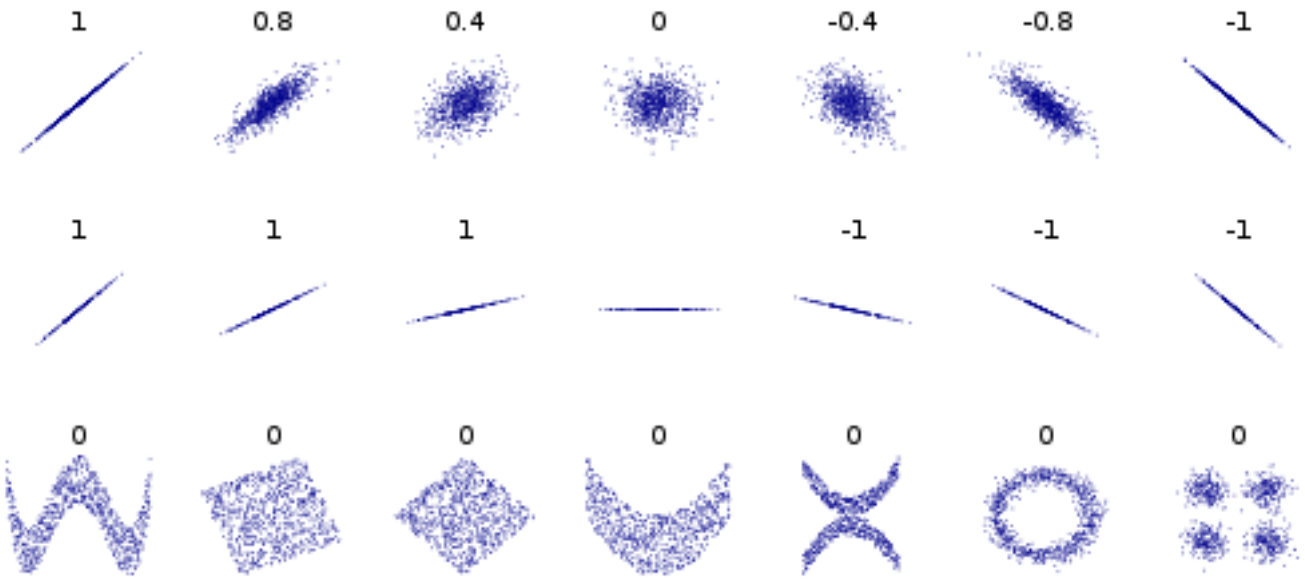
Le but de ce chapitre est d'étudier le lien qu'il existe entre X et Y et voir s'il existe une fonction $f$ qui modélise au mieux ce lien. On parle alors de **fonction d'ajustement** ou **régression**. Elle permet d'interpoler la valeur de l'une des variables connaissant l'autre et d'extrapoler certaines valeurs.
Par exemple, si la variable X désigne le temps et Y une quantité qui fluctue avec le temps (volume de ventes, production,…), on peut imaginer, à l'aide de cette fonction, l'évolution de la variable Y dans un temps futur.

On a vu que la corrélation permettait de définir un lien de dépendance (linéaire) entre deux variables. Cependant, ce lien n'est pas toujours linéaire. On peut avoir une dépendance très forte ($Y = X^2$ ou $Y = e^X$ par exemple) et une corrélation nulle.
Par la suite, on s'intéressera particulièrement aux variables présentant une dépendance linéaire importante (lorsque la corrélation se rapproche de $-1$ ou de 1).

### Remarque

On rappelle que la corrélation entre 2 variables n'implique pas nécessairement un lien de cause à effet.
Pour exemple, en 1999, l'université de Pennsylvanie publie une étude qui affirme que les jeunes enfants qui dorment avec une veilleuse ont plus de chance de devenir myopes plus tard. Peu après, l'université de l'Ohio montre un lien important entre la myopie des parents et celle de leurs enfants et remarque que les parents myopes ont plus tendance à laisser une lumière allumée pour leurs enfants durant la nuit…

Pour représenter les deux variables X et Y simultanément, on peut réaliser un **diagramme de dispersion** où les points représentés sont les points de coordonnées $(x_i; y_i)$. Selon ce diagramme, on peut prédire la relation entre X et Y.




*(source Wikipédia)*

Dans les cas où le diagramme de dispersion entre deux grandeurs montre l'existence d'une relation linéaire, on peut déterminer l'équation de la droite qui décrit cette relation.
Pour cela, on utilise en général le critère des **moindres carrés**. Selon ce critère, on cherche à **minimiser la somme des carrés des écarts entre les valeurs estimées et les valeurs observées** de la variable Y, autrement dit, on cherche à minimiser

$$E\left(Y - \widetilde{Y}\right)^2$$

où

$$\widetilde{Y} = aX + b$$

désigne la variable extrapolée. Par la suite, l'équation de la droite de régression sera notée

$$y = ax + b$$

## II. Propriétés de la droite de régression

On dispose de $n$ points de coordonnées $(x_i; y_i)$. $\overline{y_i}$ désigne l'ordonnée du point d'abscisse $x_i$ sur la droite de régression. Par définition, la somme

$$\sum_{i=1}^{n} (y_i - \overline{y_i})^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

est **minimale**.
De plus,

$$\sum_{i=1}^{n} \left(y_i - \overline{y_i}\right) = 0$$

i.e. les écarts « positifs » sont compensés par des écarts « négatifs » équivalents.

**Calcul des coefficients** $a$ **et** $b$

$$a = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{n.\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i) \times (\sum_{i=1}^{n} y_i)}{n.\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{Cov(X,Y)}{V(X)}$$

Et

$$b = \overline{y} - a.\overline{x}$$

$\overline{x}$ désigne la moyenne des $x_i$ et $\overline{y}$ désigne la moyenne des $y_i$.

**Remarque**
$\overline{y} = E(Y)$
$\overline{x} = E(X)$
Pour justifier la régression, on utilise le **coefficient de détermination** $R^2$ qui correspond, dans le cas d'un ajustement linéaire, au carré du coefficient de corrélation.

$$R^2 = \frac{\left[\left(x_1 - \overline{x}\right)\left(y_1 - \overline{y}\right) + \ldots + \left(x_n - \overline{x}\right)\left(y_n - \overline{y}\right)\right]^2}{\left[\left(x_1 - \overline{x}\right)^2 + \ldots + \left(x_n - \overline{x}\right)^2\right]\left[\left(y_1 - \overline{y}\right)^2 + \ldots + \left(y_n - \overline{y}\right)^2\right]} = \rho^2(X, Y)$$

Ce coefficient est situé entre 0 et 1.
Plus il est proche de 1, meilleure est la régression.
Plus il est proche de 0, plus la régression est mauvaise.

**Exemple**

| $x_i$ | $y_i$ | $x_i.y_i$ | $x_i^2$ | $y_i^2$ | $(x_i - \overline{x})(y_i - \overline{y})$ | $(x_i - \overline{x})^2$ | $(y_i - \overline{y})^2$ |
|---|---|---|---|---|---|---|---|
| **1** | 7 | 7 | 1 | 49 | − 12 | 16 | 9 |
| **2** | 6 | 12 | 4 | 36 | − 6 | 9 | 4 |
| **3** | 7 | 21 | 9 | 49 | − 6 | 4 | 9 |
| **4** | 4 | 16 | 16 | 16 | 0 | 1 | 0 |
| **5** | 3 | 15 | 25 | 9 | 0 | 0 | 1 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **6** | 6 | 36 | 36 | 36 | 2 | 1 | 4 |
| **7** | 0 | 0 | 49 | 0 | − 8 | 4 | 16 |
| **8** | 2 | 16 | 64 | 4 | − 6 | 9 | 4 |
| **9** | 1 | 9 | 81 | 1 | − 12 | 16 | 9 |
| **45** | 36 | 132 | 285 | 200 | − 48 | 60 | 56 |

$\overline{x} = 45:9 = 5$ et $\overline{y} = 36:9 = 4$

On en déduit les coefficients $a$ et $b$ de la droite de régression :

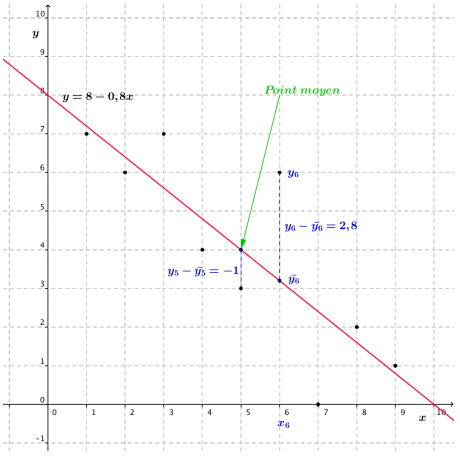
$b = \frac{-48}{60} = \frac{-9\times132-45\times36}{9\times285-45^2} = -0,8$

et $a = \overline{y} - b.\overline{x} = 4 + 0,8\times5 = 8$

La droite de régression admet donc l'équation suivante : $y = 8 - 0,8x$

Le coefficient de corrélation est :

$R^2 = \frac{[-48]^2}{60\times56} \approx 0,68571$




**Remarque**

La première formule pour b donne le résultat plus aisément à condition que les moyennes $\overline{x}$ et $\overline{y}$ soient des entiers.

## Comment obtenir les formules pour $a$ et $b$ ?

Considérons la fonction $f$ définie sur $R^2$ par

$$f(x;y) = \sum_{i=1}^{n} (y_i - xx_i - y)^2$$

Par hypothèse, la fonction $f$ admet un minimum en $(a;b)$. On en déduit que les dérivées partielles de $f$ en $(a\,;b)$ par rapport à $x$ et à $y$ sont nulles.

$$\frac{\partial f}{\partial x}(a;b) = -2\sum_{i=1}^{n} x_i\left(y_i - ax_i - b\right) = 0 \ (1)$$

$$\frac{\partial f}{\partial y}(a;b) = -2\sum_{i=1}^{n}\left(y_i - ax_i - b\right) = 0 \ (2)$$

De la relation (2), on déduit

$$\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - nb = 0$$

En remarquant que

$$\sum_{i=1}^{n} y_i = n\overline{y}$$

Et

$$\sum_{i=1}^{n} x_i = n\overline{x}$$

on peut affirmer que

$$n\overline{y} - an\overline{x} - nb = 0$$

En factorisant par $n$, on obtient $\overline{y} - a\overline{x} - b = 0$ soit $b = \overline{y} - a\overline{x}$

En remplaçant $b$ par sa valeur dans les égalités (1) et (2), on obtient

$$\sum_{i=1}^{n} x_i\left[(y_i - \overline{y}) - a(x_i - \overline{x})\right] = 0 \ (3)$$

$$\sum_{i=1}^{n}\left[(y_i - \overline{y}) - a(x_i - \overline{x})\right] = 0 \ (4)$$

En multipliant l'égalité (4) par $\overline{x}$ et retranchant membre à membre l'égalité (4) à l'égalité (3), on obtient

$$\sum_{i=1}^{n}(x_i - \overline{x})\left[(y_i - \overline{y}) - a(x_i - \overline{x})\right] = 0$$

On en déduit ainsi la valeur de $a$

$$a = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{n\times Cov(XY)}{n\times V(X)} = \frac{n^2(E(XY)-E(X)E(Y))}{n^2\left(E\left(X^2\right)-E^2(X)\right)} = \frac{n\sum_{i=1}^{n}x_i.y_i-(\sum_{i=1}^{n}x_i).(\sum_{i=1}^{n}y_i)}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}$$

après développement.

## Point moyen

**La droite de régression passe par le point moyen $(\overline{x}; \overline{y})$.**
Dans l'exemple précédent, elle passe par le point (5; 4).