Chapitre 3: Séries statistiques à deux variables .

L'analyse bivariée étudie des populations suivant deux caractères statistiques X et Y .Par exemple, on peut étudier un ensemble de salariés selon l'âge et le sexe, ou un ensemble d'individus selon leur poids et leur taille.

3-1Représentation des séries statistiques à deux variables

Les séries statistiques à deux variables peuvent être présentées de deux façons :

Présentation 1

On rassemble les données comme dans le tableau suivant

Caractère $_x$	x_1	x_2	...	x_n
Caractère $_y$	y_1	y_2	...	y_n

La représentation graphique de ce tableau est appelée le nuage de points.

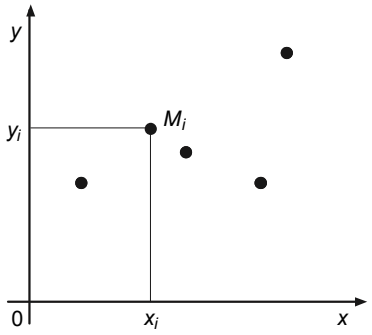


Figure 1: Représentation sous forme de nuage de points

Présentation 2

Soit La variable statistique $Z=(X,Y)$.

Soient $x_1, ..., x_k$ et $y_1,...,y_l$ les valeurs prises respectivement par X et Y .

Dans ce cas, nous définissons les valeurs de Z comme suite :

Pour $i \in \{1,....,k\}$,et pour $j \in \{1,....,l\}$ $:z_{ij}:=(x_i,y_j)$.

La variable statistique Z prend $k \times l$ valeurs.

Lors de cette étude, nous avons le tableau de contingence suivant :

X \ Y	$[b_1,b_2[$ **ou** y_1	⋯	$[b_l,b_{l+1}[$ **ou** yl	**Marginale** à**X**
$[L_1,L_2[$ **ou** x_1	$n11$ **ou**f_{11}	⋯	$n1l$ **ou**$f1l$	$n1\bullet$ **ou**$f_1\bullet$
$[L_2,L_3[$ **ou** x_2	$n21$ **ou**f_{21}	⋯	$n2l$ **ou**$f2l$	$n2\bullet$ **ou**$f_2\bullet$
$[L_3,L_4[$ **ou** x_3	$n31$ **ou**f_{31}	⋯	$n3l$ **ou**$f3l$	$n3\bullet$ **ou**$f_3\bullet$
…	…	…	…	…
$[L_k,Lk+1[$ **ou** x_k	n_{k1} **ou**f_{k1}	⋯	n_{kl} **ou**f_{kl}	$n_k\bullet$ **ou**$f_k\bullet$
Marginale à **Y**	$n\bullet 1$ ou$f\bullet_1$	⋯	$n\bullet l$ ou$f\bullet l$	N ou 1

Notations :

n_{ij} : l'effectif du couple $(xi; yj)$.
$n_i\bullet$: L'effectif marginal de x_i *est donné par* $ni\cdot = \sum_{j=1}^{k} n_{ij}$
$n_{\bullet j}$: L'effectif marginal de yi *est donné par* $n\cdot j = \sum_{i=1}^{l} n_{ij}$

f_{ij} :la fréquence du coulpe (x_i,y_i), est donnée par $f_{ij}=n_{ij}/N$

$f_{i\bullet}$: la fréquence marginale de x_i *est donné par* $f_{i\bullet} = n_{i\bullet}/N$.

$f_{\bullet j}$:la fréquence marginale de y_i *est donné par* $f_{\bullet j} = n_{\bullet j}/N$.

Remarques

$$\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = n_{11} + n_{12} + n_{13} + ... + n_{1l}$$

$$+n_{21} + n_{22} + n_{23} + ... + n_{2l}$$

$$\ddots \quad \ddots \quad \ddots \quad \ddots$$

$$+n_{k1} + n_{k2} + n_{k3} + ... + n_{kl}.$$

$$\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = 1.$$

2. Distributions marginales

Les k couples $(x_i, n_{i\bullet})$ forment la *distribution marginale* de la *variable X*.
Les l couples $(y_j, n_{\bullet j})$ forment la *distribution marginale* de la *variable Y*

Caractéristique des séries marginales

Les moyennes

Dans le cas d'une variable statistique à deux dimensions X et Y , les moyennes sont données respectivement par

$$\overline{x} := \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}x_i = \sum_{i=1}^{k} f_{i\bullet}x_i \quad \text{(moyenne } de\ X\text{)},$$

et

$$\overline{y} := \frac{1}{N}\sum_{j=1}^{l} n_{\bullet j}y_j = \sum_{j=1}^{l} f_{\bullet j}y_j \quad \text{(moyenne } de\ Y\text{)}.$$

Les variances

$$Var(X) := \overline{x^2} - (\overline{x})^2, \qquad \text{avec} \quad \overline{x^2} := \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet}x_i^2 = \sum_{i=1}^{k} f_{i\bullet}x_i^2,$$

et

$$Var(Y) := \overline{y^2} - (\overline{y})^2, \qquad \text{avec} \quad \overline{y^2} := \frac{1}{N}\sum_{j=1}^{l} n_{\bullet j}y_j^2 = \sum_{j=1}^{l} f_{\bullet j}y_j^2.$$

Les écarts-type

$$\sigma_X := \sqrt{Var(X)} \qquad \text{et} \qquad \sigma_Y := \sqrt{Var(Y)}.$$

Remarque *Dans le cas continu, xi et yj représentent respectivement le centre des classes de X etY , c'est à dire,*

$$x_i = \frac{L_{i+1} + L_i}{2} \qquad et \qquad y_j = \frac{L_{j+1} + L_j}{2}.$$

Distributions conditionnelles

Série conditionnelle par rapport à Y

Elle est notée par Y/x_j (ou Y_j) et on dit que c'est la série conditionnelle de Y sachant que $X = x_i$. Nous calculons aussi dans ce cas la fréquence conditionnelle $f_{j/i}$ (f_j sachant i), pour $j = 1, ..., l$, par

$$f_{j/i} := \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}.$$

$Y/X = x_i$	y_1	...	y_j	...	y_l	Total
Effectif	n_{i1}	...	n_{ij}	...	n_{il}	$n_{i\bullet}$

Nous avons aussi la moyenne conditionnelle $\overline{y}_i$, c'est à dire la moyenne des valeurs de Y sous la condition x_i, elle est définie par

$$\overline{y}_i := \sum_{j=1}^{l} f_{j/i} y_j = \frac{1}{n_{i\bullet}} \sum_{j=1}^{l} n_{ij} y_j \ .$$

Pour l'écart-type conditionnel, nous avons $\sigma_{Y_i} := \sqrt{Var(Y_i)}$ avec

$$Var(Y_i) := \sum_{j=1}^{l} f_{j/i}(y_j - \overline{y}_i)^2 = \overline{y^2}_i - (\overline{y}_i)^2.$$

Série conditionnelle par rapport à X

Elle est notée par X/y_j (ou X_j) et on dit que c'est la série conditionnelle de X sachant que $Y = y_j$. Nous calculons dans ce cas la fréquence conditionnelle $f_{i/j}$ (f_i sachant j), pour $i = 1, ..., k$, par

$$f_{i/j} := \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}.$$

$X/Y = y_j$	x_1	...	x_i	...	x_k	Total
Fréquence	$f_{1/j}$	...	$f_{i/j}$	...	$f_{k/j}$	1

Nous avons aussi la moyenne conditionnelle $\overline{x}_j$, c'est à dire la moyenne des valeurs de X sous la condition y_j, elle est définie par

$$\overline{x}_j := \sum_{i=1}^{k} f_{i/j} x_i = \frac{1}{n_{\bullet j}} \sum_{i=1}^{k} n_{ij} x_i.$$

Pour l'écart-type conditionnel, nous avons $\sigma_{X_j} := \sqrt{Var(X_j)}$ avec

$$Var(X_j) := \sum_{i=1}^{k} f_{i/j}(x_i - \overline{x}_j)^2 = \overline{x^2}_j - (\overline{x}_j)^2.$$

Dépendance et indépendance statistique

On dit que les variables X et Y sont **indépendantes** si Pour tout le couple (i,j) :

$$f_{ij} = f_{i\bullet} \cdot f_{\bullet j} \Leftrightarrow n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

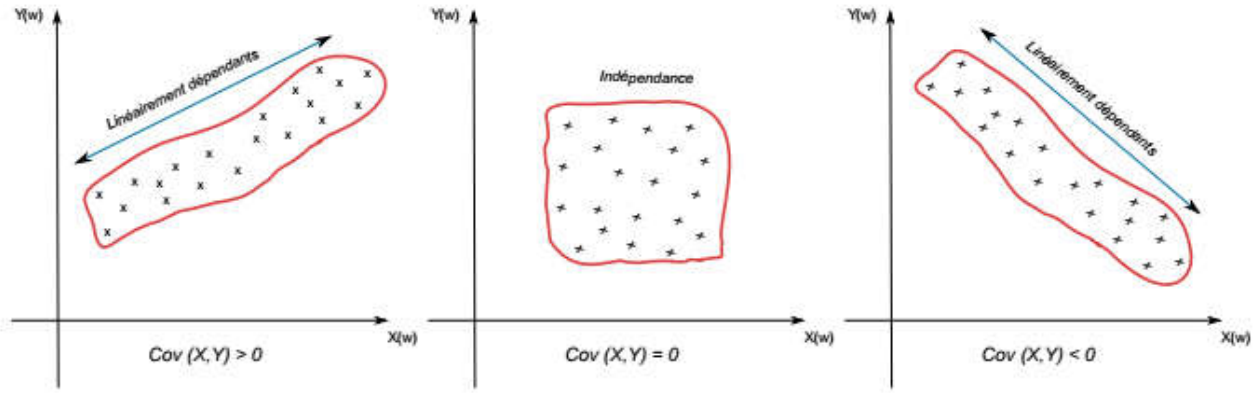
Si non On dit que sont **dépendantes**

covariance

la covariance entre les variables *X* et *Y est donnée par*

$$Cov(X,Y) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}(x_i - \overline{x})(y_j - \overline{y}).$$

Interprétation géométrique



Propriétés

$$Cov(X,Y) = \overline{xy} - \overline{x}\ \overline{y} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} x_i y_j - \overline{x}\ \overline{y}.$$

N.B *Dans le cas où nous avons un tableau des données brutes "representation 1" (nousn'avons pas d'effectifs), nous avons les formules suivantes*

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{n} x_i \qquad et \qquad \overline{y} = \frac{1}{N} \sum_{i=1}^{n} y_i.$$

$$\text{cov}(X,Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \overline{x} \cdot \overline{y}$$

Propriétés de la covariance

1. $\text{cov}(X,Y) = \text{cov}(Y, X)$

2. $\text{cov}(X, X) = \text{var}(X)$

3. $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\ \text{cov}(X, Y)$

4- si X et Y sont indépendantes alors (réciproque est fausse) Cov(X, Y) = 0.

Coefficient de corrélation linéaire

Est donné par

$$\rho_{XY} := \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

Propriété

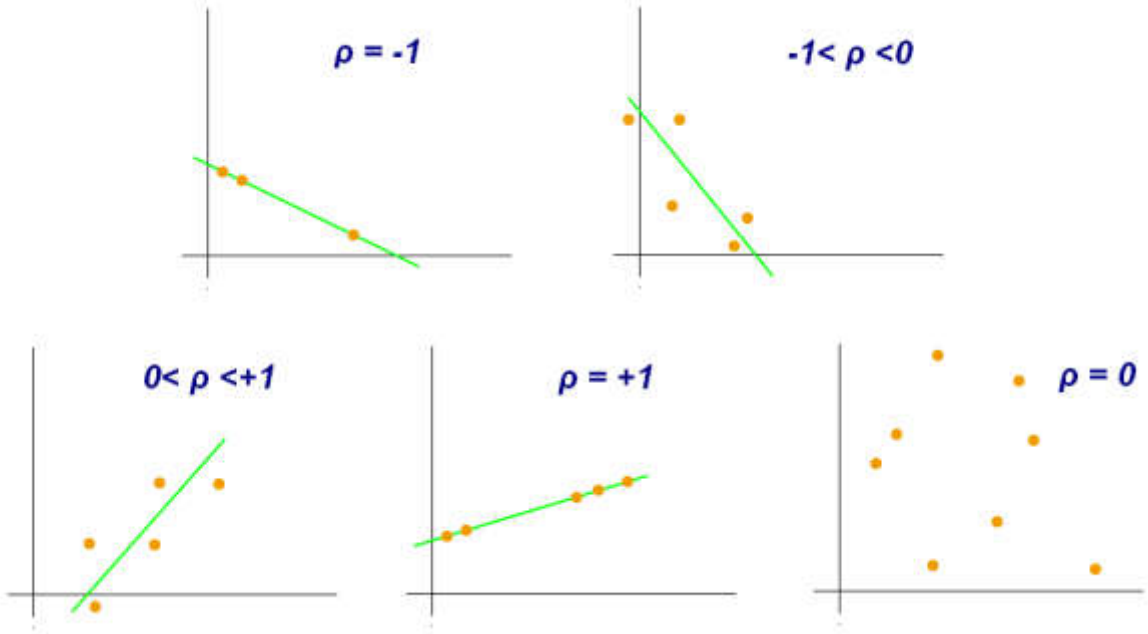
Le coefficient ρ_{XY} *est compris entre* $[-1, 1]$, *ou encore*

$$|\rho_{XY}| \leq 1.$$

Le coefficient ρ_{XY} mesure le degré de liaison linéaire entre X et Y .

Nous avons les deux caractéristiques suivantes :

Plus le module de ρ_{XY} est proche de 1 plus X et Y sont liées linéairement.
– Plus le module de ρ_{XY} est proche de 0 plus il y a l'absence de liaison linéaire entre X et Y .



Remarque

Par définition, si $\rho_{XY} = 0$, *alors* $Cov(X, Y) = 0$.