Chapitre 2

Analyse bivariée

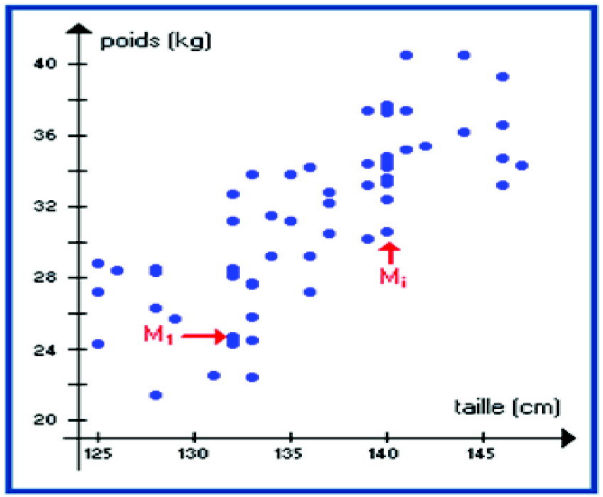
Plan

1. 1- Définitions:
2. 2- Paramètres et tableaux de calculs
   - 2.1-Données non groupées:
   - 2.2- Données groupées:
   - 2.3- Propriétés:
   - 2.4 Distribution conditionnelle, indépendance
   - 2.5- Exemple :
3. 3- Ajustement linéaire
   - 3.1- Introduction
   - 3.2- Définitions
   - 3.3- Méthode des moindres carrés:
   - 3.4-Ajustement et corrélation
   - 3.5 -Exemple (Ajustement exponentiel):
4. 4- Conclusion

1- Définitions:

On se donne une population de taille n et sur chaque élément de cette population on effectue deux observations portant sur deux caractères différents X et Y .
Pour chaque élément de l'échantillon, on peut associer un couple de valeurs (x_i, y_i) où x_i est la valeur du caractére X et y_i est la valeur du caractére Y.

On obtient aussi un nuage de n points constituant un diagramme de dispersion.



Les résultats de ces observations peuvent être présentés sous deux formes

Données non groupées :

Individu	1	2	$\cdots$	n
Valeur X	x_1	x_2	$\cdots$	x_n
Valeur Y	y_1	y_2	$\cdots$	y_n

Données groupées :

Les valeurs prises par X et Y étant respectivement $x_1, x_2, \cdots, x_p$ et $y_1, y_2, \cdots, y_q$.

n_{ij} est l'effectif des individus dont les valeurs de X et Y sont respectivement x_i et y_j.

$X \diagdown Y$	y_1	y_2	$\ldots$	y_j	$\ldots$	y_q	Totaux
x_1	n_{11}	n_{12}	$\ldots$	n_{1j}	$\ldots$	n_{1q}	$n_{1.}$
x_2	n_{21}	n_{22}	$\ldots$	n_{2j}	$\ldots$	n_{2q}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\ldots$	n_{ij}	$\ldots$	n_{iq}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_{p1}	n_{p2}	$\ldots$	n_{pj}	$\ldots$	n_{pq}	$n_{p.}$
Totaux	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.j}$	$\ldots$	$n_{.q}$	n

Ce tableau doit être lu de la façon suivante :

- l'effectif du caractère double (x_i, y_i) est n_{ij}.
- n_{ij} est l'effectif des individus présentant simultannement les modalités x_i et y_j.
- L'effectif de toute la population est $n = \sum_{i,j} n_{ij}$.
- En fréquence $f_{ij} = \dfrac{n_{ij}}{n}$ et $\displaystyle\sum_{i,j} f_{ij} = 1$.

Exemple. Soit Ω la population constituée par les quatre pays suivants : France, Allemagne, Grande bretagne et l'Italie. Notons X la production de fonte (bronze) et Y la production d'acier arrondies en millons de tonnes

Ω	Allemagne	France	G. B.	Italie
X	27.2	15.9	17.6	3.5
Y	37.2	19.8	26.7	9.8

Il est naturel de s'interroger sur la relation qui lié X et Y. On regroupe les valeurs des x_i et des y_i dans le tableau suivant :

$Y \diagdown X$	3.5	15.2	17.6	27.2	Totaux
9.8	1	0	0	0	1
19.8	0	1	0	0	1
26.7	0	0	1	0	1
37.3	0	0	0	1	1
Totaux	1	1	1	1	4

Définitions

- Effectifs marginaux.

La somme des effectifs contenus dans la ligne de x_i est égale à l'effectif des élements dont la valeur du caractère X est x_i. Elle est notée $n_{i.}$.

$$n_{i.} = n_{i1} + \cdots + n_{iq} = \sum_{j=1}^{q} n_{ij}.$$

La somme des effectifs partiels contenus dans la colonne de y_j est égale à l'effectif des élements dont la valeur du caractère Y est y_j. Elle est notée $n_{.\,j}$.

$$n_{.\,j} = n_{1j} + \cdots + n_{pj} = \sum_{i=1}^{p} n_{ij}.$$

$n_{i.}$ et $n_{.j}$: sont appelés effectifs partiels marginaux.
On a:

$$n = \sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j} = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}.$$

- Fréquences marginales

$$f_{i.} = \frac{n_{i.}}{n} \quad \text{fréquence marginale de } x_i.$$
$$f_{.j} = \frac{n_{.j}}{n} \quad \text{fréquence marginale de } y_j.$$

On a

$$\sum_{i=1}^{p} f_{i.} = \sum_{j=1}^{q} f_{.j} = \sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = 1.$$

(f_{ij} fréquence partielle correspondant à $X = x_i$ et $Y = y_j$).
Les couples $(x_i, n_{i.})_{1\leq i\leq p}$ et $\left(y_j, n.j\right)_{1\leq j\leq q}$ définnissent les distributions statistiques marginales.

2- Paramètres et tableaux de calculs

2.1-Données non groupées:

Comme dans le cas d'un seul caractère, on a :

Moyennes $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Variances

$$V(X) = \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{X}^2$$

$$\text{et } V(Y) = \sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{Y}\right)^2$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \overline{Y}^2.$$

On introduit maintenant deux nouveaux caractères qui dépendent à la fois de X et de Y.

Covariance. la Covariance de X et Y, notée $cov(X, Y)$, est définie par:

$$\sigma_{XY} = cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right).$$

On montre aisément que:

$$\sigma_{XY} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{X}\,\overline{Y}.$$

Le coefficient de corrélation linéaire du couple (X, Y) noté $\rho(X, Y)$, est définis par:

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}.$$

2.2- Données groupées:

Plus généralement et surtout lorsque l'effectif total est grand, si $x_1, \ldots, x_p$ sont les modalités de X et $y_1, \ldots, y_q$ sont les modalités de Y, on dresse le tableau suivant :

$X \diagdown Y$	y_1	$\cdots$	y_q	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$	$\sum_{j=1}^{q} n_{ij}y_j$	$x_i\sum_{j=1}^{q} n_{ij}y_j$
x_1	n_{11}	$\cdots$	n_{1q}	$n_{1.}$	$n_{1.}x_1$	$n_{1.}x_1^2$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$				
x_p	n_{p1}	$\cdots$	n_{pq}	$n_{p.}$	$n_{p.}x_p$	$n_{p.}x_p^2$		
$n_{.j}$	$n_{.1}$	$\cdots$	$n_{.q}$	n				
$n_{.j}y_j$	$n_{.1}y_1$	$\cdots$	$n_{.q}y_q$					
$n_{.j}y_j^2$	$n_{.1}y_1^2$	$\cdots$	$n_{.q}y_q^2$					
$\sum_{i=1}^{p} n_{ij}x_i$		$\cdots$						
$y_j\sum_{i=1}^{p} n_{ij}x_i$		$\cdots$						

i) **Moyennes:** $\overline{X} = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i$ et $\overline{Y} = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j$

ii) **Variances:**

$$V(X) = \sigma_X^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i.}\left(x_i - \overline{X}\right)^2$$

$$= \left(\frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i^2\right) - \overline{X}^2$$

$$\text{et} \quad V(Y) = \sigma_Y^2 = \frac{1}{n}\sum_{j=1}^{q} n_{.j}\left(y_i - \overline{Y}\right)^2$$

$$= \left(\frac{1}{n}\sum_{j=1}^{q} n_{.j}y_i^2\right) - \overline{Y}^2$$

iii) **Ecart-type:** $\sigma_X = \sqrt{V(X)}$ et $\sigma_Y = \sqrt{V(Y)}$

iv) **Covariances:**

On appelle covariance du couple (X, Y) et on le note $cov(X, Y)$ ou σ_{XY} la moyenne de $(X - \overline{X})(Y - \overline{Y})$

$$cov(X, Y) = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} (x_i - \overline{X})(y_i - \overline{Y}).$$

On montre que: $\sigma_{XY} = cov(X, Y) = \left(\frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} x_i y_j \right) - \overline{X}\,\overline{Y}.$

v) **Coefficient de correlation linéaire:**

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}.$$

N.B L'importance des paramètres $cov(X, Y)$ et $\rho(X, Y)$ apparaîtra quand on s'intéressera au lien (ou corrélation) éventuel entre X et Y .

2.3- Propriétés:

On montre que:

- $|\rho(X, Y)| \leq 1$.
- $\rho(aX + b, a'Y + b') = \dfrac{aa'}{|aa'|}\rho(X, Y)$. donc
  $\rho(aX + b, a'Y + b') = \pm\rho(X, Y)$
- $cov(aX + b, a'Y + b') = \dfrac{1}{aa'}cov(X, Y)$.

Ces formules sont utilisables pour simplifier les calculs.

En effectuant les changement de variables suivants:
$X' = \dfrac{X - c}{d}$ de et $Y' = \dfrac{Y - c'}{d'}$
avec $d, d' \neq 0$, on obtient:

$$cov(X', Y') = \frac{1}{dd'} cov(X, Y).$$

$$\rho(X', Y') = \frac{|dd'|}{dd'} \rho(X, Y)$$

donc $\rho(X', Y') = \pm\rho(X, Y)$.

Démonstration: (cf. exercice 1 ; fiche TD N°2)

2.4 Distribution conditionnelle, indépendance

La fréquence conditionnelle de x_i sachant y_j (y_j réalisé)

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

Où n_{ij} est l'effectif correspondant à $X = x_i$ et $n_{.j}$ l'effectif partiel marginal de y_j.

On a $f_{j/i} = \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{i.}}$.

Ainsi $f_{ij} = f_{i.} \times f_{j/i} = f_{.j} \times f_{i/j}$.

Définition

Deux variables statistiques X et Y sont dites statistiquement indépendantes si et seulement si, pour chacune des deux variables, les distributions conditionnelles sont identiques à la distribution marginale:

$$f_{i/j} = f_{i.} \quad ou \quad f_{j/i} = f_{.j} \forall (i,j)$$

Conséquence: Les caractères X et Y sont indépendants si et seulement si

$$\forall (i,j) \quad f_{ij} = f_{i.} \times f_{.j}$$

2.5- Exemple :

Sur le tableau suivant figure l'âge de la mère (x) et le poids de l'enfant (y) pour un échantillon de 40 naissances, présentés avec un groupement à deux dimensions en classe d'âge de 5 ans et en classe de poids de 500g

	2500	3000	3500	4000	4500	$n_{i.}$
20	1	5	4	2	-	12
25	2	3	5	1	-	11
30	1	2	2	1	-	6
35	-	3	3	1	1	8
40	-	2	-	1	-	3
$n_{.j}$	4	15	14	6	1	40

$n_{13} = 4$ signifie qu'il ya 4 enfants dont l'âge de la mère est 20 ans et dont le poids est 3500g. Il y a 6 mères dont l'âge est 30 ans. Il Y a 14 enfants dont le poids est 3500g.

3-Ajustement linéaire
3.1- Introduction

On considère une population de taille (effectif total) n sur laquelle on définit une statistique double X et Y. (On éffectue sur Ω deux observations portant sur 2 caractères différents).
Le problème qui se pose est celui qui consiste à rechercher s'il existe une relation entre X et Y.
A chaque élement ω_i de l'échantillon, on associe un couple de valeurs (x_i, y_i) qu'on représente graphiquement par un point $M_i(x_i, y_i)$ du plan. Et on obtient ainsi un nuage de n points qui constitue ce qu'on appelle un diagramme de disperssion.
Ajuster un ensemble de points consiste à déterminer une courbe (C) simple aussi proche que possible des points M_i.
L'ajustement linéaire est le cas où (C) est une droite.

3.2- Définition

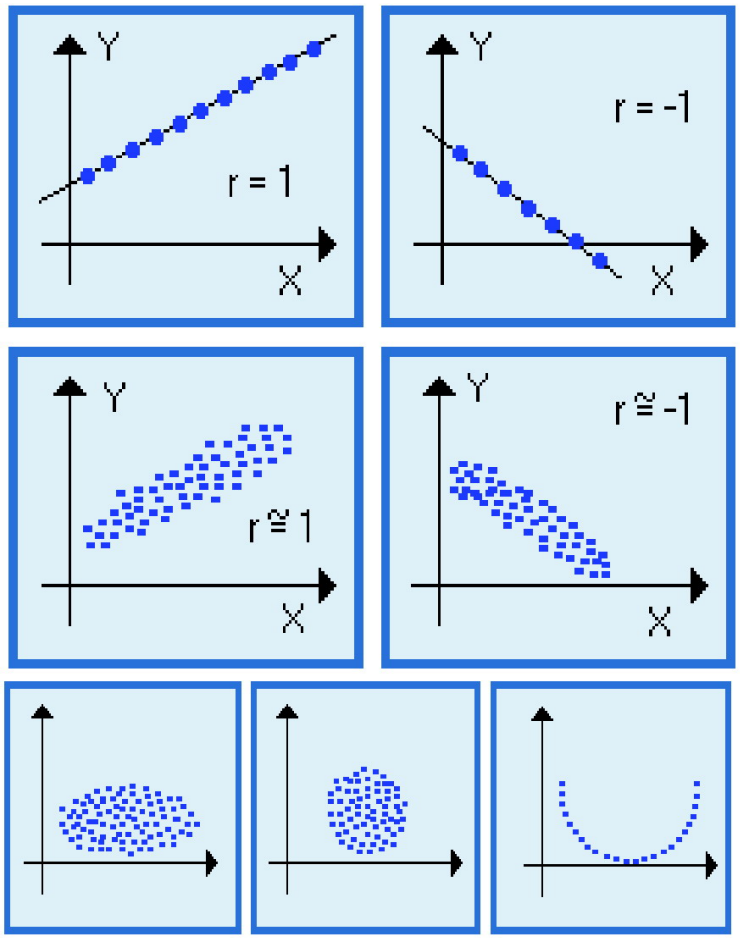
On dit qu'il y a correlation entre deux caractères observés sur une même population lorsque les variations des deux caractères se produisent dans le même sens ou lorsque les variations sont de sens contraires.

Nuage de points : Diagramme de disperssion

L'existence d'une correlation peut-être décelée (détectée) graphiquement. La forme du nuage de points formé par les points $M_i\,(x_i, y_i)$ nous permettent de constater si les caractères X et Y sont en correlation ou non.

Définition. On dit qu'une correlation (lorsqu'elle existe) qui lié 2 caractères X et Y est positive ou directe si Y croît en même temps que X. Si Y décroît lorque X croît, la correlation est dite inverse ou négative.

Exemples de différents nuages



Coefficient de correlation

Formule pratique de ρ pour le calcul

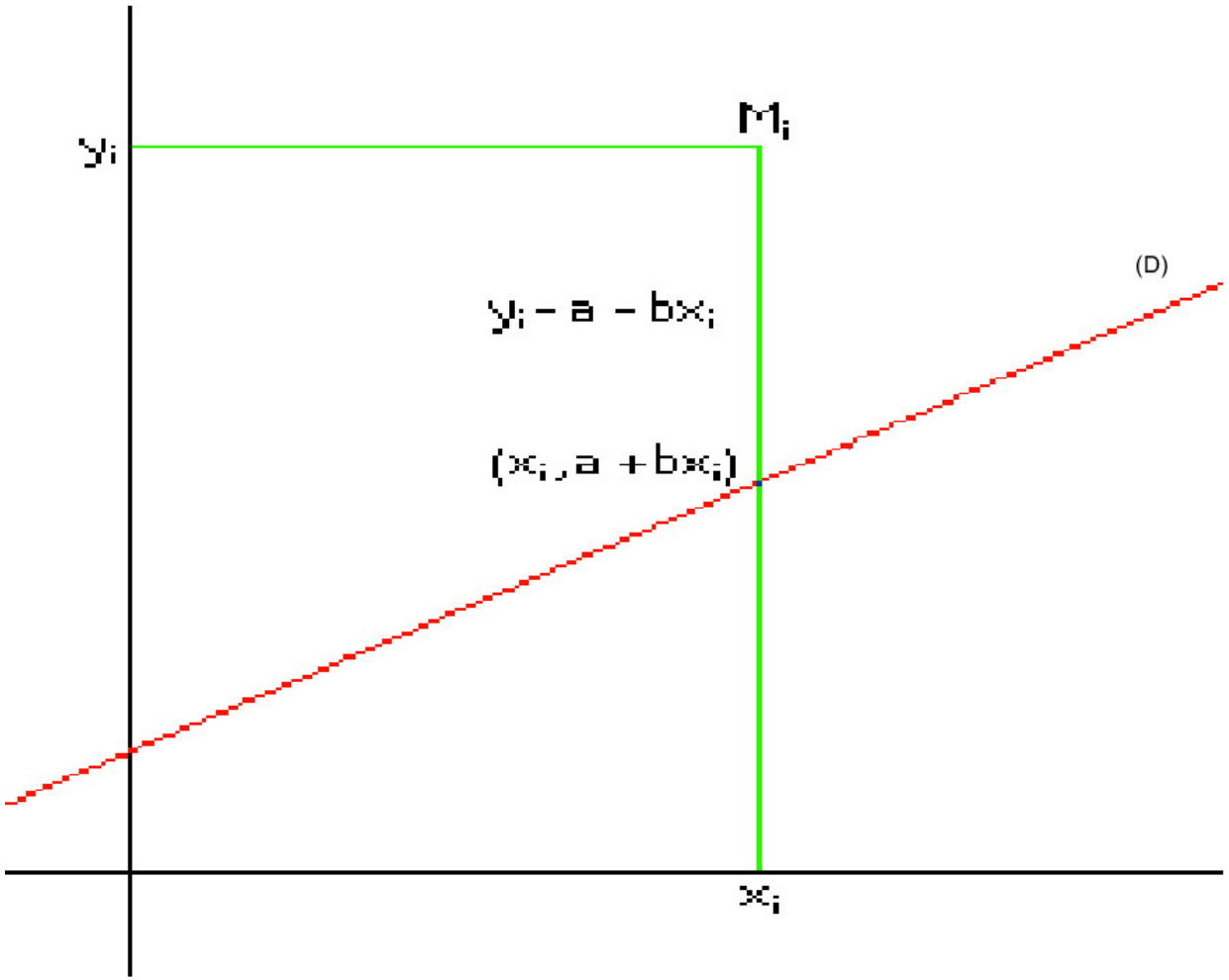
$$\rho = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sqrt{\left[\left(\sum_{i=1}^{n} x_i\right)^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]\left[\left(\sum_{i=1}^{n} y_i\right)^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}.$$

ρ est compris entre -1 et 1.

- Si les caractères X et Y sont indépendants alors $\rho = 0$. Cependant la réciproque n'est pas necessairement vraie. Si $\rho = 0$, on dit qu'il y a correlation nulle entre X et Y ; la liaison entre X et Y peut être de forme outre que linéaire.
- Si $0 < \rho < 1$, la correlation est positive (X et Y varient dans le même sens) La valeur $\rho = +1$ indique une relation linéaire parfaite $Y = aX + b$ avec $a > 0$. C'est un cas extrème très peu rencontré en pratique.
- Si $-1 < \rho < 0$, la correlation est négative (X et Y varient dans le sens contraire) La valeur $\rho = -1$ indique une relation linéaire parfaite $Y = aX + b$ avec $a < 0$. C'est un cas extrème très peu rencontré en pratique.

3.3-Méthode des moindres carrés:

Soit M_i un point de cordonnées (x_i, y_i) On appelle distance de M_i parallélernent à l'axe (oy) à la droite $(\triangle)$ d'équation $y = ax + b$, le réel positif $d_i = |y_i - ax_i - b|$ (Attention ! il ne s'agit pas des distances des points M_i à la droite $(\triangle)$.)
La méthode des moindres carrés consiste à chercher les valeures de a et b (c.à.d trouver une droite $(\triangle)$ d'équation $y = ax + b$)qui mininisent $\delta(a, b) = \sum_i (y_i - ax_i - b)^2 = \sum_i d_i^2$.


M_i
y_i
(D)
$y_i - a - bx_i$
$(x_i, a + bx_i)$
x_i

Le problème admet une solution unique solution du système linéaire issue de l'annulation des dérivées partielles premières de la fonction $\delta(a, b)$. Il s'agit de résoudre

$$\begin{cases} \dfrac{\partial \delta}{\partial a} = -2\sum_{i=1}^{n} x_i \left(y_i - ax_i - b\right) = 0 \\ \dfrac{\partial \delta}{\partial b} = -2\sum_{i=1}^{n} \left(y_i - ax_i - b\right) = 0 \end{cases}$$

Ce système en a et b admet pour solution unique

$$a = \frac{cov(X, Y)}{V(X)} \text{ et } b = \overline{y} - a\overline{x}.$$

Théorème

Soit $M_i\,(x_i, y_i)_{1\leq i\leq n}$ *un ensemble fini de points fixes du plan euclidien où* x_i *sont les modalités d'un caractère* X *et* y_i *celles d'un autre caractère* Y *définis sur une même population* Ω.

Soient $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $V(X) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$ *et*

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

La droite d'équation $y - \overline{y} = a\,(x - \overline{x})$ *où* $a = \frac{cov(X, Y)}{V(X)}$ *est la droite de régression de* y *en* x *et est notée* $D_{y/x}$.

Remarque

i) On définit de même la droite de régression de x en y notée $D_{x/y}$ et d'équation

$$x - \overline{x} = a' (y - \overline{y}) \text{ avec } a' = \frac{cov(X, Y)}{V(Y)}.$$

ii) Les droites $D_{y/x}$ et $D_{x/y}$ passent par le point $G(\overline{x}, \overline{y})$.

3.4- Ajustement et corrélation

Les droites de regression $(\triangle)$ et $(\triangle')$ ayant pour équations: $y = ax + b$, $x = a'y + b'$ ont les propriétés suivantes:

- elles passent toutes les deux par le point $G(\overline{X}, \overline{Y})$ appelé point moyen de la statistique.
- les pentes des deux droites sont de même signe celui de la covariance et sont respectivement a et $\frac{1}{a'}$
- $aa' = \frac{cov(x,y)^2}{v(x)v(y)} = (\rho(x,y))^2$

- $(\triangle)$ et $(\triangle')$ sont confondues si elles ont la même pente (car elles passent toutes les deux par $G(\overline{X}, \overline{Y})$).
- dans ce cas $a = \frac{1}{a'}$ c.à.d. $aa' = 1$ donc $\rho(x, y) = \pm 1$. les points M_{ij} sont alors alignés.
- La corrélation linéaire est d'autant bonne (ou forte) que le coefficient de corrélation ρ est proche en valeur absolue de 1 ($\rho \simeq 1 \Longleftrightarrow a \simeq \frac{1}{a'}$).
- si ρ est proche de zéro, on dit qu'il y a corrélation linéaire très mauvaise entre X et Y. il faudrait alors approcher le nuage des points M_{ij} par une courbe.

- $(\triangle)$ et $(\triangle')$ sont confondues si elles ont la même pente (car elles passent toutes les deux par $G(\overline{X}, \overline{Y})$).
- dans ce cas $a = \frac{1}{a'}$ c.à.d. $aa' = 1$ donc $\rho(x, y) = \pm 1$. les points M_{ij} sont alors alignés.
- La corrélation linéaire est d'autant bonne (ou forte) que le coefficient de corrélation ρ est proche en valeur absolue de 1 ($\rho \simeq 1 \Longleftrightarrow a \simeq \frac{1}{a'}$).
- si ρ est proche de zéro, on dit qu'il y a corrélation linéaire très mauvaise entre X et Y. il faudrait alors approcher le nuage des points M_{ij} par une courbe.

Remarque

i) La corrélation est dite forte lorsque $\rho(x,y)^2 > 0.75$*, et dans ce cas on estime qu'on peut approcher le nuage de points par une droite (la méthode des moindres carrés s'applique).*
ii) Dans le cas contraire, on ne peut pas approcher le nuage de points par une droite (l'approximation serait trop mauvaise) mais il se peut que d'autre courbe permette un bon ajustement (ajustement exponentiel par exemple)
En général, si le coefficient de corrélation est fort, on peut conclure à une corrélation entre les deux séries statistiques, mais ce n'est pas toujours vrai

3.5- Exemple (Ajustement exponentiel):

La statistique suivante indique l'évolution de la consommation d'énergie éléctrique dans un pays exprimée en TWh

Année	1949	1953	1957	1961	1965	1969	1973	1977
Cons.	30	41	56	73	97	123	165	207

La relation qui lie la consommation au temps (année) est de type exponentiel.

Déterminons la droite de régression de $Y = \log y$ en x

Exemple (suite)

On effectue le changement de variable $X = \dfrac{x - 1961}{4}$, *on a*

X_k	$Y_k = \log y_k$
-3	1.417
-2	1.613
-1	1.748
0	1.863
1	1.987
2	2.019
3	2.217
4	2.316

On en déduit $\overline{X} = \dfrac{1}{2}$, $\overline{Y} = \dfrac{15.18}{8} \simeq 1.8975$

Exemple (suite)

La droite $Y = AX + B$ est définie par

$$\begin{aligned} A &= \frac{\sigma_{XY}}{\sigma_X^2} \simeq \frac{0.64}{5.25} \simeq 0.12 \\ B &= \overline{Y} - A\overline{X} \simeq 1.836. \end{aligned}$$

Remarquons que l'on a :

$$\rho_{XY} \simeq \frac{0.64}{\sqrt{5.25 \times 0.0794}} \simeq 0.99$$

ce qui justifie la recherche d'un ajustement exponentiel.

4- Conclusion :

L'étude des séries statistiques à deux variables permet de mettre en rapport deux caractères afin de pouvoir déterminer une valeur manquante ou de prévoir une tendance.
Néanmoins, deux caractères peuvent avoir un très fort coefficient de corrélation sans pour autant être réellement liés.