CHAPITRE III

La statistique bidimensionnelle

I) Introduction

On s'intéresse au rapprochement possible entre 2 variables statistiques observées sur une même population, 3 cas sont possibles soit :

- 2 variables quantitatives
- 2 variables qualitatives

Variable quantitative/variable qualitative dans le cadre su cours les 2 variables sont de même nature. On cherche pour ces variables une étude simultanée de celles-ci à traduire au moyen de table graphique et/ou de méthode numérique, la liaison pouvant exister entre ces 2 variables.

II) Représentation graphique

1) Deux variables quantitatives
   - juxtaposition d'histogramme et/ou de diagramme en bâtons.

   - Nuage de points

     Si $(x_i ; y_i)_{i=1...n}$ ou $(y_i ; x_i)_{i=1..n}$

     Si $(x_i)_{i=1...n}$ et $(y_i)_{i=1..n}$ sont respectivement les observations de 2 variables x et y.

2) Deux variables qualitatives

   Juxtaposition de diagramme en bâtons.

III) La statistique bidimensionnelle des paramètres.

1) La série double à doubles indices.

On considère 2 variables statistiques x et y admettent (resp. p et q) modalités, classes et/ou valeur observées mesurées sur une même population de taille n.

Ex1:

a) x = poids $(x_i)_{i=1...n}$
   y = taille $(y_i)_{i=1....n}$

b) $(x_i ; y_j ; n_{ij})$
   Valeurs observées.
   Centre de classes

1.1) Tableau croisé (ou tri croisé).

Définition: déterminer le tableau croisé associé aux variables x et y, c'est déterminer le nombre d'unité statistique n_{ij} répondant simultanément aux

modalités classes ou valeurs observées x_i et y_j de x et y pour $_i \in \{1....p\}$ et $_j \in \{1....p\}$ $(n_{ij})_{i=1...p}$ constitue un tableau d'effectif croisés.

$$\begin{pmatrix} n_{11}, n_{12} \ldots\ldots\ldots n_{1q} \\ n_{21}, n_{22} \ldots\ldots\ldots n_{2q} \\ n_{i1}, n_{i2} \ldots\ldots\ldots n_{ij} .. n_{iq} \\ n_{p1}, n_{p2} \ldots\ldots\ldots n_{pq} \end{pmatrix}$$

Ex 2 : On considère une variable quantitative discrète x et y (resp. 3 et 2) valeurs observées soit :
$x_1 = 0 \quad x_2 = 1 \quad x_3 = 2$
$y_1 = 0 \quad y_2 = 1$
$n = 10$

v	X_1	X_2	X_3	Y_1	Y_2
1	0	0	1	1	0
2	1	0	0	1	0
3	1	0	0	0	1
4	0	1	0	1	0
5	0	0	1	1	0
6	0	1	0	0	1
7	0	1	0	0	1
8	1	0	0	1	0
9	1	0	0	1	0
10	0	1	0	0	1

Construisons le tableau des effectifs croisés associés à ces donnés, c'est un tableau de 3 lignes (car X prend 3 valeurs) et 2 colonnes (car Y prend 2 valeurs).
$s_{ij\ i=1....3}$ est le nombre d'unité statistique pour lesquelles on a simultanément observés x_i et y_j avec $_{i=1..3}$ et $_{j=1...2}$
n_{11} associé à X_1 et Y_1 on lit $n_{11} = 3$
n_{12} associé à X_1 et Y_2 on lit $n_{12} = 1$
n_{21} associé à X_2 et Y_1 on lit $n_{21} = 1$
n_{22} associé à X_2 et Y_2 on lit $n_{22} = 3$
n_{31} associé à X_3 et Y_1 on lit $n_{31} = 2$
n_{32} associé à X_3 et Y_2 on lit $n_{32} = 0$

Le tableau croisé est donc : $\begin{pmatrix} n_{11}....n_{12} \\ n_{21}....n_{22} \\ n_{31}.....n_{32} \end{pmatrix} = \begin{pmatrix} 3 & 1 \\ 1 & 3 \\ 2 & 0 \end{pmatrix}$

1.2) Transformation d'un tableau croisé

On considère le tableau croisé $(n_{ij})_{i=1...p\ j=1...q}$

$n_i = \sum_{i=1}^{p} n_i$ pour $_{i=1...p}$ on appel marge colonne et on note n_j la quantité définit

par $n_{ij} = \sum_{}^{p} n_{ij}$ pour $_{j=1..p}$

$i=1$

On définit le tableau des fréquences croisé par $f_{ij\ i=1\ldots p\ \ j=1\ldots q}$

Avec $f_{ij} = \frac{n_{ij}}{n}$.

On définit aussi :

Les fréquences marges lignes $f_{i.} = \sum f_{ij}$ pour $_{i=1\ldots p}$.

Les fréquences marges colonne $f_{.j} = \sum f_{ij}$ pour $_{j=1\ldots p}$.

On a les relations suivantes $n = \sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j} = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}$

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

Définition: on appelle distribution de X liée pour $Y = Y_j$ les valeurs (n_{1j}, n_{2j}, …. n_{pj} par $_{j=1\ldots p}$.

Distribution de Y liée pour $X = X_i$ la suite des valeurs (n_{i1}, n_{i2},….n_{iq}) pour $_{i=1\ldots p}$.

Ex 3 (suite ex 2) :

On a le tableau croisé $\begin{pmatrix} 3 & 1 \\ 1 & 3 \\ 2 & 0 \end{pmatrix}$.

Déterminons les effectifs des lignes et des colonnes.

$n_{1.} = \sum n_{1j} = n_{11} + n_{12} = 4$

$n_{2.} = \sum n_{2j} = n_{21} + n_{22} = 4$

$n_{3.} = \sum n_{3j} = n_{31} + n_{32} = 2$

$n_{.1} = \sum n_{i1} = n_{11} + n_{21} + n_{31} = 6$

$n_{.2} = \sum n_{i2} = n_{12} + n_{22} + n_{32} = 4$

$$n = \sum n_{i.} = \sum_{i=1\ j=1} n_{j} = \sum^{2}\sum^{2} n_{ij}$$

	Y_1	Y_2	n_i
X_1	3	1	4
X_2	1	3	4
X_3	2	0	2
n_{ij}	6	4	10

La distribution de X liée à $Y = Y_1$ est $(n_{11}, n_{21}, n_{31}) = (3 ; 1 ; 2)$.

La distribution de Y liée à $X = X_1$ est $(n_{21}, n_{22}) = (1 ; 3)$.

Pour les fréquences on a le tableau des fréquences croisées.

$f_{ij} = \frac{n_{ij}}{n} = \frac{n_{ij}}{10}$ pour $_{i} = 1, 2, 3.$

$_{j} = 1, 2$

Les fréquences marginales s'écrivent $f_{i.} = \sum_{j=1}^{2} f_{ij}$ pour $_{i=1\ldots3}$

$f_{.j} = \sum f_{ij}$ pour $_{i=1\ldots2}$

De manière générale on à une présentation de la fréquence.

● Propriété d'indépendance des distributions à 2 variables qualitatives.

Définition : distribution théorique

On appel distribution théorique et on note :

$(V_{ij})_{\substack{i=1\ldots p \\ j=1\ldots q}}$

La distribution définie par $V_{ij} = \dfrac{n_{ij} \cdot n_{ij}}{n}$ pour $_{i=1\ldots p \quad j=1\ldots q}$

Problème a-t-on indépendance des 2 caractères observés.

A-t-on donc pour cela

$_i \in \{1\ldots.p\}$ et $_j \in \{1\ldots q\}$.

$n_{ij} = V_{ij} = \dfrac{n_{i.} \cdot n_{.j}}{n}$ ou de manière équivalente $f_{ij} = n \cdot b_i \cdot f_{ij}$?

● Une mesure de l'indépendance de 2 variables :

On a les effectifs observés

n_{ij} $_{\substack{i=1\ldots p \\ j=1\ldots q}}$

On a les effectifs théoriques

$(Y_{ij})_{\substack{i=1\ldots p \\ j=1\ldots q}}$

$Y_{ij} = \dfrac{n_{i.} \cdot n_{.j}}{n}$ où $\begin{cases} n_{i.} = \sum n_{ij} \\ n_{.j} = \sum n_{ij} \end{cases}$

On s'intéresse à la différence $n_{ij} - y_{ij}$

Si pour tout $(_i ; _j)$ $n_{ij} - Y_{ij} = 0$ alors les variables sont indépendantes.

Si pour tout $(_i ; _j)$ donné on a $n_{ij} - Y_{ij} > 0$ alors on dit que le couple $(x_i ; y_j)$ est surreprésenté.

Si $n_{ij} - Y_{ij} < 0$, alors on dit que $(x_i ; y_j)$ est sous-représenté.

Déterminons les effectifs théoriques :

$Y_{ij} = \dfrac{n_{i.} \cdot n_{.j}}{n}$ où $_i = 1, 2, 3$

$_j = 1, 2.$

effectif	y_1	y_2
x_1	2.4	1.6

x_2	2.4	1.6
x_3	1.2	0.8

$$y_{11} = \frac{n_{1.} \cdot n_{.1}}{n} = \frac{4 \cdot 6}{10} = 2.4$$

$$y_{12} = \frac{n_{1.} \cdot n_{.2}}{n} = \frac{4 \cdot 4}{10} = 1.6$$

$$y_{21} = \frac{n_{2.} \cdot n_{.1}}{n} = \frac{4 \cdot 6}{10} = 2.4$$

$$y_{22} = \frac{n_{2.} \cdot n_{.2}}{n} = \frac{4 \cdot 4}{10} = 1.6$$

$$y_{31} = \frac{n_{3.} \cdot n_{.1}}{n} = \frac{2 \cdot 6}{10} = 1.2$$

$$y_{32} = \frac{n_{3.} \cdot n_{.2}}{n} = \frac{2 \cdot 4}{10} = 0.8$$

Par exemple le couple (1 ; 2) tel que $n_{12} - y_{12} = 1 - 16 = 0.6 < 0$.
Donc le couple $(x_i ; y_{12})$ est sous population.

2) Paramètres associés à 2variables quantitatives double à double indice.

1) Les données

On considère la série $(x_i, y_j, n_{ij})_{i=1\ldots p}$ associée à 2 variables quantitatives X et Y observée sur une population de taille n et ayant respectivement p et q valeurs observées.

2) Les résumés numériques

a) Les moyennes et les variances marginales

Définition : La moyenne marginale de x est x

$$x = 1/n \sum_{i=1}^{p} n_{i.}\, x_i$$

La moyenne marginale de y est y

$$y = 1/n \sum_{i=1}^{q} n_{.j}\, y_j$$

$i=1$

Définition : La variance marginale de x est :

$$V(x) = 1/n \sum_{i=1}^{p} n_{i.} (x_i - x)^2$$

$$= 1/n \sum_{i=1}^{p} n_{i.} x_i^2 - x^2$$

La variance marginale de y est :

$$V(y) = 1/n \sum_{i=1}^{q} n_{.j} (y_j - y)^2$$

$$= 1/n \sum_{i=1}^{q} n_{.j} y_j^2 - y^2$$

b) Moyenne et variance conditionnelle

Définition : La moyenne conditionnelle de X liée par $Y = y_j$ est :

$$x_j = 1/n_{.j} \sum_{i=1}^{p} n_{ij} x_i$$

La moyenne conditionnelle de Y liée par $X = x_i$ est :

$$y_i = 1/n_{i.} \sum_{i=1}^{q} n_{ij} y_j$$

Définition : La variance conditionnelle de X liée par $Y = y_j$ est :

$$V_j(x) = 1/n_{.j} \sum_{i=1}^{p} n_{ij} (x_i - x_j)^2$$

$$= 1/n_{.j} \sum_{i=1}^{p} n_{ij} x_i^2 - x_j^2$$

La variance conditionnelle de Y liée par $X = x_i$ est :

$$V_i(y) = 1/n_{i.} \sum_{i=1}^{q} n_{ij} (y_j - y_i)^2$$

$$V_i(y) = 1/n_{i.} \sum_{i=1}^{q} n_{ij} y_j^2 - y_i^2$$

$i=1$

c) Relation entre moyenne marginale et moyenne conditionnelle

Propriété :

La moyenne arithmétique pondéré des moyennes conditionnelle de Y (resp. de Y) est égale et la moyenne marginale de X (resp. de Y) soit :

$$x = 1/n \sum_{i=1}^{p} n_{.j}\, x_j \quad (\text{resp. } y = 1/n \sum_{i=1}^{q} n_{i.}\, y_i)$$

d) Décomposition de la variance marginale

⇨ Relation entre les variances de X.

$$V(x) = 1/n \sum_{i=1}^{p} n_{.j} \cdot V_j(x) + 1/n \sum_{i=1}^{q} n_{.j} (x_j - x)^2$$

Variance résiduelle + variance expliquée

⇨ Relation entre les variances de Y.

$$V(x) = 1/n \sum_{i=1}^{p} n_{i.} \cdot V_i(Y) + 1/n \sum_{j=1}^{q} n_{i.} (y_i - y)^2$$

	Y_1	Y_2	$n_{i.}$
x_1	3	1	4
x_2	1	3	4
x_3	2	0	2
$n_{.j}$	6	4	10

Calculons les moyennes conditionnelles de X liée par $y = y_j$ pour $_{j=1..2}$

On a x_1 moyenne conditionnelle

$$x_1 = 1/n \sum_{i=1}^{3} n_{.1}\, x_i$$

$$= 1/6 \cdot (3 \cdot 0 + 1 \cdot 1 + 2 \cdot 2)$$

$$= 5/6.$$

$$x_2 = 1/n \sum_{i=1}^{3} n_{.2}\, x_i$$

$$= ¼ \cdot (1 \cdot 0 + 3 \cdot 1 + 0 \cdot 2) = ¾ = 0.75$$

La moyenne marginale de x est donc :

$$x = 1/n \sum_{j}^{2} n_{.j}\, x_j$$

$i=1$

$= 1/n \cdot (n._1 x_1 \cdot n._2 x_2)$

$= 7/10 \cdot (6 \cdot 5/6 + 6 \cdot ¾) = 8/10$

$= 4/5 = 0.8$

Calculons les variances conditionnelles de X liée par $y = y_{j \quad j=1..2}$
On a :

$$V_1(x) = 1/n._1 \sum_{i=1}^{3} n._1 (x_i - x_1)^2$$

$$= 1/n._1 \sum_{i=1}^{3} n._1 x_i^2 - x_1)^2$$

$$= 1/6 ((3 \cdot 0^2 + 1 \cdot 1^2 + 2 \cdot 2^2) - (5/6)^2$$

$$= 9/6 - 35/36 = 29/36.$$

$$V_2(x) = 1/n._2 \sum_{i=1}^{3} n._2 \ x_i^2 - (x_2)^2$$

$$= ¼ ((1 \cdot 0^2 + 3 \cdot 1^2 + 0 \cdot 2^2) - (3/4)^2 = ¾ - (3/4)^2 = ¾ (1/4)$$

$$= ¾ - (3/4)^2 = ¾ (1/4) = 3/16.$$

$$\text{On obtient } V(x) = 1/n \sum_{i=1}^{2} n._j V_j(x) + 1/n \sum_{o=1}^{2} n._j (x_j - x)^2$$

$$= 1/10 (6 \cdot 29/36 + 4 \cdot 3/16) + 1/10 [6 \cdot (5/6)^2 + 4 \cdot (3/4)^2 - (0.8)^2]$$

$$= 0.56.$$

Paramètre de liaison entre 2 caractères quantitatifs.

1) La covariance

Définition :

La série observée (x_i, y_i, n_{ij}) associé aux variances x et y est tel que la variance de x et y est définit par définit par :

$$\text{Cos}(x ; y) = 1/n \sum_{i=1}^{p} \sum_{j=1}^{q} (x_i - x)(y_j - y)$$

Propriété 2: on a

$$\text{Cos}(x ; y) = 1/n \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} x_i y_j - xy$$

Propriété 3 : on a

La covariance est un paramètre symétrique.

Cov (x ; y) = Cov (y ; x).

Propriété 4 :
Pour (a, b, c) $\in R^4$ on a :
Cov (ax + b, cy + d) = a · c cov (x ; y).

Exercice 7 (suite ex 2)
La covariance de x est de y s'écrit:

$$\text{cov}(x ; y) = 1/n \sum_{i=1}^{3} \sum_{j=1}^{2} n_{ij} (x_i - x) (y_j - y)$$

$$= 1/n \sum_{i=1}^{3} \sum_{i=1}^{2} n_{ij} x_i y_j - xy$$

$$= 1/n \sum_{i=1}^{3} \sum_{j=2}^{2} (n_i x_i y_1 + n_{i2} x_i y_2) - xy$$

$$= 1/n (n_{11} x_1 y_1 + n_{12} x_2 y_1 + n_{22} x_2 y_2 + n_{31} x_3 y_1 + n_{32} x_3 y_2)$$

= 3/10 – 0.8 · 0.4 (calculs fait à l'ex 5 y = 0.4 x = 0.8)
= 0.3 – 0.32 = -0.02

La corrélation
Définition : La corrélation entre les variables x et y est définit par :

$$x (x ; y) = p (x ; y) = \frac{\text{cov}(x ; y)}{\sqrt{V(x) - V(y)}}$$

$$= \frac{\text{cov}(x ; y)}{\sigma(x) \cdot \sigma(y)}$$

Le coefficient de corrélation est le coefficient de corrélation de Bravers Pearson.
Propriété : le coefficient de corrélation est un paramètre symétrique.

3) Série statistique à indice simple.
On considère la série statistique double à indice simple (x_i, y_i n_i) associé aux variance X et Y observée sur une même population de taille n réécrivons les paramètres du § III) 2) pour cette série.
Propriété : Les moyennes arithmétique de X et de Y s'écrivent :

$$x = 1/n \sum_{i=1}^{p} n_i x_i \text{ et } y = 1/n \sum_{j=1}^{q} n_i y_j$$

La variance de X et de Y S'écrivent :

$$V(x) = 1/n \sum_{i=1}^{p} n_i (x_i - x)^2 = 1/n \sum_{i=1}^{p} n_i x_i^2 - x^2$$

$$V(y) = 1/n \sum_{i=1}^{q} n_i (y_i - y)^2 = 1/n \sum_{i=1}^{q} n_i y_j^2 - y^2$$

$j=1$ $j=1$

La variance entre X et Y s'écrit :

$$Cos(x ; y) = 1/n \sum n_i (x_i - x)(y_i - y)$$

$$= 1/n \sum n_i x_i y_i - xy.$$

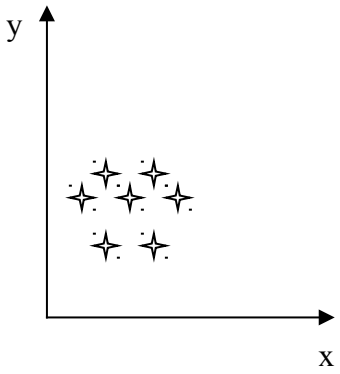
La propriété énoncé dans le § III) 2.2) (les résumés numériques) restent vraies pour les séries double à indice simple.

IV) Etude de la liaison entre 2 variables quantitatives

La régression
On considère la série statistique double $(x_i, y_i, n_i)_{i=1...p}$ observée pour les variables doubles quantitative X et Y sur une même population de taille n.

1) Approche graphique le nuage de points.

Le nuage de points est constitué des points $(x_i\ y_i)$ pour $_{i=1....7}$

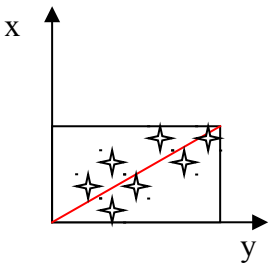


2) Régressions linéaires droites des moindres carrées.

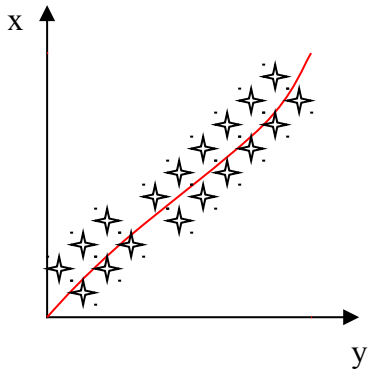
1) Le principe

Définition : la régression linéaire de Y en X consiste à déterminer la fonction affine et les valeurs $(a ; b) \in R^2$ qui rendait minimum la quantitative.

$$\sum_{i=1}^{p} n_i (y_i - (ax_1 + b))^2$$



Régression de X en Y $\sum n_i (x_i - (ay_i + b))^2$



2) Définition et propriété

La régression linéaire de Y en X par la méthode des moindres carrée conduisent aux coefficient approchés pour a et b.

$$\hat{a} = \frac{\operatorname{cov}(x\ y)}{V(x)}$$

$$b^{\wedge} = y - \bar{a}\ x$$

Propriété : Pour $(a, b, c, d) \in R^4$ tel que $(a, b) \in R^*$
On a :
$|2\ (ax + b\ ;\ cy + a)| = |r\ (x\ ;\ y)|$.
En effet on a :

$$r\ (ax + b)\ cy + d) = \frac{\operatorname{cov}(ax + b\ ;\ cy + d)}{\sigma\ (ax + b) \cdot (cy + d)}$$

$$= \frac{a\ c\ \operatorname{cov}(\ y\ x\ )}{|a|\ \sigma\ (x) \cdot |c|\ \sigma\ (y)}$$

$$= \frac{a\ c}{|ac|} \cdot n\ (x\ ;\ y) = \pm\ n\ (x\ ;\ y)$$

en fonction des signes de ac

Exercice 8
Le coefficient de corrélation de X et de Y est :

$$r\ (x\ ;\ y) = \frac{\operatorname{cov}(x\ ;\ y)}{\sqrt{V\ (x)\ V\ (y)}} = \frac{-0.02}{\sqrt{0.56 \cdot 0.24}} = \frac{-1}{4\sqrt{2}} < 0$$

Le rapport de corrélation :
Définition :
Le rapport de corrélation de X en Y est définit par :

$$\frac{\text{Variance expliquée de x}}{\text{Variance total de x}} = 1/n \sum_{i=1}^{q} n_{.j} (x_j - x)^2$$

Le rapport de corrélation de Y en X est définit par :

$$\frac{\text{Variance expliquée de y}}{\text{Variance total de y}} = \frac{1/n \sum_{i=1}^{p} n_{i.} (y_i - y)^2}{V(y)}$$

Propriété :
Quand les variables sont indépendantes, le rapport est nul (les variances des moyennes card. est réelle). Quand il existe une liaison fonctionnelle entre X et Y, la variance résiduelle est nulle et le rapport vaut 1.

Rq : Le rapport de corrélation correspond donc à la proportion de variance expiqué

Ex 9 (suite ex2) :
Le rapport de corrélation de X en Y est donné par :

$$\frac{1/n \sum n_{ij} (x_j - x)^2}{V(x)}$$

$$= \frac{1/10 (n_{.1} x_1^2 + n_{.2} x_2^2) - x^2}{V(x)}$$

$$= \frac{1/10 (6 (5/6)^2 + 4 (3/4)^2 - (0.8)^2}{0.56}$$

$$= \frac{1/10 (25/6 + 9/6) - 0.64}{0.56}$$

$$= \frac{\frac{50 + 27}{120} - 0.64}{0.56}$$

$= 0.2/120 \approx 0.00016.$

La droite de régression à pour équation :

$\hat{y} = \bar{a}x + b^{\wedge}$

Les valeurs estimées sont pour $_{i=1...p}$

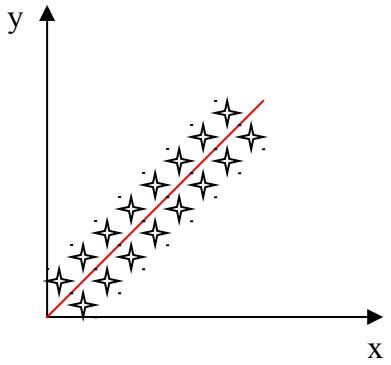
Propriété : $\hat{y}$ est une nouvelle variable et on a :

$$\begin{cases} \hat{y} = y \\ V(y) = 1/n \sum_{i=1}^{p} n_i (\hat{y}_i - y)^2 \end{cases}$$

$$= 1/n \sum_{i=1}^{p} n_i (y_i - y)^2$$

Définition : on appelle point moyen dans le cadre de la régression de Y en X le point de coordonnées (x ; y).

Propriété : Le point moyen est un point de la droite de régression de Y en X.

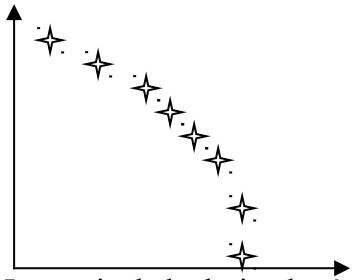


$y = âx + b^$

On a en effet $âx + b^ = âx + y - âx = y$ (x ; y) est donc un point de la droite de régression de Y en X.

Propriété : Les coefficients de la régression de Y en X et le coefficients de la régression de X en Y sont tel que :

â â' = R^2 (x ; y)
On a â (resp. â') est l'estimation du coefficient de l'équation y = ax + b (resp. x = ay + b').



Calcul de l'indice moyen sur 1 an

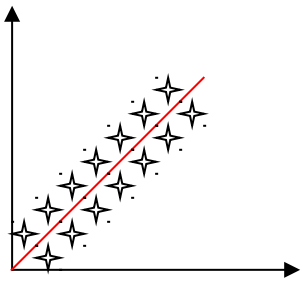
$$y = 1/n \sum_{i=1}^{p} n_i y_i$$

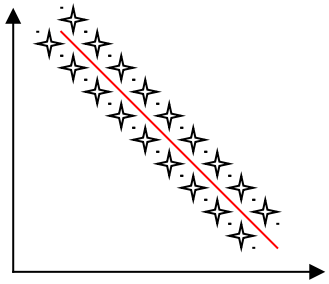
$$= 1/n \sum_{i=1}^{p} y_i \quad = 2170/10 = 217$$

Rq : La partie de la droite de régression est liée aux signes de la covariance :

$$â = \frac{cov(xy)}{\_\_\_\_\_} \quad \geq 0 \text{ si } cov(xy) \geq 0$$

V (x) ≤ 0 si cov (xy) ≤ 0.





â > 0 â < 0

3) Mesure de la qualité de la régression.

a) Variance expliquée et variance résiduelle dans le cadre de la régression.

Définition :
On appelle variance résiduelle ou variance des écarts de la quantité :

$$V_R(y) = 1/n \sum n_i (y_i - y_i)^2$$

On appelle variance expliquée la quantité :

$$V_C = (y) = 1/n \sum n_i (\hat{y} - y_i)^2$$

Propriété :
on a :

$$V_R(y) = \sigma(y)\, 1 - r^2(x ; y)$$
$$= V(x)\, 1 - r(x ; y)$$

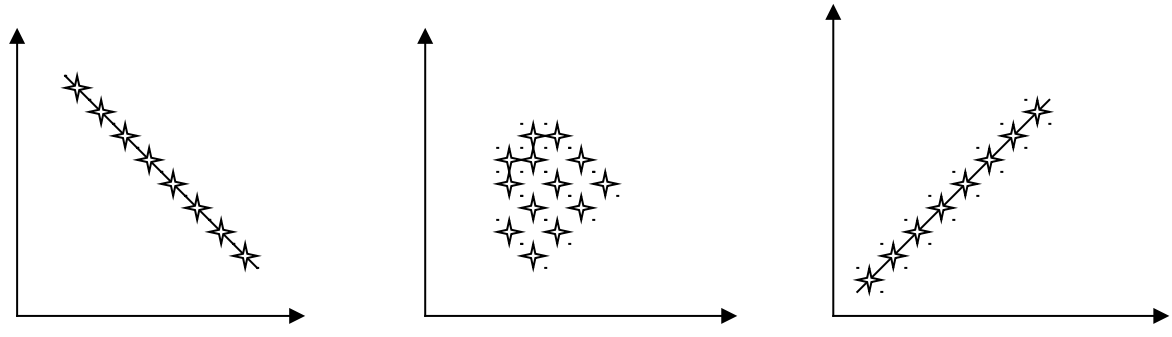
$$V_E(y) = V(y) - V_R(x) = V(y)\, r^2(x ; y)$$

b) Coefficient de corrélation

Propriété :
Pour 2 variances X et Y de coefficient de corrélation :

$$r(x ; y) = \frac{\text{cov}(xy)}{\sqrt{V(x)\, V(y)}}$$ est tel que $r(x ; y) \in [-1 ; 1]$.

La régression est considérée ĉ justifié si :
$|r(x ; y)| \approx 1$ c'est-à-dire $r(x ; y) \approx -1$ ou $r(x ; y) \approx 1$



$r(x ; y) \approx -1$ $\qquad$ $r(x ; y) \approx 0$ $\qquad$ $r(x ; y) \approx 1$

$$\hat{a} = \frac{\text{cov}(x ; y)}{V(x)} < 0 \text{ pente négative}$$

$$\hat{a} = \frac{\text{cov}(x ; y)}{V(x)} > 0 \text{ pente positive}$$

c) Erreur absolue

Définition : l'erreur absolue est liée à la variance résiduelle soit e tel que :

$$e^2 = 1/n \sum n_i (y_i - \hat{y}_i)^2$$
$$= V_R(y)$$

d) Erreur relative

Définition : l'erreur absolu est définit par :

$$\frac{\sigma(\hat{y})}{\sigma(y)}$$

Propriété :
on a :

$$\frac{\sigma(\hat{y})}{\sigma(y)} = |r(x ; y).$$

e) Coefficient de détermination

Définition :

On appelle coefficient de détermination le carré du coefficient de corrélation. on la note en général R^2.

Ex 12 (suite ex 10)

On a :

$$\frac{r(x ; y)}{\sqrt{V(x)\, V(y)}}$$

Avec

$$\text{cov}(x ; y) = 1/n \sum^{p} x_i y_i - xy$$

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

$$V(x) = 1/n \sum_{i=1}^{p} x_i^2 - x^2$$

$$V(y) = 1/n \sum_{i=1}^{p} y_i^2 - y_i^2$$

Avec les données on a

$$y = \frac{2170}{10} = 217$$

$$\text{cov}(x; y) = \frac{11275}{10} - 217 \cdot 5.5 = -66$$

$$V(x) = \frac{325}{10} = (5.5)^2 = 8.25$$

$$V(y) = \frac{476532}{10} = (217)^2 = 564.2 \text{ d'ou}$$

$$r(x; y) = \frac{-66}{\sqrt{8.25 \cdot 564.2}} \approx -0.96$$

ĉ $|r(x; y) \approx 1$ l'argument linéaire est envisageable.

Le coefficient de la régression sont $\hat{a} = \frac{\text{cov}(x ;y)}{V(x)}$ et $b^{\wedge} = y - \hat{a}x$

soit $\hat{a} = \frac{-66}{8.25} = -8$ et $b^{\wedge} = 217 + 8 \cdot 5.5 \approx 26$

L'équation de la droite de régression est $\hat{y} = -8x + 26$

Rq : On aurait pu considérer $e^2 = V_R(y) = V(y) - V_E(y)$.

• L'existence d'une corrélation n'implique pas toujours un lien de causalité.

4) Régression non linéaire

1) Ajustement d'une fonction puissance

On a $Y = \beta x^{a}$ (E_p)

Avec $\beta > 0$ et $x > 0$ or linéaire (E_p) et on a $\ln y = \ln \beta + \alpha \ln x$, on pose $Y' = \ln y$ et $x' = \ln x$.

$Y' = ax' + b$

Avec $a = \alpha$

$b = \ln \beta$

a et b sont estimé par $\begin{cases} \hat{a} = \dfrac{\text{cov}(x ; y)}{V(x)} \\ b^{\wedge} = y' - \hat{a}x' \end{cases}$

On déduit de â et $b^{\wedge}$ des estimation de α et β $\begin{cases} \alpha = \hat{a} \\ \beta = e^{b^{\wedge}} \end{cases}$

Ex 12

$Y = ax + b$

On estime a par :

$$\hat{a} = \frac{\text{cov}(x ; y)}{V(x)}$$

et b par $b^{\wedge} = y - \hat{a}x$

On obtient :

$x = 13.5 \quad y = 180$

$\text{cov}(x ; y) = 325.375$

$V(x) = 17.25$

$V(y) = 6928.75$

D'où $\begin{cases} \hat{a} \approx 18.86 \\ b^{\wedge} \approx 134.64 \end{cases}$

On a $\hat{y} = 18.86x + 134.64$

On considère le modèle puissance $y = bx^{a}$

On linéaire et on a $\ln y = \ln b + a \ln x$

On pose :

$\begin{cases} y' = \ln y \\ b' = \ln b \\ a' = \ln a \\ x' = \ln x \end{cases}$

D'où $y' = a'x + b'$

On estime a' et b' par

$$\hat{a}' = \frac{\text{cov}(x' ; y')}{V(x)} \quad \text{et} \quad b^{\wedge}{}' = y - \hat{a}'x'$$

On obtient $x' = 1/n \sum_{i=1} x_i = 2.55$

$$y' = 1/n \sum_{i=1}^{P} y_i = 4.5$$

$$\text{cov}(x' ; y') = 1/n \sum_{i=1}^{p} x_i y_i - xy \approx 0.2231$$

$$V(x') = 1/n \sum_{i=1}^{P} x_i'^2 - x_i'^2 = 0.12$$

$V(y') \approx 0.48$

On aboutit $\begin{cases} \hat{a} \approx 2.017 \\ b^{\wedge} \approx -0.59 \end{cases}$

Comme $\begin{cases} a = a' \\ b = b' \end{cases}$

On a $\begin{cases} a = a' \\ b = e^{b} \end{cases}$

D'où les estimations :
$\hat{a} \approx 2.017$
$b^{\wedge} \approx e^{b^{\wedge}'} \approx e^{-0.59} \approx 0.55$

On a :

$\hat{Y} = b^{\wedge} \cdot \hat{a} = 0.55 \cdot 2.017$

Pour déterminer le meilleur modèle, on considère l'erreur absolue.

Ces linéaires $e^2 = 1/n \sum_{i=1}^{p} (y_i - y_i)^2$ avec

$\hat{y}_i = \hat{a} x_i + b^{\wedge}$
$\hat{a} \approx 18.86$ $b^2 \approx -134.44$

Ces puissances $e^2 - 1/n \sum_{i=1}^{p} (y_i - y_i)^2$

avec $\hat{y}_i = b^{\wedge} x_i^{\hat{a}}$
$\hat{a} \approx 2.017$
$b^{\wedge} \approx 0.55$

On obtient $e^2 \approx \dfrac{6331.384}{8}$ et $e^2 \approx \dfrac{4372.59}{8}$

Soit $e'^2 < e^2$ la meilleur modèle est le modèle puissance.
Utilisons le modèle puissance pour donner une valeur ou a $\hat{y}_0 = b\hat{} \; x_0 \; \hat{a} \approx 0.55 \cdot 26^{2.017}$ ≈ 2

2) Ajustement d'une fonction exponentielle.

$Y = e^{\alpha x + \beta}$
$Y' = \ln y = \alpha^x + \beta$

3) Ajustement d'une fonction logarithme
$Y = \ln (\alpha x + \beta)$
$e^y = \alpha x + \beta$ avec $ax + b > 0$

$Y = ax + b$ $Y = ax^{0.5} + b$
$Y = a \ln x + b$

Modèle 1

$$\hat{a} = \frac{cov(x;y)}{V(x)} \qquad b\hat{} = \hat{y} = \hat{a}x$$

Modèle 2

$$\hat{a} = \frac{cov(x;y)}{V(x^{0.5})} \quad \text{avec } X' = x^{0.5}$$
$b\hat{} = y - \hat{a}x'$

Modèle 3

$$\hat{a} = \frac{cov(x''y)}{V(x'')} \quad \text{avec } x'' = \ln x$$

On obtient :

$\hat{a} = 0.31$ $\quad b\hat{} = 533.5$
$\hat{a}' \approx 27.68$ $\quad b\hat{}' = 1208.35$
$\hat{a}'' \approx 614.88$ $\quad b\hat{}'' = -4644.38$

⇨ Recherche du meilleur modèle

$$e^2 = 1/n \sum_{i=1}^{p} (y_i - \hat{y}_i)^2 \text{ avec } \hat{y}_i = \hat{a}x_i + b^2$$

$$e^{2'} = 1/n \sum_{i=1}^{p} (y_i - \hat{y}_1')^2 \text{ avec } \hat{y}_2' = \hat{a}x_i^{0.5} + b\hat{}'$$

$$e^{2''} = 1/n \sum_{i=1}^{P} (y_i - y_i'')^2 \text{ avec } \hat{y}_i'' = \hat{a}x + b\hat{}$$

2.18 2.18 2.07

On obtient $e^2 \approx \frac{\quad}{6}$ $e^{2'} \approx \frac{\quad}{6}$ $e^{2''} \approx \frac{\quad}{6}$

Le meilleur modèle est n_o Y = a ln x + b

⇨ Prévision pour $x_0 = 2020$

$\hat{y}_0'' = \hat{a}'' \ln(x_0) + \hat{b}'' \approx 35.39$