Chapitre 3

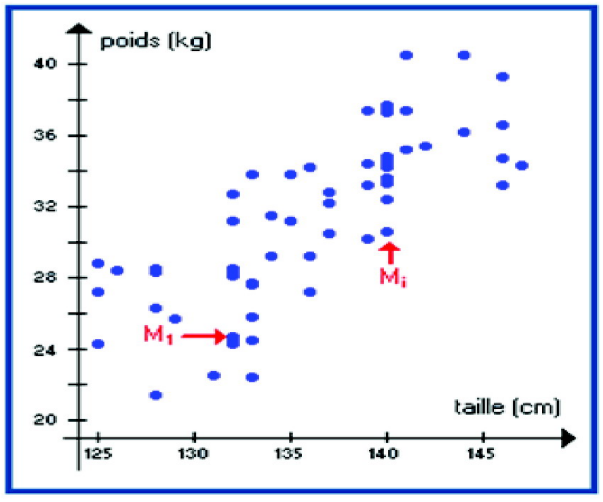
Analyse bivariée

Plan

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

Ici, on a ce qu'on appelle données groupées.

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:

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

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:

$X \diagdown Y$	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				
$n_{.j}y_j$	$n_{.1}y_1$	$\cdots$	$n_{.q}y_q$					
$n_{.j}y_j^2$	$n_{.1}y_1^2$	$\cdots$	$n_{.q}y_q^2$					
$\sum_{i=1}^{p} n_{ij}x_i$		$\cdots$						
$y_j\sum_{i=1}^{p} n_{ij}x_i$		$\cdots$						

Calculs

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:**

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

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

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 $\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)$

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

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:
$X' = \frac{X - c}{d}$ de et $Y' = \frac{Y - c'}{d'}$
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

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

fréquence conditionnelle de x_i sachant y_j (y_j réalisé) 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

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

3- 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.