Tableau de contingence

yi

xi nij (effectif partiel)

Distribution marginales :

$$ni. = \sum_{j=1}^{q} nij$$

effectifs marginaux

$$n.j = \sum_{i=1}^{p} nij$$

Exemple : On considère une distribution avec 10 000 jeunes salariés en fonction de l'âge, variable x, et selon le salaire mensuel noté y et en milliers d'euros.

On a le tableau de contingence suivant:

y x	[0,8-1,2[	[1,2-1,6[	[1,6-2,0[	Σ(distributions marginales de x)
19-21	700	300	100	1100 **(n1.)**
21-23	1400	2000	300	3700
23-25	1000	2900	1300	5200
Σ (distributions marginales de y)	3100	5200	1700	**n=**10 000

On a : n12 = 300 ⟶ jeunes qui ont entre 19 et 21 ans et dont le salaire est compris [1,2-1,6[

…

n21 = 1400

n32 = 2900 (effectifs partiels)

On a aussi les fréquences partielles :

$$f32 = \frac{n32}{n} = \frac{2900}{10\,000} = 29\%$$

Distribution marginale de x :

n1. = 1100 nombre de jeunes qui ont entre 19 et 21 ans

n2.= 3700

n3. = 5200

Fréquence marginale :

$$f1. = \frac{n1.}{n} = \frac{1100}{10\,000}$$

Distribution marginale de y :

n.1 = 3100 nombre de jeunes dont le salaire est compris entre 0,8 et 1,2 milliers d'euros

n.2= 5200

n.3 = 1700

$$f.1 = \frac{3100}{10\,000} = 31\%$$

Remarques :

$$\sum_{i=1}^{p}\sum_{j=1}^{p} fij = 1$$

car $fij = \frac{nij}{n}$

$$\sum_{i=1}^{p} fi. = 1$$

car $\sum_{i=1}^{p} ni. = n$

$$\sum_{j=1}^{p} f.j = 1$$

car $\sum_{j=1}^{q} n.j = n$

c. Distribution conditionnelles :

Ce sont les distributions d'un caractère pour une modalité fixée de l'autre.
Par exemple :

- distribution conditionnelle de x pour y=yj , ce sont (xi, effectif nij) qui correspondent à la "jème" colonne du tableau.
- distribution conditionnelle de y pour x=xj , ce sont (yi, effectif nij) qui correspondent à la "ième" ligne du tableau.
- Fréquence conditionnelle :

$$fi/_j = \frac{nij}{n.j}$$

nij : nombre d'unités ayant le caractère (xi, yj)
n.j : nombre d'unités ayant le caractère yj
et aussi :

$$fj/_i = \frac{nij}{ni.}$$

On a fi/j = proportion d'unités ayant xi parmi ceux ayant yj
= fréquence conditionnelle de x selon y

Exemple des salariés :

$$fi = 1/_j = 2 = \frac{n12}{n.2} = \frac{300}{5200} = 5{,}8\%$$

58% de ceux qui gagnent entre 1,2 et 1,6 milliers d'euros ont entre 19 et 21 ans
Propriété :

$$fi/_j = \frac{fij}{f.j} \quad ou \quad fj/_i = \frac{fij}{fi.}$$

Remarque:

$$\sum_{i=1}^{p} fi/_j = 1 \quad et \quad \sum_{j=1}^{q} fj/_i = 1$$

2. Caractéristiques des séries à 2 caractères :

a. Caractéristiques des séries marginales : (xi , ni.) et (yj , n.j)

Moyenne de x :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} xi\ ni. = \frac{1}{n}\sum_{j=1}^{q} xi\ nij$$

Moyenne de y :

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{q} yj\ n.j = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{p} yj\ nij$$

Exemple : tableau des 10 000 jeunes salariés

$$\bar{x} = \frac{1}{10\,000}(20 * 1100 + 22 * 3700 + 24 * 5200) = 22{,}82$$

20 , 22 et 24 : centres des classes

$$\bar{y} = \frac{1}{10\,000}(1{,}0 * 3100 + 1{,}4 * 5200 + 1{,}8 * 1700) = 1{,}344$$

Variance de x :

$$\sigma x^2 = \frac{1}{n}\sum_{i=1}^{p} ni.(xi - \bar{x})^2 = \overline{x^2} - (\bar{x})^2$$

$$\sigma y^2 = \frac{1}{n}\sum_{j=1}^{q} n.j(yi - \bar{y})^2 = \overline{y^2} - (\bar{y})^2$$

Exemple : tableau des 10 000 jeunes salariés

$$\overline{x^2} = \frac{1}{10\,000}\,(20^2 * 1100 + 22^2 * 3700 + 24^2 * 5200) = 522{,}6$$

d'où

$$\sigma x^2 = 522{,}6 - (22{,}82)^2 = 1{,}85$$
$$\sqrt{\sigma x^2} = \sqrt{1{,}85} = 1{,}36$$

$$\overline{y^2} = \frac{1}{10\,000}\,(1{,}0^2 * 3100 + 1{,}4^2 * 5200 + 1{,}8^2 * 1700) = 1{,}88$$

d'où

$$\sigma y = 0{,}27$$

b. Caractéristiques des distributions conditionnelles de x si yj et de y si xi :

Moyennes :

de x si yj, $\bar{x}j = \frac{1}{n.j}\sum_{i=1}^{p} xi\ nij$ d'après "jème" colonne

de y si yj, $\bar{y}\imath = \frac{1}{ni.}\sum_{j=1}^{q} yj\ nij$ d'après "ième" colonne

Variance de x si yj :

$$\sigma j^2(x) = \frac{1}{n.j}\sum_{i=1}^{p} nij(xi - \bar{x}j)^2 = \overline{xj^2} - (\bar{x}j)^2$$

Variance de y si xj :

$$\sigma i^2(y) = \frac{1}{ni.}\sum_{j=1}^{q} nij(yj - \bar{y}\imath)^2 = \overline{y\imath^2} - (\bar{y}\imath)^2$$

⇒ Démonstration en TD

Exemples : tableau des salariés

$$\overline{x1} = \frac{1}{3100}\,(20 * 700 + 22 * 1400 + 24 * 1000) = 22{,}194$$

Moyenne d'âge de ceux qui gagnent entre 0,8 et 1,2 milliers d'euros

$$\sigma 1^2(x) = \frac{1}{3100}\,(20^2 * 700 + 22^2 * 1400 + 24^2 * 1000) - (22{,}194)^2 = 2{,}15$$
$$\sigma 1(x) = 1{,}47$$

σ1(x) représente la dispersion d'âge de ceux qui gagnent entre 0,8 et 1,2 milliers d'euros (j=1)

$$\overline{y2} = \frac{1}{3700}\,(1{,}0 * 1400 + 1{,}4 * 200 + 1{,}8 * 300) = 1{,}281$$

i=200 , x=x2

Il s'agit du salaire moyen de ceux qui ont entre 21 et 23 ans.

$$\sigma 2^2(y) = \frac{1}{3700}\,(1{,}0^2 * 1400 + 1{,}4^2 * 2000 + 1{,}8^2 * 300) - (1{,}281)^2 = 0{,}0596$$
$$\sigma 2(y) = 0{,}24$$

σ2(y) représente la dispersion du salaire de ceux qui ont entre 21 et 23 ans

c. Relations entre les caractéristiques marginales et conditionnelles

Sur les moyennes :

On a calculé les $\bar{x}j$ pour y=yj fixé et d'effectif n.j alors :

$$\boxed{\bar{x} = \frac{1}{n}\sum_{j=1}^{q} \bar{x}j\ \ n.j}$$

Idem pour y

d. Covariance de x et y

Définition :

$$cov\,(x,y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} nij\,(xi-\bar{x})(yj-\bar{y})$$

Cette notion de covariance sert à savoir si les caractères x et y sont dépendants ou indépendants.

Propriété :

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} nij\,xiyj - \bar{x}\bar{y}$$

$$= \bar{x}\bar{y}$$

donc :

$$cov(x,y) = \overline{xy} - \bar{x}\bar{y}$$

Remarque :

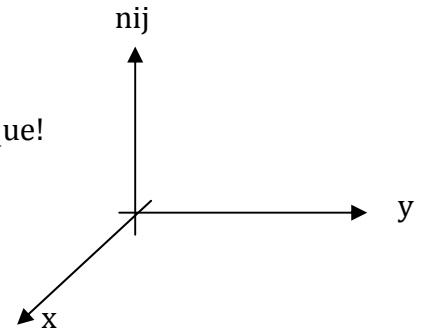
$$cov(x,x) = \overline{x^2} - (\bar{x})^2 = \sigma x^2$$

3. Représentation graphique des séries à 2 caractères :

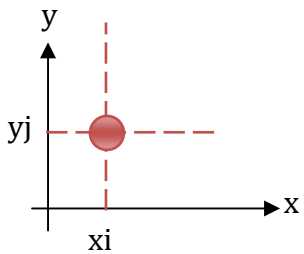
On considère une série (xi, yj) d'effectifs nij.

- Graphe à 3 dimensions:

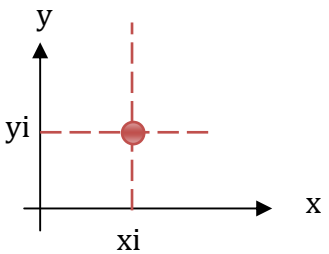
Représentation peu pratique!



- Graphe à 2 dimensions:



point dont la grosseur est proportionnelle à l'effectif ni j

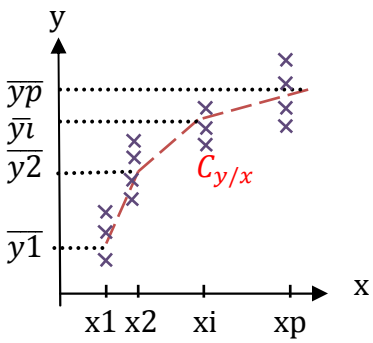


avec effectif plus faible

⟶ courbe de régression

ou encore lignes polygonales

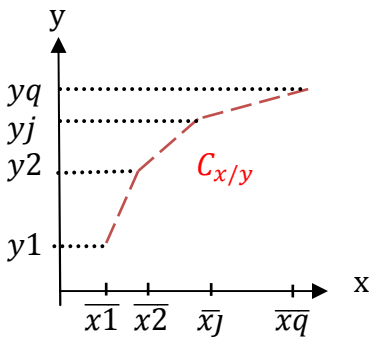
moyennes conditionnelles



Valeurs de y

Ici , $\overline{yi}$ syntétise l'information des (yj,nij)

ET



Ici , $\overline{xj}$ syntétise l'information des (xi,nij)

Que représente ces courbes de régression?
On peut dire que la Σ du carré des distances des points à $C_{y/x}$ ou $C_{x/y}$ est minimale!

Si y est en fonction de x , × , on obtient alors un nuage de points.
On va chercher la meilleure courbe passant au milieu de ce nuage de points.
Cette courbe sera notée $y'i = f(xi)$

Donc, on doit avoir la somme S des distances au carré aux points yj minimale (distance y'i , yj minimale).

$$S = \sum_i \sum_j nij \ (yj - y'i)^2$$

On veut que S soit le plus petit possible!
D'où : ... développement (la flemme) ...
S est minimale si :

$$y'i = \overline{yi}$$

⇒ courbe de régression.

4. Série à caractères x et y indépendants :

Définition : Les caractères sont indépendants si une variation de l'un n'entraîne pas une variation de l'autre.

Conséquences : Les fréquences conditionnelles $f_{i/j}$ ne dépendent pas j ($f_{i/j} = \frac{nij}{n.j}$)
Les fréquences conditionnelles $f_{j/i}$ ne dépendent pas i ($f_{j/i} = \frac{nij}{n.i}$)
$f_{i/j} = f_{i.}$ et $f_{j/i} = \boldsymbol{f_{.j}}$
⟶fréquence marginale

Démonstration :

$$f_{i/j} = \frac{nij}{n.j} = \frac{ni1}{n.1} = \frac{ni2}{n.2} = \frac{ni3}{n.3} = \cdots = \frac{niq}{n.q} = \frac{ni.}{n} = f_{i.}$$

On a vu que

$$f_{i/j} = \frac{fij}{f.j}$$

donc

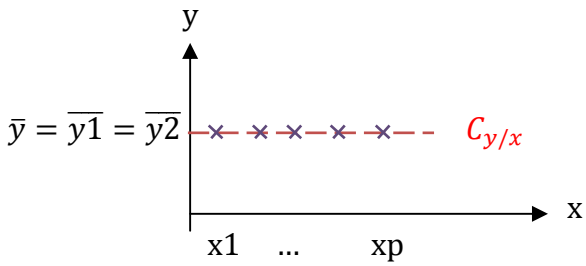
$$\boxed{fij = fi. \ * f.j}$$

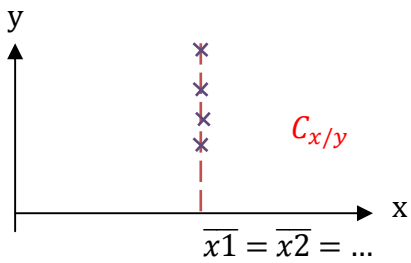
Si indépendant.

Connaissance de fij ⇒ connaissance de fi. et f.j
⇐ réciproque juste si x et y indépendants

On a :
$\bar{x} = \bar{x}j$ et $\bar{y} = \bar{y}i$

Conséquence pour les courbes de régression :





Les courbes de régression sont des droites //

On a $\overline{xy} = \bar{x}\,\bar{y}$
d'où $\text{cov}(x,y) = 0$

Remarque : Si $\text{cov}(x,y) = 0$, cela n'entraîne pas forcément l'indépendance de x et de y

Exemple : Soit la relation $y=x^2$ (pas dépendante)
où $x = -1$, $x = 1$, $y = +1$

x/y	+1	Σ
-1	1	1
+1	1	1
Σ	+2	2

$$\bar{x} = -1 * \frac{1}{2} + 1 * \frac{1}{2} = 0$$

$$\overline{xy} = \frac{1}{n}\sum_i \sum_j nij\ xi\ yi$$

$$= \frac{1}{2}(-1 * (+1) + 1 * 1) = 0$$

donc $\overline{xy} - \bar{x}\,\bar{y} = 0$
$\Rightarrow$ $cov(x,y) = 0$

5. Ajustement, corrélation et régression

Objectifs :

- évaluer quantitativement le degré d'indépendance entre x et y
- savoir si il existe une dépendance linéaire entre x et y $(y = ax + b)$

a. Ajustement linéaire – Droites de régression

Problème :

- si visuellement le "nuage" de points fait apparaître une droite, les courbes de régression ("meilleures" courbes représentant le nuage) sont les lignes brisées autour de droites que l'on va déterminer avec la méthode des moindres carrés (MMC).

On va parler de droites de régression

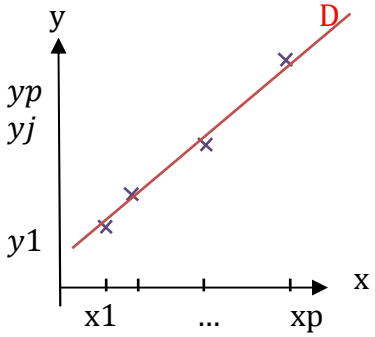
- On pourra alors determiner l'angle entre ces deux droites de régression qui va quantifier le degré de dépendance entre x et y :

Si angle de 90° : x et y indépendantes

Si angle de 0° : on vera que y est fonction de x

d'où les graphes dans les deux cas:

1er cas :

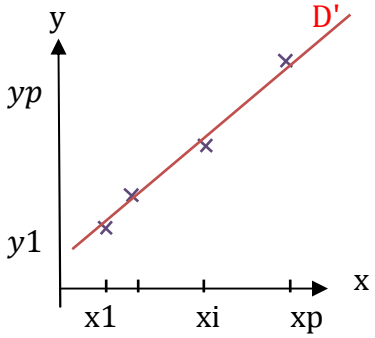


On va chercher y =ax + b

Cela va remplacer la courbe de régression $C_{x/y}$

D : droite de régression de y en x

2ème cas :



On va chercher x =a'y + b'

D' : droite de régression de x en y

- MMC dans le cas 1 :

On veut minimiser la grandeur :

$$S = \sum_i \sum_j nij\ (yj - axi - b)^2$$

$\frac{dS}{db} = 0 \quad \Rightarrow$

$$\sum_i \sum_j nij\ (yj - axi - b) = 0$$

$$\sum_i \sum_j nij\ yj = a \underbrace{\sum_i \sum_j nij\ xi}_{=n\bar{x}} + b \underbrace{\sum_i \sum_j nij}_{=n}$$

$$n\bar{y} = a\, n\, \bar{x} + nb$$

$$\bar{y} = a\, \bar{x} + b$$

On sait que la droite de régression passe par $(\bar{x}, \bar{y})$

$\frac{dS}{da} = 0 \quad \Rightarrow$

$$\sum_i \sum_j nij\ (yj - axi - b)\, xi = 0$$

$$n\, \overline{xy} = a\, n\, \overline{x^2} + nb\bar{x}$$

$$\overline{xy} = a\, \overline{x^2} + b\bar{x}$$

d'où

$$a = \overline{xy} - \frac{\bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{cov(x,y)}{\sigma x^2}$$

et

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

- MMC dans le cas 2 :

x en fonction de y

On obtient pour D' :
x = a'y + b'
On a alors

$$a' = \frac{cov(x,y)}{\sigma y^2}$$

et

$$b' = \bar{x} - a'\bar{y}$$