série de n couples (x_i, y_j) avec $i = 1$ à p et $j = 1$ à q : - d'effectifs partiels notés n_{ij}

(si caractères continus : x_i et y_j = centres de classes) - de fréquences partielles : $f_{ij} = \dfrac{n_{ij}}{n}$

1. Distributions marginales - distributions conditionnelles

a) *tableau de contingence*

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij}$$

$$n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$$

$x \backslash y$	y_1	y_j	y_q	somme
x_1	n_{11}	n_{1j}	n_{1q}	$n_{1\bullet}$
x_i	n_{i1}	n_{ij}	n_{iq}	$n_{i\bullet}$
x_p	n_{p1}	n_{pj}	n_{pq}	$n_{p\bullet}$
somme	$n_{\bullet 1}$	$n_{\bullet j}$	$n_{\bullet q}$	n

b) *distributions marginales*

- distribution marginale de x : valeurs x_i, effectif marginal de $x_i = n_{i\bullet}$
- distribution marginale de y : valeurs y_j, effectif marginal de $y_j = n_{\bullet j}$
- fréquence marginale de x_i : $f_{i\bullet} = \dfrac{n_{i\bullet}}{n}$ et fréquence marginale de y_j : $f_{\bullet j} = \dfrac{n_{\bullet j}}{n}$

$$n = \sum_{i=1}^{p} n_{i\bullet} = \sum_{j=1}^{q} n_{\bullet j} = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} \qquad \text{et} \qquad \sum_{i=1}^{p} f_{i\bullet} = \sum_{j=1}^{q} f_{\bullet j} = \sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = 1$$

c) *distributions conditionnelles*

- distribution conditionnelle de x pour $y = y_j$: valeurs x_i, effectif = n_{ij} (j^ème^ colonne du tableau)
- distribution conditionnelle de y pour $x = x_i$: valeurs y_j, effectif = n_{ij} (i^ème^ ligne du tableau)
- fréquence conditionnelle de x_i si $y = y_j$: $f_{i|j} = \dfrac{n_{ij}}{n_{\bullet j}}$ (% d'unités ayant x_i parmi ceux ayant y_j)
- fréquence conditionnelle de y_j si $x = x_i$: $f_{j|i} = \dfrac{n_{ij}}{n_{i\bullet}}$ (% d'unités ayant y_j parmi ceux ayant x_i)
- *propriétés* : $f_{ij} = f_{i|j} \times f_{\bullet j}$, $f_{ij} = f_{j|i} \times f_{i\bullet}$, et $\displaystyle\sum_{i=1}^{p} f_{i|j} = \sum_{j=1}^{q} f_{j|i} = 1$

2. Caractéristiques des séries à 2 caractères

a) *caractéristiques des distributions marginales*

- moyenne de x : $\bar{x} = \displaystyle\sum_{i=1}^{p} x_i\, f_{i\bullet} = \sum_{i=1}^{p}\sum_{j=1}^{q} x_i\, f_{ij}$ moyenne de y : $\bar{y} = \displaystyle\sum_{j=1}^{q} y_j\, f_{\bullet j} = \sum_{i=1}^{p}\sum_{j=1}^{q} y_j\, f_{ij}$
- variance de x : $\sigma_x^2 = \displaystyle\sum_{i=1}^{p} (x_i - \bar{x})^2\, f_{i\bullet} = \overline{x^2} - \bar{x}^2$ et de y : $\sigma_y^2 = \displaystyle\sum_{j=1}^{q} (y_j - \bar{y})^2\, f_{\bullet j} = \overline{y^2} - \bar{y}^2$

b) *caractéristiques des distributions conditionnelles*

- moyenne de x si y_j : $\bar{x}_j = \sum_{i=1}^{p} x_i \; f_{i|j}$ moyenne de y si x_i : $\bar{y}_i = \sum_{j=1}^{q} y_j \; f_{j|i}$

variance de x si y_j: $\sigma_j^2(x) = \sum_{i=1}^{p} (x_i - \bar{x}_j)^2 \; f_{i|j} = \overline{x_j^2} - \bar{x}_j^{\,2}$ et de y si x_i : $\sigma_i^2(y) = \sum_{i=1}^{p} (y_j - \bar{y}_i)^2 \; f_{j|i} = \overline{y_i^2} - \bar{y}_i^{\,2}$

c) *relations entre les caractéristiques marginales et conditionnelles*

- sur les moyennes : $\bar{x} = \sum_{j=1}^{q} \bar{x}_j \; f_{\bullet j}$ $\bar{y} = \sum_{i=1}^{p} \bar{y}_i \; f_{i\bullet}$
- sur les variances : $\sigma_x^2 = \sum_{j=1}^{q} \sigma_j^2(x) \; f_{\bullet j} + \sum_{j=1}^{q} (\bar{x}_j - \bar{x})^2 \; f_{\bullet j}$ et $\sigma_y^2 = \sum_{i=1}^{p} \sigma_i^2(y) \; f_{i\bullet} + \sum_{i=1}^{p} (\bar{y}_i - \bar{y})^2 \; f_{i\bullet}$

(moyenne des variances conditionnelles + variance des moyennes conditionnelles)

d) *covariance de x et y*

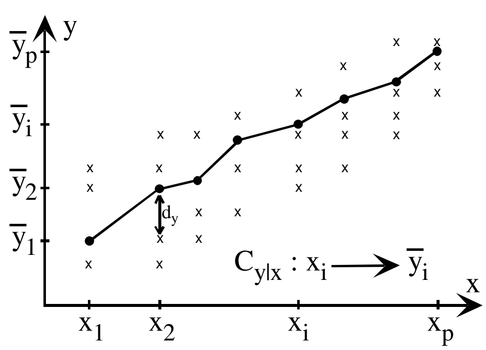
$$\mathrm{Cov}(x, y) = \sum_{i=1}^{p} \sum_{j=1}^{q} (x_i - \bar{x})(y_j - \bar{y}) \; f_{ij} = \sum_{i=1}^{p} \sum_{j=1}^{q} x_i \; y_j \; f_{ij} - \bar{x} \; \bar{y} = \overline{xy} - \bar{x} \; \bar{y}$$

$$\mathrm{Cov}(x, x) = \overline{x^2} - \bar{x}^2 = \sigma_x^2$$

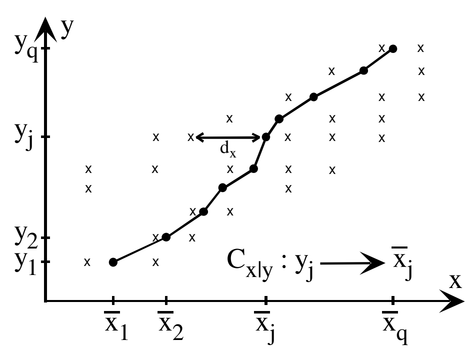
3. <u>Représentation graphique des séries à 2 caractères</u>

courbes de régression
$C_{y|x}$ et $C_{x|y}$:

"meilleures" courbes (polygonales) passant au milieu des points



somme des d_y^2 minimale



somme des d_x^2 minimale

4. <u>Séries à caractères *x* et *y* indépendants</u>

$$f_{i|j} = f_{i\bullet} \text{ et } f_{j|i} = f_{\bullet j}$$

d'où les conséquences :

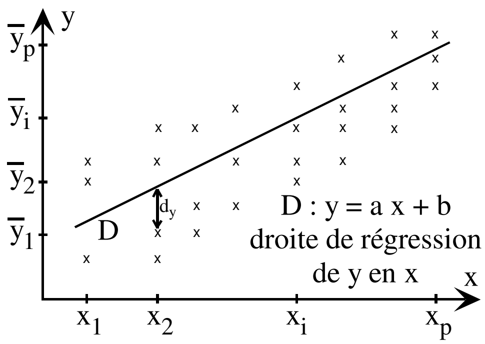
- $f_{ij} = f_{i\bullet} \times f_{\bullet j}$ les fréquences marginales de x et y donnent la fréquence partielle de (x_i, y_j)
- $\bar{x} = \bar{x}_j$ et $\bar{y} = \bar{y}_i$ la courbe de régression $C_{x|y}$ est verticale, $C_{y|x}$ est horizontale
- $\overline{xy} = \bar{x} \; \bar{y}$ d'où Cov(x,y) = 0 (attention : si Cov(x,y) = 0, x et y peuvent être dépendants)

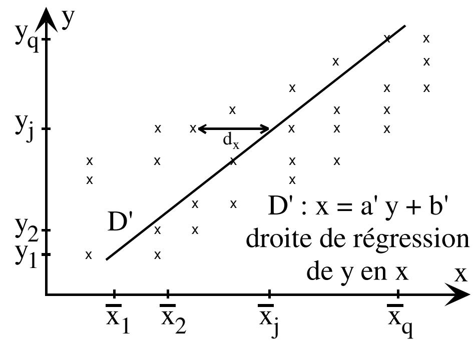
5. <u>Ajustement, corrélation, régression</u>

a) *ajustement linéaire - droites de régression*

- *droites de régression* $C_{y|x}$ et $C_{x|y}$:

"meilleures" droites passant au milieu des points





$$a = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{Cov}(x, y)}{\sigma_x^2}\ ,\ b = \bar{y} - a\bar{x}$$

a = coefficient de régression de y en x

$$a' = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{y^2} - \bar{y}^2} = \frac{\mathrm{Cov}(x, y)}{\sigma_y^2}\ ,\ b' = \bar{x} - a'\,\bar{y}$$

a' = coefficient de régression de x en y

- angle entre D et D' : $\theta = \mathrm{Arc\,tan}\dfrac{1}{a'} - \mathrm{Arc\,tan}\, a$
- θ mesure le degré de corrélation linéaire entre x et y :
  - $\theta = 0$: D et D' confondues : x et y liés par une fonction linéaire
  - $\theta = \pi/2$: D et D' perpendiculaires : x et y indépendants
- règle : $0 \leq a\,a' \leq 1$ a pour conséquence que la droite D' est toujours plus pentue que D

b) *coefficient de corrélation linéaire*

- $R = \dfrac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x\,\sigma_y} = \dfrac{\mathrm{Cov}(x, y)}{\sigma_x\,\sigma_y}$ avec $-1 \leq R \leq 1$, R mesure le degré de corrélation linéaire entre x et y
- $R = a\,a'$: donc l'angle entre D et D' mesure aussi ce degré de corrélation linéaire

c) *corrélation non linéaire*

- rapport de corrélation de y en x :

$$\eta^2_{y|x} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(y_j - \bar{y})^2}$$

avec : $0 \leq \eta^2_{y|x} \leq 1$

- rapport de corrélation de x en y :

$$\eta^2_{x|y} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(\bar{x}_i - \bar{x})^2}{\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(x_i - \bar{x})^2}$$

avec : $0 \leq \eta^2_{x|y} \leq 1$

- ces rapports mesurent le degré de corrélation directement sur les courbes de régression $C_{y|x}$ et $C_{x|y}$
- plus ces rapports sont proches de 1, plus la liaison est forte entre x et y
- si $\eta^2_{y|x} = 0$: alors $\bar{y} = \bar{y}_i$, $C_{y|x}$ est horizontale, y ne dépend pas de x (en général)
- si $\eta^2_{x|y} = 0$: alors $\bar{x} = \bar{x}_j$, $C_{x|y}$ est verticale, x ne dépend pas de y (en général).