INSA 3TC

Statistique descriptive

Aimé Lachal 1999/2000

Ordonnancement des données

individu i \ caractère j	1	$\dots$	j	$\dots$	p
1	x_{11}	$\dots$	x_{1j}	$\dots$	x_{1p}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	x_{i1}	$\dots$	x_{ij}	$\dots$	x_{ip}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	x_{n1}	$\dots$	x_{nj}	$\dots$	x_{np}

Tableau individus $\times$ caractères

i	1	$\dots$	i	$\dots$	n
x_i	x_1	$\dots$	x_i	$\dots$	x_n

Cas univarié

i	1	$\dots$	i	$\dots$	n
x_i	x_1	$\dots$	x_i	$\dots$	x_n
y_i	y_1	$\dots$	y_i	$\dots$	y_n

Cas bivarié

Analyse univariée

- Observations : $x_1, \ldots, x_n$

individu i	1	$\ldots$	i	$\ldots$	n
caractère x_i	x_1	$\ldots$	x_i	$\ldots$	x_n

Individus $\times$ *caractères*

- Observations ordonnées : $x_{(1)} \leqslant \cdots \leqslant x_{(n)}$ (*avec* répétition), $x'_1 < \cdots < x'_p$ $(p \leqslant n)$ (*sans* répétition)

effectif k	n_1	$\ldots$	n_k	$\ldots$	n_p
caractère x'_k	x'_1	$\ldots$	x'_k	$\ldots$	x'_p

Série statistique

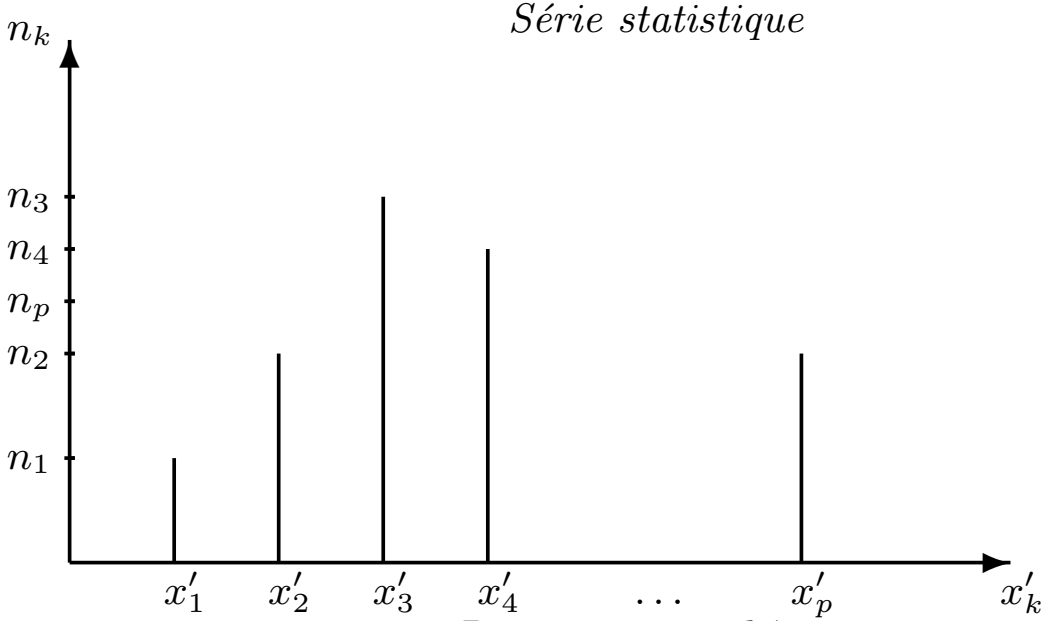
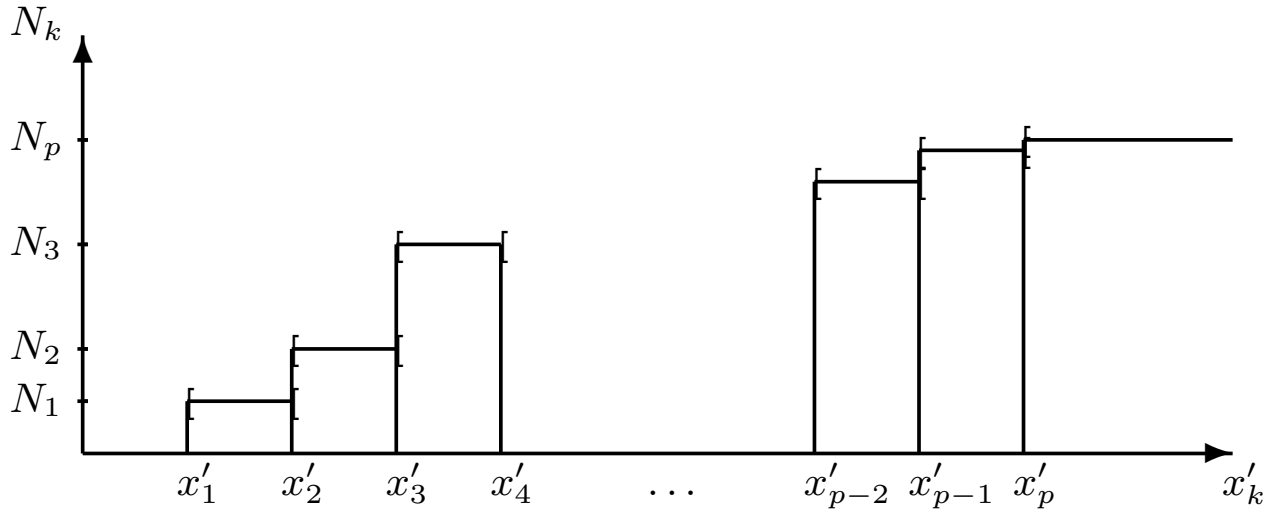


Diagramme en bâtons

Effectifs cumulés, fréquences cumulées

- Effectifs cumulés à gauche : $\boxed{N_k = \sum_{\ell=1}^{k} n_\ell}$

- Fréquences cumulées à gauche : $\boxed{F_k = \sum_{\ell=1}^{k} \frac{n_\ell}{n}}$

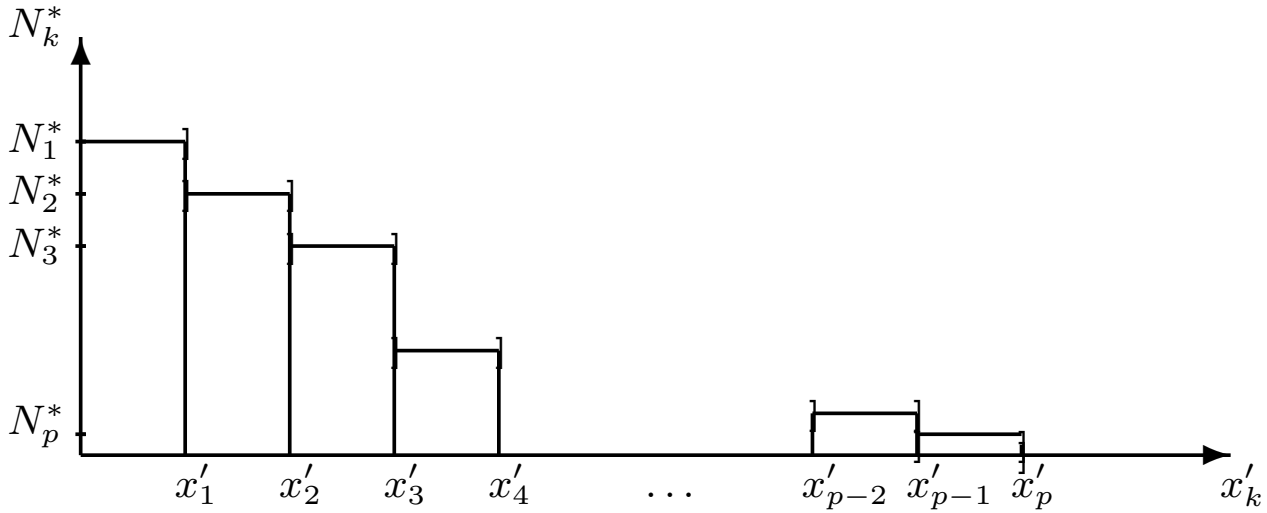


Courbe cumulative à gauche $y = N(x)$

Effectifs cumulés, fréquences cumulées

- Effectifs cumulés à droite : $\boxed{N_k^* = \sum_{\ell=k}^{p} n_\ell}$

- Fréquences cumulées à droite : $\boxed{F_k^* = \sum_{\ell=k}^{p} \frac{n_\ell}{n}}$

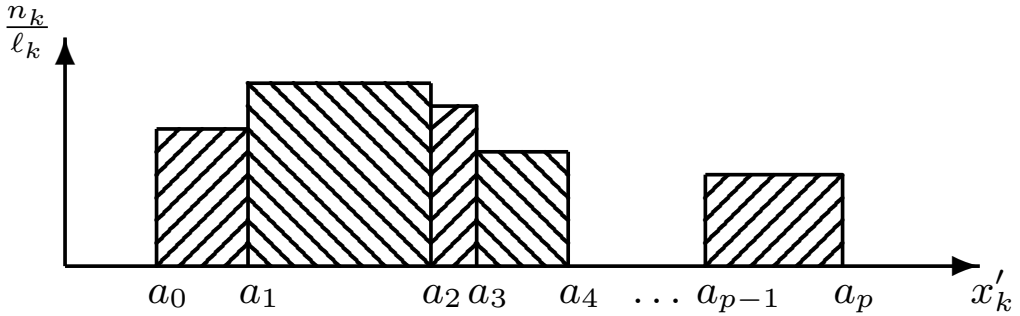


Courbe cumulative à droite $y = N^*(x)$

$$\boxed{N_k(x) + N_k^*(x) = \begin{cases} n & \text{si } x \notin \{x'_1, \ldots, x'_p\} \\ n + n_k & \text{si } x = x'_k \end{cases}}$$

Observations groupées

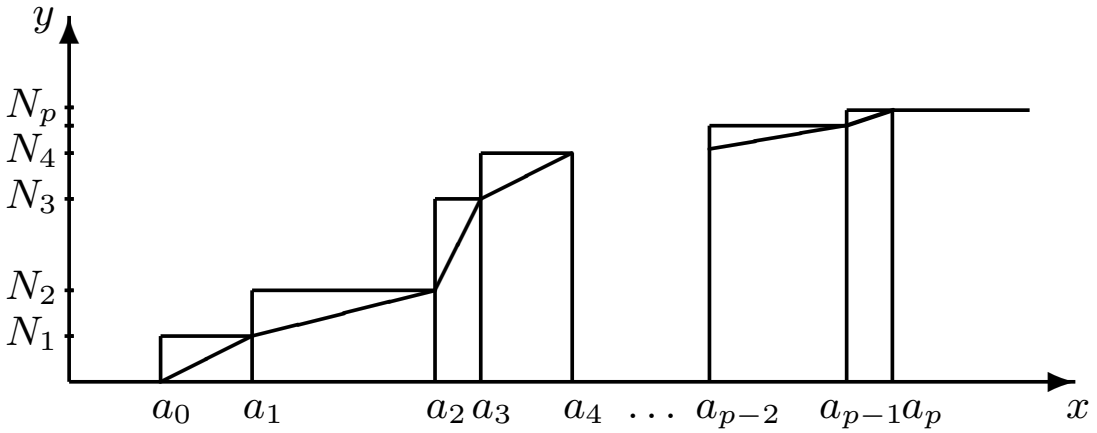
- Données réparties en classes $[a_0, a_1[, [a_1, a_2[, \ldots, [a_{p-2}, a_{p-1}[, [a_{p-1}, a_p]$
- Effectif et longueur de la classe $[a_{k-1}, a_k[,\ 1 \leqslant k \leqslant p : n_k$ et ℓ_k
- Densité d'effectif : $\dfrac{n_k}{\ell_k},\ 1 \leqslant k \leqslant p$



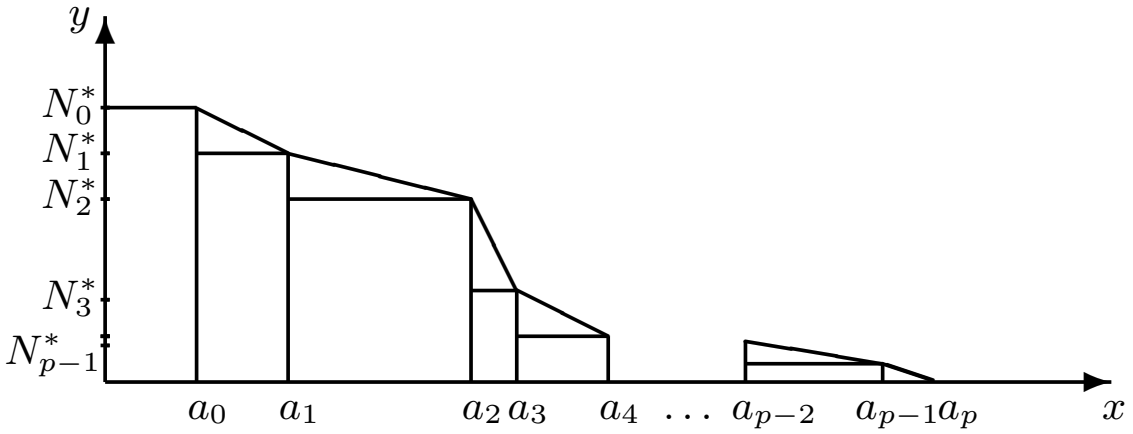
Histogramme

$$N(x) = \begin{cases} 0 & \text{si} \quad x < a_0 \\ \frac{n_1}{\ell_1}(x - a_0) & \text{si} \quad a_0 \leqslant x < a_1 \\ \vdots & \vdots & \vdots \\ N_{k-1} + \frac{n_k}{\ell_k}(x - a_{k-1}) & \text{si} \quad a_{k-1} \leqslant x < a_k \\ \vdots & \vdots & \vdots \\ n & \text{si} \quad x \geqslant a_p \end{cases}$$

Observations groupées



Courbe cumulative à gauche $y = N(x)$



Courbe cumulative à droite $y = N^*(x)$

$$\boxed{\forall x, N(x) + N^*(x) = n.}$$

Moyenne

$$\boxed{\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{k=1}^{p} n_k x'_k}$$

- Pour les observations groupées : x'_k est le centre de la classe $[a_{k-1}, a_k[$ soit $x'_k = \frac{1}{2}(a_{k-1} + a_k)$

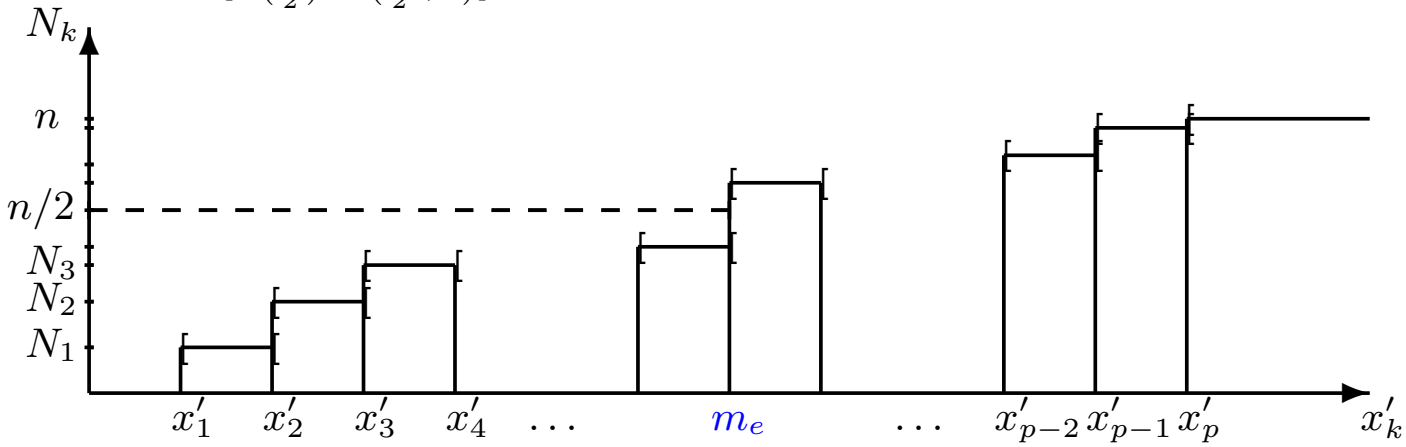
- Cas de deux séries statistiques $\{x_i, 1 \leqslant i \leqslant n_x\}$ et $\{y_j, 1 \leqslant j \leqslant n_y\}$:
  la série regroupée $\{z_k, 1 \leqslant k \leqslant n_z\} = \{x_i, 1 \leqslant i \leqslant n_x\} \cup \{y_j, 1 \leqslant j \leqslant n_y\}$ a pour moyenne

$$\boxed{\overline{z} = \frac{n_x\overline{x} + n_y\overline{y}}{n_x + n_y}}$$

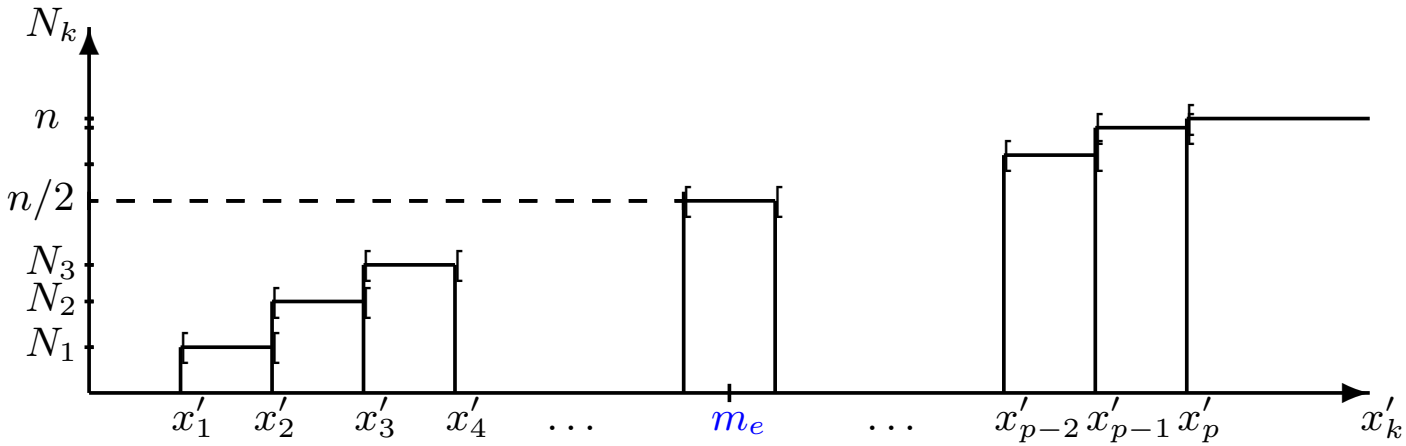
Médiane

$$m_e = \begin{cases} x_{(\frac{n+1}{2})} & \text{si } n \text{ est impair} \\ \frac{1}{2}[x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}] & \text{si } n \text{ est pair} \end{cases}$$

- Cas n pair : intervalle médian $[x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}]$



Courbe cumulative à gauche $y = N(x)$ (n impair)



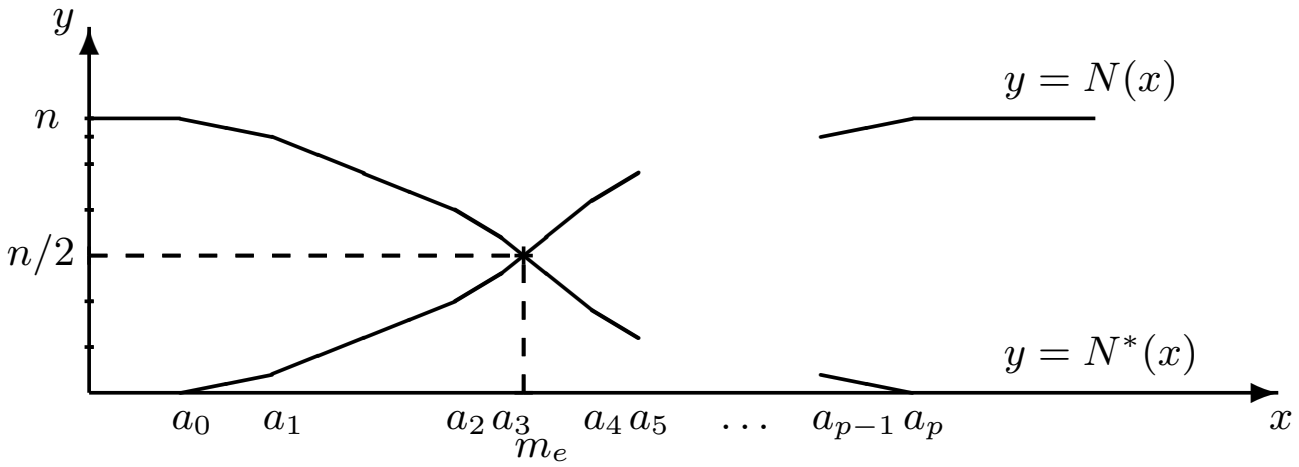
Courbe cumulative à gauche $y = N(x)$ (n pair)

Médiane

- Cas d'observations groupées :

$$\boxed{m_e = a_{i-1} + \frac{\ell_k}{n_k}\left(\frac{n}{2} - N_{k-1}\right)}$$

k étant l'indice de la classe contenant m_e



Courbes cumulatives $y = N(x)$ *et* $y = N^*(x)$

Quantiles

Soit $p \in]0,1[$

$$\boxed{Q_p = \begin{cases} x_{([np]+1)} & \text{si } np \text{ n'est pas entier} \\ \frac{1}{2}\left[x_{([np])} + x_{([np]+1)}\right] & \text{si } np \text{ est entier} \end{cases}}$$

$[np]$ étant la partie entière de np

- Médiane : $p = \frac{1}{2} \longrightarrow m_e = Q_{\frac{1}{2}}$
- Quartiles : $p \in \{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\} \longrightarrow$ intervalle interquartile : $Q_{\frac{3}{4}} - Q_{\frac{1}{4}}$
- Déciles : $p \in \{\frac{1}{10}, \frac{2}{10}, \ldots, \frac{9}{10}\}$
- Percentiles : $p \in \{\frac{1}{100}, \frac{2}{100}, \ldots, \frac{99}{100}\}$

Variance, écart-type

$$v_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{k=1}^{p} n_k (x'_k - \overline{x})^2 \qquad \sigma_x = \sqrt{v_x}$$

- Propriétés :

$$v_x = \overline{x^2} - \overline{x}^2 \text{ avec } \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

$$v_x = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2$$

$$\forall m \in \mathbb{R},\ \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 = v_x + (\overline{x} - m)^2$$

- Cas de deux séries statistiques $\{x_i, 1 \leqslant i \leqslant n_x\}$ et $\{y_j, 1 \leqslant j \leqslant n_y\}$:
  la série regroupée $\{z_k, 1 \leqslant k \leqslant n_z\} = \{x_i, 1 \leqslant i \leqslant n_x\} \cup \{y_j, 1 \leqslant j \leqslant n_y\}$ a pour variance

$$v_z = v_{\text{intra}} + v_{\text{inter}}$$

avec
$$v_{\text{intra}} = \frac{n_x v_x + n_y v_y}{n_x + n_y} \qquad v_{\text{inter}} = \frac{n_x(\overline{z}-\overline{x})^2 + n_y(\overline{z}-\overline{y})^2}{n_x + n_y} = \frac{n_x n_y}{(n_x+n_y)^2}(\overline{x} - \overline{y})^2$$

v_{intra} est la variance *dans* les groupes, v_{inter} est la variance *entre* les groupes

Analyse bivariée

- Observations : $(x_1, y_1), \ldots, (x_n, y_n)$

individu i	1	$\ldots$	i	$\ldots$	n
caractère x_i	x_1	$\ldots$	x_i	$\ldots$	x_n
caractère y_i	y_1	$\ldots$	y_i	$\ldots$	y_n

Individus × caractères bivariés

- Observations ordonnées : $x'_1 < \cdots < x'_p$ et $y'_1 < \cdots < y'_q$ $(p, q \leqslant n)$
- Effectifs marginaux : $n_{i.} = \sum_{j=1}^{q} n_{ij}$ et $n_{.j} = \sum_{i=1}^{p} n_{ij} \longrightarrow \sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j} = n$

caractère i \ caractère j	y'_1	$\ldots$	y'_j	$\ldots$	y'_q	effectifs marginaux
x'_1	n_{11}	$\ldots$	n_{1j}	$\ldots$	n_{1q}	$n_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x'_i	n_{i1}	$\ldots$	n_{ij}	$\ldots$	n_{iq}	$n_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x'_p	n_{p1}	$\ldots$	n_{pj}	$\ldots$	n_{pq}	$n_{p.}$
effectifs marginaux	$n_{.1}$	$\ldots$	$n_{.j}$	$\ldots$	$n_{.q}$	n

Tableau de contingence

Covariance, corrélation

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{k=1}^{p}\sum_{\ell=1}^{q} n_{k\ell}(x'_k - \overline{x})(y'_\ell - \overline{y}) \qquad \rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

- Propriétés :

$$\sigma_{xx} = \sigma_x^2 \qquad \sigma_{xy} = \overline{xy} - \overline{x}\,\overline{y} \text{ avec } \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \qquad |\rho_{xy}| \leqslant 1$$

- Les données $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$ sont *non-corrélées* lorsque $\sigma_{xy} = 0$ (ou encore $\rho_{xy} = 0$)
- Matrice de covariance : $\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$