STATISTIQUE DESCRIPTIVE

I Historique

- Questions de dénombrement (problèmes d'inventaire, de recensement, de sondage) remontant à la préhistoire.
- Plus récemment : * Probabilités : 17^e^ siècle, Pascal et Fermat
  puis 19^e siècle, famille Bernoulli, Gauss, Laplace.
  début 20^e siècle : axiomatisation ; Kolmogorov ...
  * Statistiques : 19^e siècle, Quételet (recensement de Bruxelle qui fut un exemple à suivre) (Adolphe Quételet 1796-1874)
  puis 20^e siècle, grand essor avec Pearson, Gosset (Student) Neyman, Fisher (de 1890 à 1945).

II Traitement des données

Présentation : tableau IC (individus x caractères)

i : individus n individus
j : variables p variables

$i \backslash j$	$1 \cdots j \cdots p$
1	$x_{11} \cdots x_{1j} \cdots x_{1p}$
$\vdots$	
i	$x_{i1} \cdots x_{ij} \cdots x_{ip}$
$\vdots$	
n	$x_{n1} \cdots x_{nj} \cdots x_{np}$

x_{ij} : caractères
n : effectif total

cas univarié :

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

, cas bivarié :

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

, etc...

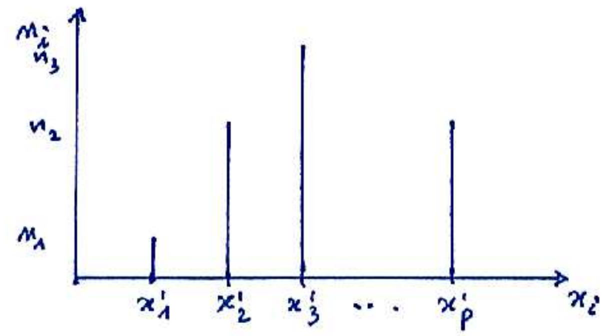
Organisation : série observée - série ordonnée

observations : $\{x_1, \dots, x_n\}$
on ordonne : $\{x_{(1)}, \dots, x_{(n)}\}$ avec $i \leq j \Rightarrow x_{(i)} \leq x_{(j)}$.

ordre strict : $\{x'_1, \dots, x'_p\}$ avec $i < j \Rightarrow x'_i < x'_j$
et nombre d'observations identiques x'_i : n_i (effectif de x'_i)

$\{(x'_i, n_i), 1 \leq i \leq p\}$ est appelé série statistique. $\sum_{i=1}^{p} n_i = n$.

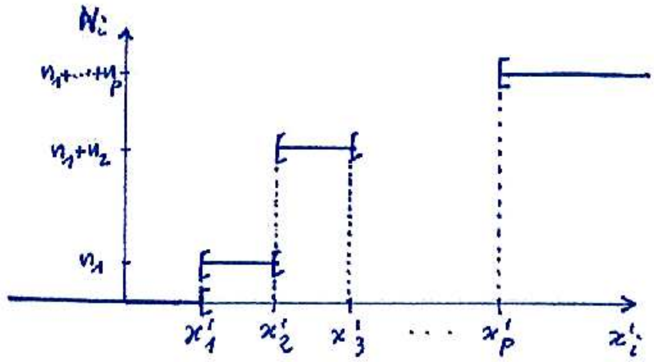
diagramme en bâtons :



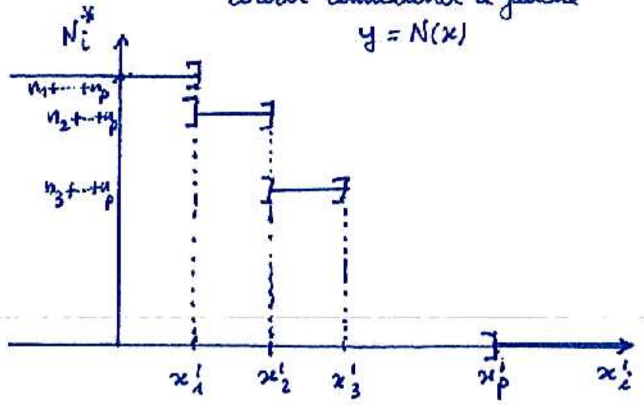
fréquences, effectifs cumulés, fréquences cumulées

$f_i = \dfrac{n_i}{n}$, $\sum_{i=1}^{p} f_i = 1$

$\begin{cases} \text{effectifs cumulés à gauche } N_i = \sum_{k=1}^{i} n_k \\ \text{fréquences cumulées à gauche } F_i = \dfrac{N_i}{n} \end{cases}$



courbe cumulative à gauche
$y = N(x)$



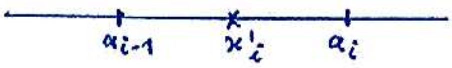
$\begin{cases} \text{effectifs cumulés à droite } N_i^* = \sum_{k=i}^{p} n_k \\ \text{fréquences cumulées à droite } F_i^* = \dfrac{N_i^*}{n} \end{cases}$

On a $N(x) + N^*(x) = \begin{cases} n & \text{si } x \notin \{x'_1, \dots, x'_p\} \\ n + n_i & \text{si } x = x'_i \end{cases} \geqslant n$

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

regroupement des données en classes

Si on dispose de trop de données, le diagramme en bâtons est surchargé.
On regroupe alors les données en classes $[a_0, a_1[\,, [a_1, a_2[\,, \dots, [a_{p-1}, a_p]$.



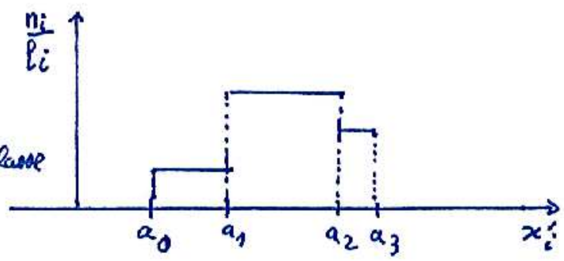
centre de la classe : x'_i
longueur de la classe : $l_i = a_i - a_{i-1}$
effectif de la classe : n_i

histogramme - polygône des effectifs (ou fréquences)

- histogramme des effectifs (ou fréquences)

$\dfrac{n_i}{l_i}$ = densité d'effectif = effectif par unité de longueur de classe

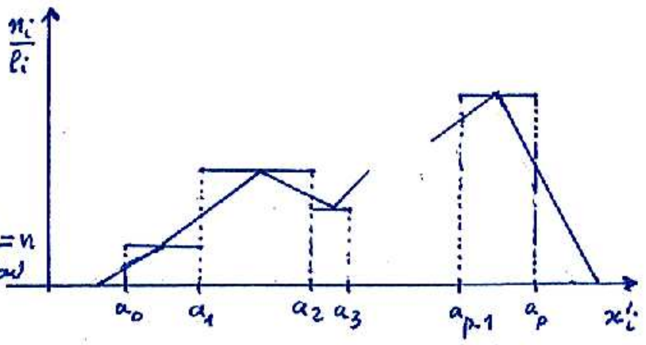
(idem avec les fréquences, $\dfrac{f_i}{l_i}$)



aire de chaque rectangle : $\dfrac{n_i}{l_i} \times l_i = n_i$ = effectif de la classe $[a_{i-1}, a_i[$.

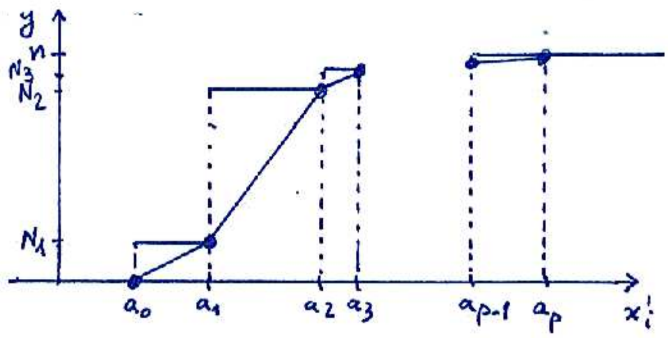
aire sous l'histogramme : n = effectif total.

- "image plus continue" de l'histogramme



polygône des effectifs = interpolation linéaire entre les centres des classes.
On suppose l'uniformité des répartitions dans chaque classe.
L'aire sous la courbe polygonale est inchangée $= n$
(les 2 bouts extrêmes sont placés pour vérifier cette condition)
Lorsque les l_i deviennent petits, on obtient un "graphe de densité locale".

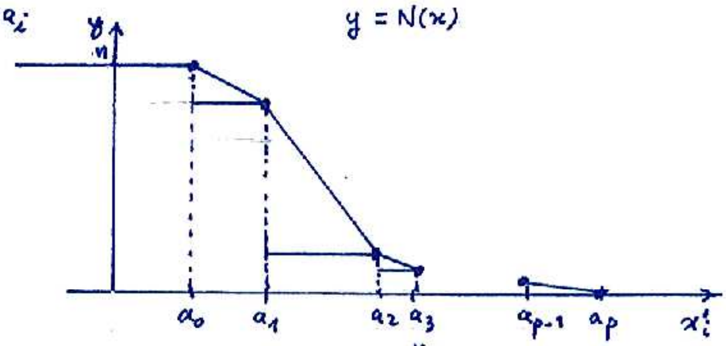
- courbes cumulées



$y = N(x)$

courbe cumulée à gauche = ligne brisée joignant les points (a_i, N_i)

$$N(x) = \begin{cases} 0 & \text{si } x < a_0 \\ \frac{n_1}{l_1}(x - a_0) & \text{si } a_0 \leq x < a_1 \\ \vdots & \\ N_{i-1} + \frac{n_i}{l_i}(x - a_{i-1}) & \text{si } a_{i-1} \leq x < a_i \\ \vdots & \\ n & \text{si } x \geq a_p \end{cases}$$



$y = N^*(x)$

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

III Paramètres de position

moyenne (pour les effectifs quantitatifs) : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{p} n_i x'_i = \sum_{i=1}^{p} f_i x'_i$

Pour les observations groupées : $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x'_i$, x'_i centre de la classe $[a_{i-1}, a_i[$.

C'est le paramètre le plus répandu, mais pas nécessairement le plus adéquate.
Un inconvénient : e.g. s'il y a une donnée aberrante, elle affecte sensiblement la moyenne. On peut donc être amené à tronquer la moyenne, ou à la pondérer :

$$\frac{1}{p-2}\sum_{i=2}^{p-1} n_i x'_i, \quad \frac{\sum_{i=1}^{p} w_i x'_i}{\sum_{i=1}^{p} w_i} \quad \text{etc} \ldots$$

Si on a deux séries statistiques $\{x_i, 1 \leq i \leq n_x\}$ et $\{y_j, 1 \leq j \leq n_y\}$, la série complète $\{z_k, 1 \leq k \leq n_x + n_y\} = \{x_i, 1 \leq i \leq n_x\} \cup \{y_j, 1 \leq j \leq n_y\}$ a pour moyenne

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

<u>médiane</u> : m_e est la valeur "centrale" de la série ordonnée $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ (avec répétition)

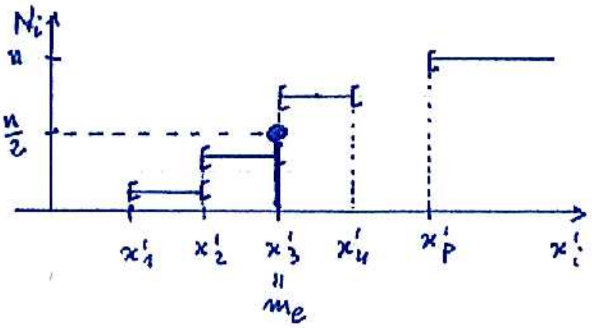
- si n est impair $\boxed{m_e = x_{(\frac{n+1}{2})}}$
- si n est pair, il y a un <u>intervalle médian</u> $[x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}]$. On pose $m_e = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$.

ex : série ordonnée 1 2 2 3 4 4 4 5 6 $\rightarrow n = 9$ $m_e = x_{(5)} = 4$

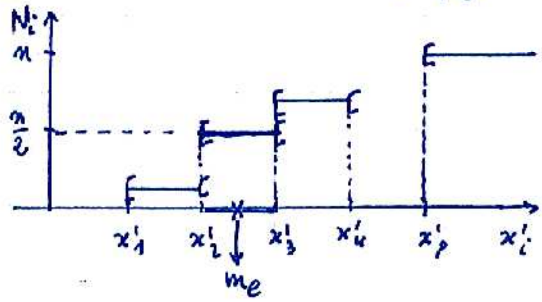
(le 4 du milieu : m_e)

1 2 2 3 4 4 4 5 $\rightarrow n = 8$ $m_e = \frac{x_{(4)} + x_{(5)}}{2} = 3{,}5$

Détermination de m_e à l'aide des courbes cumulatives : on a $N(m_e) \geq \frac{n}{2}$ et $N^*(m_e) \geq \frac{n}{2}$
(i.e. le nombre de valeurs de la série $< m_e$ est égal au nombre de valeurs $> m_e$: $[\frac{n}{2}]$)

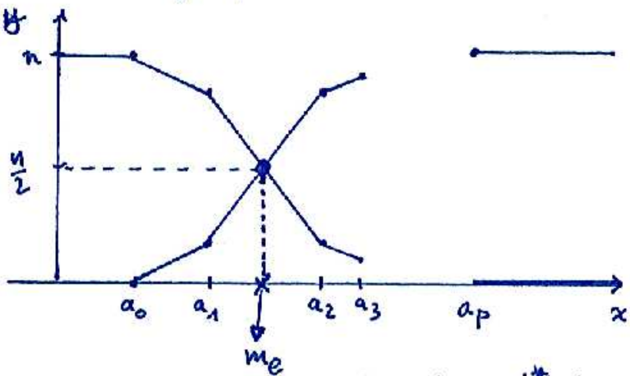


ou



S'il existe un i tel que $N_{i-1} < \frac{n}{2} < N_i$ alors $m_e = x'_i$
S'il existe un i tel que $N_i = \frac{n}{2}$ alors $[x'_i, x'_{i+1}]$ est l'intervalle médian et $m_e = \frac{x'_i + x'_{i+1}}{2}$.

Cas des observations groupées : on utilise les courbes cumulatives continues :



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

$N + N^* = n$

$N(m_e) = N^*(m_e) = \frac{n}{2}$.

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

i étant l'indice de la classe contenant m_e.

<u>quantiles (ou fractiles)</u> : Soit $p \in]0,1[$. Q_p : p-quantile (ou quantile d'ordre p)

Q_p est la valeur du caractère telle que $N(Q_p) \geq np$ et $N^*(Q_p) \geq n(1-p)$.
(i.e. le nombre de valeurs de la série $< Q_p$ est égal à $[np]$, et $> Q_p$: $[n(1-p)]$)

si $np \notin \mathbb{N}$ $\boxed{Q_p = x_{([np]+1)}}$; si $np \in \mathbb{N}$ choisir $\boxed{Q_p \in [x_{(np)}, x_{(np+1)}]}$

ex : médiane : $p = \frac{1}{2}$, $m_e = Q_{1/2}$

cas d'observations groupées :

quartiles : $p = \frac{1}{4}, \frac{2}{4}, \frac{3}{4}$ $Q_{1/4}$, $Q_{2/4} = m_e$, $Q_{3/4}$

$$\boxed{Q_p = a_{i-1} + \frac{l_i}{n_i}(np - N_{i-1})}$$

déciles : $p = \frac{1}{10}, \frac{2}{10}, \dots, \frac{9}{10}$

percentiles : $p = \frac{1}{100}, \frac{2}{100}, \dots, \frac{99}{100}$

mode : valeur observée qui a le plus grand effectif
distribution unimodale (plurimodale) : qui contient un seul (plusieurs) mode(s).

Cas des observations groupées : si les classes ont même longueur, la classe modale est celle d'effectif maximal. Sinon, considérer l'histogramme des densités $\frac{n_i}{l_i}$.

autres paramètres de position : $c_1 = \frac{x_{(1)} + x_{(n)}}{2}$, $c_2 = \frac{Q_{1/4} + Q_{3/4}}{2}$, $c_3 = \frac{Q_{1/4} + 2Q_{2/4} + Q_{3/4}}{4}$ etc ...

remarque : tous les paramètres considérés ici ont même dimension que les caractères.

IV Paramètres de dispersion

étendue (empan) : $x_{(n)} - x_{(1)}$. Inconvénient : l'étendue ne tient pas compte de toutes les observations.

ex :

ces trois séries ont même étendue.

écart interquantile : $[Q_p, Q_{1-p}]$ pour $0 < p < \frac{1}{2}$. Cet intervalle contient un effectif égal ou juste supérieur à $n(1-2p)$: $\begin{cases} n(1-2p) & \text{si } np \in \mathbb{N} \\ [n(1-2p)]+1 & \text{si } np \notin \mathbb{N} \end{cases}$.

ex : intervalle interquartile $[Q_{1/4}, Q_{3/4}]$: contient $\approx 50\%$ d'observations
intervalle interdécile $[Q_{1/10}, Q_{9/10}]$: contient $\approx 80\%$ d'observations.

ces intervalles sont utilisés pour éliminer les valeurs extrêmes (éventuellement) aberrantes.

écart moyen absolu : $e_m = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| = \frac{1}{n} \sum_{i=1}^{p} n_i |x'_i - \bar{x}| = \sum_{i=1}^{p} f_i |x'_i - \bar{x}|$

Pour les observations groupées : $e_m = \frac{1}{n} \sum_{i=1}^{p} n_i |x'_i - \bar{x}|$, x'_i centre de classe

écart médian absolu : $e^*_m = \frac{1}{n} \sum_{i=1}^{n} |x_i - m_e|$. m_e minimise $m \mapsto \sum_{i=1}^{n} |x_i - m|$.

Inconvénient : le calcul de e_m et e^*_m est peu maniable à cause de la valeur absolue, et ils ont peu de propriétés mathématiques.

écart-type et variance : $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^{p} n_i (x'_i - \bar{x})^2 = \sum_{i=1}^{p} f_i (x'_i - \bar{x})^2$

Théorème de König-Huygens : $\frac{1}{n} \sum_{i=1}^{n} (x_i - m)^2 = \sigma^2 + (\bar{x} - m)^2$

Donc σ^2 minimise $m \mapsto \sum_{i=1}^{n} (x_i - m)^2$ et $\sigma^2 = \overline{x^2} - \bar{x}^2$.
(→ évite de calculer tous les écarts $x_i - \bar{x}$)

autre écriture : $\sigma^2 = \frac{1}{2n^2} \sum_{1 \leq i,j \leq n} (x_i - x_j)^2$

Valeurs centrées réduites : $y_i = \frac{x_i - \bar{x}}{\sigma}$, $\bar{y} = 0$, $\sigma_y = 1$ (y est sans dimension) si on réunit deux séries statistiques $\{x_i, 1 \le i \le n_x\}$ et $\{y_j, 1 \le j \le n_y\}$ en une seule $\{z_k, 1 \le k \le n_x + n_y\}$, les écarts-type sont reliés par

$$\boxed{\sigma_z^2 = \frac{n_x \sigma_x^2 + n_y \sigma_y^2}{n_x + n_y} + \frac{n_x (\bar{z}-\bar{x})^2 + n_y (\bar{z}-\bar{y})^2}{n_x + n_y}}$$

$$\underbrace{\frac{n_x \sigma_x^2 + n_y \sigma_y^2}{n_x + n_y}}_{\text{variance DANS les groupes}} \qquad \underbrace{\frac{n_x (\bar{z}-\bar{x})^2 + n_y (\bar{z}-\bar{y})^2}{n_x + n_y}}_{= \frac{n_x n_y}{(n_x+n_y)^2}(\bar{x}-\bar{y})^2 \text{, variance ENTRE les groupes}}$$

démonstration : utiliser König : $\sigma_z^2 = \frac{1}{n_x+n_y}\left[\sum_{i=1}^{n_x}(x_i - \bar{z})^2 + \sum_{j=1}^{n_y}(y_j - \bar{z})^2\right] = \frac{n_x}{n_x+n_y}\left[\sigma_x^2 + (\bar{z}-\bar{x})^2\right] + \frac{n_y}{n_x+n_y}\left[\sigma_y^2 + (\bar{z}-\bar{y})^2\right]$

<u>Inégalité de Bienaymé - Tchebycheff</u> : $\left\{\begin{array}{l}\text{la fréquence des } x_i \text{ tels que } |x_i - \bar{x}| \ge k \text{ est} \le \text{ à } \frac{\sigma^2}{k^2}, \\ \text{la fréquence des } x_i \text{ tels que } |x_i - \bar{x}| < k \text{ est} \ge \text{ à } 1 - \frac{\sigma^2}{k^2}.\end{array}\right.$

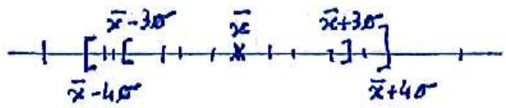
démonstration : $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|^2 = \frac{1}{n}\sum_{i : |x_i - x| \ge k}|x_i - x|^2 + \frac{1}{n}\sum_{i : |x_i - x| < k}|x_i - x|^2$

$\ge \frac{k^2}{n}\operatorname{card}\{i : |x_i - \bar{x}| \ge k\}$.

Puis $\frac{1}{n}\operatorname{card}\{i : |x_i - \bar{x}| < k\} = 1 - \frac{1}{n}\operatorname{card}\{i : |x_i - \bar{x}| \ge k\} \ge 1 - \frac{\sigma^2}{k^2}$. $\square$

application : $k = 3\sigma$: la fréquence des x_i tels que $|x_i - \bar{x}| \ge 3\sigma$ est $\le$ à $\frac{1}{9} \approx 11,2\%$

$k = 4\sigma$: la fréquence des x_i tels que $|x_i - \bar{x}| \ge 4\sigma$ est $\le$ à $\frac{1}{16} \approx 6,3\%$.



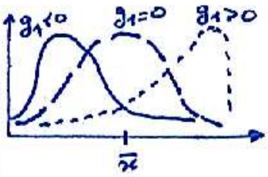
<u>Autres paramètres de dispersion</u> :

- coefficient de variation : $CV = \frac{\sigma}{\bar{x}}$: nombre sans dimension, se mesure en %

  (mesure relative de dispersion)

- paramètres de forme : * $m_3 = \overline{(x - \bar{x})^3} = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i' - \bar{x})^3$ moment centré d'ordre 3

  c'est un coefficient de dissymétrie.



  coefficient de Fisher $g_1 = \frac{m_3}{\sigma^3}$ (sans dimension)

  coefficient empirique de Pearson : $S_k = \frac{\bar{x} - me}{\sigma}$

  coefficient empirique de Yule-Kendall : $Y_k = \frac{Q_{1/4} + Q_{3/4} - 2Q_{1/2}}{Q_{3/4} - Q_{1/4}}$

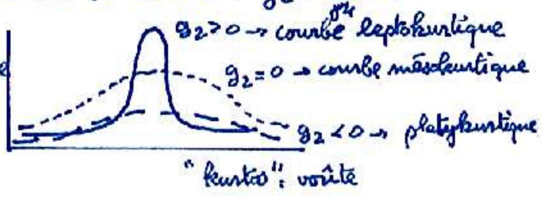
  Attention : l'étude des signes de S_k et Y_k peut donner des contradictions.

  * $m_4 = \overline{(x - \bar{x})^4} = \frac{1}{n}\sum_{i=1}^{p} n_i (x_i' - \bar{x})^4$

  coefficient d'aplatissement de Pearson : $b_2 = \frac{m_4}{\sigma^4}$

  coefficient d'aplatissement de Fisher : $g_2 = \frac{m_4}{\sigma^4} - 3$

  $g_2 = 0$ pour la cloche



V Analyse bivariée

On part de deux caractères x_i, y_i $1 \le i \le n$. On ordonne strictement $x'_1 < \dots < x'_p$ et $y'_1 < \dots < y'_q$:

Série statistique double $\{(x'_i, y'_j, n_{ij}),\ 1 \le i \le p,\ 1 \le j \le q\}$

Tableau de contingence

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

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

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

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

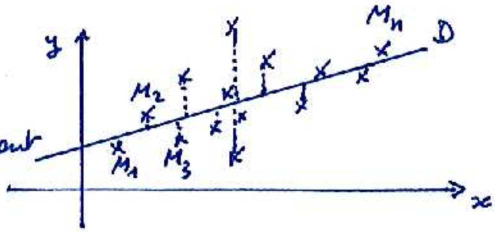
covariance : $\sigma_{xy} = \overline{(x-\bar{x})(y-\bar{y})} = \dfrac{1}{n} \sum\limits_{1 \le i,j \le n} n_{ij} (x'_i - \bar{x})(y'_j - \bar{y}) = \overline{xy} - \bar{x}\,\bar{y}$

coefficient de corrélation de Bravais-Pearson : $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y} \in [-1,1]$

matrice de covariance $\begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$

régression linéaire

nuage de points $\{(x_i, y_i),\ 1 \le i \le n\}$



méthode des moindres carrés : droite d'ajustement $y = ax + b$ minimisant $\sum\limits_{i=1}^{n} d(M_i, D)^2$

régression de y par rapport à x :
$d_V(M_i, D)$: distance "verticale" de M_i à D.

$$\begin{cases} a = \dfrac{\sigma_{xy}}{(\sigma_x)^2} & \text{: coefficient de regression (pente a)} \\ b = \bar{y} - a\bar{x} \end{cases} \qquad y = \bar{y} + \frac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$$

variance résiduelle : $\sigma_{y,x}^2 = \frac{1}{n}\sum (y_i - ax_i - b)^2 = \sigma_y^2 (1 - \rho_{xy}^2)$

régression de x par rapport à y : $x = a'y + b'$ minimisant $\sum\limits_{i=1}^{n} d_H(M_i, D)^2$, d_H distance "horizontale".

$$\begin{cases} a' = \dfrac{\sigma_{xy}}{\sigma_y^2} & \left(\text{pente } \dfrac{1}{a'}\right) \\ b' = \bar{x} - a'\bar{y} \end{cases} \qquad x = \bar{x} + \frac{\sigma_{xy}}{\sigma_y^2}(y - \bar{y})$$

En général ces deux droites sont distinctes. $a a' = \dfrac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \rho_{xy}^2$.

Les pentes se rapprochent de plus en plus lorsque $|\rho_{xy}|$ est proche de 1.
On considère l'ajustement valable pour $0,7 \le |\rho_{xy}| \le 1$, excellent si $|\rho_{xy}| \ge 0,95$.

autres ajustements : $y = a x^{\alpha}$ → régression linéaire en $(\ln x, \ln y)$: $\ln y = \alpha \ln x + \ln a$

$y = a e^{bx}$ → régression linéaire en $(x, \ln y)$: $\ln y = bx + \ln a$.

Bibliographie : J.-J. Droesbeke : éléments de statistiques 1988, chapitres 2, 3, 10
H. Egon : statistique et probabilités 1992, chapitre A
G. Demengel : probabilités, statistique inférentielle, fiabilité, 1997, chapitres 1, 2
M. Gaultier : statistique 1997, chapitre 2 - 7