

3. Caractéristiques numérique d'une variable statistique

Lorsque l'on observe une représentation graphique d'une série statistique, on peut retirer deux impressions:

- 1) L'ordre de grandeur de la v.s. caractérisée par les valeurs situées au centre de la distribution: c'est ce qu'on appelle "Caractéristiques de tendance centrale (ou de position) de la série.
- 2) Les fluctuations des observations autour de la valeur centrale. C'est ce qu'on appelle "Caractéristiques de dispersion".

3.1 / Caractéristiques de tendance centrale

Elles sont destinées à définir les valeurs centrales de la série statistique, ce sont: le mode, la médiane et la moyenne arithmétique.

3.1.1) Le mode

a) Cas discret:

On appelle "mode" noté par " M_0 " d'une v.s. discrète la valeur de cette variable qui a la fréquence la plus élevée.

Exemple: Dans l'exemple 2 "nbre de pièces par logement".

$M_0 = 4$ (car la valeur 4 a le plus grand effectif (30))

Rq: Sur le diagramme en bâtons, le mode correspond à la valeur de la variable qui correspond au plus long bâton.

b) Cas continu: Dans ce cas on parle plutôt de "classe modale" qui est la classe dont la fréquence par amplitude unité est la plus élevée.

Exemple : Dans l'exemple 3 "salaire mensuel ^{des entreprises}" la classe modale est $[5000 \text{ DA}, 6000 \text{ DA}]$.

Dans l'ensemble 4 : "Taille d'i gpe d'individus" la classe modale est $[170, 175]$.

Rqre : Dans le cas continu, on peut appeler mode le centre de la classe modale.

• Pour un calcul approché, on utilise la formule suivante : soit la classe modale $[e_{i-1}, e_i]$.

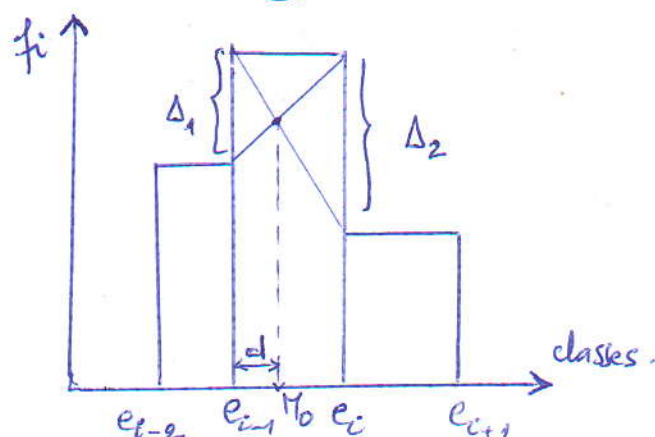
on a $M_0 \in [e_{i-1}, e_i]$

$$M_0 = e_{i-1} + a_i \frac{\Delta_1}{\Delta_1 + \Delta_2}$$

avec : a_i l'amplitude de la classe $[e_{i-1}, e_i]$.

Δ_1 : l'excès de la classe modale par rapport à la classe précédente.

Δ_2 : L'excès de la classe modale par rapport à la classe suivante.



• Graphiquement : $M_0 = e_{i-1} + d$ (d calculée en utilisant l'échelle)

Rqre : 1) Pour le calcul du mode, et dans le cas des classes inégales, il faut corriger les fréquences.

2) Une série statistique peut-être unimodale (1 seul mode) bimodale (2 modes), ... plurimodale (plusieurs modes).

3.1.2) La médiane : La médiane (M_e) est la valeur de la v.s. qui partage les individus supposés rangés dans l'ordre croissant (ou décroissant), de la variable, en deux effectifs égaux.

Exemple : soit la série : $\{12, 26, 6, 3, 32, 15, 21\}$.

on l'ordonne $\{3, 6, 12, 15, 21, 26, 32\}$: La médiane est $M_e = 15$.

• Si la taille médiane des individus d'un échantillon est de 1,70 m, cela signifie qu'il y a autant d'individus mesurant moins de 1,70 m que d'individus mesurant plus de 1,70 m.

• D'une manière générale, la médiane d'une v.s. est la valeur notée Me telle que $F(Me) = \frac{1}{2} = 0,5$. avec F est la fct de répartition empirique.

* Calcul de la médiane:

a) Cas discret:

soit $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ les n observations d'une v.s. discrète rangées dans l'ordre croissant:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

• 1er cas: n impair : $n = 2p + 1$.

$$\underbrace{x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(p)}}_{p \text{ observations}} \leq \boxed{x_{(p+1)}} \leq \underbrace{x_{(p+2)} \leq \dots \leq x_{(2p+1)}}_{p \text{ observations}}$$

• $Me = x_{(p+1)}$: La valeur de la variable située à l'ordre $(p+1)$ [$(p+1)^{\text{ème}}$ observation]

• 2^{ème} cas: n pair : $n = 2p$.

$$\underbrace{x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(p)}}_{p \text{ observations}} \leq \underbrace{x_{(p+1)} \leq \dots \leq x_{(2p)}}_{p \text{ observations}}$$

$$Me = \frac{x_{(p)} + x_{(p+1)}}{2}$$

$[x_{(p)}, x_{(p+1)}]$ est appelé intervalle médian.

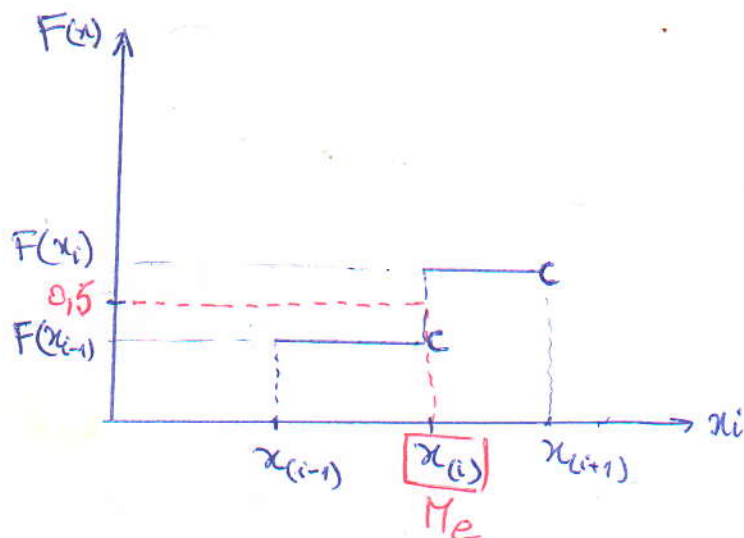
Dans l'exemple 2: "nbre de pièces par logement".

$$n = 100 = 2 \times 50 \Rightarrow Me = \frac{x_{(50)} + x_{(51)}}{2} = \frac{4 + 4}{2} = 4$$

4 1 2 2 3 3 4 4 5 ...
 $\uparrow \quad \uparrow$
 $x_{(50)} \quad x_{(51)}$

Détermination graphique de la médiane :

- 1) aucun palier horizontal de la courbe cumulative n'a pour ordonnée (0,5).



$$F(x_{(i)}) < 0,5 < F(x_{(i+1)})$$

$$\Rightarrow \boxed{Me = x_{(i)}}$$

Dans l'exemple 2 : "nbre de pièces" :

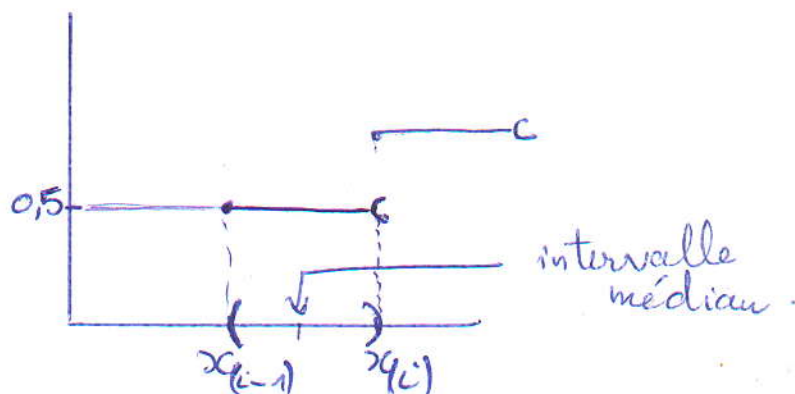
$$0,35 < 0,5 < 0,65$$

$$F(3) < 0,5 < F(4)$$

$$\Rightarrow \boxed{Me = 4} \text{ pièces}$$

- 2) si un palier horizontal de la courbe cumulative a pour ordonnée 0,5, la médiane est indéterminée entre deux valeurs possibles successives.

$$Me = \frac{x_{(i-1)} + x_{(i)}}{2}$$



b) Cas continu :

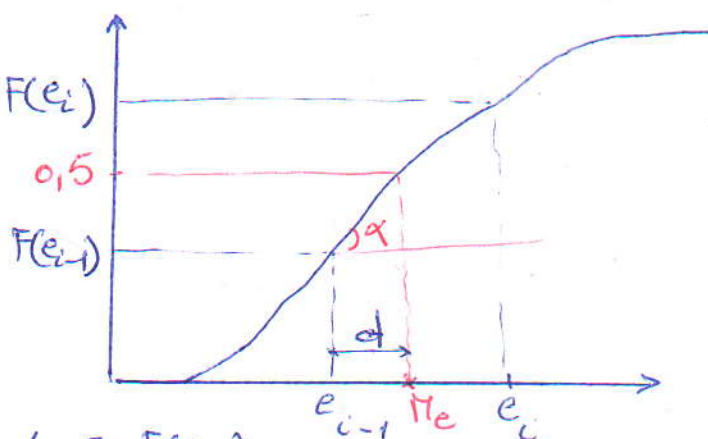
Comme la fct F est continue, monotone ($\nearrow$) entre 0 et 1, la médiane est l'unique solution de l'équation $F(x) = \frac{1}{2}$.

• Détermination de la médiane (méthode d'interpolation).
soit d'abord la classe médiane $[e_{i-1}, e_i[$:

$$F(e_{i-1}) < 0,5 < F(e_i)$$

$$\text{tg } \alpha = \frac{F(e_i) - F(e_{i-1})}{e_i - e_{i-1}} = \frac{0,5 - F(e_{i-1})}{\pi_e - e_{i-1}} \cdot F(e_i)$$

$$\Rightarrow \frac{b_i}{a_i} = \frac{0,5 - F(e_{i-1})}{\pi_e - e_{i-1}}$$



D'où : $b_i (\pi_e - e_{i-1}) = a_i (0,5 - F(e_{i-1}))$

par la suite :

$$\pi_e = e_{i-1} + a_i \frac{0,5 - F(e_{i-1})}{b_i}$$

a_i : amplitude
 b_i : fréquence de la classe médiane

• graphiquement : $\pi_e = e_{i-1} + d$. (d : calculée à partir de l'échelle).

3.1.3) Moyenne arithmétique : c'est la somme de toutes les observations divisée par le nbre total d'observations.

a) Cas discret :

Soit X une v.s. discrète pouvant prendre les valeurs $x_1, x_2, \dots, x_k$ auxquels correspondent les effectifs $n_1, n_2, \dots, n_k$ avec $\sum_{i=1}^k n_i = n$.

La moyenne arithmétique notée $\bar{x}$ est définie par :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i = \frac{n_1 x_1 + n_2 x_2 + \dots + n_k x_k}{n}$$

$$\bar{x} = \sum_{i=1}^k b_i x_i \quad (\text{moyenne pondérée})$$

b/ Cas continu : les modalités sont regroupées par classes.
 par convention, on choisit les x_i , les centres de
 chaque classe et on utilise la même formule.

$$\bar{x} = \sum_{i=1}^k f_i x_i \quad (x_i \text{ centre de la } i^{\text{ème}} \text{ classe})$$

$$x_i = \frac{e_{i+1} + e_i}{2}$$

Propriétés algébriques de la moyenne arithmétique:

Propriété (1) : chgt d'origine;

soit $x_1, x_2, \dots, x_k$ les valeurs observées d'une v.s. X
 et $n_1, n_2, \dots, n_k$ leurs effectifs.

soit x_0 : une nouvelle origine.

On définit une nouvelle variable $y_i = x_i - x_0 \quad \forall i=1, k$

Alors $\boxed{\bar{x} = \bar{y} + x_0}$ avec $\bar{y} = \sum_{i=1}^k f_i y_i$

en effet : $x_i = y_i + x_0$

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^k n_i x_i = \frac{1}{n} \sum_{i=1}^k n_i (y_i + x_0) \\ &= \frac{1}{n} \sum_{i=1}^k n_i y_i + \frac{1}{n} \sum_{i=1}^k n_i x_0 = \bar{y} + x_0 \end{aligned}$$

Rq : on utilise ce ~~calcul~~ chgt de variable
 pour simplifier les calculs et on pratique

on choisit $x_0 = M_0$ ou M_e .

Propriété (2) : chgt d'échelle et d'origine;

si on choisit : $y_i = \frac{x_i - x_0}{a}$, x_0 a ces -

Alors $\boxed{\bar{x} = a \bar{y} + x_0}$.

Preuve : exercice.

Rq : En pratique, on choisit x_0 : médiane, mode, centre de la classe modale.
 $a = \text{pgcd}(n_i)$ (cas continu)
 $= \text{écart entre les } x_i$ (cas discret)

Preuve : exercice.

Proposition 1 : La somme des écarts à la moyenne arithmétique est nulle : $\sum_{i=1}^k n_i (x_i - \bar{x}) = 0$

en effet : $\sum_{i=1}^k n_i (x_i - \bar{x}) = \sum_{i=1}^k n_i x_i - \sum_{i=1}^k n_i \bar{x}$
 $= n\bar{x} - n\bar{x} = 0$

Proposition 2 : La somme des carrés des écarts à la moyenne arithmétique est minimale :

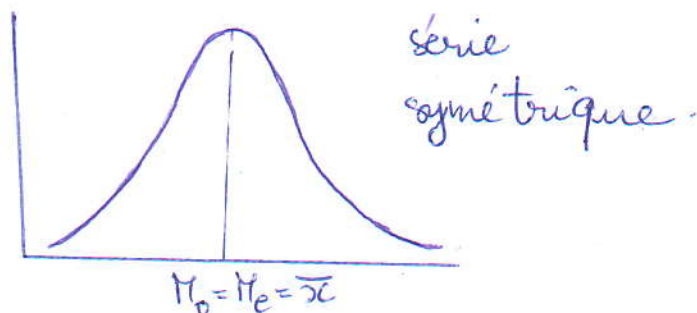
$\varphi(a) = \sum_{i=1}^k n_i (x_i - a)^2$ est minimale pour $a = \bar{x}$.

Preuve : exercice.

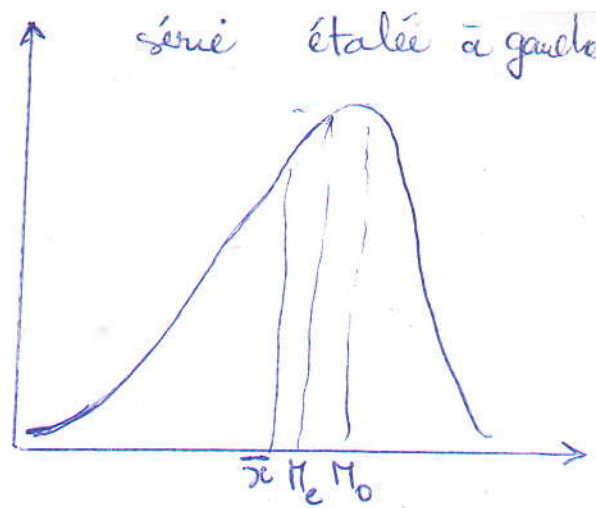
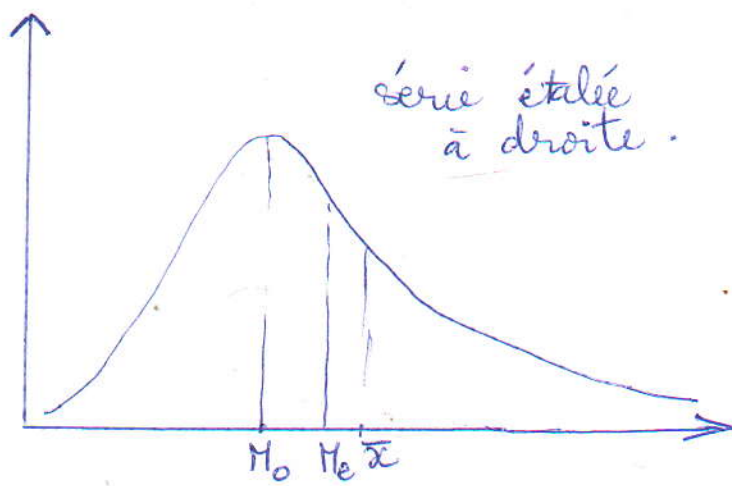
- Position respective du mode, de la médiane et de la moyenne.

Considérons une série statistique unimodale.

- 1) Lorsque la série statistique est symétrique, les trois paramètres de tendance centrale sont confondus.



- 2) Lorsque la série est asymétrique, la médiane est généralement comprise entre le mode et la moyenne et elle est plus proche de cette dernière.



3.2) Caractéristiques de dispersion:

Exemple: soit les 2 séries statistiques:

$$X = \{6, 6, 7, 7, \boxed{8}, 9, 9, 10, 10\}; Y = \{1, 2, 4, 6, \boxed{8}, 10, 12, 14, 15\}.$$

on remarque que X et Y ont même moyenne et même médiane $\bar{x} = \bar{y} = Me = 8$ mais elle sont différentes; La 1^{ère} X est moins dispersée que la seconde Y .

• Considérons la série statistique X de k valeurs rangées dans l'ordre croissant: $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(k)}$.

3.2.1) L'étendue: on appelle étendue, notée par "e" la différence entre les deux valeurs extrêmes: la plus petite et la plus grande valeur observée.

$$e = \max_{1 \leq i \leq k} (x_i) - \min_{1 \leq i \leq k} (x_i) = x_{(k)} - x_{(1)}.$$

3.2.2) variance et écart-type:

• on appelle variance empirique de la variable statistique X de valeurs x_i , $1 \leq i \leq k$, d'effectifs n_i : $1 \leq i \leq k$,

$$\sum_{i=1}^k n_i = n$$

$$V(X) = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2 = \sum_{i=1}^k f_i (x_i - \bar{x})^2.$$

• on appelle écart-type empirique noté σ_x :

$$\sigma_X = \sqrt{V(X)} = \sqrt{\frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2} = \sqrt{\sum_{i=1}^k f_i (x_i - \bar{x})^2}$$

• Propriétés de la variance:

1) $V(X) \geq 0$

2) $V(X) = \frac{1}{n} \sum_{i=1}^k n_i x_i^2 - \bar{x}^2$

$$= \sum_{i=1}^k f_i x_i^2 - \bar{x}^2 \quad (\text{exercice})$$

3) soit $\bar{x}$ et $V(X)$ la moyenne et la variance ~~(respectivement)~~ de la variable statistique X .

• On définit une nouvelle variable X' de moyenne $\bar{x}'$ et de variance $V(X')$ telle que:

$$x'_i = \frac{x_i - x_0}{a} \quad i=1, k \quad \text{et } x_0, a, \text{ sont des}$$

Constantes; alors: $V(X) = a^2 V(X')$ et $\sigma_X = a \sigma_{X'}$

• Règle: lorsqu'on compare deux séries statistiques de même nature celle qui a σ_X le plus élevé est la plus dispersée.

3.2.3) Le coefficient de variation:

C'est un coefficient de dispersion défini par:

$$CV = \frac{\sigma_X}{|\bar{x}|}$$

• Propriétés:

1) CV est une quantité sans unité.

2) CV est indépendant des unités choisies.

3) CV permet de comparer 2 séries exprimées dans deux unités différentes.

3.2.4) écart-inter-quantiles :

a) les quantiles : ces quantiles généralisent la médiane.

• Un quantile d'ordre α ($0 \leq \alpha \leq 1$); noté x_α , est la racine de l'équation $F(x) = \alpha$.
ie : une proportion égale à α des individus possède un caractère x inférieur à x_α .

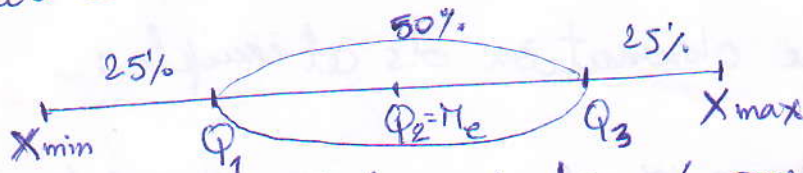
• On utilise couramment les quartiles.

Ce sont les valeurs du caractère x_i qui partage la série en 4 sous-ensembles égaux. Il y en a donc 3, on les note Q_1 , Q_2 et Q_3 .

avec Q_1 est le quantile d'ordre $\frac{1}{4}$, Q_2 : quantile d'ordre $\frac{1}{2}$ et Q_3 : quantile d'ordre $\frac{3}{4}$.

ie : $F(Q_1) = \frac{1}{4}$; $F(Q_2) = \frac{1}{2} = F(M_e)$ et $F(Q_3) = \frac{3}{4}$.

• L'intervalle inter-quartiles : contient 50% de la population laissant à droite 25% et à gauche 25%.



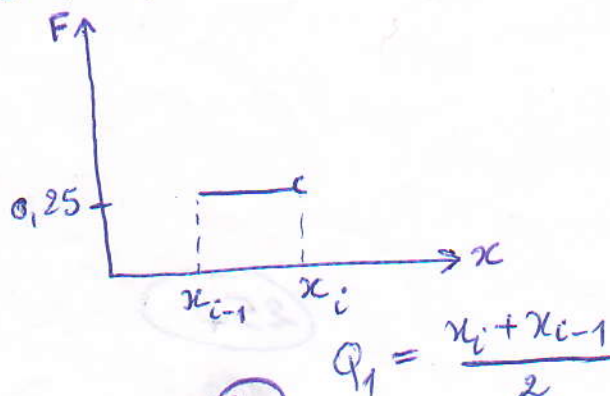
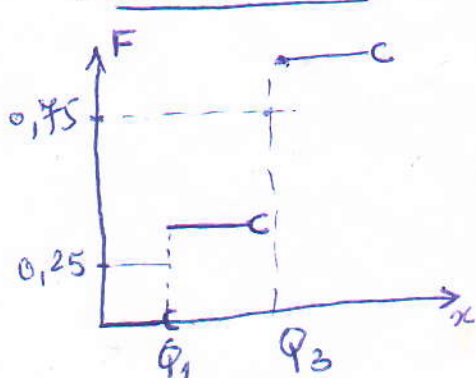
$$\Rightarrow I_Q = [Q_1, Q_3]$$

• l'écart-interquartiles est donné par $E_Q = Q_3 - Q_1$.

Détermination pratique:

Pour déterminer $Q_3 - Q_1$, il faut d'abord calculer Q_1 et Q_3 en utilisant la même méthode que celle utilisée pour déterminer la médiane.

• Cas discret : (graphiquement)



Détermination pratique des quartiles

Reprenons l'exemple 1: "nombre d'enfants de 10 ménages".

$$\underbrace{0 \ 0 \ 1 \ 1}_{\text{Me}} \ 1 \ 1 \ 1 \ 2 \ 4 \ 4 \quad ; \quad n=10 = 2 \times 5. \text{ (pair)}$$

$$Q_2 = M_e = \frac{x_{(5)} + x_{(6)}}{2} = \frac{1+1}{2} = 1 \text{ enfant.}$$

La 1^{ère} moitié de la série :

$$0 \ 0 \ 1 \ 1 \ 1 \quad p = 5 = 2 \times 2 + 1.$$

$$Q_1 = x_{(3)} = 1 \text{ enfant.}$$

La 2^{ème} moitié de la série :

$$1 \ 1 \ 2 \ 4 \ 4 \quad p = 5 = 2 \times 2 + 1$$

$$Q_3 = x_{(3)} = 2 \text{ enfants.}$$

Si on ajoute une observation ds cet exemple :

$$\underbrace{0 \ 0 \ 1 \ 1 \ 1}_{\text{Me}} \ 1 \ 1 \ 2 \ 2 \ 4 \ 4 \quad n=11 = 2 \times 5 + 1 \text{ (impair).}$$

$$Q_2 = M_e = x_{(6)} = 1.$$

$$0 \ 0 \ 1 \ 1 \ 1 \quad 1 \ 2 \ 2 \ 4 \ 4$$

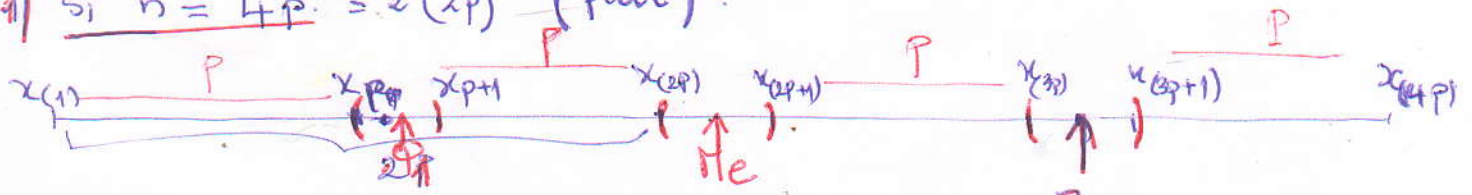
$$Q_1 = 1 \text{ enfant,} \quad Q_3 = x_{(3)} = 2 \text{ enfants.}$$

$= x_{(3)}.$

• Détermination pratique des quartiles :

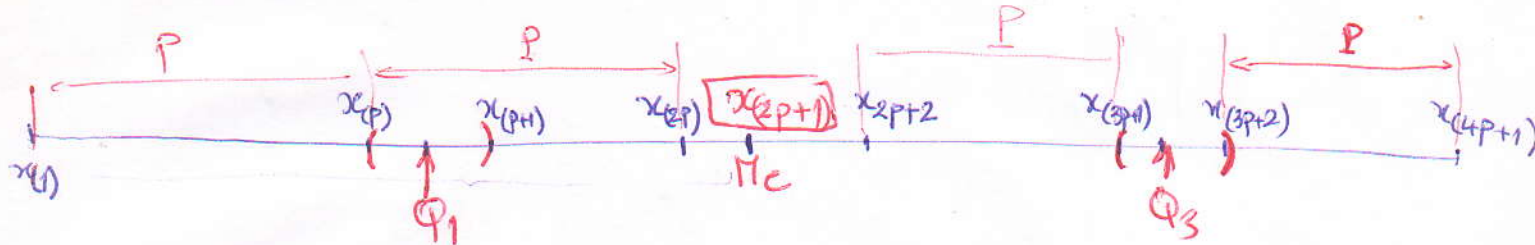
* méthode algébrique (cas discret) :

1) si $n = 4p = 2(2p)$ (pair).



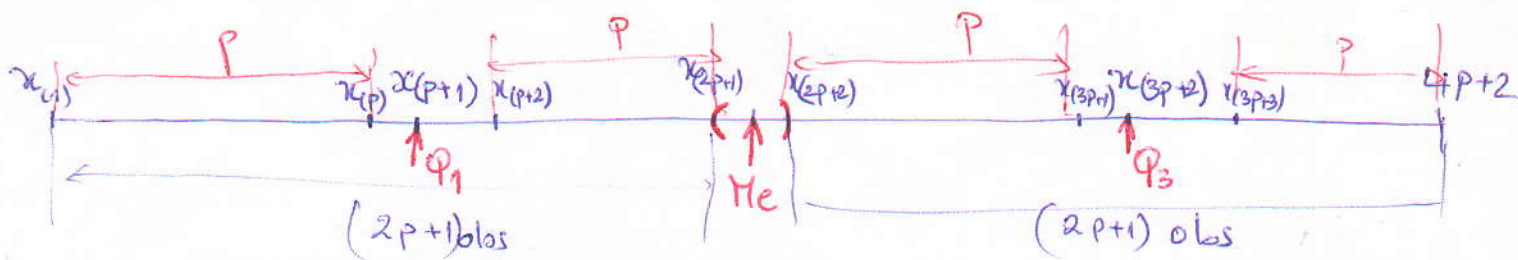
$$\left(\begin{aligned} Me &= \frac{x_{(2p)} + x_{(2p+1)}}{2} \\ Q_1 &= \frac{x_{(p)} + x_{(p+1)}}{2} \end{aligned} \right. \text{ et } Q_3 = \frac{x_{(3p)} + x_{(3p+1)}}{2}$$

2) si $n = 4p + 1 = 2(2p) + 1$ (impair).



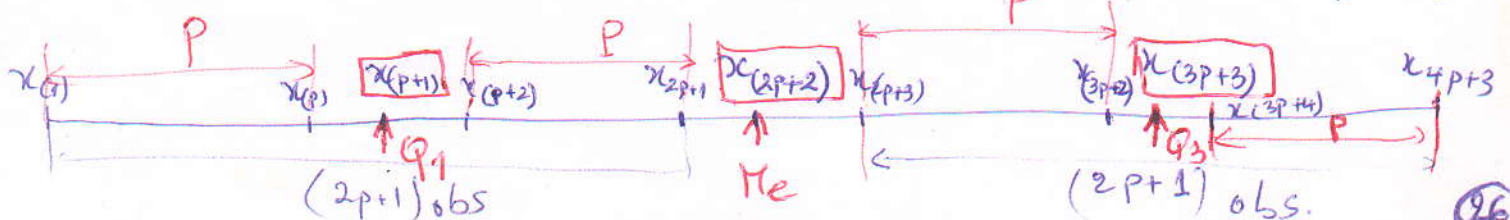
$$\left(\begin{aligned} Me &= x_{(2p+1)} \\ Q_1 &= \frac{x_{(p)} + x_{(p+1)}}{2} \end{aligned} \right. , \text{ et } Q_3 = \frac{x_{(3p+1)} + x_{(3p+2)}}{2}$$

3) si $n = 4p + 2 = 2(2p + 1)$ (pair).

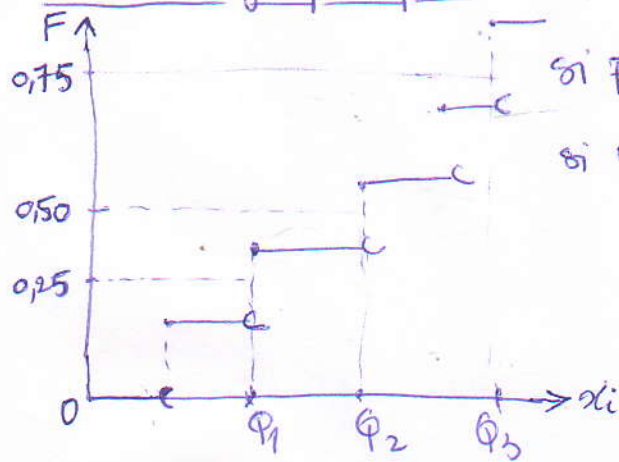


$$\left(\begin{aligned} Me &= \frac{x_{(2p+1)} + x_{(2p+2)}}{2} \\ Q_1 &= x_{(p+1)} \end{aligned} \right. \text{ et } Q_3 = x_{(3p+2)}$$

4) si $n = 4p + 3 = 2(2p + 1) + 1$ (impair). $Me = x_{(2p+2)}$, $Q_1 = x_{(p+1)}$, $Q_3 = x_{(3p+3)}$



* Méthode graphique: si $F(x_{i-1}) < 0,5 < F(x_i) \Rightarrow M_e = x_i$.



si $F(x_{i-1}) < 0,25 < F(x_i) \Rightarrow Q_1 = x_i$

si $F(x_{i-1}) < 0,75 < F(x_i) \Rightarrow Q_3 = x_i$

si $\forall x \in]x_{i-1}, x_i[$.

$$F(x) = 0,50 \Rightarrow M_e = \frac{x_i + x_{i-1}}{2}$$

• Cas Continu:

même méthode que la médiane.

si $F(e_i) = 0,50$ (resp. 0,25) (resp. 0,75).

alors $e_i = M_e$ ou $e_i = Q_1$ ou $e_i = Q_3$.

si $F(e_{i-1}) \leq 0,50 \leq F(e_i)$.

alors $M_e = e_{i-1} + a_i \frac{0,50 - F(e_{i-1})}{f_i}$ avec $[e_{i-1}, e_i[$ classe médiane.

$Q_1 = e_{i-1} + a_i \frac{0,25 - F(e_{i-1})}{f_i}$ avec $Q_1 \in [e_{i-1}, e_i[$

$Q_3 = e_{i-1} + a_i \frac{0,75 - F(e_{i-1})}{f_i}$ avec $Q_3 \in [e_{i-1}, e_i[$

— FIN .

