

Chap 2 / Introduction à la statistique descriptive

2-1 / Généralités sur la statistique :

On appelle statistique l'ensemble des méthodes (ou encore des techniques) permettant d'analyser des ensembles d'observations (ou de données).

Les méthodes en question relèvent essentiellement des mathématiques et font largement appel à l'outil informatique pour leur mise en œuvre.

La statistique descriptive a pour but de présenter les données sur une forme telle qu'on puisse en prendre connaissance et les exploiter facilement.

Elle peut conserver une seule variable (série statistique 1D) ou plusieurs variables (série statistique à plusieurs dimensions).

2-2 / Vocabulaire de base.

a) Population et individu

- On appelle population statistique, tout ensemble d'unités statistiques (ou d'unités observées).
- On appelle individu (ou unité stat), tout élément de la population à étudier.

NB: la détermination avec précision de la population (ou un échantillon de la population) et des individus qui la composent conditionne la fiabilité des résultats.

b) Variable statistique (V-S)

- On appelle variable stat (ou caractère), une application qui associe à chaque individu de la population une observation particulière.

NB: Il existe 2 types de V.S.:

① Variable qualitative:

une variable est qualitative si elle est liée à un ensemble d'observations non mesurables.

Exemple: la catégorie professionnelle dans une entreprise: (cadres, employés, ouvriers, ----)

La nature qualitative d'une variable s'exprime donc par l'appartenance à une catégorie (ou rubrique) d'un ensemble fini.

② Variable quantitative:

c'est une variable associée à un ensemble d'observations numériques.

Exemple: le pH d'une solution chimique.

Il existe des variables discrètes (VSD) et des variables continues (VSC).

c) Effectifs

- On appelle effectif d'une valeur (ou rubrique) donnée x_i le nombre de fois où cette valeur apparaît dans la population à étudier. Ce nombre est noté n_i .

$$n_i \longrightarrow n_i$$

- On appelle effectif total, noté n la somme des p effectifs particuliers n_i .

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

d) Fréquences:

- On appelle fréquence de la valeur (ou modalité) x_i , notée f_i le rapport:

$$f_i = n_i / n \quad (i = 1, p).$$

(2)

NB: $0 \leq f_i \leq 1$, $\forall i = \overline{1, p}$

1) f_i peut être exprimée en %.

2) $\sum_{i=1}^p f_i = 1$.

- on appelle fréquence cumulée croissante en n_i le nombre F_i tel que:

$$F_i = f_i^{cum} = \sum_{j=1}^i f_j$$

NB: on peut aussi définir l'effectif cumulé N_i :

$$N_i = n_i^{cum} = \sum_{j=1}^i n_j$$

1) $F_i = \sum_{j=1}^i f_j = \sum_{j=1}^i \frac{n_j}{n} = \frac{N_i}{n}$

2) Il existe aussi la freq-cumulée Δ .

1-3/ Définition d'une série statistique à une variable (SS-1D).

Une série statistique correspond aux différentes modalités d'un caractère sur un échantillon d'individus appartenant à une population donnée.

NB: Il existe plusieurs niveaux de description statistique: la présentation brute des données, des présentations par tableaux numériques, des représentations graphiques, ...

1-3-2/ Tableaux Statistiques:

le tableau de distribution de fréquences et un mode synthétique de représentation.

a) Caractère qualitatif :

	Modalités (Numéro i)	Effectif n_i	Fréquence f_i
	1	n_1	f_1
	2	n_2	f_2
	$\vdots$	$\vdots$	$\vdots$
	p	n_p	f_p
Total		n	1.

b) Caractère quantitatif discret : VSD.

Valeurs obs (n_i)	n_i	f_i	f_i^{cum}
n_1	n_1	f_1	f_1
n_2	n_2	f_2	$f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$	
n_p	n_p	f_p	1
Total	n	1.	

c) Caractère quantitatif continu : (VSC)

Classes $[b_i, b_{i+1}[$	Centre c_i	n_i	f_i	f_i^{cum}
$[b_1, b_2[$	c_1	n_1	f_1	f_1
$[b_2, b_3[$	c_2	n_2	f_2	$f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$		
$[b_p, b_{p+1}[$	c_p	n_p	f_p	1
Total		n	1.	

NB:

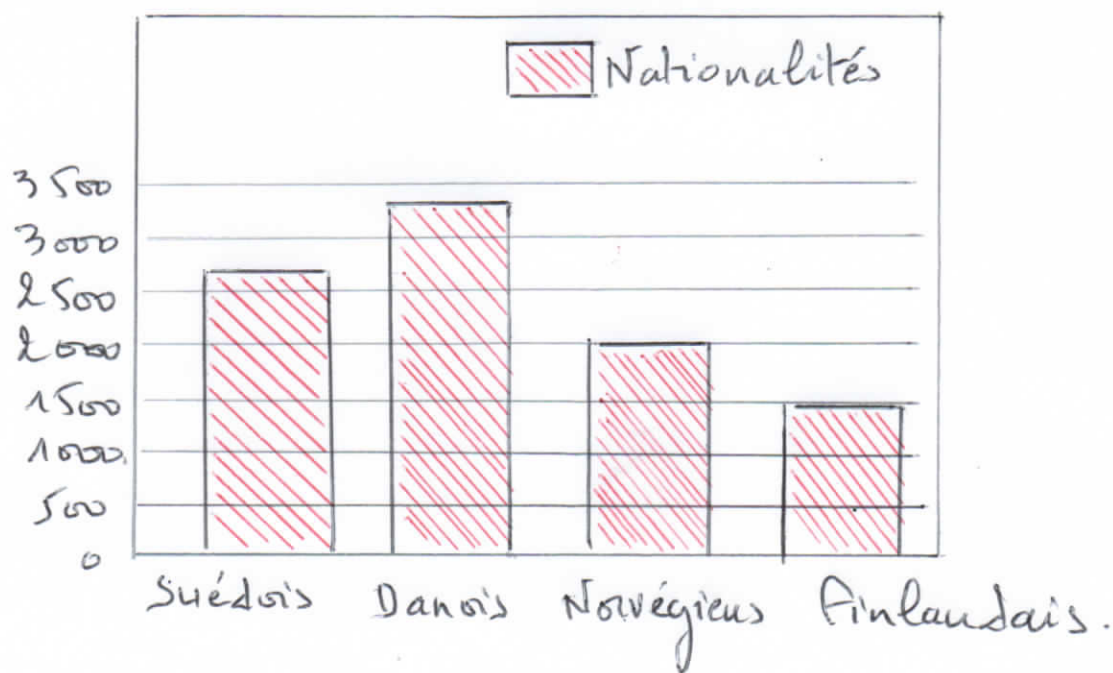
- Une classe et un intervalle de type $[b_i, b_{i+1}[$
- le centre de classe c_i est: $c_i = \frac{b_i + b_{i+1}}{2}$
- l'amplitude d'une classe est: $2_i = b_{i+1} - b_i$

1-3-2/ Représentations graphiques:

a/ Caractère qualitatif

a-1/ Diagramme à bandes: c'est un diagramme qui à chaque modalité de la variable associe un rectangle de base constante et dont la hauteur est proportionnelle à l'effectif.

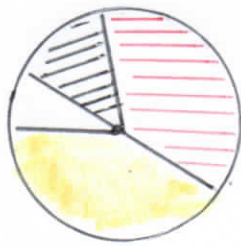
Exemple: nombre d'arrivées aux frontières de scandinaves par nationalités pour l'année 1985.



a-2/ Diagramme à secteurs: c'est un graphique qui divise un disque en secteurs angulaires dont les angles au centre sont proportionnels aux effectifs de chaque modalité.

Pour une modalité donnée M_i d'effectif n_i , l'angle au centre α_i correspondant est donné (en degré) par: $\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360$.

Exemple: (Précédent)



 Norwegiens.
 Danois.
 Suédois.
 Finlandais.

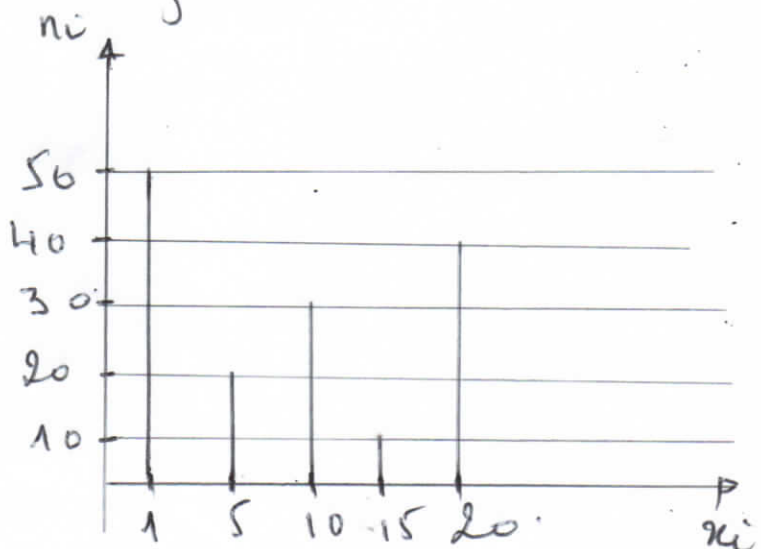
b) Caractère quantitatif discret (VSD).

b-1/ Diagramme en batons: c'est un graphique, qui associe à chaque valeur de la variable un segment (baton) dont la hauteur est proportionnelle à l'effectif.

NB: on suppose les valeurs observées de la VSD ordonnées par ordre croissant.

Exemple: soit la série statistique suivante:

x_i	1	5	10	15	20
n_i	50	20	30	10	40



b-2/ Diagramme cumulé:

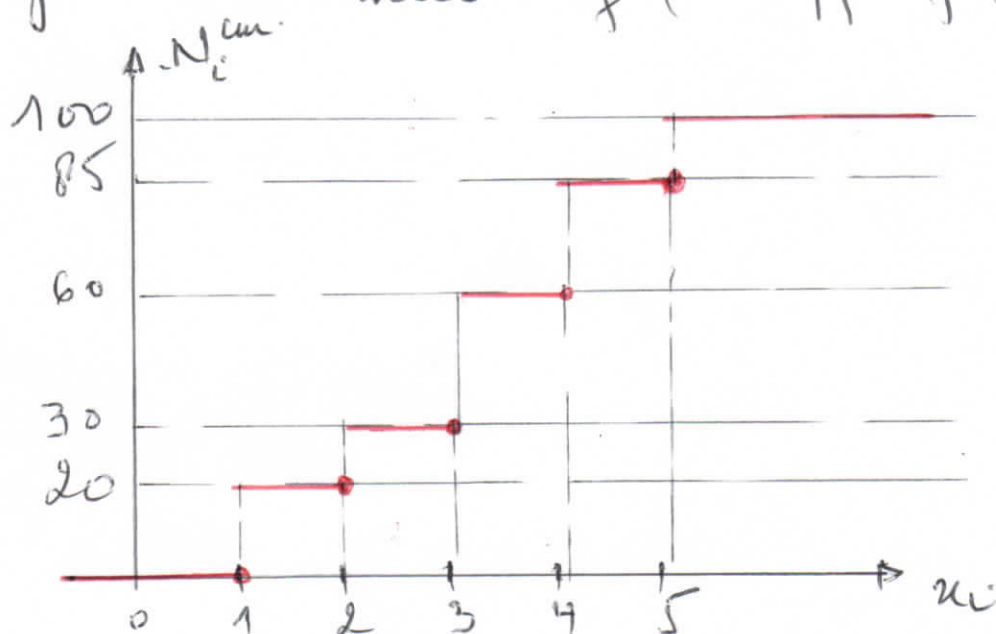
c'est le diagramme représentatif de la fonction de répartition observée c'est à dire la fréquence cumulée F_i ou N_i^{cum} .

Exemple: Soit la série stat donnée par le tableau:

x_i	1	2	3	4	5
n_i	20	10	30	25	15
N_i^{cum}	20	30	60	85	100

$$N_i^{cum} = \sum_{j=1}^i n_j \quad | \quad \text{on a } n = \sum_{i=1}^5 n_i = 100 \text{ (Effectif Total)}$$

On a le diagramme cumulatif (des effectifs) suivant:



NB: le diagramme cumulatif d'une VSD est une fonction en escalier.

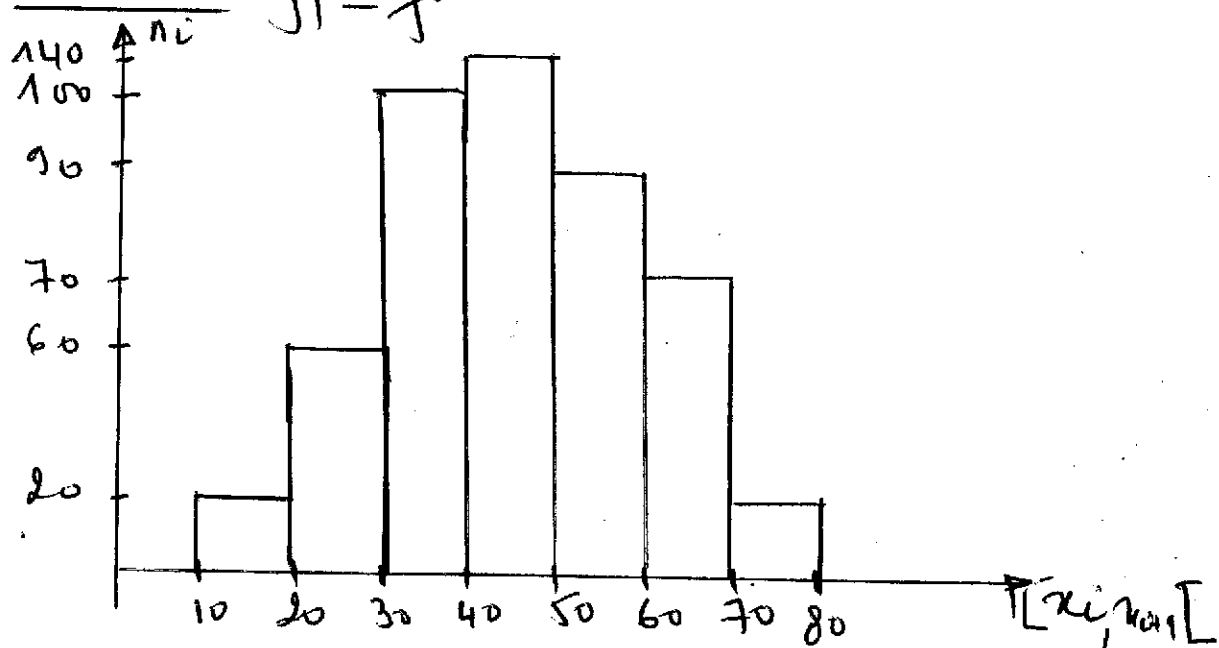
c) Caractère continu (vs c)

c-1) Histogramme: c'est un diagramme composé de rectangles adjacents, chaque rectangle, associé à chaque classe, ayant une surface proportionnelle à l'effectif (ou à la fréquence) de cette classe.

Exemple: Soit la série donnée par le tableau suivant.

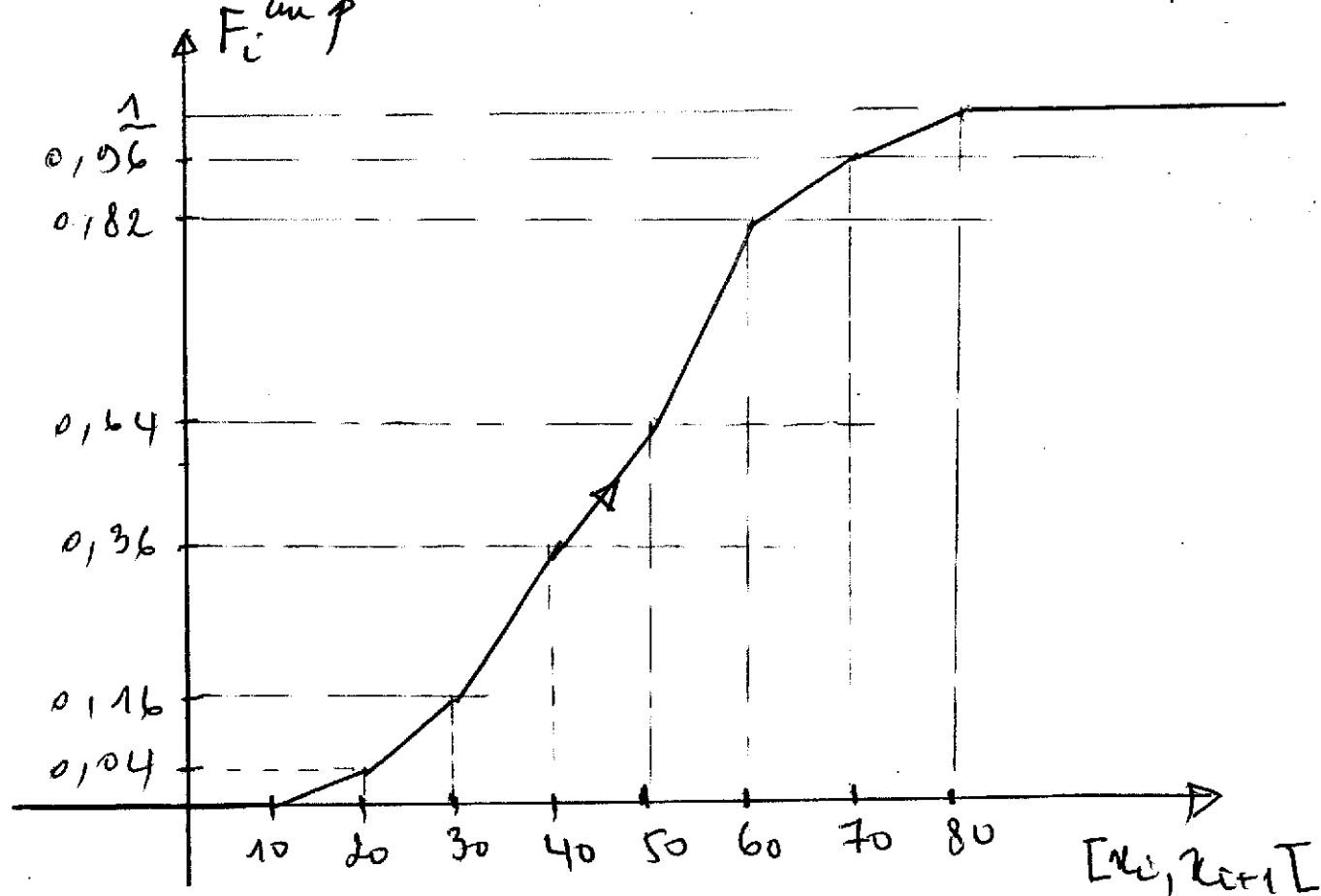
$[x_i, x_{i+1}[$	n_i	z_i	f_i	$F_i^{cum} \neq$
$[10, 20[$	20	15	0,04	0,04
$[20, 30[$	60	25	0,12	0,16
$[30, 40[$	100	35	0,2	0,36
$[40, 50[$	140	45	0,28	0,64
$[50, 60[$	90	55	0,18	0,82
$[60, 70[$	70	65	0,14	0,96
$[70, 80[$	20	75	0,04	1
Total	500			

* Histogramme des effectifs:



c-2/ Diagramme cumulé: (polygone des fréquences)
 c'est le diagramme représentatif de $F_i^{cum} \neq$ ou
 de $N_i^{cum} \neq$ (formé à l'aide de segment
 de droite).

**/ Diagramme des fréquences cumulées $\neq$
 - Exemple précédent - (8) $F_i^{cum} \neq = \sum_{j=0}^i f_j$



1-3-3/ Paramètres caractéristiques d'une variable statistique:

1-3-3-1/ Paramètres de position:

Les paramètres de position (ou de tendance centrale) permettent de savoir autour de quelles valeurs se situent les valeurs d'une variable statistique.

a) Mode (M_0)

- Pour une VSD, le mode est la modalité qui représente le plus grand des effectifs.
- Pour une VSC, nous parlons de classe modale dont l'effectif est maximum.

$$(n_i)_{\max} \longrightarrow x^* ([x_i^*, x_{i+1}^*[)$$

b) Moyenne:

- la moyenne arithmétique

$$\text{Si } X \sim \text{VSD} / X \in \{x_1, x_2, \dots, x_p\} \Rightarrow$$

(9)

la moyenne arithmétique de X est:

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_p}{n}$$

- la moyenne pondérée: lorsque les valeurs sont affectées de coefficients (ici d'effectifs), on parle de moyenne pondérée.

*/ Donnée VSD: $\bar{X} = \frac{1}{n} \sum_{i=1}^p n_i x_i$
(n = effectif total)

*/ Donnée VS C: $\bar{X} = \frac{1}{n} \sum_{i=1}^p n_i c_i$
(c_i = centre de classe)

NB: Soient deux séries (S_1) et (S_2), telles que:

n_i	x_1	x_2	...	x_p
n_i	n_1	n_2	...	n_p

(S_1)

n_i	n_1	n_2	...	n_p
y_i	y_1	y_2	...	y_p

si $\forall i = 1, p: y_i = a x_i + b \Rightarrow \boxed{\bar{Y} = a \bar{X} + b}$

c) Médiane (Me)

- c'est le nombre qui permet de couper la population étudiée en deux groupes contenant le même nombre d'individus.



NB: Pour calculer Me, il faut d'abord ordonner la série (A v B).

- Calcul de Me pour donnée VSD

On distingue les cas suivants:

*1/ si le nombre de valeurs prises par X est impair: $p = 2k+1 \Rightarrow Me$ est parfaitement déterminée: $Me = x_{k+1}$

Exemple: Soit la série:

x_i	0	1	2	3	4	5	6
n_i	30	16	14	5	10	15	2

on a $p = 7 = 2 \times 3 + 1 \Rightarrow$

$$Me = x_{3+1} = x_4 = 3$$

*1/ si le nombre de valeurs prises par X est pair ($p = 2k$) $\Rightarrow$ on ne peut pas définir parfaitement la médiane. On définit un intervalle médian.

Exemple: Soit la série

x_i	4	4	6	7	8	12	15	18	20	30
n_i	10	15	5	7	20	30	15	10	5	2

on a $p = 10$, l'intervalle médian est $]9, 12[\Rightarrow$

$$Me \approx \frac{9+12}{2} = 10,5$$

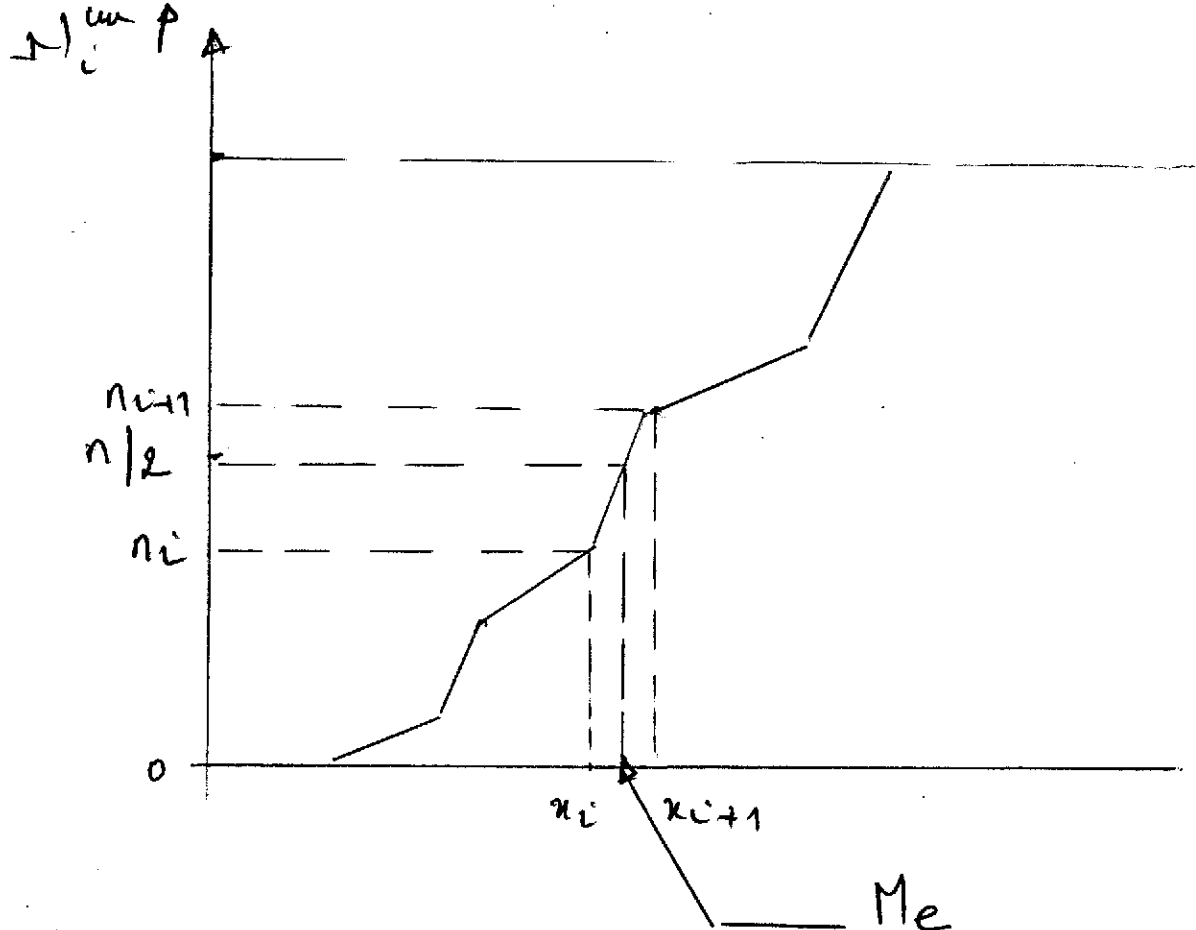
- Calcul de Me pour une VSC

Pour une VSC, on détermine la classe médiane de même façon que pour une VSD en utilisant les effectifs cumulés.

NB: On détermine d'abord la classe médiane. puis le calcul de Me se fait par des méthodes d'interpolation. (dans ce chapitre, on utilisera l'interpolation linéaire).

- Calcul de Me par interpolation linéaire

On trace d'abord le graphique des N_i^{cum} puis, on calcule Me .
(11)



On détermine d'abord l'équation de la droite qui passe par les points (n_i, x_i) et (n_{i+1}, x_{i+1}) : $y = ax + b$.

on a : $n/2 = a Me + b \Rightarrow Me \approx \frac{\frac{n}{2} - b}{a}$.

1-3-3-2/ Paramètres de dispersion

Les paramètres de dispersion nous renseignent sur la dispersion des valeurs autour de la valeur centrale de référence.

a) Étendue (E_t).

L'étendue d'une série statistique quantitative est la différence entre la plus grande valeur de la variable et la plus petite valeur.

$$E_t = x_{\max} - x_{\min}$$

b) Variance et Écart-type.

* Pour me VSD: $\text{Var}(X) = \frac{1}{n} \sum_{i=1}^p n_i n_i^2 - \bar{X}^2$

* Pour me VSC: $\text{Var}(X) = \frac{1}{n} \sum_{i=1}^p n_i c_i^2 - \bar{X}^2$

* l'écart-type est $\sigma_x = \sqrt{\text{Var}(X)}$ ($\sigma_x^2 = \text{Var}(X)$)

c) Quartiles:

- le premier quartile ($Q1$) d'une série stat est la plus petite valeur prise par la série, telle que 25% au moins des valeurs de la série soient inférieures ou égales à $Q1$. ($n_i^* = Q1$)
- le troisième quartile ($Q3$): c'est la plus petite valeur n_i^{**} prise par la série, telle que 75% au moins des valeurs de la série soient $\leq$ à $Q3$ ($n_i^{**} = Q3$).

NB: l'intervalle interquartile est: $[Q1, Q3]$

.) l'écart interquartile est: $\Delta = Q3 - Q1$.

d) Déciles:

- le premier décile $\rightarrow (D1)$ d'une série-stat est la plus petite valeur y_i^* prise par la série, telle que 10% au moins des valeurs de la série soient $\leq$ à y_i^* . ($D1 = y_i^*$)
- le neuvième décile ($D9$) d'une S-S est la plus petite valeur $D9$ prise par la série, telle que 90% au moins des valeurs de la série soient $\leq$ à $D9$.

NB!

- l'intervalle interdecile est: $[D_1, D_9]$
- l'écart interdecile est $D = D_9 - D_1$.