Partie II = Statistique

Chapitre I: Generalité et notions de base.

Definition = La statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, organisé puis à analyser, à commenter, à critiquer ces données et à Tirer des conclusions et prendre la décision qu'il faut.

Vocabulaire general:

Population = C'est l'ensemble des objets ou des personnes retenues dans notre étude statistique on appelle aussi univers

Individus = les éléments de la population seront appelés individus.

L'echantillon = le plus souvent il est difficile d'observer toutes les données on lieu examiner l'ensemble totale, on examine une partie, tout sous ensemble de la population est une échantillon.

Le caractère c'est le critaire retenu pour mener l'étude statistique, chaque individu présente un caractère un caractere peut être:

qualitatif = Par exp: cause d'un accident ou =

quantitatif = Par exp: une durée de vie.

- Un caractère quantitatif est continu, s'il peut prendre toutes les valeurs d'un intervalle (par exp = une longueur).
- Il est discontenu ou discret s'il ne peut prendre que des valeurs isolées (par exp : le nombre d'enfants dans la famille).

Remarque : Quand le caractère est continu, ou quand il est discret avec beaucoup de valeurs possibles, on effectue des regroupements en classes statistiques.

Définition d'une série statistique :

Une série statistique est l'ensemble des valeurs numériques resultants de l'observation : la série statistique est du même type que le caractère qui la constitue.

exp : si le caractère est quantatif ⇒ la serie est quantative

- Le nbre total des valeurs numériques de la population ou l'échantillon est appelé l'effectif total et on note n
- L'effectif partiel (fréquence absolue)

c'est le nombre de fois on l'on a rencontré cette valeur du caractère dans la série statistique, la distribution est noté n_i et on a : $\sum_{i=1}^{+} n_i = n$ / n_i = l'effectif partiel

- la fréquence (relative) :

si n_i est l'effectif partiel de la modalité (ou valeurs)

x_i, le rapport = $f_i = \frac{n_i}{n}$ s'appelle la frequence.

• En multipliant la fréquence f_i par 100 on obtient le pourcentage = $p_i = 100 f_i$ de la valeur x_i.

Remarque = $0 \leq n_i \leq n \Leftrightarrow 0 \leq \frac{n_i}{n} \leq \frac{n}{n} \Rightarrow 0 \leq f_i \leq 1$.

$\sum_{i=1}^{p} n_i = n$, $\sum_{i=1}^{n} f_i = 1$, $\sum_{i=1}^{p} p_i = 100$

1] Un caractère discret = les données $(x_i, n_i) / 1 \leq i \leq p$ ou (x_i, f_i) $1 \leq i \leq p$, s'appelle une série statistique discrète ou discontenues esc: on a compté dans une classe de 30 élèves. le nombre de dents atteints par la carie, les resultats sont données par:

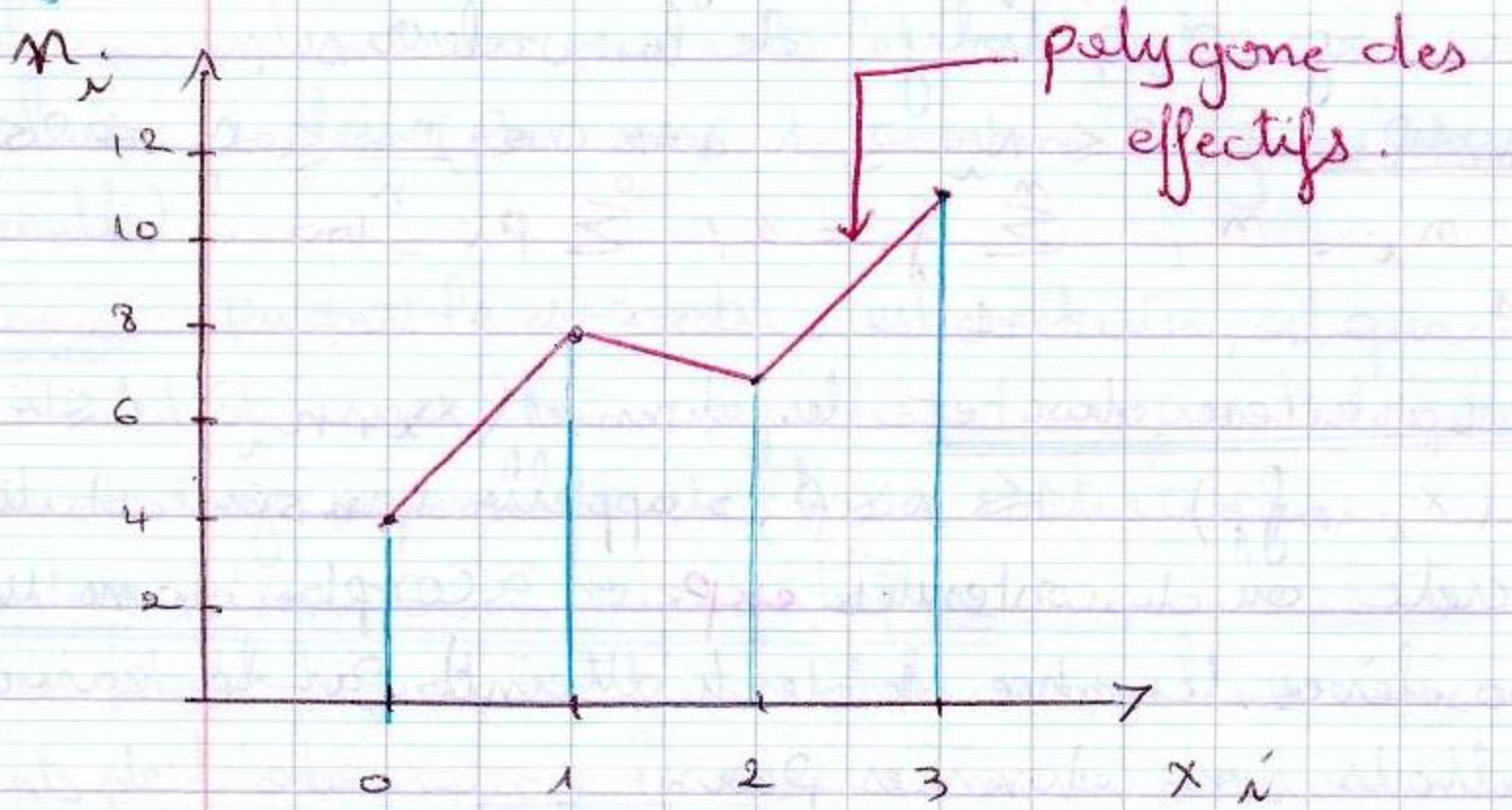
1, 3, 3, 0, 1, 2, 3, 3, 0, 2, 2, 3, 3, 2, 3, 2, 1, 0, 2, 1, 1, 2, 3, 1, 3, 3, 3, 0, 1. Nous interprétons les resultats suivants on dissout que: le 1er élève à un dent atteint

" 2ème " " 3 " " "

" 3ème " " 3 " "

x_i	n_i	f_i	p_i
0	4	0,1333	$\simeq$ 13,33
1	8	0,2666	$\simeq$ 26,66
2	7	0,2333	$\simeq$ 23,33
3	11	0,3666	$\simeq$ 36,66
Σ	30	1	100

La representation graphique de cette serie donnée un diagramme en bâtons



2/ Un caractère continu :

Si les données de la statistique ont été groupées en classes $[a_1, a_2[\dots [a_p, a_{p+1}[$ d'effectif $n_1, \dots, n_p$ la famille $([a_i, a_{i+1}[, n_i)\ 1 \leq i \leq p$ s'appelle une série statistique groupée ou continue.

Définitions pour la classe est $[a_i, a_{i+1}[$:

- Les nbres a_i, a_{i+1} s'appellent les **limites**, ou **les bornes** de la classe.
- la demi somme $\frac{a_i + a_{i+1}}{2} = C_i$ s'appelle **le centre** de la classe.
- La difference $a_{i+1} - a_i = e_i$ s'appelle **l'endue** ou

L'amplitude de la classe.

- L'effectif de la classe est la somme des effectifs partiels des valeurs appartenant à la classe.

Choix du nbre de classes :

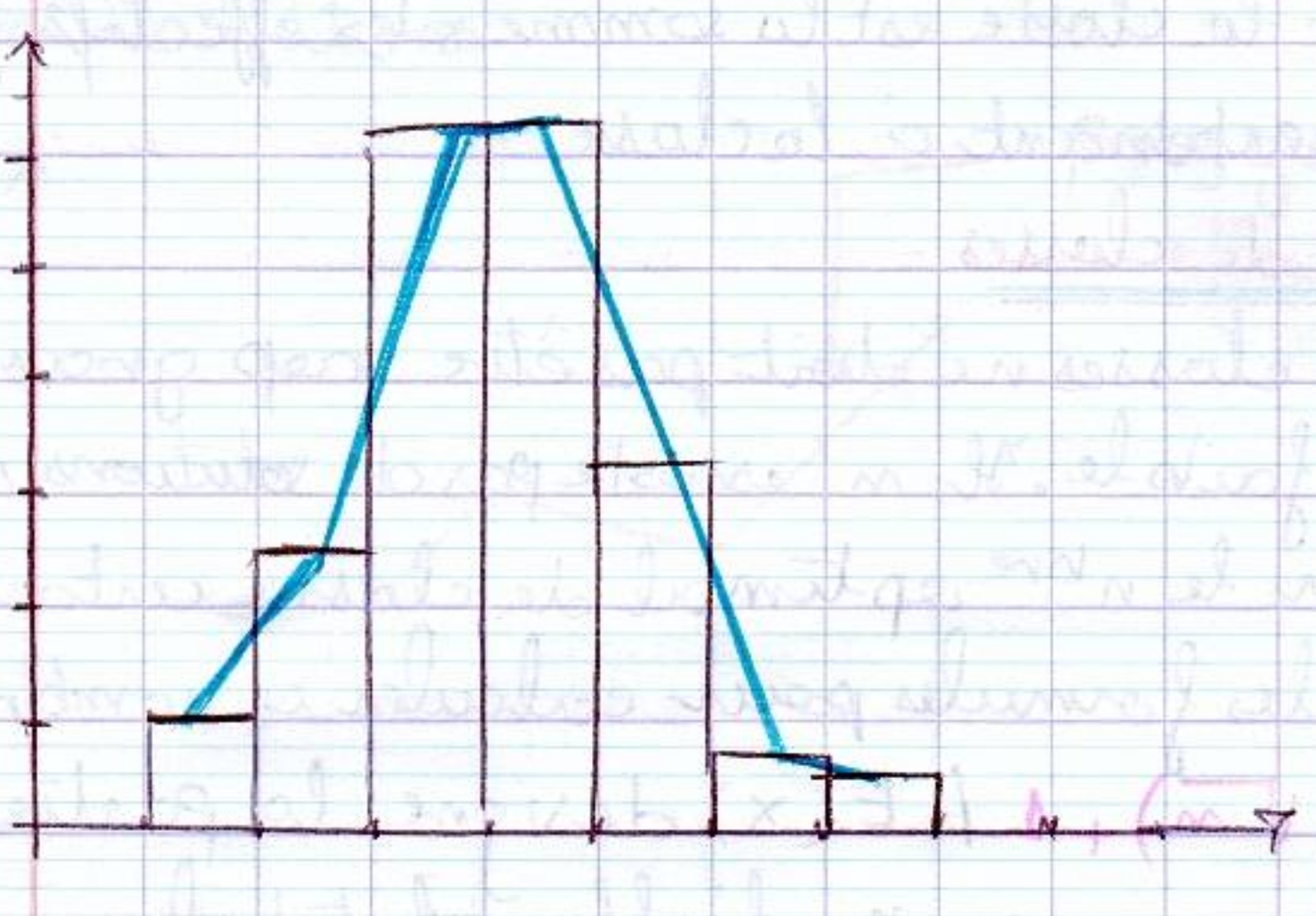
Le nbre des classes ne doit pas être trop grand et ne doit pas être trop faible. Il n'existe pas de solutions mathématique pour déterminer le nbre optimal de classes certains auteurs ont proposé les formules pour calculer ce nombre, voici, un exp. $C = E(\sqrt{n}) + 1$ / $E(x)$ désigne la partie entière de x.

n l'effectif total

exp : On a étudié le poids dans une groupe de 100 adultes normaux de sexe féminin.

Classe	C_i	n_i	f_i	P_i
39,5 - 44,5	42	5	0,05	5
44,5 - 49,5	47	12	0,12	12
49,5 - 54,5	52	31	0,31	31
54,5 - 59,5	57	31	0,31	31
59,5 - 64,5	62	16	0,16	16
64,5 - 69,5	67	3	0,03	3
69,5 - 74,5	72	2	0,02	2
Σ	/	100	1	100 %

La representation graphique de la serie se fait par histogr-
amme.



3/ Un caractère qualitatif :

Lorsqu'il s'agit d'un caractère qualitatif tel que : le sexe, la couleur des yeux ou des cheveux, on classe les resultats en catégories ou modalités et on indique le nombre ni de cas qui retient dans chaque catégorie

pour la representation graphique On peut realiser une série qualitatif on plusieurs methodes :-

Un diagramme à secteurs.

Un diagramme en bandes. (Tuyaux d'orgue)

* Effectif et Frequences cumulées croissantes :

Soit (x_i, n_i) $1 \leq i \leq p$ une série statistique

discrète prenant les valeurs $x_1, x_2, \dots x_p$ classés par ordre croissant.

• On appelle effectif cumulé croissante de la valeur x_i la somme des effectifs partiels des valeurs $x_1, x_2, \dots, x_i$.

$$N_i = n_1 + n_2 + \dots + n_i = \sum_{k=1}^{i} n_k = n_i \nearrow$$

Ainsi l'effectif cumulé croissante d'une variable x_i est le nbre d'individus de la population sur lesquels la série X prend une valeur inferieur ou égal x_i ($\leq x_i$).

On appelle frequence cumulée croissante de la valeur x_i la somme des fréquences des valeurs $x_1, x_2 \dots, x_i$.

$$F_i = f_1 + f_2 + \dots + f_i = \sum_{k=1}^{i} f_k = f_i \nearrow$$

• Effectif et fréquences cumulés décroissantes :

On appelle effectif cumulé décroissante de la valeur x_i le nombre

$$n_{i\searrow} = n - [n_1 + n_2 + \dots + n_{i-1}] = n - \sum_{k=1}^{i} n_k$$

Ainsi l'effectif cumulé décroissante d'une valeur x_i est le nbre d'individus de la population

sur les quels la serie prend une valeur superieur ou égale x_i, $(\geq x_i)$.

* On appelle fréquence cumulée décroissante de la valeur x_i le n^bre

$$f_{i\downarrow} = 1 - \left[f_1 + f_2 + \dots + f_i \right]$$

$$= 1 - \sum_{k=1}^{i-1} f_k$$

Exemple --

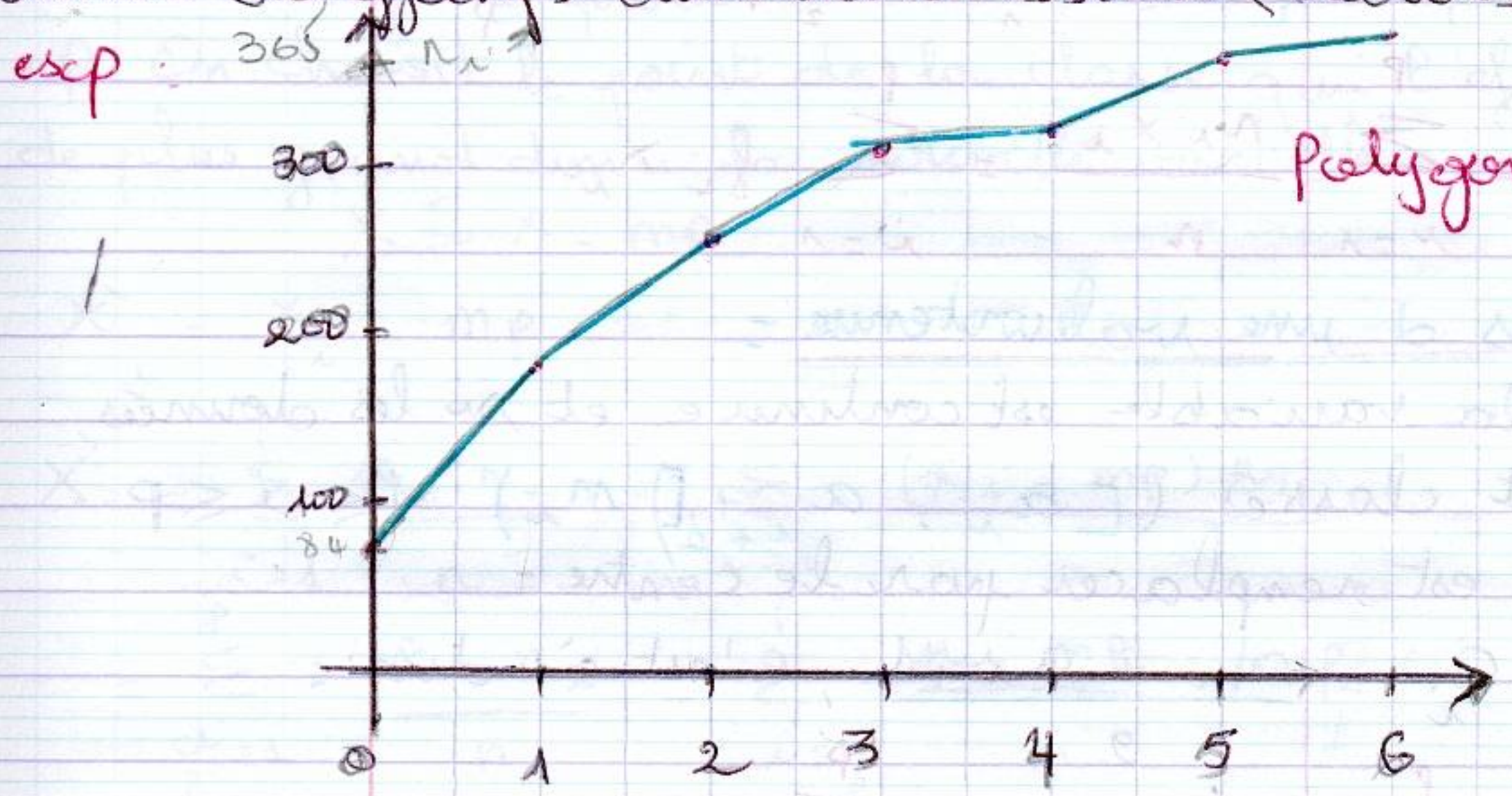
le n^bre d'intervention par jour des pompiers dans une caserne est distribué comme indiqué le tableau (pendant une année)

x_i : le nobre de sortie par jour.

x_i	n_i	f_i	$N_i \uparrow$	$f_i \uparrow$	$n_{i\downarrow}$	$f_{i\downarrow}$
0	84	0,230	84	0,230	365	1
1	105	0,287	189	0,517	281	0,77
2	72	0,197	261	0,714	176	0,483
3	59	0,161	320	0,875	104	0,286
4	28	0,076	348	0,951	45	0,125
5	15	0,041	363	0,992	17	0,049
6	2	0,005	365	1	2	0,008
$\sum$	365	1				

La représentation graphique de la série (x_i, n_i) $1 \leq i \leq p$.

$(x_i, n_{i\nearrow})$ $1 \leq i \leq p$ donne un diagramme appelé courbe des effectifs cumulées croissante (décroissante)



* Les paramètres de position :

1/ La moyenne arithmétique

la moyenne arithmétique d'une série de valeurs d'une variable statistique est égale à la somme de ces valeurs divisée par leur nombre :

* Cas d'une V. discrète :

Soit (X_i, n_i) $1 \leq i \leq p$ où (x_i, f_i) $1 \leq i \leq p$ une série statistique discrète. On appelle moyenne

arithmétique de cette série le nombre :

$$\bar{X} = \frac{n_1 x_1 + n_2 x_2 + n_3 x_3 + \dots + n_p x_p}{n_1 + n_2 + \dots n_p}$$

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

* Cas d'une v. continue :

Si la variable est continue et si les données sont classés $([a_i, a_{i+1}[\ n_i)$ $1 \leq i \leq p$.

x_i est remplacer par le centre :

$C_i = \frac{a_i + a_{i+1}}{2}$, c'est à dire :

$$\bar{X} = \sum_{i=1}^{p} \frac{n_i \cdot C_i}{n} = \sum_{i=1}^{p} f_i C_i$$

Remarque : Si on a pas de pondération (répétition) $\bar{X}$ est donnée par :

$$\bar{X} = \frac{x_1 + x_2 + \dots + x_p}{1 + 1 + \dots + 1} = \sum_{i=1}^{p} \frac{x_i}{p}$$

La moyenne provisoire :

Lorsque les données sont nombreuses on peut simplifier le calcul de la moyenne arithmétique en adoptant le procède dit **La moyenne provisoire** mp. On choisit le point de la classe qui l'effectif de plus grand donce la nouvelle variable.

$$X' = X - mp$$

$$X'_i = X_i - mp$$

$$\bar{X}' = \sum_{i=1}^{p} \frac{x'_i n_i}{n} = \sum_{i=1}^{p} \frac{(x_i - mp) n_i}{n}$$

$$= \sum_{i=1}^{p} \frac{x_i n_i}{n} - \sum_{i=1}^{p} \frac{n_i mp}{n} \quad / mp = c^{te}$$

$$= \sum_{i=1}^{p} \frac{x_i n_i}{n} - mp \sum_{i=1}^{p} \frac{n_i}{n}$$

$$\bar{X}' = \bar{X} - mp \implies \bar{X} = \bar{X}' + mp.$$

exp - Revenons à la distribution de poids. / $\hat{mp} = 52$

Classes	c_i	n_i	$n_i \cdot c_i$	$c'_i = c_i - mp$	$c'_i \cdot n_i$
] ,]	42	5	210	-10	-50
] ,]	47	12	564	-5	-60
] ,]	52	31	1612	0	0
] ,]	57	31	1767	5	155
] ,]	62	16	992	10	160
] ,]	67	3	201	15	45
] ,]	72	2	144	20	40
$\sum$		100	5490		290

$$\bar{X} = \frac{\sum n_i C_i}{n} = \frac{5490}{100} = 54,9 \text{ Kg}.$$

$$\bar{X'} = \frac{\sum n_i C_i}{n} = \frac{290}{100} = 2,9 \text{ Kg}.$$

$$\bar{X'} + mp = 2,9 + 52 = 54,9 = \bar{X}$$

2/ Le mode (la valeur dominante) M_0 =
Le mode ou la valeur modale est la valeur que la variable prend plus fréquemment

* Variable discrète : On appelle mode d'une série discrète la valeur x_i correspondante à n_i maximale.

* variable continue : Pour déterminer le mode d'une v.c. en 1er lieu on détermine la classe modale

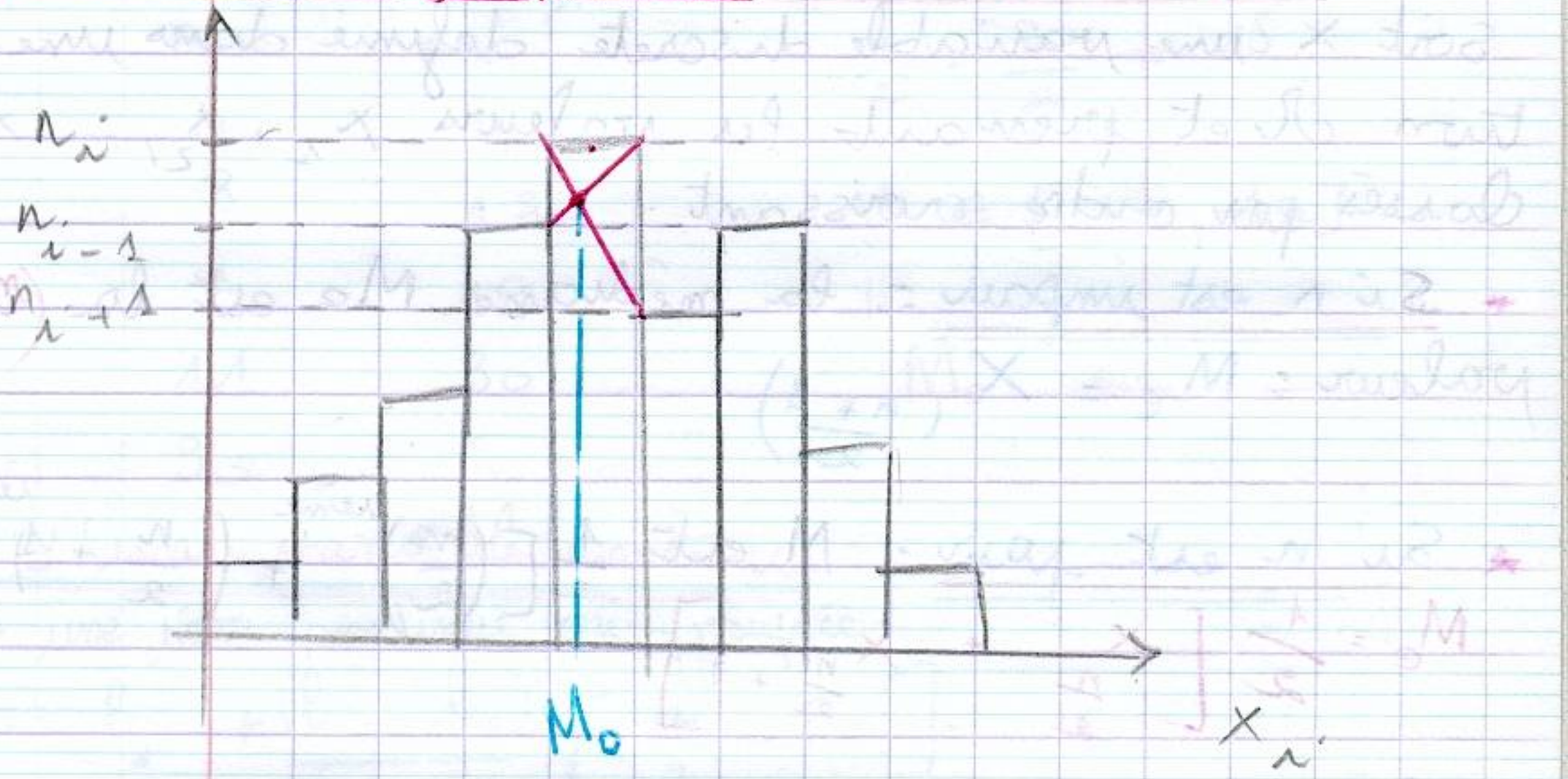
On appelle classe modale d'une série continue, toute classe d'effectif maximal.
le mode est donnée par la formule

$$M_0 = l_1 + (l_2 - l_1)\left[\frac{n_i - n_{i-1}}{2n_i - (n_{i-1} + n_{i+1})}\right]$$

$M_0 \in [l_1, l_2[$ ← classe modale.

l_1 : la limite inferieure de la classe modale.
l_2 : // // superieure // // // //
n_i : l'effectif de la classe modale.
n_{i-1} : // // // // precedente
n_{i+1} : // // // // suivante.

Representation graphique (Histogramme)



Remarque :

1/ le mode n'est pas unique
2/ Une serie statistique peut avoir 2 modes on dira dans ce cas qu'elle est bimodale ou 3 modes (trimodale), ou en generale plurimodale

3/ La médiane : (M_e) :

La médiane est la valeur x_i de la variable statistique qui correspond à l'effectif cumulés $\frac{N}{2}$ ou à la fréquence cumulée 0,5 (50%)

la médiane est la valeur x_i qui divise une série statistique en 2 sous groupe de même effectif.

Cas d'une variable discrete:

Soit x une variable discrete definié dans une population N et prenant les valeurs $x_1, x_2, \dots x_n$ classés par ordre croissant.

* **Si n est impair** : la médiane Me est la $\left(\frac{n+1}{2}\right)^{ème}$ valeur : $M_e = X_{\left(\frac{n+1}{2}\right)}$

* **Si n est pair** : M est $\frac{1}{2}\left[\left(\frac{n}{2}\right)^{ième} + \left(\frac{n}{2}+1\right)^{ième}\right]$

$$M_e = \frac{1}{2}\left[x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right]$$

exp :

On a les 2 série suivantes

1/ 2 2 3 3 3 4 [5] 6 6 7 8 8 9.

$n = 13 \Rightarrow M_e = X_{\frac{13+1}{2}} = X_7$.

$M_e = 5$

2/ 2 2 3 3 3 (4) ↓ (5) 6 6 7 8 8.

$n = 12 \Rightarrow M_e = \frac{1}{2}\left[x_{\frac{12}{2}} + x_{\frac{12}{2}+1}\right] = \frac{1}{2}\left[x_6 + x_7\right]$

$M_e = \frac{1}{2}[4 + 5] = 4,5$.

exemple 2 : si les données sont classés dans un tableau

x_i	n_i	$n_i\uparrow$
0	4	4
1	8	12
2	7	19
3	11	30
Σ	30	

$\frac{n}{2} = \frac{30}{2} = 15$.

$15 \in [12, 19]$.

$M_e = 2$

Cas d'une variable continue :

Pour une série continue ou groupée

$$M_e = l_1 + \frac{(l_2 - l_1)}{n}\left[\frac{n}{2} - F\uparrow\right].$$

- l_1, l_2 : sont les bornes de la classe contenant la médiane.
- n : est l'effectif de cette classe médiane.
- $F\uparrow$: c'est l'effectif cumulé de toutes les classes précédant la classe médiane.

esd:

* Considerons la série suivantes :

classes	n_i	$n_{cum\,i}$ ↗
[10, 15[	10	10
[15, 20[	11	21
[20, 25[	17	38
[25, 30[	06	44
[30, 35[	12	56
[34, 40[	04	60
Σ	60	

$n = 60$

$\Rightarrow \frac{n}{2} = 30$

$\frac{n}{2} = 30 \in [21, 38]$

$M_e \in [20, 25[$.

$$M_e = 20 + \frac{(25 - 20)}{17}[30 - 21] = 22,64$$

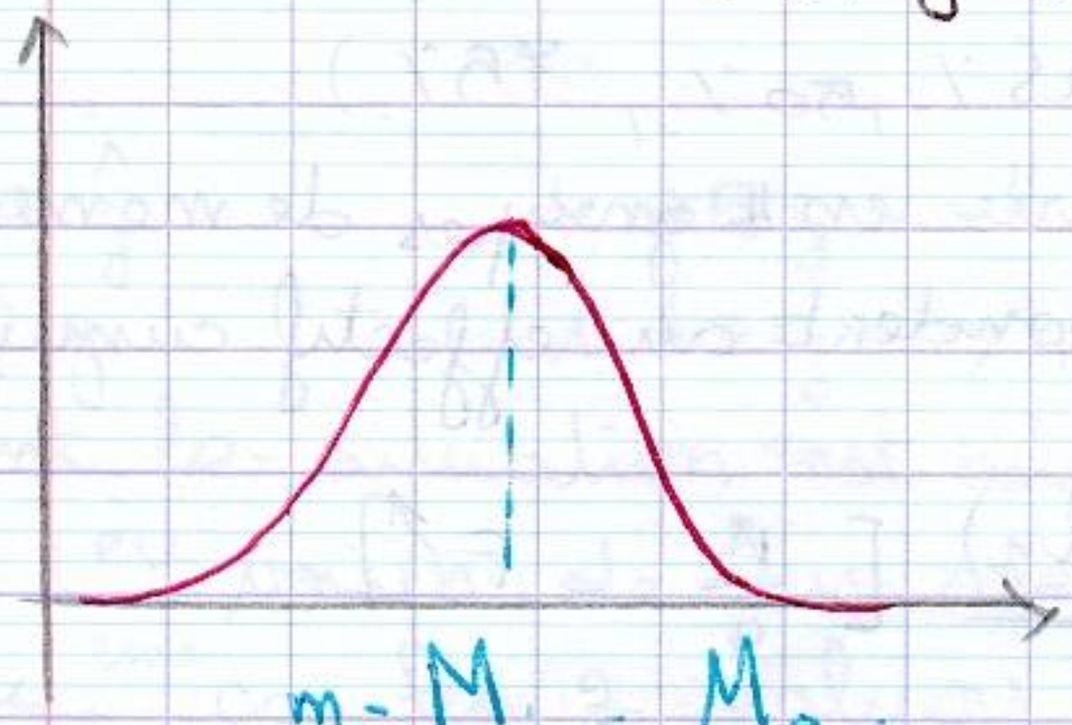
Determination graphique de la mediane :

la determination graphique de la médiane se fait, on trace le polygone des effectifs cumulés (des fréquences cumulés) de rechercher l'abscisse du point de ce polygone d'ordonnée egale $\frac{n}{2}$ ou $(O, \vec{j})$. Une autre méthode : on trace les 2 polygones $n_{i\,\nearrow}$, $n_{i\,\searrow}$

- M_e est l'abscisse du point d'intersection

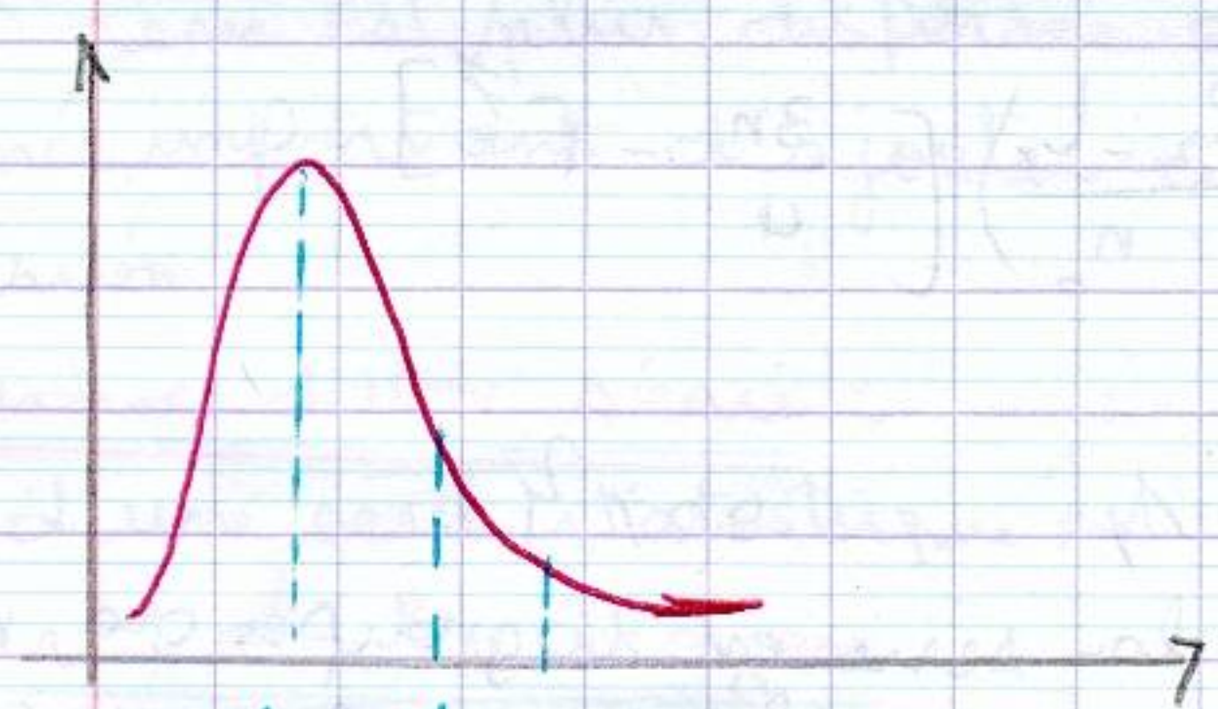
Remarque :

* Si mode = médiane = moyenne.



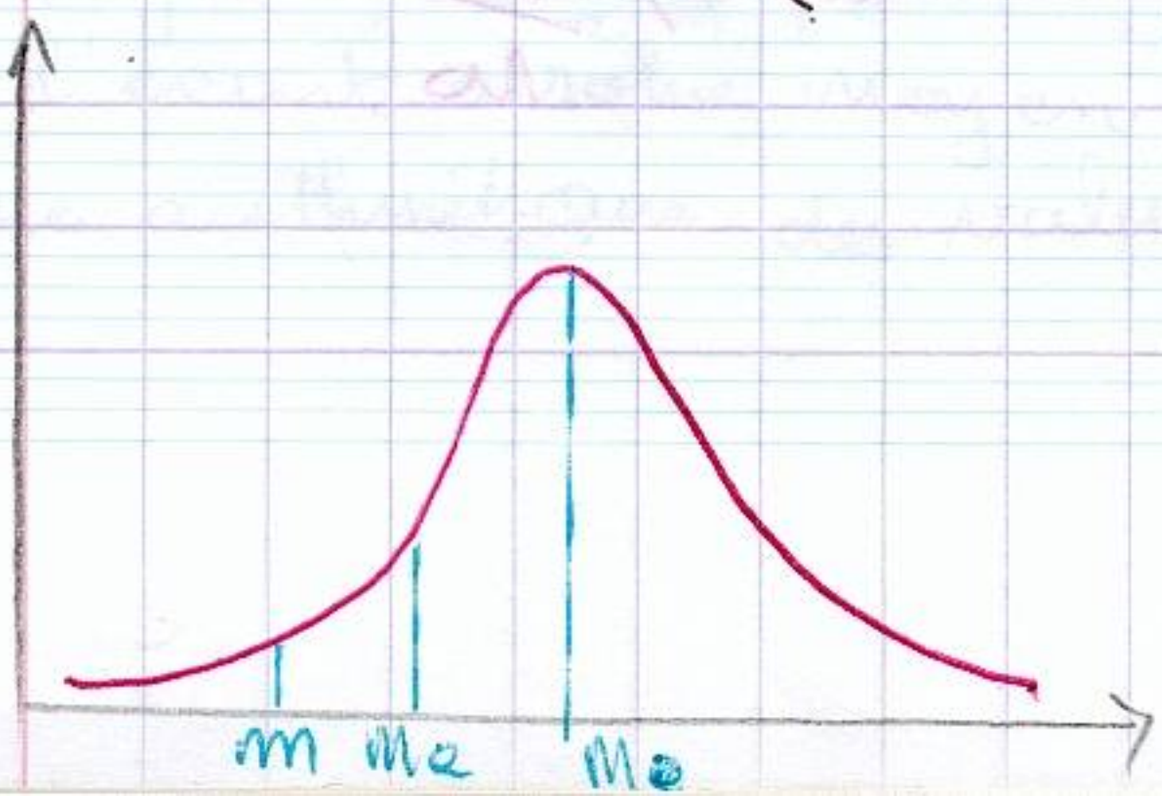
La série est dite symétrique.

* Si Mode < Médiane < Moyenne.



on dit que la série est biaisée à droite ou positivement

* Si moyenne < mediane < Mode



On dit que la série est biaisée à gauche ou négativement

* Les quartiles : (25%, 50%, 75%)

Ce qui divise la série en 4 groupes de même effectif. Le

- 1ère quartile correspondent au effectif cumulé $\frac{n}{4}$
- Le cas continue :

$$Q_1 = l_1 + \frac{(l_2 - l_1)}{n_0}\left[\frac{n}{4} - F^{\uparrow}\right]$$

- le 2ème quartile est la médiane. $Q_2 = M_e$
- le 3ème cartile correspondent aux $\frac{3n}{4}$.
- Cas continue :

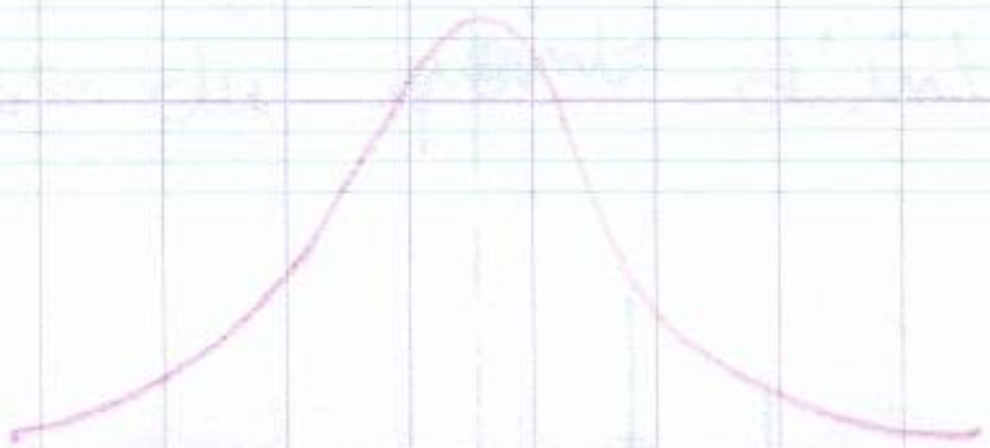
$$Q_3 = l_1 + \left(\frac{l_2 - l_1}{n_0}\right)\left[\frac{3n}{4} - F^{\uparrow}\right]$$

* les déciles :

$K = \{10\%, 20\%, \ldots 90\%\}$.

, ce qui divise la série en 10 groupes de même effectif. Il s'agit des valeurs de la variable qui correspondent au effectifs cumulées

$\frac{n}{10}, \frac{2n}{10}, \ldots\ldots, \frac{9n}{10}$.



2] Les paramètres de dispersion

Considerons les 2 séries suivantes qui donnent les glycémies :

0,95 g . 1,05 g . / m = 1 , M_e = 1 g .

0,50 g 1,50 g . / m = 1 , Me = 1 g .

et pourtant la situation est tout à fait differente dans la 1ère voisines de leur moyenne.

dans le 2ème cas les 2 valeurs très éloignées de leur moyenne, c'est ce qu'en statistique on nomme.

dispersion.

La 2ème série est plus dispersée que la 1ère

Il est donc important d'ajouter autres paramètres de dispersion

1/ L'étendue d'une série :

L'étendue d'une série statistique est : $E = x_{max} - x_{min}$

2/ L'écart absolu moyen :

Soit une série statistique prenant les valeurs $x_1 \dots x_p$. On considère sont écart à la moyenne.

$e_1 = x_1 - \overline{x}$, $e_2 = x_2 - \overline{x}$, $\dots e_p = x_p - \overline{x}$.

On définie l'écart absolue moyen comme étant la moyenne arithmétique des valeurs absolue des écarts.

$$\bar{e} = \sum_{i=1}^{p} \frac{e_i n_i}{n} = \sum_{i=1}^{p} \frac{|x_i - \bar{x}| n_i}{n}$$

Dans le cas continue x_i est remplacer por c_i centre de classe $[a_i, a_{i+1}]$. $i \leq i' \leq p$.

$$\bar{e} = \sum_{i=1}^{p} \frac{|c_i - \bar{x}| n_i}{n}$$

3/ L'écart type et la variance.

le caracteristique de dispersion la plus usuelle est en effet l'écart type

- **la variance** : Soit la serie statistique prenant les valeurs $x_1, x_2, \ldots x_p$ / $e_1 = x_1 - \bar{x}, \ldots e_p = x_p - \bar{x}$.

On définie la variance par la moyenne arithmetique des carrés des écarts à la moyenne.

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

L'écart type d'une série

c'est la moyenne cadratique des écarts à la moyenne autrement dit:

$$\epsilon = \sqrt{V} = \sqrt{\sum_{i=1}^{p} \frac{n_i (x_i - \bar{x})^2}{n}}.$$

Pour le cas continue on utilise la même formule remplaçant x_i par C_i centre de classe

Méthode de calcul :

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

$$= \sum_{i=1}^{p} \frac{1}{n} \left[n_i (x_i^2 - 2\bar{x}x_i + \bar{x}^2) \right]$$

$$= \sum_{i=1}^{p} \frac{1}{n} \left[n_i \cdot sc_i^2 - 2 n_i \bar{x} x_i + n_i \cdot \bar{x}^2 \right]$$

$$= \sum_{i=1}^{p} \frac{1}{n} n_i x_i^2 - \sum_{i=1}^{p} \frac{2 n_i}{n} \bar{x} x_i + \bar{x}^2 \sum \frac{n_i}{n}$$

$$= \frac{1}{n} \sum n_i x_i^2 - \frac{2\bar{x}}{n} \sum n_i x_i + \bar{x}^2 \sum \frac{n_i}{n}$$

$$= \sum \frac{n_i x_i^2}{n} - 2\bar{x}\bar{x} + \bar{x}^2$$

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

Signification de l'écart type :

On se fait une idée simple de la signification de l'écart type lorsque l'on compare 2 series statistiques de même nature.

Celle qui a l'écart type le plus grand est le plus dispersée.

exp: Considerons les 2 séries:

1/ 95 97 100 103 105.

2/ 50 75 100 125 150.

$\bar{X}_1 = \sum_{i=1}^{5} \frac{X_i \cdot i}{5} = 100$. , $\bar{X}_2 = \sum_{i=1}^{5} = \frac{x_i \cdot i}{5} = 100$.

$n(1) = n(2) = 5 \Rightarrow M_{e_1} = X_{\frac{n+1}{2}} = X_3 = 100$.

$M_{e_2} = 100$.

$V_1(X_1) = \sum_{i=1}^{5} \frac{X_{i_1}^2}{n} = (\bar{X}_1)^2 = 13,6$.

$\Rightarrow 6_1 = \sqrt{X_1} = 3,68$.

$V(X_2) = \sum_{i=1}^{5} \frac{(X_{i_2})^2}{n} - (\bar{X}_2)^2 = 12,50$.

$6_2 = \sqrt{V_2} = 35,35$

$6_2 >> 6_1$ donc la 2ème série est plus dispersée que la 1ère.

L'intervalle interquartile (L'écart interquartile) :

La dispersion d'une série statistique est parfois mesurée par l'intervalle interquartile ou la semi interquartile

* L'intervalle interquartile est $[Q_1 ; Q_3]$ est-l'interquartile est la difference entre le 1er quartile et le 3ème quartile $Q_i = Q_3 - Q_1$.
* le semi interquartile : $\frac{Q_i}{2} = \frac{Q_3 - Q_1}{2}$
* le coefficient interquartile est : $\frac{Q_3 - Q_1}{Q_2} = \frac{Q_i}{M_e}$.

On définit de la même manière l'intervalle interdécile $[D_9 ; D_1]$. de l'interdécile est : $D_i = D_9 - D_1$

le coefficient de variation

Le coefficient de variation permet de comparer les dispersion de divers séries statistique, le coefficient de variation est le rapport entre la moyenne et l'écart type. $C_V = \frac{\sigma}{\bar{x}}$.