CHAPITRE III - VARIABLE STATISTIQUE DISCRETE

Remarquons tout d'abord que seulement le caractère quantitatif peut être appelé ***variable statistique***. Cette appellation n'est pas correcte quand il s'agit d'un caractère qualitatif.

Nous avons déjà deux exemples qui traitent d'une variable statistique discrète ; l'exemple 1 où il est question du nombre de fruits sur les arbrisseaux et l'exemple 3 où on s'intéresse au nombre de doses nécessaires pour l'anesthésie des cobayes.

Nous avons affaire à une variable statistique discrète quand on étudie, par exemple, un ensemble de familles suivant le nombre de leurs enfants ou un ensemble d'entreprises suivant le nombre de leurs employés ou un groupes d'avions d'une compagnie aérienne selon le nombre de vols effectués au cours d'une année par chacun d'eux ... etc.

L'étude descriptive d'une variable statistique (aussi bien discrète que continue) se fait en quatre étapes :

- Description préliminaire
- Caractéristiques de position centrale
- Caractéristiques de dispersion
- Caractéristiques de symétrie et d'aplatissement.

III – A – Description préliminaire

Nous allons adopter ce même ordre dans le développement de ce chapitre et dans celui du chapitre suivant (qui traitera le cas d'une variable statistique continue).

Lorsqu'on a noté nos observations, fait nos mesures ou achevées nos enquêtes nous aurions terminé la phase de la collecte des données. Commence alors la phase de traitement. Cette phase débute nécessairement par une description préliminaire dont une partie est numérique et l'autre graphique.

La partie numérique consiste à dresser le tableau statistique et à lui ajouter :

- des colonnes qui font état de l'aspect **différentiel** (le poids des modalités une à une pour faire ressortir les différences des unes par rapport aux autres et faire distinguer celles qui sont importantes de celles qui le sont moins).
- des colonnes qui font état de l'aspect **intégral** (la vision globale du phénomène par la considération de l'effet cumulé des différentes modalités. Ceci afin d'avoir une idée de son développement quand la variable évolue).

la raison d'être de la partie graphique est, essentiellement, d'offrir une meilleure commodité de lecture. Elle consiste, tout simplement, à visualiser par des courbes certaines colonnes du tableau qui se rapportent aux deux aspects, nous aurons donc :

- des diagrammes différentiels pour le premier aspect
- des diagrammes intégraux pour le second aspect.

a) Tableau statistique

Dans le cas d'une variable statistique discrète les modalités sont les différentes valeurs de la variable. On les classe par ordre croissant dans la première colonne du tableau. Dans la

deuxième colonne, et en face de chaque valeur, on rapporte l'effectif correspondant. Nous avons ainsi,

Valeurs de la variable statistique X_i	**Effectifs correspondants aux différentes valeurs** n_i
X_1	n_1
X_2	n_2
...	...
X_i	n_i
...	...
X_k	n_k
Total	**N**

avec $X_1 < X_2 < \ldots < X_i < \ldots < X_k$

b) Fréquences relatives et diagramme différentiel

Nous avons déjà vu que la fréquence relative est le rapport de la fréquence absolue à l'effectif total, c'est à dire : $f_i = n_i / N$

La troisième colonne du tableau statistique sera celle des fréquences relatives. C'est cette colonne qui sera représentée par le diagramme différentiel.

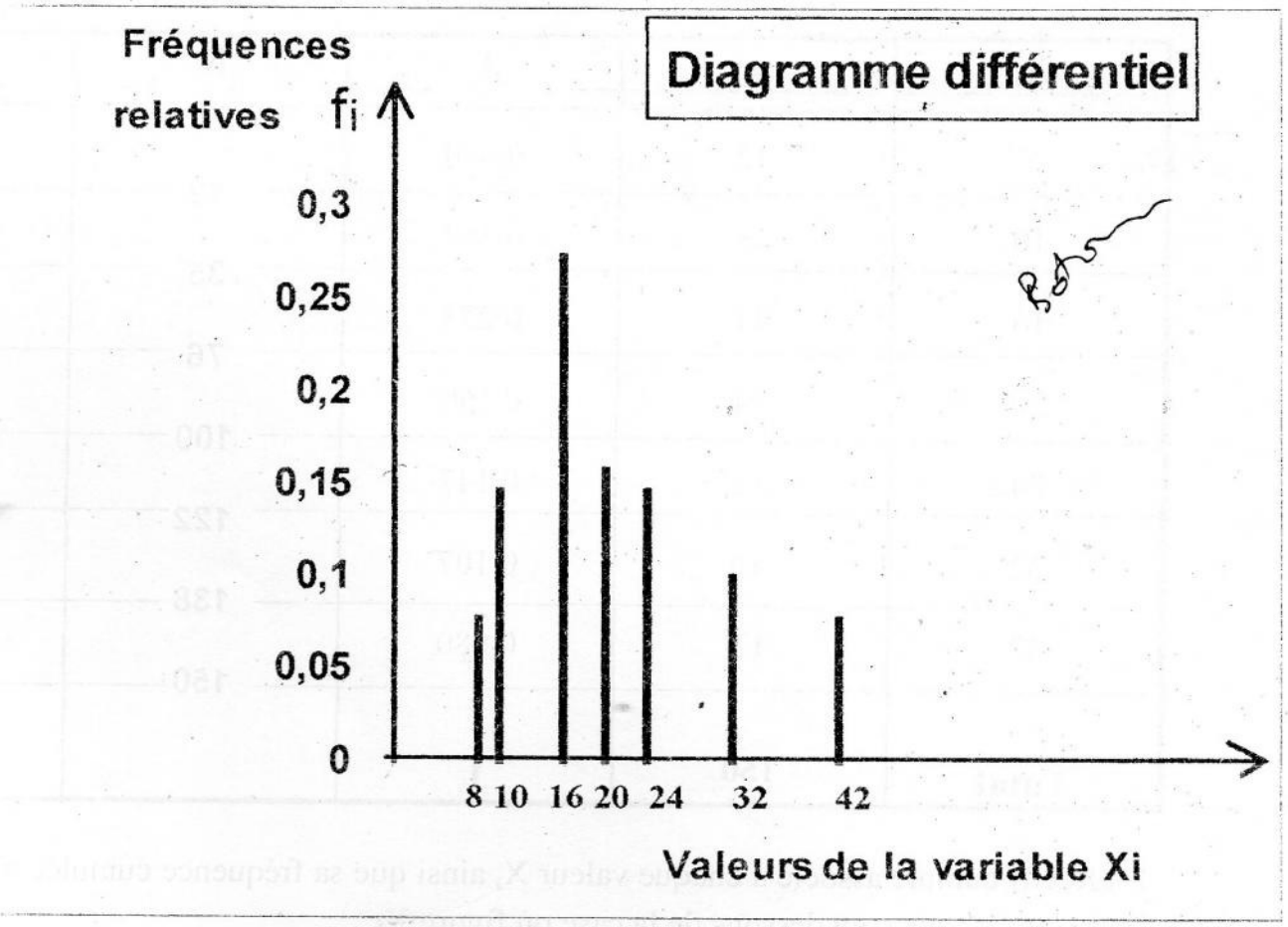
***Dans le cas d'une variable statistique discrète, le diagramme différentiel est appelé le* diagramme en bâtons.**

Dans un repère orthogonal, on associe à chaque valeur X_i un segment de droite dont la longueur est la fréquence relative f_i.

Nous obtenons un diagramme qui ressemble à une collection de bâtons (d'où son nom).

Nous avons, ainsi, dans le cas de l'exemple 1 :

Valeurs de la variable statistique X_i (Nombre de fruits)	**Fréquences absolues** n_i (Nombre d'arbrisseaux)	**Fréquences relatives** f_i
8	12	**0.080**
10	23	**0.153**
16	41	**0.273**
20	24	**0.160**
24	22	**0.147**
32	16	**0.107**
42	12	**0.080**
Total	**150**	**1**



Remarques :

1. Pour les longueur des segments, on peut très bien utiliser les effectifs au lieu des fréquences relatives. Le changement ne sera qu'une variation d'échelle.
2. Lorsqu'on joint les sommets des bâtons par des segments de droite on obtient une ligne brisée qu'on appelle communément polygone des fréquences. Bien que ce polygone soit d'un usage fréquent, il est injustifié car la variable statistique n'existe pas entre deux valeurs successives et ne peut donc avoir d'ordonnées en des points où elle n'est pas définie.

c) Effectifs cumulés, fréquences relatives cumulées et diagramme intégral

L'effectif cumulé correspondant à la valeur X_i est le nombre des individus ayant une valeur inférieure ou égale à X_i.

Autrement dit, c'est la somme des effectifs qui se sont accumulés en atteignant cette valeur. Ce qui s'écrit :

$$N_i = \sum_{p=1}^{i} n_p$$

De même,

La fréquence cumulée correspondant à la valeur X_i est la fréquence des individus ayant une valeur inférieure ou égale à X_i.

C'est-à-dire :

$$F_i = \sum_{p=1}^{i} f_p$$

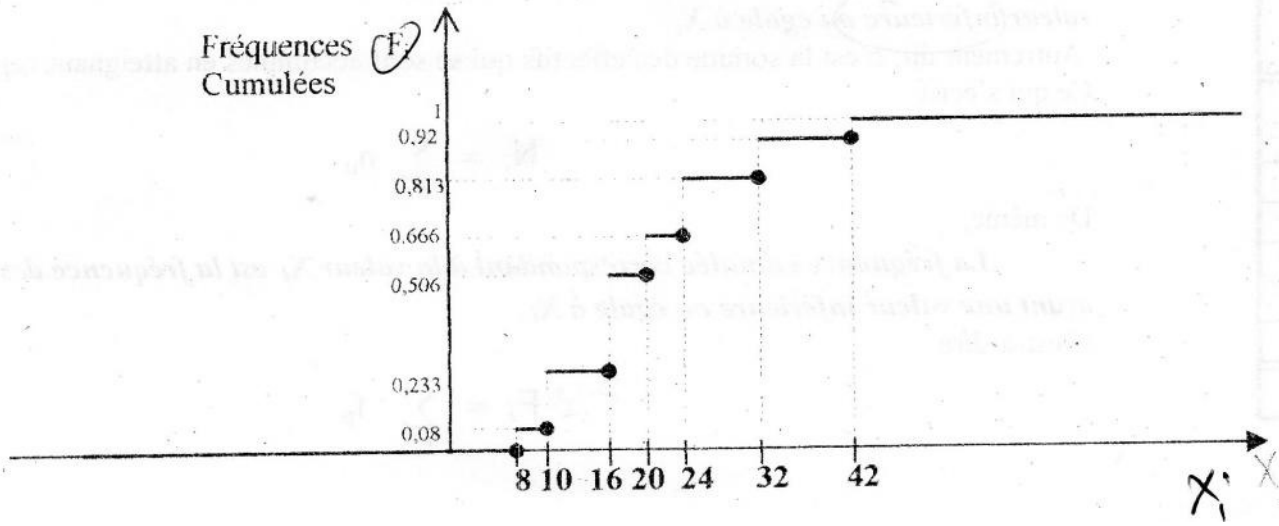
Pour l'exemple 1, nous obtenons le tableau suivant :

X_i	n_i	f_i	N_i	F_i
			0	0
8	12	0.080		
			12	0,080
10	23	0.153		
			35	0,233
16	41	0.273		
			76	0,506
20	24	0.160		
			100	0,666
24	22	0.147		
			122	0,813
32	16	0.107		
			138	0,920
42	12	0.080		
			150	1,000
Total	**150**	**1**		

L'effectif cumulé associé à chaque valeur X_i ainsi que sa fréquence cumulée s'écrivent sur la ligne qui se trouve en dessous de la case où figure X_i.
Cette écriture (conventionnelle) sur les lignes plutôt que dans les cases est pour préciser que les quantités en question sont atteintes juste quand on dépasse la valeur X_i.
La représentation graphique intégrale nécessite l'introduction d'une fonction appelée fonction cumulative dont voici la définition :
La fonction :

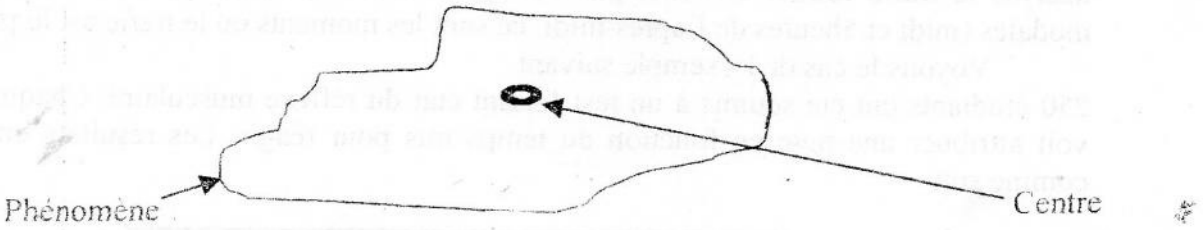
$$F : R \longrightarrow R$$
$$x \in]X_{i-1} - X_i] \longrightarrow F_i$$

S'appelle fonction de répartition de la variable X.
Sa courbe représentative s'appelle **courbe cumulative** et c'est le diagramme intégral de la variable statistique discrète X. Elle a la forme d'une courbe en escalier.



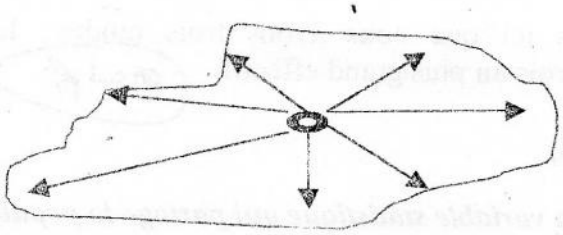
III – B – Caractéristiques de position centrale:

Quand on se dispose à étudier et à décrire un phénomène on commence par s'intéresser aux valeurs autour desquelles il se développe. On cherche son "centre".
Il nous faut donc définir des quantités qui peuvent jouer ce rôle. Certaines de ces quantités nous renseignerons sur la position (mode, médiane, quartiles). D'autres, en plus de cette information, serviront à résumer les données en notre possession (moyennes).

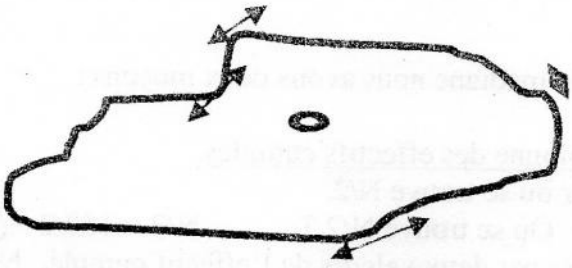


Une fois que le "centre" est cerné, il s'agit de voir comment se développe le phénomène autour de ce centre. C'est-à-dire :

- comment il se disperse ?



- Et sous quelle forme ?



Les caractéristiques de position centrale sont :

- Le mode
- La médiane et les quantiles
- Les moyennes (arithmétique, géométrique, harmonique).

a) Le mode : (noté M_o)

Le mode est la valeur la plus fréquente de la variable statistique, c'est-à-dire celle qui correspond au plus grand effectif.

Dans le cas de l'exemple 1 le plus grand effectif est 41 (troisième ligne), le mode est donc :

$$M_o = 16$$

Remarque :

Le mode peut ne pas être unique. Certaines distributions peuvent présenter plusieurs valeurs dont les effectifs sont égaux et qui sont les plus grands de la distribution. Si on analyse le trafic routier en ville, par exemple, on va constater qu'il possède deux valeurs modales (midi et 5heures de l'après-midi, ce sont les moments où le trafic est le plus intense).

Voyons le cas de l'exemple suivant :

250 étudiants ont été soumis à un test faisant état du réflexe musculaire. Chaque étudiant se voit attribuer une note en fonction du temps mis pour réagir. Les résultats ont été classés comme suit :

Nombre de points X_i	1	2	3	4	5	6	7	8
Nombre d'étudiants n_i	17	30	45	38	45	45	28	2

Nous observons ici que nous avons trois modes ; les valeurs 3 ; 5 et 6. Il correspondent tous les trois au plus grand effectif.

b) La médiane : (notée M)

C'est la valeur de la variable statistique qui partage la population en deux populations d'effectifs égaux.

Autrement dit, les individus appartenant à la première moitié de la population ont de valeurs inférieures à la médiane. Les individus appartenant à la deuxième moitié lui ont de valeurs supérieures.

Ainsi, pour déterminer la médiane nous avons deux moyens :

- Utiliser la colonne des effectifs cumulés.

Il suffit de situer où se trouve N/2.

Reprenons l'exemple 1 : Où se trouve N/2 ? $N/2 = 150/2 = 75$

Nous le trouvons encadré par deux valeurs de l'effectif cumulé : N_{i-1} et N_i

Dans notre exemple, $N_{i-1} = 35$ et $N_i = 76$

Ces effectifs cumulés sont écrits sur des lignes. La valeur de la variable inscrite dans la cas comprise entre ces deux lignes est la médiane.

Pour notre exemple M = 16.

- Utiliser la colonne des fréquences relatives cumulées.

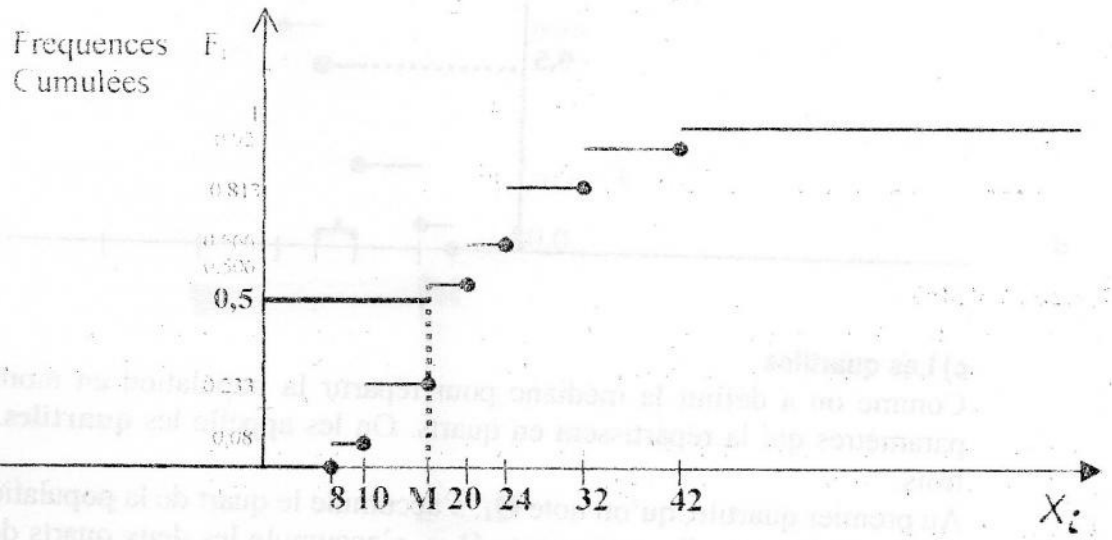
Quand on se réfère à la définition de la médiane M alors la fonction cumulative donne :

$$F(M) = 0,5$$

Ainsi, nous devons chercher les valeurs F_{i-1} et F_i qui encadrent 0,5.
Dans notre exemple, $F_{i-1} = 0,233$ et $F_i = 0,506$.
Donc, et de la même manière qu'au point précédent, on trouve M = 16.

X_i	N_i	F_i
	0	0
8		
	12	0,080
10		
	35	0,233
M=16 ◄	**75** ◄	**0,5**
	76	0,506
20		
	100	0,666
24		
	122	0,813
32		
	138	0,920
42		
	150	1,000
Total		

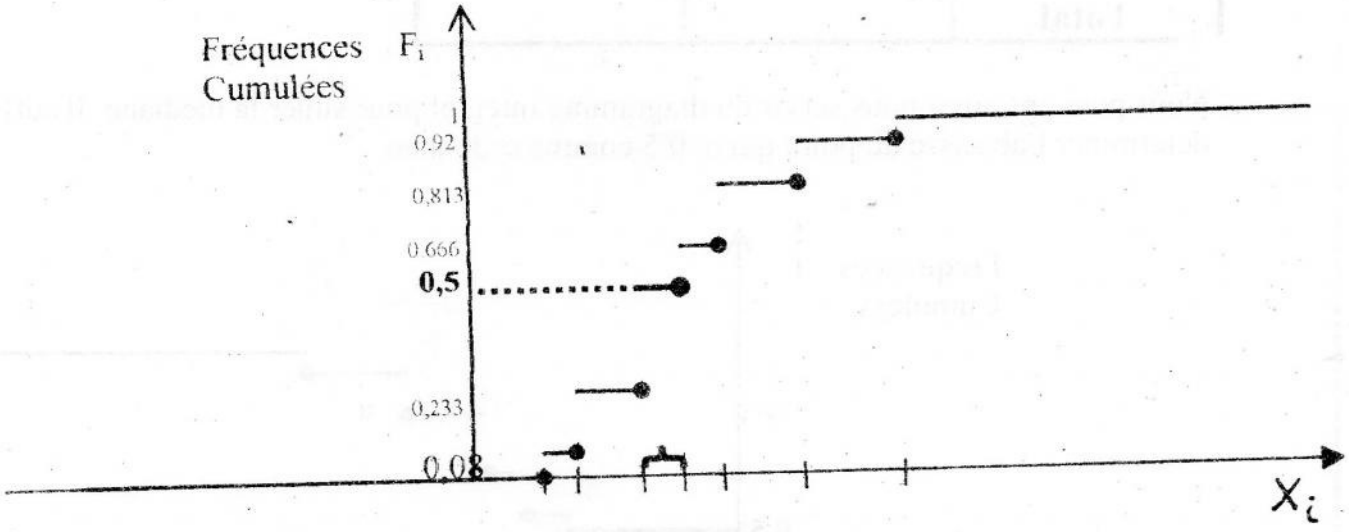
Nous pouvons aussi nous servir du diagramme intégral pour situer la médiane. Il suffit de déterminer l'abscisse du point qui a 0,5 comme ordonnée.



Remarque :

Il peut arriver que N/2 soit precisement sur une ligne (et donc 0.5 aussi, dans la colonne des fréquences relatives cumulées). Dans ce cas, ce n'est plus une seule valeur qui est mediane mais c'est tout un intervalle. On l'appelle **intervalle médian**.
Si dans l'exemple précédent on apporte quelques légers changements pour aboutir à cette situation on aura [16 - 20[comme intervalle médian.

X_i	N_i	F_i
	0	0
8	12	0,080
10	35	0,233
16	**75**	**0,5**
20	100	0,666
24	122	0,813
32	138	0,920
42	150	1,000
Total		



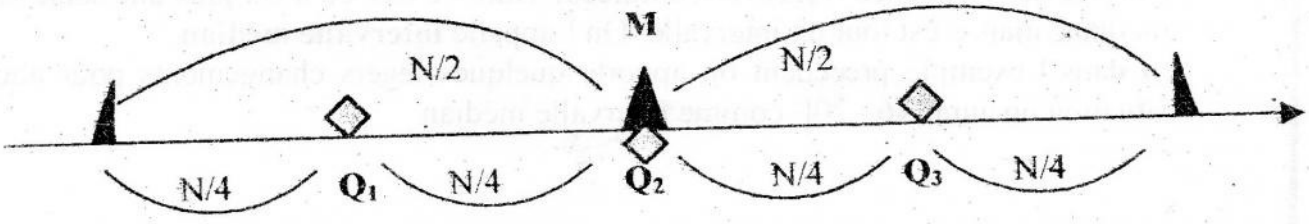
c) Les quartiles

Comme on a définit la médiane pour répartir la population en moitiés on peut définir des paramètres qui la répartissent en quarts. On les appelle les **quartiles**. Ils sont au nombre de trois.

Au premier quartile, qu'on note Q_1, s'accumule le quart de la population.

Au deuxième quartile, qu'on note Q_2 , s'accumule les deux quarts de la population (et c'est donc la médiane M).

Au troisième quartile, qu'on note Q_3 , s'accumule les trois quarts de la population.



Pour déterminer les quartiles on procède de la même manière que pour la médiane.
Il suffit, pour le premier quartile, de situer où se trouve N/4 sur la colonne des effectifs cumulés (ou 0,25 sur la colonne des fréquences relatives cumulées).
Pour le troisième quartile, nous avons a situer 3N/4 comme effectif cumulé (ou 0,75 comme fréquence relative cumulée).
Pour notre exemple, nous avons :

X_i	N_i	F_i
	0	0
8		
	12	0,080
10		
	35	0,233
$Q_1=$ **16**	**37,5**	**0,25**
	76	0,506
20		
	100	0,666
$Q_3=$ **24**	**112,5**	**0,75**
	122	0,813
32		
	138	0,920
42		
	150	1,000
Total		

Remarque :

De la même manière qu'on a définit les quartiles on peut définir les déciles (qui divisent la population en dixièmes) ou les centiles (qui divisent la population en centièmes).
D'une manière générale, tous ces paramètres s'appellent les **quantiles**.

d) La moyenne arithmétique

Nous savons que la moyenne arithmétique de N nombres est le rapport de leur somme à leur nombre. C'est-à-dire : Si $X_1 ; X_2 ; \ldots ; X_i ; \ldots ; X_N$ sont les N valeurs alors :

$$X = \frac{X_1 + X_2 + \ldots + X_i + \ldots + X_N}{N}$$

C'est-à-dire :

$$X = (1/N) \sum_{i=1}^{N} X_i$$

Ainsi, par exemple, la moyenne des nombres suivants :

3 ; 3 ; 4 ; 4 ; 4 ; 4 ; 4 ; 6 ; 6 ; 6 ; 6 ; 10

est :

$$\mathbf{X} = \frac{3+3+4+4+4+4+4+6+6+6+6+10}{12} = 5$$

C'est une somme de 12 termes qu'on divise par 12.
Or, remarquons que n'avons, en fait, que 4 valeurs différentes : 3 ; 4 ; 6 et 10. Chacu se répétant un certain nombre de fois.
Nous pouvons écrire :

$$\mathbf{X} = \frac{\mathbf{2}.3 + \mathbf{5}.4 + \mathbf{4}.6 + \mathbf{1}.10}{12}$$

C'est une somme de 4 termes qu'on divise par 12. Le nombre de termes est le nombre valeurs différentes. Chacune étant multipliée par le nombre de fois où elle figure (s effectif).
Ainsi, d'une manière générale, nous pouvons écrire la moyenne arithmétique sous la forme

$$\mathbf{X} = (1/N) \sum_{i=1}^{k} n_i X_i \quad .$$

où X_1 ; X_2 ; ... ; X_i ; ... ; X_k sont les différentes valeurs et n_1 ; n_2 ; ... ; n_i ; ... ; n_k le effectifs correspondants.

Nous pouvons aussi nous servir des fréquences relatives pour calculer la moyen arithmétique. En effet, nous avons :

$$\mathbf{X} = \sum_{i=1}^{k} f_i X_i$$

Dans le cas de l'exemple 1, nous avons :

X_i	n_i	f_i	$n_i X_i$	$f_i X_i$
8	12	0.080	96	0.64
10	23	0.153	230	1.53
16	41	0.273	656	4.368
20	24	0.160	480	3.2
24	22	0.147	528	3.528
32	16	0.107	512	3.424
42	12	0.080	504	3.36
Total	150	1	3006	**20.05**
Total/N			**20.04**	

Ainsi, la moyenne arithmétique pour cet exemple est :

$$\mathbf{X} = 20.04$$

(La légère différence entre les deux valeurs 20.04 et 20.05 provient des erreurs d'arrondis sur les fréquences relatives f_i).

e) Remarques

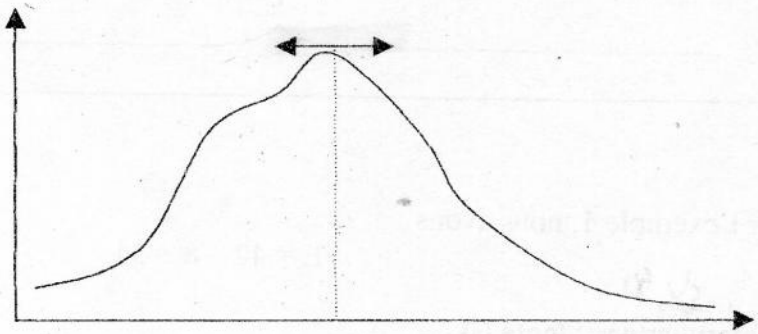
Le mode, la médiane et la moyenne arithmétique ont été, tous les trois, classés comme caractéristiques de position centrale. Néanmoins, chacune de ces caractéristiques a sa spécificité et chacune éclaire un aspect particulier du phénomène qu'on étudie.

D'abord, concernant le mode, on est en droit de poser la question suivante :

Le mode provient de l'idée d'intensité (le plus grand effectif), qu'est ce qui le prédispose donc à être classé comme indicateur de centralité ?

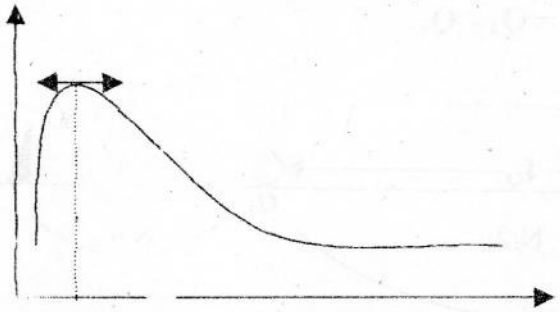
La réponse est, qu'en réalité, la majorité des phénomènes que l'on rencontre dans la nature se développent suivant le schéma :

Naissance ⟶ croissance ⟶ décroissance ⟶ disparition

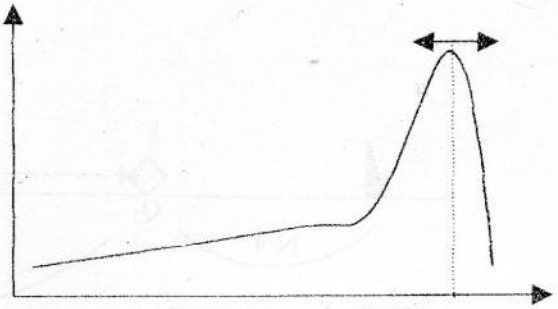


Ainsi, l'apogée se trouve généralement au "milieu".

Ne s'écarte de ce schéma que très peu de cas. Ils sont de deux sortes :



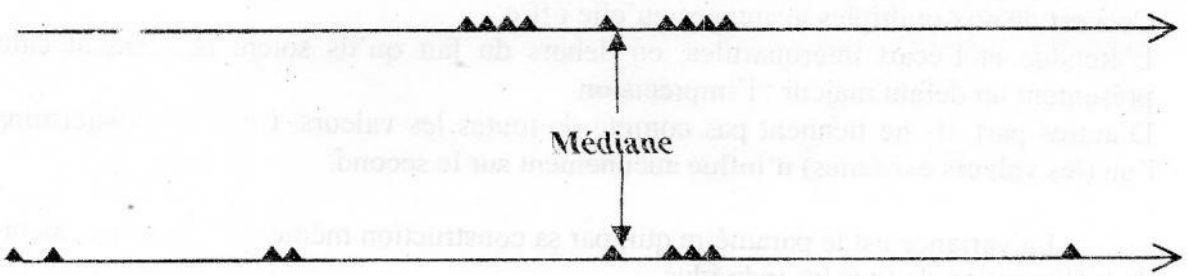
N ssance explosive



Disparition brutale

2. La médiane est liée à l'effectif et non aux valeurs de la variable statistique. Elle détermine le partage de la population en effectifs égaux indépendamment des valeurs que possèdent les individus.

De cette manière, elle est insensible aux variations touchant les valeurs extrêmes qui sont souvent aberrantes. C'est ce qui la fait, parfois, préférer à la moyenne arithmétique.



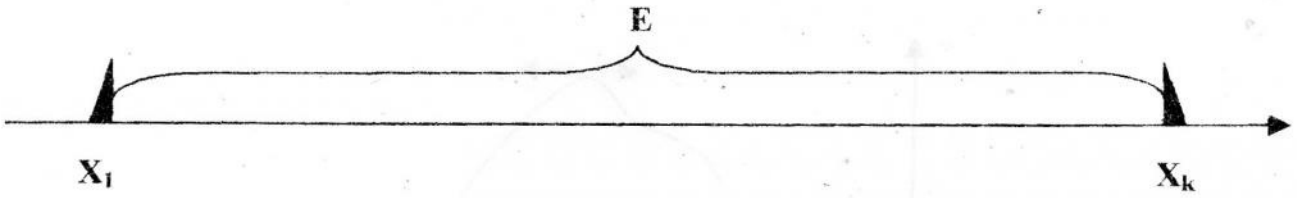
Dans la figure ci-dessus, la médiane est la même dans les deux cas bien que les points soient plus dispersés sur la deuxième droite que sur la première. La moyenne aurait été tout à fait différente.

III – C – Caractéristiques de dispersion:

a) L'étendue : (noté E)

C'est la longueur de l'intervalle sur lequel se disperse la variable. C'est-à-dire :

$$E = X_k - X_1$$



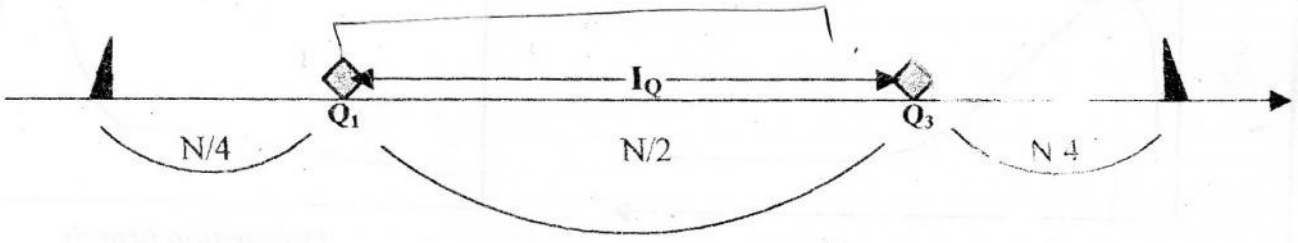
Ainsi, pour l'exemple 1, nous avons :

$$E = 42 - 8 = 34$$

b) L'écart interquartiles : (noté I_Q)

C'est la différence entre le troisième quartile et le premier quartile. C'est donc l'intervalle où se trouve la moitié centrale de la population.

$$I_Q = Q_3 - Q_1$$



Pour l'exemple 1, nous avons :

$$I_Q = 24 - 16 = 8$$

c) La variance : (notée Var X)

La caractéristique qui est réellement utilisée pour mesurer la dispersion est la variance. Ceci est dû aux multiples avantages qu'elle offre.

L'étendue et l'écart interquartiles, en dehors du fait qu'ils soient rapidement calculables, présentent un défaut majeur : l'imprécision.

D'autres part, ils ne tiennent pas compte de toutes les valeurs. Ce qui est déterminant pour l'un (les valeurs extrêmes) n'influe aucunement sur le second.

La variance est le paramètre qui, par sa construction même, effectivement compte de la dispersion de tous les individus.

En effet, observons les deux distributions suivantes :

Ainsi,

$$1/N \sum_{i=1}^{N} (X_i - \mathbf{X})^2 \qquad (1)$$

est la moyenne des carrés des distances entre la moyenne arithmétique **X** et tous les points X_i

Il reste à préciser un détail :
Remarquons que la somme a été faite sur N termes, i = 1, 2, … , N
Or, nous savons que les individus sont groupés sur certaine valeurs seulement i = 1, 2, … , k
Les individus ayant la même valeur sont à la même distance de la moyenne **X**
Nous avons

n_1 individus sur la première valeur, et donc correspondant au terme $(X_1 - \mathbf{X})^2$
n_2 individus sur la deuxième valeur, et donc correspondant au terme $(X_2 - \mathbf{X})^2$
….
D'une manière générale,
n_i individus sur la première valeur, et donc correspondant au terme $(X_i - \mathbf{X})^2$

Par conséquent, la quantité (1) peut s'écrire

$$1/N \sum_{i=1}^{k} n_i(X_i - \mathbf{X})^2$$

Cette quantité s'appelle la **variance** de la variable statistique X. On la note **Var(X)**.

Remarques :

1. En développant l'expression entre parenthèses nous aboutissons à la formule suivante

$$\mathrm{Var}(X) = \left[1/N \sum_{i=1}^{k} n_i X_i^2\right] - \mathbf{X}^2$$

2. Nous obtenons également, en faisant entrer 1/N sous le signe somme,

$$\mathrm{Var}(X) = \left[\sum_{i=1}^{k} f_i X_i^2\right] - \mathbf{X}^2$$

C'est l'une ou l'autre de ces deux expressions qu'on utilise pour calculer la variance.

Nous allons voir cela sur l'exemple 1.

Lorsque nous arrivons au point où nous devons calculer la variance cela suppose que nous avons déjà calculé la moyenne arithmétique **X**. Nous avons donc à notre disposition un tableau statistique suffisamment construit.

Il suffit de lui ajouter la colonne $(n_i X_i^2)$ (en multipliant la colonne $(n_i X_i)$ par la colonne (X_i) lorsqu'on veut utiliser la première formule.

Si on veut utiliser la deuxième formule on ajoute au tableau, de la même manière, la colonne $f_i X_i^2$.

X_i	n_i	f_i	$n_i X_i$	$f_i X_i$	$n_i X_i^2$	$f_i X_i^2$
8	12	0.080	96	0,64	768	5,12
10	23	0.153	230	1,53	2300	15,3
16	41	0.273	656	4,368	10496	69,888
20	24	0.160	480	3,2	9600	64
24	22	0.147	528	3,528	12672	84,672
32	16	0.107	512	3,424	16384	109,568
42	12	0.080	504	3,36	21168	141,12
Total	150	1	3006	**20.05**	73388	**489,668**
Total/N			**20.04**		**489,253**	

Nous avons donc :

$\text{Var}(X) = 489{,}253 - (20{,}04)^2 =$ **87,6514**

par la première formule.

$\text{Var}(X) = 489{,}668 - (20{,}05)^2 =$ **87,6655**

par la deuxième formule.

(La légère différence entre les deux valeurs de la variance provient des erreurs d'arrondis sur les fréquences relatives f_i. Cette différence est à négliger).

d) L'écart-type : (noté σ_X)

Lorsque nous avons établi la formule pour le calcul de la variance les distances ont été prises au carré. De cette façon l'unité de la variable statistique figurera au carré dans la quantité représentant la variance.

pour revenir à l'unité initiale nous allons introduire un nouveau paramètre appelé écart-type.

C'est la racine carrée de la variance.

$$\sigma_X = \sqrt{\text{Var}(X)}$$

Pour notre exemple, nous avons

$$\sigma_X = \sqrt{87{,}6514} \approx 9{,}36$$

$$\sigma_X = \sqrt{87{,}6655} \approx 9{,}36$$

CHAPITRE IV

VARIABLE STATISTIQUE CONTINUE

Dans la première partie de ce cours (définitions et terminologie) nous avons exposé six exemples dont deux déjà traitent de variables statistiques continues :

L'exemple 2 où on s'intéresse au temps de réaction au son d'un certain nombre de personnes.

L'exemple 4 où il est question de pesées effectuées sur les éléments d'un échantillon de poissons prélevé dans lac.

Dans ces deux exemples, le caractère est de par sa nature (temps ou poids) une variable continue.

D'une manière générale, l'essence même du caractère indique qu'il est continu ou pas. Ainsi, ce qui a rapport au temps, à la vitesse, à la longueur, à la température, ...etc. est continu. On peut bien imaginer un mobile partir du repos et atteindre une vitesse V en étant passé continûment par toutes les valeurs de l'intervalle [0 ; V]. On peut également imaginer l'épuisement d'une durée déterminée t d'une manière continue, de la valeur t jusqu'à la valeur zéro.

Il est possible donc que la variable prenne n'importe quelle valeur de l'intervalle considéré. Ce qui n'est pas le cas de certaines variables (les variables discrètes) qu'on a étudié précédemment. Le nombre d'enfants d'une famille, par exemple, ne peut prendre que certaines valeurs isolées.

Ainsi, nous pouvons distinguer naturellement une variable continue d'une autre qui ne l'est pas. Mais en statistique ce n'est pas ceci qui constitue la véritable préoccupation. La distinction entre variables discrètes et variables continues se fait surtout en fonction de la manière dont on veut les traiter

Lorsqu'on considère les différentes valeurs de la variable valeur par valeur et qu'on les traite isolément nous avons, de cette manière admis, que la variable est discrète. Mais si on répartit l'étendue en intervalles et que l'on adopte que les modalités des individus ne sont plus leurs valeurs exactes mais leurs appartenances à tel ou tel intervalle alors la variable est considérée continue.