Statistique Descriptive

Chapitre1 : Introduction: vocabulaires statistiques

1)- Définitions:

*** statistique :** la statistique est un ensemble des méthodes qui servent à organiser les épreuves fournissant des observations, à analyser celles-ci et à interpréter les résultats.

L'analyse statistique se subdivise en deux parties

Statistique descriptive : a pour but de décrire c-à -d de résumer ou représenter les données :

*Représentation graphique

*Paramètres de position, de dispersion, de relation.

Statistique inférentielle : l'ensemble des méthodes permettant de formuler un jugement ; Elle nécessite des outils mathématiques plus pointus (théorie des probabilités).

2)- Notions de bases :

***POPULATION :** La collection d'objets ou de personnes étudiées (élèves, habitants, voitures…).

***INDIVIDU :** élément de la population étudiée. Exemple :(un élève, un habitant, une voiture,…).

***ECHANTILLON** : partie de la population étudiée. Exemple : Nombre d'individus dans un échantillon Noté n est appelé taille de l'échantillon.

***VARIABLE (CARACTERE) :** propriété commune aux individus de la population, que l'on veut étudier ; Un caractère peut être :

a)-qualitatif : un caractère est dit qualitatif si ses modalités ne sont pas mesurable. Dans ce cas, les modalités sont aussi appelées catégories. Si les modalités du caractère qualitatif ne sont pas naturellement ordonnées on dira qu'il est nominal, par contre si les modalités sont organisées selon un ordre hiérarchique on dira qu'il est ordinal.

Le caractère "couleur des yeux" est nominal alors que le caractère " mention du Bac" est ordinal.

b)- quantitatif : peut prendre des valeurs numérique (poids, longueur) un caractère quantitatif peut être :

***Continue :** peut prendre toutes les valeurs numériques d'un intervalle déterminé (Taille, poids…).

***Discontinue (discrète) :** ne peut prendre que des valeurs numériques isolées (nombre de pièces d'habitations, nombre de fruits endommagés…).

***MODALITE :** l'une des formes particulières d'un caractère. La couleur des yeux est un caractère, ses modalités sont : bleu, vert, marron,…

***EFFECTIF OU FREQUENCE ABSOLUE (noté n_i) :** nombre d'apparitions de la valeur associé à un caractère dans un échantillon.

***FREQUENCE RELATIVE (noté f_i) f_i = ou fréquence d'une valeur est** le quotient de l'effectif de cette valeur par l'effectif de la population (effectif total) ; c'est un nombre compris entre 0 et 1 ; la somme des fréquences est toujours égale à 1.

*On appelle **LES STATISTIQUES (au pluriel)** des collections de nombres présentées sous forme de tableaux ou de graphique groupant des observations relatives à un phénomène considéré.

Exemple 1 :

Nombre d'enfant	0	1	2	3	4	5	TOTAL
Nombre de famille ou effectif : ni	16	18	14	11	3	2	64
Fréquence relative : fi	0,250	0,281	0,218	0,172	0,047	0,031	1

Population étudié : les familles

L'échantillon sur lequel porte l'étude : familles d'un immeuble ; n=64.

Le caractère étudié est le nombre d'enfants par famille. C'est un caractère quantitatif discret.

3) Traitement d'une série statistique :

***Série ordonnée :** les valeurs obtenues peuvent être rangées par ordre de grandeur par

Exemple croissante. On obtient une série statistique ordonnée.

***Etendue de la série :** la différence entre les deux valeurs extrêmes est appelée étendue

De la série.

***Classe :** quand le caractère étudié est quantitatif continu, la série statistique est répartie

En classes ou intervalles semi ouverts. Le nombre de classes, k est calculé par l'une des deux formules :

La règle de Sturges $k=1+3.3\log(n)$

La règle de Yule $k=2.5(n)^{0.25}$

Centre de classe : on appelle centre de classe, la demi-somme des valeurs extrêmes de la classe. On note C_i le centre de la classe.

***Effectif cumulé :** la somme des effectifs des i première classe est appelé effectif cumulé

de la ième classe on le note **ni cum** .

***Fréquence cumulée :** le rapport **ni cum/N** est appelé fréquence cumulé de la ième classe (N est la taille de l'échantillon).

Exemple2 : Le taux de glucose sanguin (glycémie) déterminé chez 32 sujets est donné ci-dessous en g/l

Série ordonnée :

0,85 0,95 1,00 1,06 1,11 1,19

0,87 0,97 1,01 1,07 1,13 1,20

0,90 0,97 1,03 1,08 1,14

0,93 0,98 1,03 1,08 1,14

0,94 0,98 1,03 1,10 1,15

0,94 0,99 1,04 1,10 1,17

Etendue de la série : 1,20 – 0,85 en g/l = 0,35 g/l.

On a N = 32 et la formule de Yule donne

$K = 2.5(32)^{1/4} = 5.94 \approx 6$.

Classe en g/l	C_i	n_i	f_i	n_i cum
[0, 85 ; 0, 91]		3		
[0, 91 ; 0, 97]		4		
[0, 97 ; 1, 03]		7		
[1,03 ; 1, 09]		8		
[1, 09 ; 1, 15]		6		
[1, 15 ; 1, 21]		4		

Au bas de la colonne n_i , on indique la somme de tous les n_i , $\sum n_i$ qui n'est autre que l'effectif total N de l'échantillon.

De la même façon au bas de la colonne des f_i on indique leur somme $\sum f_i$ qui doit être égal à 1.

La dernière colonne, dite des effectifs cumulés croissants a la signification suivante :

Pour la classe [0,85 ; 0,91[: n_i cum = 3, on dit qu'il ya 3 valeurs inferieur à 0,91 g/l.

-Pour la classe [0,91 ; 0,97[: n_i cum = 3 + 4 = 7

il ya 7 valeurs inferieures à 0,97 (3 inferieures à 0,91 et 4 comprises entre 0,91 et 0,97).

-Pour la dernière classe on a donc n_i cum = N .

On appelle fréquence cumulée croissante pour la ième classe le rapport

$f_i = n_i$ cum/N

On a donc :

Pour la 1è classe $F_1 = 3/32$.

-Pour la 2è classe $F_2 = 7/32$.

-Pour la dernière $F_6 = 32/32 = 1$.

On peut de la même façon concevoir des effectifs cumulés décroissants et des fréquences cumulés décroissantes

Chapitre 2 : Les représentations graphiques :

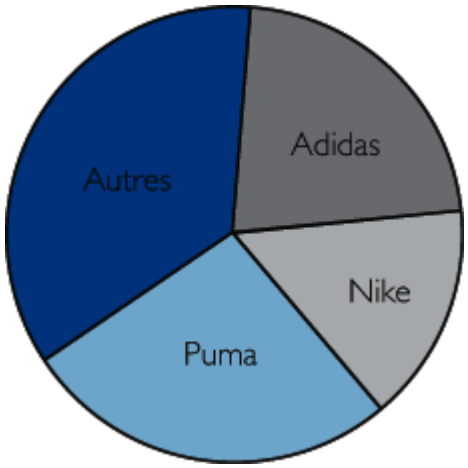
1) **Diagramme circulaire :** Un diagramme circulaire est une représentation graphique de données statistiques sous la forme d'un disque partagé en secteurs circulaires. À chaque valeur du caractère étudié correspond un secteur. Les mesures des secteurs sont proportionnelles aux effectifs représentés.

Exemple : Voici la répartition de 450 collégiens selon la marque de leurs baskets.

Marque	Adidas	Nike	Puma	Autres marques	Total
Effectif	100	70	120	160	450

On veut représenter cette répartition sous la forme d'un diagramme circulaire. À chaque marque correspond un secteur circulaire. Les mesures des secteurs sont proportionnelles aux effectifs représentés, le coefficient de proportionnalité étant ici égal à **360/450.**

On obtient le diagramme suivant :



Remarque :

Un diagramme semi-circulaire se présente sous la forme d'un demi-disque.

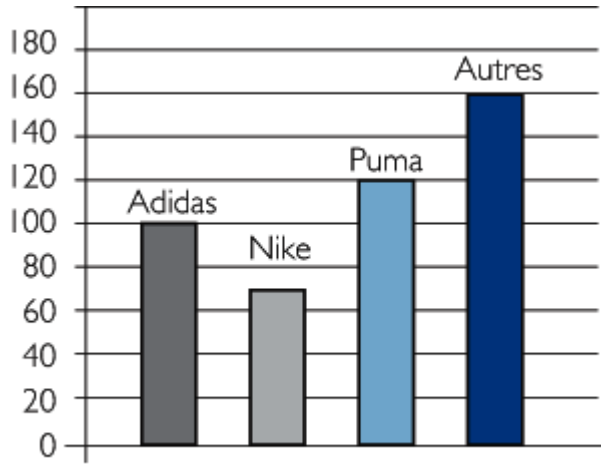
2) Diagramme en barres :

Un diagramme en barres est une représentation graphique de données statistiques à l'aide de rectangles de même largeur.
Les valeurs du caractère étudié sont représentées sur l'axe horizontal, les effectifs sur l'axe vertical.
À chaque valeur correspond une barre. Les hauteurs des barres sont proportionnelles aux effectifs représentés.

Exemple : Voici la répartition de 450 collégiens selon la marque de leurs baskets.

Marque	**Adidas**	**Nike**	**Puma**	**Autres marques**	**Total**
Effectif	100	70	120	160	450

On veut représenter cette répartition sous la forme d'un diagramme en barres.
À chaque marque correspond une barre. Les hauteurs des barres sont proportionnelles aux effectifs représentés.
On obtient le diagramme suivant :



3) Diagramme en bâtons :

Un diagramme en bâtons est une représentation graphique de données statistiques à l'aide de segments.
Les valeurs du caractère étudié sont représentées sur l'axe horizontal, les effectifs sur l'axe vertical. À chaque valeur correspond un bâton.

Les hauteurs des bâtons sont proportionnelles aux effectifs représentés.

Exemple :

Voici la répartition de 450 collégiens selon la marque de leurs baskets.

Marque	**Adidas**	**Nike**	**Puma**	**Autres marques**	**Total**
Effectif	100	70	120	160	450

On veut représenter cette répartition sous la forme d'un diagramme en bâtons. À chaque marque correspond un bâton. Les hauteurs des bâtons sont proportionnelles aux effectifs représentés.
On obtient le diagramme suivant :

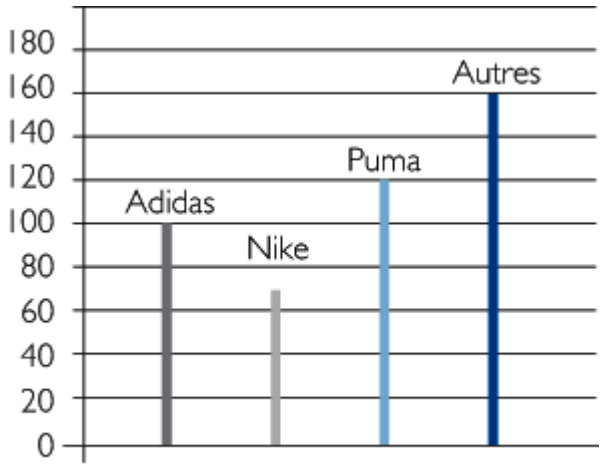


Diagramme statistique

Un diagramme statistique est une représentation graphique de **données statistiques.**
On distingue : les diagrammes circulaires, les diagrammes en barres, les diagrammes en bâtons.

4) L'histogramme: est adapté pour représenter des caractères statistiques dont les valeurs sont réparties en classes, c'est-à-dire que les valeurs sont réparties dans des intervalles.
Il est très utilisé pour représenter les effectifs ou les fréquences de ces classes.

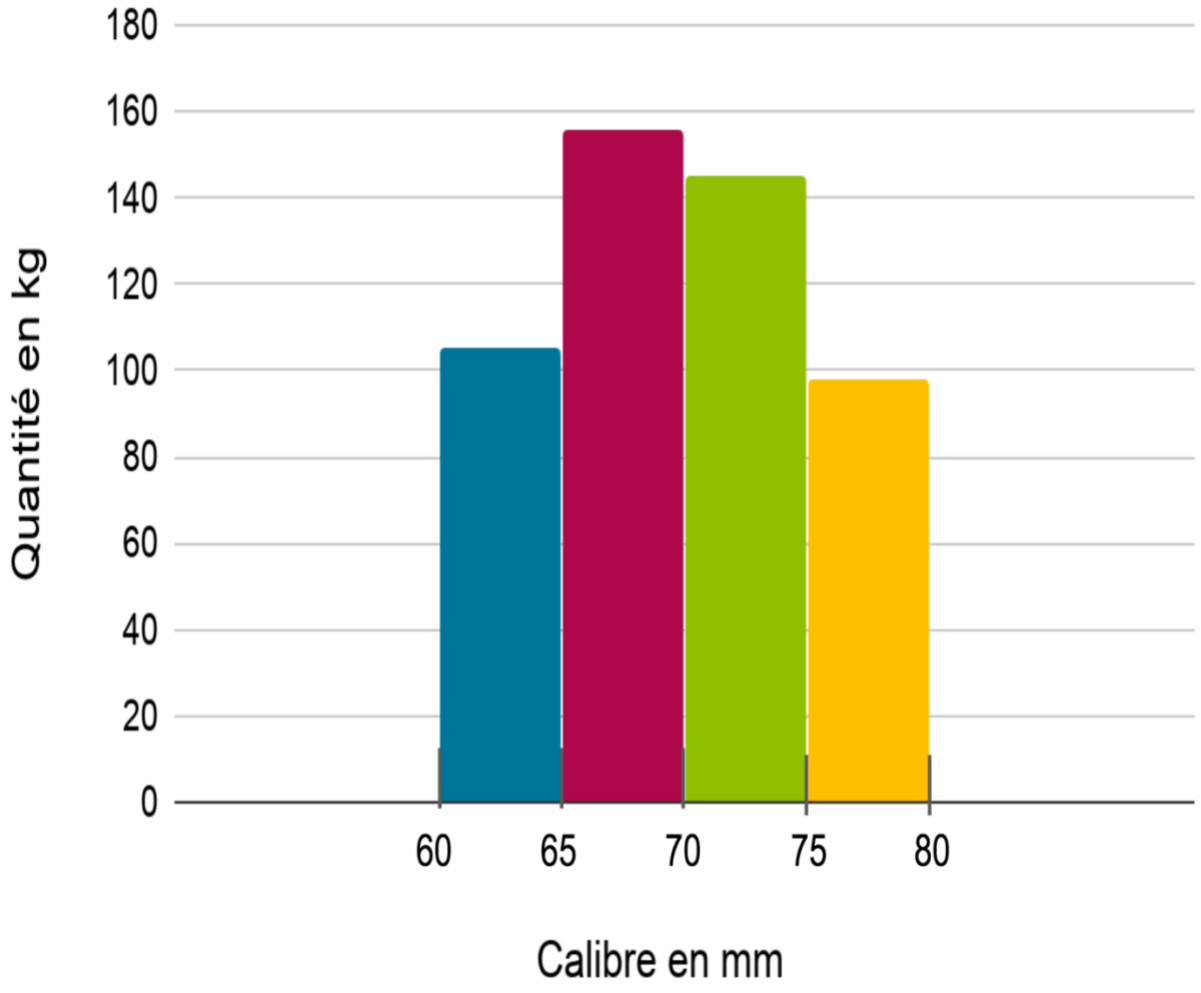
Pour le construire : On met généralement le caractère étudié en abscisse et les effectifs correspondants en ordonnée. Il faut bien choisir les unités pour que le diagramme soit le plus lisible possible.

Exemple

Dans une entreprise, pour être vendues, les pommes de terre sont triées selon leur calibre. On a regroupé les résultats dans le tableau suivant.

Calibre en mm	[60 ; 65[	[65 ; 70[	[70 ; 75[	[75 ; 80[
Quantité en kg	105	156	145	98

Sur l'axe des abscisses, on peut choisir 2 cm pour 5 mm et commencer à 55 mm.
Sur l'axe des ordonnées, on peut choisir 1 cm pour 10 kg.
On obtient l'histogramme correspondant.



Répartition de la quantité de pommes de terre en fonction de leur calibre

Pour le lire : Généralement, le caractère étudié est en abscisse et l'effectif correspondant en ordonnée. Dans ce cas, la hauteur de chaque barre est proportionnelle à l'effectif.
Si le caractère étudié est en ordonnée et l'effectif en abscisse, c'est la longueur de chaque barre (de gauche à droite) qui est proportionnelle à l'effectif.

Exemple :

Dans l'histogramme ci-dessus, le caractère étudié est le calibre de pommes de terre (en abscisse) et l'effectif correspondant est en ordonnée.

La barre verte signifie qu'environ 145 kg de pommes de terre ont un calibre compris entre 70 et 75 mm.

5) Le diagramme en boite :

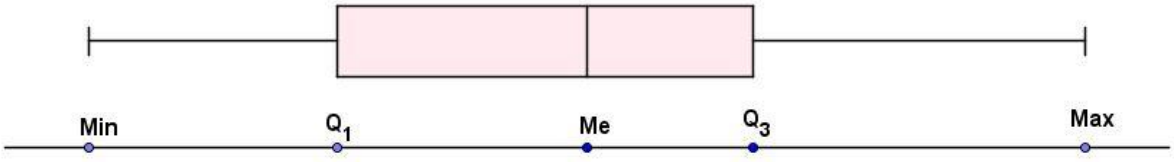
Le **diagramme en boite**, appelé également **boite à moustaches**, est en quelque sorte une représentation symbolique comportant plusieurs paramètres de la série. Sa définition peut varier mais on y trouvera toujours :

- la médiane Me ;
- les quartiles Q_1 et Q_3 ;
- les minimales et maximales du caractère (Min et Max).

Il est possible d'y trouver également les déciles D_1 et D_9 (et même des centiles...).

Pour le construire et pour le lire

Il se construit de la manière suivante :



En réalité, sur l'axe gradué qui est sous la boite ne doivent figurer que les valeurs des graduations.

Le diagramme en boite permet de visualiser les quatre quarts de la série :
- 25 % de la population a une valeur de caractère inférieure à Q_1 ;
- 50 % de la population a une valeur de caractère inférieure à Me ;
- 75 % de la population a une valeur de caractère inférieure à Q_3 ;
- 100 % de la population a une valeur de caractère inférieure à Max.

Le diagramme indique aussi l'étendue de la série (la longueur du diagramme) et l'écart interquartile (la longueur de la boite).
On a ainsi 50 % de la population « dans la boite ».

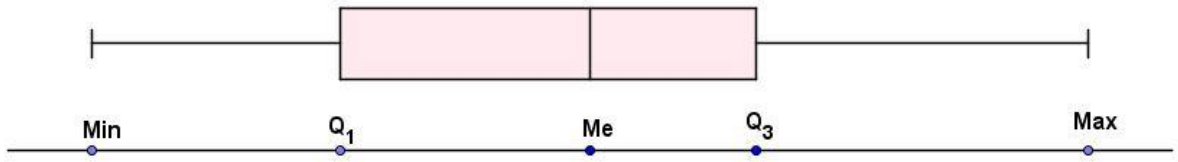
Exemple:

On considère la série statistique de 50 valeurs présentées dans le tableau ci-dessous. On a pris soin de les ordonner de la plus petite à la plus grande.

5	15	19	21	29
7	16	19	21	30
11	16	19	23	33
14	16	19	24	33
15	16	19	24	33
15	16	20	25	33
15	16	20	25	33
15	17	20	26	34
15	18	20	27	34
15	18	21	28	35

La plus petite valeur est Min = 5, la plus grande Max = 35.
Il y a 50 valeurs, la médiane est la moyenne de la 25^e et de la 26^e :
Me = (19 + 20) ÷ 2 = 19,5.
Pour Q_1 : 25 % de 50, cela fait 12,5. Il faut laisser au moins 25 % en dessous de Q_1, on prend la 13^e valeur : $Q_1 = 16$.
Pour Q_3 : 75 % de 50, cela fait 37,5. Il faut laisser 75 % au moins en dessous de Q_3, on prend la 38^e valeur: $Q_3 = 26$.

L'écart interquartile est donc égal à $Q_3 - Q_1 = 26 - 16 = 10$.
Voici une représentation de la boite à moustache de cette série.



6) Fonction de répartition et diagramme cumulatif

On appelle fonction de répartition d'une variable statistique quantitative toute Application définie par :

$F : \mathbb{R} \rightarrow [0, 1]$

$x \rightarrow F(x) = P(X \leq x)$

F(x) proportion des individus dont la valeur de la variable est strictement inférieure Ou égale à x, c'est-à-dire $X \leq x$.

6.1 : Cas de la variable statistique quantitative discrète

$F(x)$ = fréquence de $(X \leq x) = f1 + f2 + \ldots + fp = Fp$ tel que : f_1 , f_2 , …, fp sont les fréquences des valeurs de la variable $\leq x$, si non $F(x) = 0$. Donc :

$F(x) = 0$ Si $x < x_1$; $F(x) = F_i$ Si $x_i \leq x < x_{i+1}$; $F(x) = 1$ Si $x_r \leq x$. tel que r désigne l'ordre de la dernière valeur (modalité).

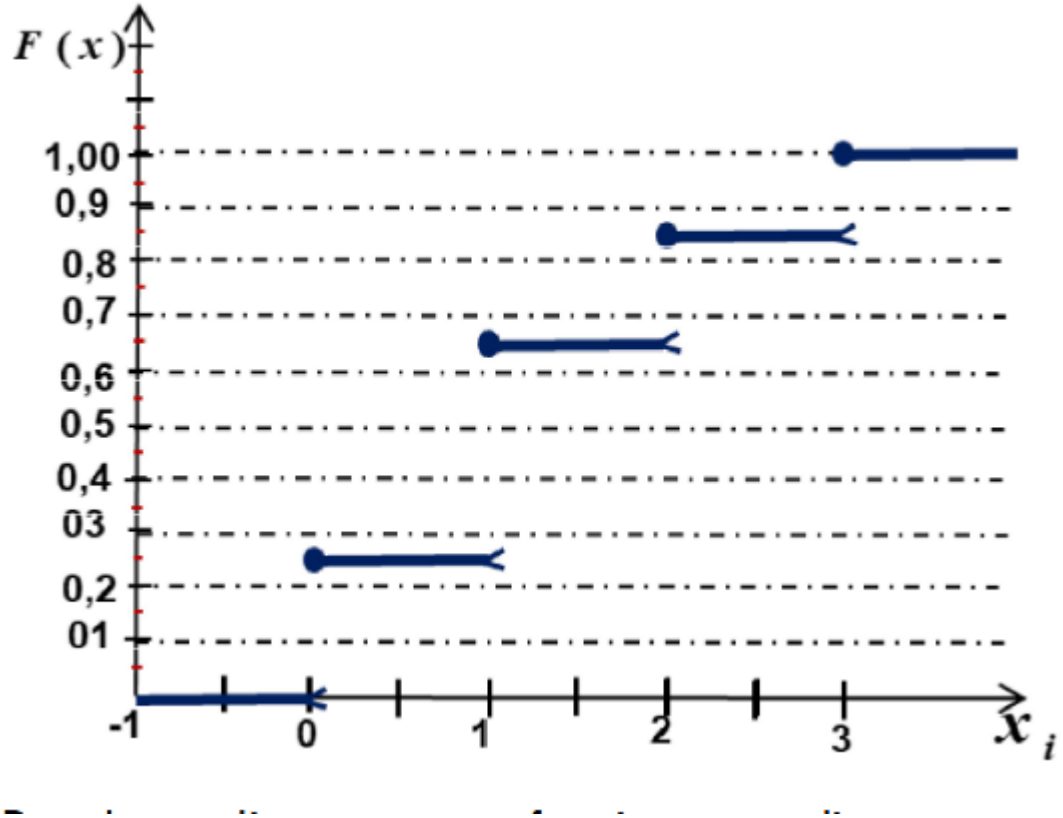
Exemple 6.1 : **Le tableau suivant, donne le nombre d'absences des étudiants au module d'analyse.**

Nombre d'absences x_i	**Effectifs n_i**	**f_i**	**F(x)**
0	5	0.25	0.25
1	8	0.4	0.65
2	4	0.2	0.85
3	3	0.15	1
TOTAL	20	1	

Donc la fonction de répartition correspondante est

$$F(x) = \begin{cases} 0 & si \quad x < 0 \\ 0{,}25 & si \quad 0 \leq x < 1 \\ 0{,}65 & si \quad 1 \leq x < 2 \\ 0{,}85 & si \quad 2 \leq x < 3 \\ 1 & si \quad 3 \leq x \end{cases}$$

Ainsi on obtient la représentation de la fonction de répartition, appelée diagramme cumulatif ou diagramme intégral.



Remarque : Dans le cas discret on a une fonction en escalier.

6.2 : Cas de la variable statistique quantitative continu

Dans ce cas on commence par la technique d'obtention de la courbe de la fonction de Répartition qui est appelée courbe cumulative.

Donc la fonction de répartition dans le cas quantitative continu est définit de la même

Façon que dans le cas quantitatif discret :

$F : \mathbb{R} \rightarrow [0 , 1]$

$x \rightarrow F(x) = P(X \leq x)$

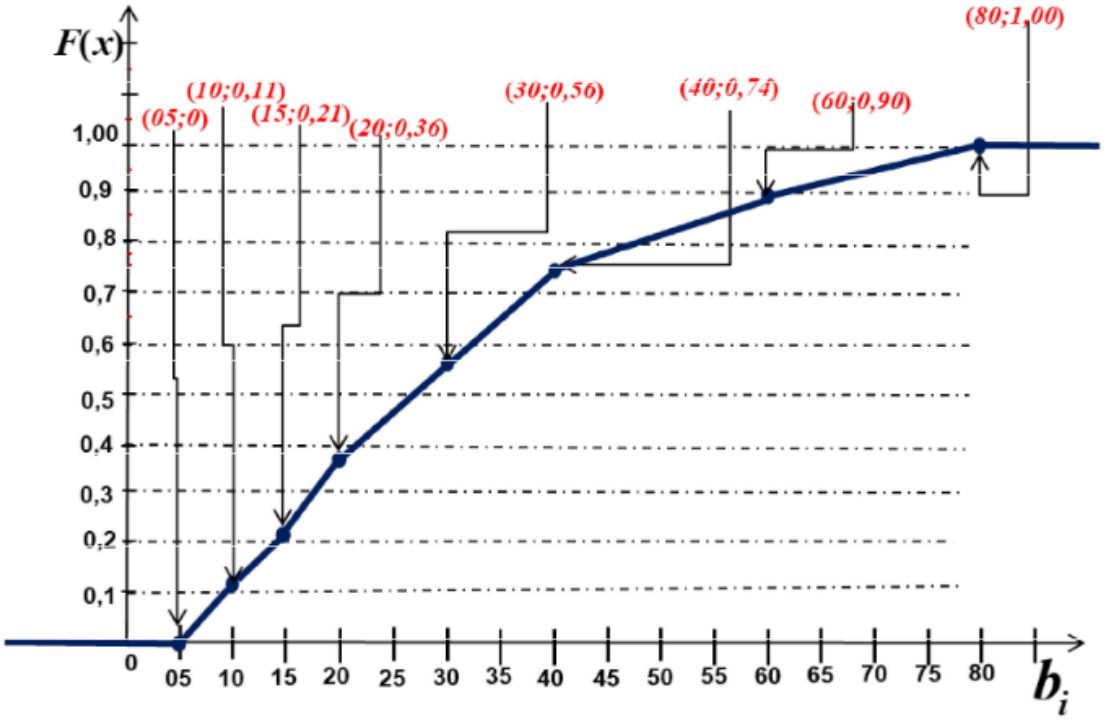
F(x) proportion des individus dont la valeur de la variable est strictement inférieure

Ou égale à x, c'est-à-dire $X \leq x$.

Exemple 6.2 :

$[b_{i-1}, b_i[$	n_i	N_i	f_i	F_i
[5 , 10[	11	11	0 ,11	0,11
[10 , 15[	10	21	0 ,10	0,21
[15, 20[	15	36	0 ,15	0,36
[20, 30[	20	56	0 ,20	0,56
[30 , 40[	18	74	0 ,18	0,74
[40 , 60[	16	90	0 ,16	0,90
[60, 80[	10	100	0 ,10	1,00

Ainsi la courbe de la fonction de répartition, appelée courbe cumulative se dessine comme suit :



Chapitre 3 : LES PARAMETRES DE POSITION.

1. Le Mode

Définition : En statistique, le mode désigne le nombre qui apparaît le plus souvent dans un ensemble de nombres le **mode**, **Il correspond au sommet de la distribution, graphiquement,** il est dénoté par M_O. Autrement dit le mode (ou valeur modale), est la valeur que la variable statistique prend le plus souvent (la valeur qui a le plus grand effectif).

Le mode peut être calculé pour les caractères qualitatifs comme pour les caractères quantitatifs.
Dans ce cours on va apprendre comment calculer le mode dans le cas d'une variable statistique qualitative, quantitative discrète et quantitative continue. Aussi on va apprendre comment trouver le mode par la méthode mathématique.

Par exemple, si un ensemble de nombres contient les chiffres suivants **:**

1, 1, 3, 5, 6, 6, 7, 7, 7, 8, le mode sera 7, car il apparaît le plus souvent parmi tous les nombres de l'ensemble.

Calculer le mode dans le cas d'une variable qualitative

Le mode est la modalité correspondante à l'effectif le plus important.

EXEMLE : Répartition des salariés de l'entreprise X selon le contrat de travail :

CSP	Cadres supérieurs	Contremaitres	Employer	Ouvriers spécialisés	Autres catégories
Effectifs des salariés	10	5	20	40	5

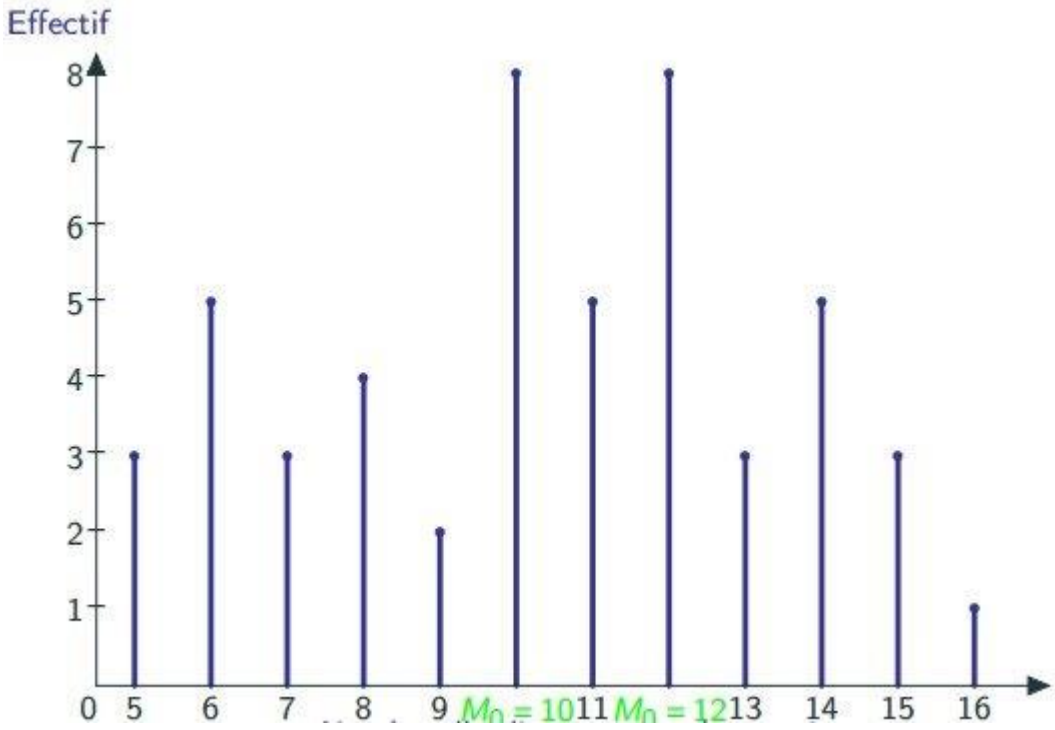
Le mode de ce caractère est la modalité « ouvriers spécialisés »

Calculer le mode dans le cas d'une variable quantitative discrète

Le mode est la modalité correspondante à l'effectif le plus important.

x_i	n_i	N_i	$f_i(\%)$	$F_i(\%)$
5	3	3	6	6
6	5	8	10	16
7	3	11	6	22
8	4	15	8	30
9	2	17	4	34
M_0=10	8	25	16	50
11	5	30	10	60
M_0=12	8	38	16	76
13	3	41	6	82
14	5	46	10	92
15	3	49	6	98
16	1	50	2	100
Total	50	-	100	-

A partir du tableau statistique ci-dessus (ou de son diagramme en bâtons correspondant), on a deux modes :
M0 = x6 = 10 et M0 = x8 = 12.



La classe modale dans le cas d'une variable quantitative continue

Dans le cas de variable continue, on parle plutôt de classe modale ; c'est la classe qui a la plus grande densité d'effectif (ou la plus grande densité de fréquence).

Si les classes sont toutes de même amplitude, la classe modale est la classe d'effectif le plus élevé ou de fréquence la plus élevée.

Si les classes ne sont pas toutes de même amplitude, la classe modale est la classe dont l'effectif corrigé ou la fréquence corrigée est maximum.

EXEMPLE 1 :

On considère la distribution statistique d'une population d'étudiants selon leur taille (en cm):

Taille (cm)	<160	[160;170[	[170;180[	[180;190[	≥ 190	total
effectif	6	7	8	2	1	24
Fréquences en %	25	29,1	33,3	8,3	4,3	100

L'effectif ou la fréquence les plus élevés montrent que le classe modale est [170;180[

Soit la distribution d'une population des étudiants répartis suivant leur poids (en kg) :

L'effectif ou la fréquence les plus élevés montrent que le classe modale est [170;180[

Soit la distribution d'une population des étudiants répartis suivant leur poids (en kg) :

Poids (en kg)	Effectif (n_i)	Fréquence (f_i) en %	Amplitude (a_i)	fréquence corrigée $f_{ico}=f_i/a_i$
<55	2	8,33	5	1,67
[55;60[	3	12,5	5	2,5
[60;70[	4	16,67	10	1,67
[70;75[	5	20,83	5	4,17
[75;85[	6	25	10	2,5
≥85	4	16,67	10	1,67
Total	24	100		

La classe modale, à laquelle est associée la fréquence corrigée la plus grande, est la classe [70; 75[

Détermination du Mode : On peut également déterminer la valeur du mode, à l'intérieur de la classe modale [2.09; 2.24 [.

$[b_{i-1}, b_i[$	n_i	N_i	f_i	F_i
[1.94, 2.09[	8	8	0.18	0.18
classe modale=[2.09, 2.24[	10	18	0.23	0.41
[2.24, 2.39[	9	27	0.20	0.61
[2.39, 2.54[	8	35	0.18	0.79
[2.54, 2.69[	5	40	0.11	0.90
[2.69, 2.84[	4	44	0.10	1.00
Total	44	-	1.00	-

On peut calculer le mode pour un caractère quantitatif continu par la formule suivante:

$$M_O = B_i + A_m \frac{\Delta 1}{\Delta 1 + \Delta 2}$$

où
- B_i est la borne inférieur de la classe modale
- A_m est l'amplitude de la classe modale
- Δ1= : différence entre la fréquence de la classe modale et la fréquence de la classe précédente
- Δ2= : différence entre la fréquence de la classe modale et la fréquence de la classe suivante

On peut aussi dire:

- Δ1: différence entre l'effectif de la classe modale et l'effectif de la classe précédente
- Δ2 : différence entre l'effectif de la classe modale et l'effectif de la classe suivante

Pour l'exemple précédent nous pourrons déterminer le mode plus précisément par cette formule dans ce même exemple :

On a $\Delta 1 = n2 - n1 = 10 - 8 = 2$ et $\Delta 2 = n2 - n3 = 10 - 9 = 1$,

donc : $M0 = Bi + (\Delta 1 / (\Delta 1 + \Delta 2)) A_m = 2.09 + 2/3 \times 0.15 = 2.19$

Exemple 2 : Soit le tableau suivant décrivant la distribution des salaires de75 employés:

Salaires	**[215,235[**	**[235,255[**	**[255,275[**	**[275,295[**
Effectifs	4	6	13	22
Fréquences	5,33	8	17,33	29,33
Salaires	**[295,315[**	**[315,335[**	**[335,355[**	**[355,375[**
Effectifs	15	6	5	4
Fréquences	20	8	6,67	5,33

Le mode calculé par la relation précédente est : 275+20((29,33-17,33)/(29,33-17,33)+(29,33-20)) = 286,25

2. La Médiane :

Médiane d'une série statistique

1) Définition :

La médiane d'une série ordonnée est la valeur qui partage cette série en deux séries de même effectif.

Il y a donc autant de valeurs supérieures à la médiane que de valeurs inférieures.

Interprétation de la médiane : Au moins 50% des valeurs de la série sont inférieures ou égales à la médiane et au moins 50% des valeurs de la série sont supérieures ou égales à la médiane.

Pour déterminer la médiane d'une série :

● On range les valeurs de la série dans l'ordre croissant

● On cherche la valeur qui partage la série en deux séries de même effectif.

Il y a deux cas, selon que l'effectif total est pair ou impair

2) Détermination de la médiane

a) Cas où l'effectif total est impair

Calculer la médiane de la série suivante :

6 ; 8 ; 10 ; 11 ; 12 ; 13 ; 15 ; 18 ; 20

● La série doit être rangée dans l'ordre croissant :

● Pour avoir deux groupes de même effectif on divise l'effectif total par 2 : $9 \div 2 = 4{,}5$

6 ; 8 ; 10 ; 11 ; 12 ; 13 ; 15 ; 18 ; 20

4 valeurs 4 valeurs

Valeur de la médiane Dans ce cas la médiane est 12.

b) Cas où l'effectif total est pair

Calculer la médiane de la série suivante :

1 ; 2 ; 5 ; 8 ; 9 ; 12 ; 14 ; 16 ; 17 ; 19

● La série doit être rangée dans l'ordre croissant :

● Pour avoir deux groupes de même effectif : $10 \div 2 = 5$

1 ; 2 ; 5 ; 8 ; 9 ; 12 ; 14 ; 16 ; 17 ; 19

5 valeurs 5 valeurs

Valeur de la médiane

Dans ce cas la médiane est comprise entre 9 et 12

En général on choisit comme médiane la moyenne des deux valeurs centrales :

$(9 + 12) \div 2 = 10{,}5$

La médiane est : 10,5

Médiane d'une série continue

Si la variable est continue (regroupement par intervalle des résultats) le calcul de la médiane se fait autrement :

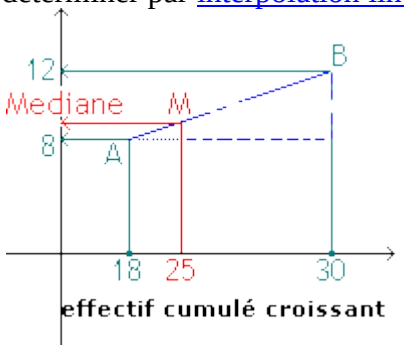
Notes	Effectifs	Effectifs cumulés
[0 ; 5[	10	10
[5 ; 8[	8	18
[8 ; 12[	12	30
[12 ; 15	11	41
[15 ; 20	9	50
	50	

Utilisons la colonne des effectifs cumulés pour déterminer la médiane : il y a 50 notes, 50 % de l'effectif total c'est 25, la médiane est ici la note correspondant à l'effectif cumulé 25.

D'après la colonne "effectif cumulé" :

- 18 personnes ont moins de 8
- 30 personnes ont moins de 12

La médiane se trouve donc dans l'intervalle [8;12[(appelée classe médiane) on va la déterminer par interpolation linéaire.



Les points A, M, B sont alignés ce qui se traduit par les droites (AM) et (AB) ont même coefficient directeur (ou on utilise le théorème de Thalès dans le triangle bleu) :

$$\frac{Me-8}{25-18}=\frac{12-8}{30-18}$$

$$\frac{Me-8}{7}=\frac{4}{12}$$

$$Me-8=\frac{4}{12}\times 7$$

$$Me=8+\frac{4}{12}\times 7\approx 10,33$$

La médiane est environ 10,33

50 environ des personnes ont eu moins de 10,33 et 50 % plus de 10,33 .

Les différentes moyennes

A) **Moyenne arithmétique :** **La *moyenne arithmétique* est la moyenne « ordinaire », c'est-à-dire la somme des valeurs numériques (de la liste) divisée par le *nombre* de ces valeurs numériques.**

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

i) Variable quantitative discrète

La moyenne arithmétique pondérée , est égale à la somme des valeurs distinctes de la variable multipliées par leurs effectifs respectifs divisée par la somme des effectifs.

$$\overline{x} = \frac{\sum_i n_i x_i}{\sum_i n_i} = \frac{\sum_i n_i x_i}{N}$$

et comme $f_i = \frac{n_i}{N}$ on a aussi $\overline{x} = \sum_i f_i x_i$

ii) Variable quantitative continue

La moyenne arithmétique notée toujours x, est égale à la somme des centres des classes de la variable multipliées par leurs effectifs respectifs divisée par la somme des effectifs.

$$\overline{x} = \frac{\sum_i n_i c_i}{\sum_i n_i} = \frac{\sum_i n_i c_i}{N}$$

où c_i est le centre de de la classe associée à l'effectif n_i.
et comme $f_i = \frac{n_i}{N}$ on a aussi $\overline{x} = \sum_i f_i c_i$

B) Moyenne quadratique :

i) Variable quantitative discrète : La moyenne quadratique notée x q , est égale à la somme des carrés des valeurs distinctes de la variable multipliées par leurs effectifs respectifs divisée par la somme des effectifs

$$\overline{x}_q = \frac{\sum_i n_i\, x_i^2}{\sum_i n_i} = \frac{\sum_i n_i\, x_i^2}{N} = \sum_i f_i\, x_i^2 \quad (\text{ car } f_i = \frac{n_i}{N})$$

ii) Variable quantitative continue

La moyenne quadratique notée toujours $\mathbf{x_q}$, est égale à la somme des carrés des centres des classes de la variable multipliées par leurs effectifs respectifs divisée par la somme des effectifs.

$$\overline{x}_q = \frac{\sum_i n_i\, c_i^2}{\sum_i n_i} = \frac{\sum_i n_i\, c_i^2}{N} = \sum_i f_i\, c_i^2 \quad (\text{ car } f_i = \frac{n_i}{N})$$

où c_i est le centre de de la classe associée à l'effectif n_i.

C) Moyenne géométrique :

i) Variable quantitative discrète

La moyenne géométrique notée x G, d'une variable quantitative discrète est donnée par :

$$\overline{x}_G = \sqrt[N]{\prod_i x_i^{n_i}} \text{ où } N = \sum_i n_i$$

D) Moyenne harmonique

i) Variable quantitative discrète

C'est l'inverse de la moyenne arithmétique des inverses des valeurs de la variable. On la note x H ,

$$\overline{x}_H = \frac{N}{\sum_i n_i/x_i}$$

ii) Variable quantitative continue dans ce cas La moyenne harmonique est donnée par :

$$\overline{x}_H = \frac{N}{\sum_i n_i/c_i}$$

Remarque :

$$\overline{x}_H \leq \overline{x}_G \leq \overline{x} \leq \overline{x}_q$$

Chapitre 4 : LES PARAMETRES DE DISPERSION

4.1 INTRODUCTION :

Le résumé d'une distribution que donne une valeur centrale ne nous renseigne pas sur la dispersion des valeurs autour de cette valeur centrale, c'est-à-dire sur la tendance qu'elles-ont à se concentrer ou se disperser autour de celle-ci.

Exemple : Si l'on considère deux professeurs X et Y chargés de noter 9 élèves, peut-on apprécier leur manière de noter simplement en regardant la moyenne, la médiane ou le mode de leurs notes ?

Notation de 9 étudiants par les professeurs X et Y :

Etudiant	Notes du Pr X	Notes du Pr Y
A	7	0
B	8	5
C	9	9
D	10	10
E	10	10
F	10	10
G	11	11
H	12	15
I	13	20
mode	10	10
moyenne	10	10
médiane	10	10

=> A s'en tenir à l'analyse des valeurs centrales, on serait amené à conclure que les deux Pr X et Y notent rigoureusement de la même manière (moyenne=médiane=mode=10) mais on sent bien intuitivement que ce n'est pas le cas et qu'il existe une différence dans leur style de notation. Cette différence tient au fait que le professeur X "concentre" ses notes autour de 10 alors que le professeur Y "disperse" davantage ses notes autour de la valeur de référence.

Il est donc utile de compléter les valeurs centrales par un paramètre de dispersion absolue qui donne un ordre de grandeur de l'écart des valeurs entre elles ou, ce qui revient au même, de l'écart des valeurs à la valeur centrale de référence.
On appelle **dispersion statistique**, la tendance qu'ont les valeurs de la distribution d'un caractère à s'étaler de part et d'autre d'une valeur centrale et/ou à s'éloigner les unes des autres. Ce calcul n'a évidemment de sens que pour les caractères quantiatifs.

On distingue les paramètres de dispersion absolue (mesurée dans l'unité de mesure du caractère) et les paramètres de dispersion relative (mesurée par un nombre sans dimension).

4.2 LES PARAMETRES DE DISPERSION ABSOLUE

Les paramètres de dispersion absolue indiquent de combien les valeurs d'une distribution s'écartent en général de la valeur de référence. Un paramètre de dispersion absolue s'exprime toujours dans l'unité de mesure de la variable considérée. Ainsi, si l'on étudie la densité de population des régions européennes, l'unité de mesure de la dispersion de ce caractère sera exprimée en habitants par km2.

Les quatre paramètres de dispersion absolue les plus courants sont **l'étendue, l'intervalle interquartiles, l'écart absolu moyen et l'écart type**.

4.2.1 Etendue

Définition : l'étendue d'une distribution est égale à la différence entre la plus grande et la plus petite valeur de la distribution :

Etendue de X = X_{max}-X_{min}

Exemple : Notation des professeurs X et Y :
- L'étendue des notes données par le professeur X est de (13-7)=6, ce qui signifie que l'écart maximum entre deux notes du professeur X est de 6.
- L'étendue des notes données par le professeur Y est de (20-0)=20 ce qui signifie que l'écart maximum entre deux notes du professeur Y est de 20
=> La dispersion des notes du professeur Y est donc beaucoup plus forte que celle des notes du professeur X.

L'inconvénient de l'étendue est qu'elle dépend uniquement des deux valeurs les plus extrêmes de la distribution. **Elle indique donc la différence maximum entre deux valeurs mais pas la différence typique.**

4.2.2 Quartiles

Pour remédier aux inconvénients de l'étendue, on peut retirer les valeurs les plus extrêmes et calculer l'intervalle des valeurs restantes : c'est la base de la méthode des quartiles. On appelle **quantiles** les bornes d'une partition en classes d'effectifs égaux. **Attention, lorsqu'on utilise les quantiles, ce sont les effectifs qui sont égaux et non pas les amplitudes.**

- Les **quartiles** par exemple, sont les trois valeurs qui permettent de découper la distribution en quartes classes d'effectifs égaux, on les note Q_1, Q_2 et Q_3.

Classes	fréquence simple
[Xmin ; Q1 [	25 %
[Q1 ; Q2 [	25 %
[Q2 ; Q3 [	25 %
[Q3 ; Xmax]	25 %

- **L'intervalle interquartile (Q_3-Q_1)** est un paramètre de dispersion absolue qui correspond à l'étenude de la distribution une fois que l'on a retiré les 25% des valeurs les plus faibles et les 25% des valeurs les plus fortes. 50% des observations sont donc concentrées entre Q_1 et Q_3.en quatres classes d'effectifs égaux, on les note Q_1, Q_2 et Q_3.

4.2.3 Ecart absolu moyen / Ecart absolu médian

Définition : **l'écart absolu moyen** est la moyenne de la valeur absolue des écarts à la moyenne. Autrement dit, c'est la distance moyenne à la moyenne. Bien qu'il soit moins utilisé, on peut calculer de la même manière **l'écart absolu médian** qui est la moyenne des écarts à la médiane.

Formules de l'écart absolu moyen et de l'écart absolu médian

$$Ecart\,absolu\,moyen = \frac{1}{N}\sum_{i=1}^{N}\left| x_i - \bar{x} \right|$$

$$Ecart\,absolu\,médian = \frac{1}{N}\sum_{i=1}^{N}\left| x_i - Médiane \right|$$

L'intérêt de ces deux valeurs centrales est d'être facile à calculer et simples à interpréter.

Exemple : Notation des professeurs X et Y :

Le calcul de l'écart absolu moyen des notes du Pr X est obtenu en effectuant la moyenne de la valeur absolue des écarts à la moyenne :

L'écart absolu moyen de la notation du professeur X est donc de 1.3, ce qui signifie que les notes s'écartent en moyenne de 1.3 de la moyenne. Il n'y a donc pas, en moyenne, de gros écarts à la moyenne. Si on effectue le même calcul pour le professeur Y, on trouve un écart absolu moyen de 3.6, ce qui signifie que ses notes s'écartent généralement beaucoup plus de la moyenne. On peut donc conclure que la dispersion des notes du Pr Y est plus forte que celle du Pr X.

4.2.4 Ecart-type

Définition : l'écart-type est la racine carrée de la variance, elle-même définie comme la moyenne du carré des écarts à la moyenne :

Formules de la variance et de l'écart-type

$$Variance: (\sigma_X)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

$$écart-type: \sigma_X = \sqrt{Variance} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

La variance n'est pas à proprement parler un paramètre de dispersion absolue mais plutôt une mesure globale de la variation d'un caractère, c'est-à-dire de la **quantité moyenne d'information** contenue dans les différentes valeurs de ce caractère : cette quantité d'information serait évidemment nulle si toutes les valeurs étaient égales et elle est d'autant plus élevée que ces valeurs sont différentes les unes des autres.

L'écart-type est le paramètre de dispersion absolue le plus utilisé en statistique.

4..2.5 Les paramètres de dispersion relative : Un paramètre de dispersion relative est une mesure de l'écart absolu des valeurs d'une distribution à une valeur centrale. C'est donc un rapport :

$$\text{Paramètre de disp. relative} = \frac{\text{param. de disp. absolue}}{\text{valeur centrale}}$$

Les plus courants sont :

- **le coefficient de variation (C.V.)** = $\sigma_X / \overline{x}$
- **l'écart moyen relatif** = écart absolu moyen / $\overline{x}$
- **le coefficient interquartile relatif** = $(Q_3-Q_1)/Q_2$

Dans tous les cas, le paramètre de dispersion est un nombre sans dimension (on a fait le rapport de deux nombres ayant la même unité de mesure) qui exprime de combien les valeurs s'écartent de la valeur centrale en valeur relative. On note souvent les paramètres de dispersion en %.

Chapitre 5 : Les paramètres de forme d'une série statistique

Caractéristiques de forme : Il est intéressant de repérer la forme par des mesures de son asymétrie et de son Aplatissement. Une variable statistique est symétrique si ses valeurs sont réparties de manière symétrique autour de la moyenne c'est à dire si le polygone des fréquences a la forme d'une clôche comme dans la figure suivante :

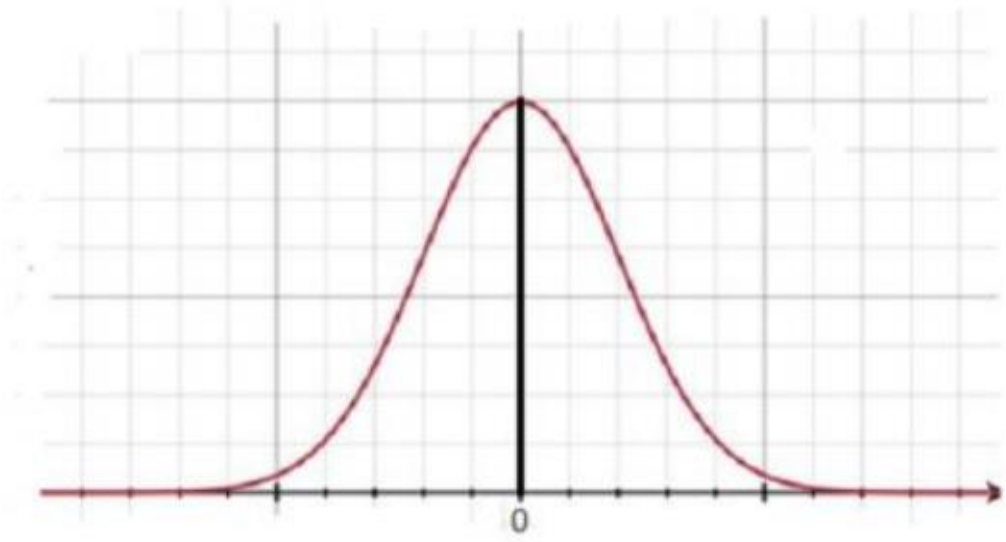


Figure. Clôche

- **Symétrie : moyenne =médiane = mode**
- **Asymétrie à droite : mode < médiane < moyenne**
- **Asymétrie à gauche : moyenne < médiane < mode**

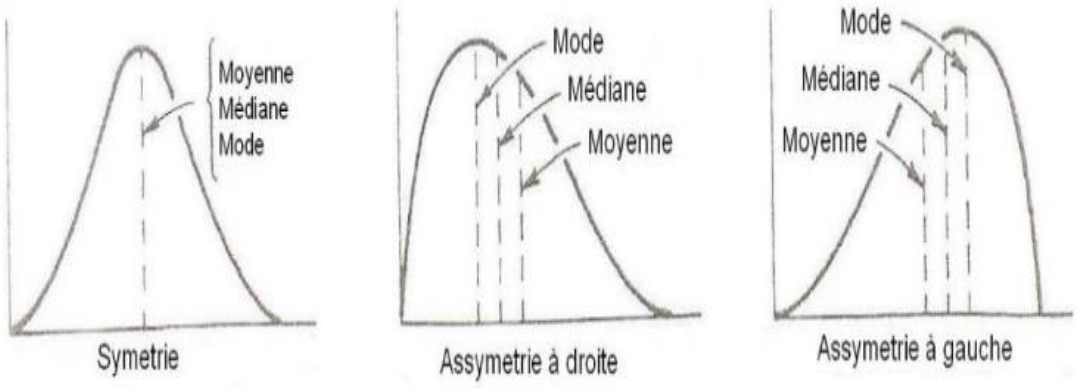


Figure. Symétrie et asymétrie

Les Coefficients d'asymétrie :

1. **Le coefficient de dissymétrie de Pearson** : Le coefficient de dissymétrie est basé sur les écarts entre les mesures de tendance centrale.

$$\beta_1 = \frac{3(\bar{x} - M)}{\sigma}$$

L'interprétation de ces coefficients est directe

- si le coefficient est **nul**, la distribution est **symétrique**
- si le coefficient est **négatif**, la distribution est **déformée à gauche** de la médiane (sur-représentation de valeurs faibles, à gauche)
- si le coefficient est **positif**, la distribution est **déformée à droite** de la médiane (sur-représentation de valeurs fortes, à droite)

2° Le coefficient de dissymétrie de YULE : Dans une distribution symétrique, les quartiles sont situés à égale distance de chaque coté de la médiane.

Par conséquent : (Q3 -Q2)- (Q2 -Q1)=0

N.B : Si la distribution est dissymétrique, l'égalité ci-dessus n'est plus vraie.

Le coefficient dissymétrie de Yule mesure l'asymétrie à partir de la position relative des quartiles par rapport à la médiane :

$$C_Y = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

Interprétations

La valeur du coefficient de Yule est toujours comprise entre -1 et +1 et son signe indique le sens de l'asymétrie :

$-1 \leq C_\gamma < 0$ distribution dissymétrique à gauche

$C_\gamma = 0$ distribution symétrique

$0 < C_\gamma \leq 1$ distribution dissymétrique à droite

3.Le coefficient de dissymétrie de FISHER **:** Ce coefficient est basé sur les écarts par rapport à la moyenne des valeurs en utilisant le moment centré d'ordre 3.

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \qquad \text{où } \mu_3 = \sum_{i=1}^{k} f_i (x_i - \bar{x})^3$$

$$\text{et } \sigma = \sqrt{\sigma^2} = \sqrt{\mu_2} = \sqrt{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2} \text{ écart-type}$$

Interprétations

Le signe du coefficient de Fischer indique le sens de la dissymétrie :

$\gamma_1 < 0$ distribution dissymétrique à gauche

$\gamma_1 = 0$ distribution symétrique

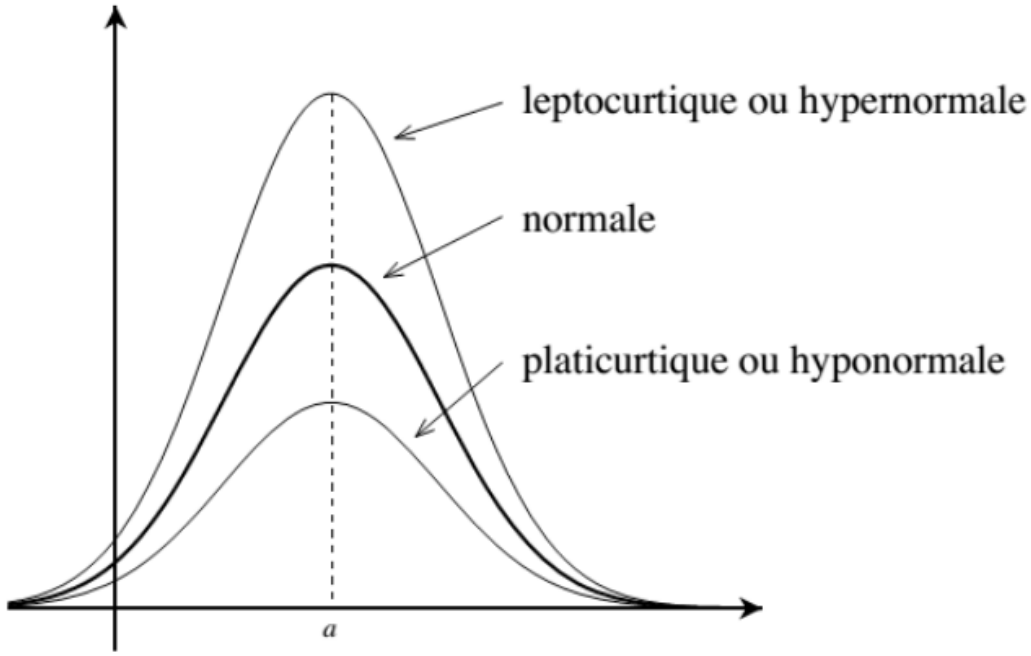
$\gamma_1 > 0$ distribution dissymétrique à droite

Mesures de forme: coefficients d'aplatissement

Les mesures d'aplatissement font partie des mesures qui caractérisent la forme d'une distribution.

Elles caractérisent le degré d'aplatissement de la distribution par rapport à l'aplatissement de la distribution normale («courbe en cloche»2).Il est alors utile de pouvoir mesurer si la forme de la distribution présente une déviation par rapport à l'aplatissement de la distribution normale.

Une distribution est platicurtique ou hyponormale si la courbe est plus aplatie que la courbe normale; elle est leptocurtique ou hypernormale si la courbe est plus pointue que la courbe normale.



Interprétations

$\beta_2 > 3$ courbe leptocurtique ou hypernormale

$\beta_2 = 3$ courbe normale

$\beta_2 < 3$ courbe platicurtique ou hyponormale

Coefficients d'aplatissement :

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} \qquad \gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\sigma^4} - 3$$

Pour mesurer l'aplatissement de la courbe, on utilise **le coefficient β_2 de Pearson** basé sur le moment centré d'ordre 4:

1. Le coefficient d'aplatissement de Pearson :

$$\beta_2 = \frac{\mu_4}{\sigma^4}$$

où $\mu_4 = \sum_{i=1}^{k} f_i (x_i - \bar{x})^4$ et $\sigma^2 = \mu_2 = \sum_{i=1}^{k} f_i (x_i - \bar{x})^2$ variance

Interprétations

$\beta_2 > 3$ courbe leptocurtique ou hypernormale

$\beta_2 = 3$ courbe normale

$\beta_2 < 3$ courbe platicurtique ou hyponormale

Coefficients d'aplatissement :

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} \qquad \gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\sigma^4} - 3$$

Interprétations

On interprète le kurtosis 2 de la manière suivante :

- Si $\gamma_2 = 0$, la courbe de fréquences est comparable à celle de la loi normale. On dit qu'elle est mésokurtique.
- Si $\gamma_2 > 0$, la courbe de fréquences est plus pointue que celle de la loi normale. On dit qu'elle est leptokurtique.
- Si $\gamma_2 < 0$, la courbe de fréquences est plus aplatie que celle de la loi normale. On dit qu'elle est platykurtique.