Première partie

Chapitre I.

1- Statistiques descriptive à une dimension

1.1. Introduction:

Recueillir et analyser les données sont les deux objectifs fondamentaux de la statistique. Pour y parvenir, il faut suivre les étapes suivantes :

La collecte des données : définir l'objet étudié, les variables statistiques mises en cause, le questionnaire et construire l'échantillon représentatif (sondage recensement, plan d'expériences….).

Une fois les données collectées et corrigées, les visualiser sous forme de tableaux et (ou) de graphiques et les résumer grâce à des paramètres qui permettent de dégager les caractéristiques essentielles du phénomène étudié (statistique descriptive, analyse des données).

L'étape de la modélisation (statistique mathématique) est de fournir des résultats relatifs à une population à partir de mesures statistiques réalisées sur des échantillons. La statistique mathématique fournit des éléments permettant de spécifier du mieux possible le modèle probabiliste qui a engendré les données.

Les méthodes statistiques sont utilisées dans de nombreux domaines tels que l'ingénierie (contrôle de qualité de fabrication), la médecine (expérimentation de nouveaux médicaments ….), les sciences sociales et économiques (l'économétrie, la démographie….) et bien dans d'autres domaines.

La statistique est une branche des mathématiques appliquées, qui consiste à réunir des données chiffrées sur des ensembles nombreux, à analyser et à critiquer ces données. Elle désigne l'ensemble des techniques d'interprétation mathématique appliquée à des phénomènes pour lesquels une étude exhaustive de tous les facteurs est impossible. Elle fournit de la manière la plus rigoureuse possible des éléments d'appréciation utiles à l'explication ou à la prévision de ces phénomènes. Par contre Le mot ″ **statistiques** ″ au pluriel, désigne l'ensemble des données chiffrées qui regroupent toutes les observations faites sur des faits relatifs à un même phénomène qui concerne un groupe d'individus ou d'objets. Ces données sont essentiellement extraites des recensements de la population, des déclarations du registre d'état civil ou

d'enquêtes appropriées et sont groupées sous forme de tableaux, de graphiques et d'indicateurs statistiques.

La statistique, en tant que méthode d'analyse des données quantitatives et qualitatives, comporte deux phases:

- La statistique descriptive, qui consiste en la collecte et la présentation de données, ainsi que leur première analyse. Le but est de représenter d'une manière compréhensible et utilisable l'information fournie par les données.
- La statistique mathématique, qui cherche à trouver les caractéristiques de la population mère à partir des observations faites sur un échantillon. Elle prend la suite de la statistique descriptive et fait appel au calcul des probabilités.

1.2. Concepts de base des statistiques

1.2.1. Population statistique : C'est l'ensemble de référence constitué des unités observées, cela ne signifie pas exclusivement un ensemble de personnes physiques, mais peut concerner des personnes morales ou des objets (entreprises, exploitation agricoles, ampoules, voitures….). Une population doit être bien définie.

1.2.2. Caractères et modalités : Le **caractère** désigne une grandeur ou un attribut observable sur un individu, c'est donc un aspect particulier de l'élément que l'on désire étudier qui est susceptible de varier en prenant différents états appelés **modalités.**

On distingue deux types de caractères : caractères **qualitatifs** et caractères **quantitatifs.**

1.2.2.1. Caractère qualitatif : Un caractère est dit **qualitatif** lorsqu'il est lié à une observation qui n'est pas mesurable. Les modalités du caractère qualitatif rangent les unités de la population étudiée en catégories. Tout individu appartient, sans ambiguïté à une seule catégorie.

Exemple : Lors de l'étude de la population estudiantine, on s'intéresse à quelques caractéristiques telles que : La mention du baccalauréat, le milieu de résidence, le sexe, l'état matrimonial, la couleur des yeux etc…….

Les modalités d'un caractère qualitatif sont simplement les différentes rubriques d'une nomenclature définie à priori et associées à ce caractère. La nomenclature désigne la liste des modalités d'un caractère précédées d'un numéro.

Tableau statistique associé à un caractère qualitatif.

$Modalités \quad M_i$	$Effectif \quad n_i$	$Fréquence \quad f_i$
M_1	n_1	f_1
M_2	n_2	f_2
M_3	n_3	f_3
…..	……	…….
M_p	n_p	f_p
$Somme$	N	1

N : Effectif total (nombre total d'éléments observés)

n_i: Effectif d'une modalité, appelé aussi fréquence absolue (c'est le nombre de fois ou la modalité numéro i a été observée).

f_i **:** Fréquence relative d'une modalité est égale au rapport de l'effectif partiel n_i à l'effectif total N.

$$f_i = \frac{n_i}{f_i} \quad \text{et} \quad \sum_{i=1}^{k} f_i = 1$$

Remarque : les fréquences relatives peuvent être exprimées en pourcentage.

1.2.2.2. Caractère quantitatif : Lorsque les observations relatives à un caractère sont mesurables, le caractère est dit **quantitatif** (taille, poids, nombre d'enfants par familles, surface d'un logement…….). A chaque modalité correspond un nombre différent. On distingue deux types de caractères quantitatifs :

- **Caractère quantitatif discret :** Auquel, les valeurs possibles de la variable sont des nombres isolés, généralement des nombres entiers tels que (nombre d'enfants par famille, nombres d'années d'études, nombres de pièces par logement, … etc.

Tableau statistique associé à un caractère quantitatif discret.

x_i	n_i	$N_i \uparrow$*(croissantes)*	f_i	$F_i \uparrow$*(croissantes)*
x_1	n_1	$N_1 = n_1$	f_1	$F_1 = f_1$

x_2	n_2	$N_2 = n_1 + n_2$	f_2	$F_2 = f_1 + f_2$
x_3	n_3	$N_3 = n_1 + n_2 + n_3$	f_3	$F_3 = f_1 + f_2 + f_3$
……	……	……	……	……
x_p	n_p	$N_p = \sum_{i=1}^{r} n_i$	f_p	$F_p = 1$
Total	*N*	……	1	……

F_i : Fréquences cumulées croissantes : Le cumule des fréquences associées aux valeurs du caractère inférieures strictement à la valeur x_i.

- **Caractère quantitatif continu :** Auquel les valeurs possibles de la variable sont à priori en nombre infini dans un intervalle de valeurs telle que (l'âge, la taille, la moyenne des notes aux examens, etc.….).

Tableau statistique associé à un caractère quantitatif continu.

Classes $[b_{i-1};\ b_i\ [$	c_i	n_i	f_i	$F_i\uparrow$
$[b_0\ ;\ b_1\ [$	C_1	n_1	f_1	$F_1 = f_1$
$[b_1\ ;\ b_2\ [$	C_2	n_2	f_1	$F_2 = f_1 + f_2$
$[b_2\ ;\ b_3\ [$	C_3	n_3	f_3	$F_3 = f_1 + f_2 + f_3$
…	…	…	…	…
$[b_{p-1}\ ;\ b_p\ [$	C_p	n_p	f_p	*1*
		N	*1*	

Remarque :

- Par convention, les classes sont fermées à gauche et ouvertes à droites.
- Le centre d'une classe est : $c_i = \frac{b_{i-1}+b_i}{2}$
- L'amplitude d'une classe est : $a_i = b_i - b_{i-1}$

1.3. Représentations graphiques.

D'une manière générale, il est plus commode d'observer un graphique que de lire un tableau. La synthèse visuelle fournie autant d'information qu'un alignement de chiffres. C'est pour cette raison, qu'en statistique descriptive, il est recommandé d'utiliser les représentations graphiques.

1.3.1. Caractère qualitatif.

Pour un caractère **qualitatif** il existe un nombre important de représentations qui aboutissent au même résultat ; mais les plus répandus et utilisés sont les représentations en tuyaux d'orgues et les diagrammes secteurs angulaires.

a- Représentation en tuyaux d'orgues.

Le diagramme en tuyaux d'orgues est représenté dans un repère formé d'un axe vertical gradué et d'un axe horizontal non gradué. A chaque modalité correspond un rectangle, parallèle à l'axe des ordonnées, **dont la hauteur est proportionnelle à la fréquence ou à l'effectif** associée à cette modalité. Les rectangles ont des largeurs égales et sont espacés les uns des autres par des distances égales, et en abscisse sont portées, les modalités de la variable de façon arbitraire.

Remarque :

- Etant donné que les modalités sont qualitatives il n ya pas d'ordre entre elles.
- Ce type de graphique permet de comparer les résultats obtenus sur différentes populations.

b- Représentations par secteurs angulaires. (Diagramme circulaire).

Les modalités sont représentées par un secteur angulaire d'un disque ou d'un demi-disque dont l'angle au centre est proportionnel à l'effectif ou à la fréquence, l'angle de chaque secteur, pour un disque plein, est donné par la formule : $\boldsymbol{\alpha_i° = 360° \times f_i}$

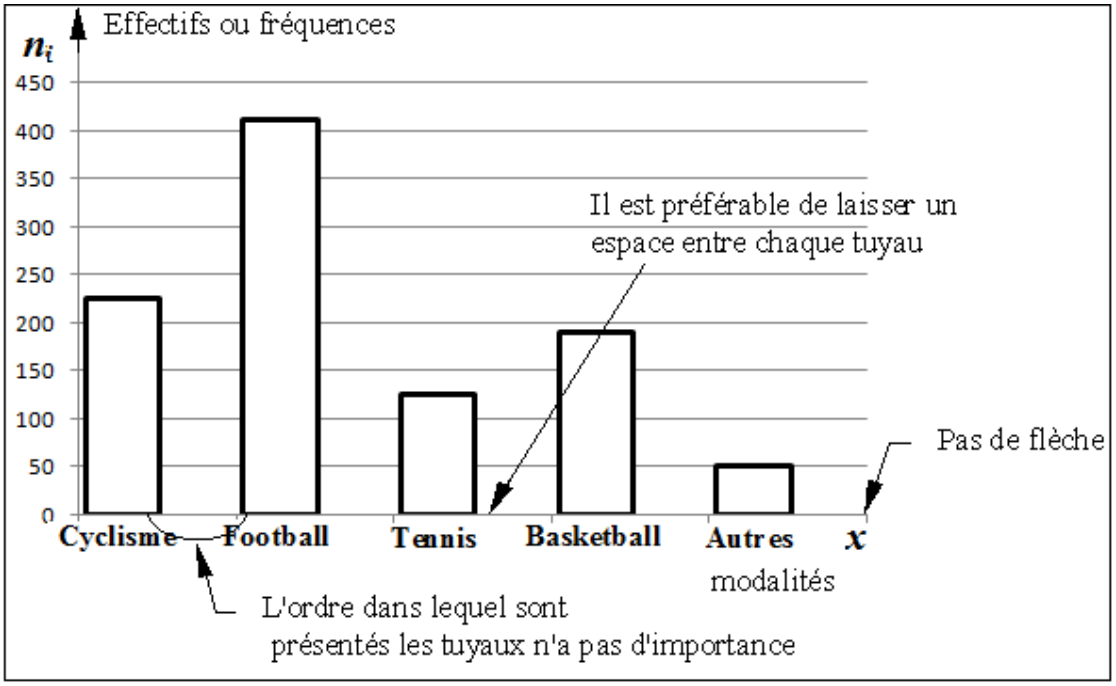
Le diagramme circulaire ou semi circulaire, consiste à partager un disque ou un demi disque, en tranches ou secteurs, correspondants aux modalités observées et dont la **surface** est **proportionnelle aux effectifs ou aux fréquences de chaque modalité**.

Exemple: On étudie la répartition de 1000 étudiants pratiquant une discipline sportive à l'université. Les résultats sont distribués comme suit :

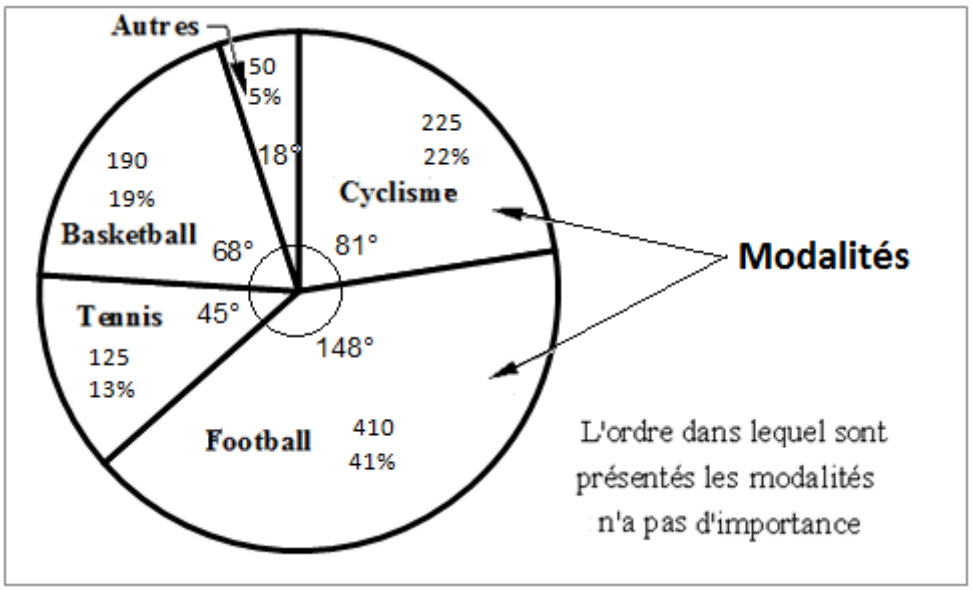
Variable X	Cyclisme (1)	Football (2)	Tennis (3)	Basketball (4)	Autres (5)	Total
Effectif n_i	225	410	125	190	50	1000
Fréquence f_i	0,225	0,410	0,125	0,190	0,05	1,000
Angle $\alpha_i°$	81°	148°	45°	68°	18°	360°

Représentation graphique la série statistique.

1- Représentation en tuyaux d'orgues.



2- Représentations par un diagramme circulaire.



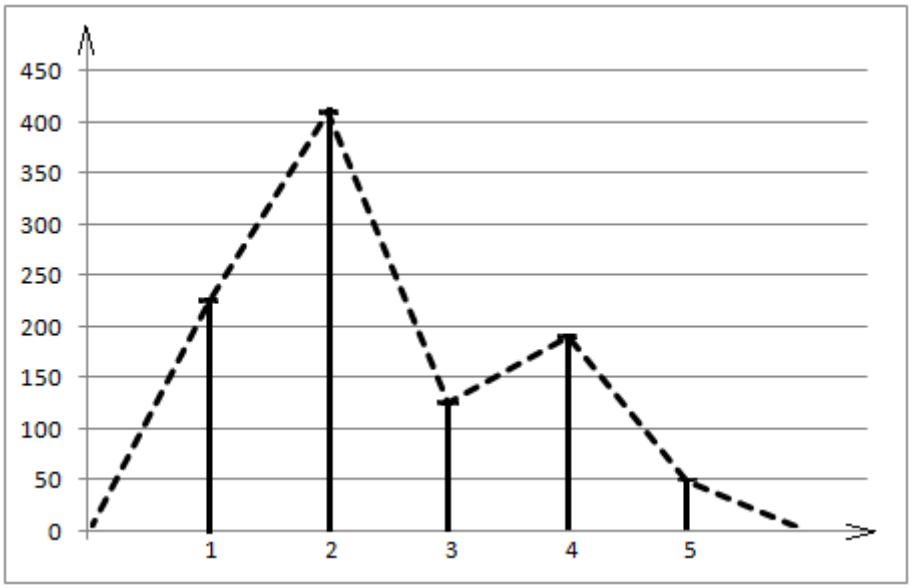
1.3.2. Cas d'un caractère quantitatif.

1.3.2.1. Caractère quantitatif discret. Les représentations les plus utilisées sont:

a- Diagramme en bâtons. Il s'agit de la figure obtenu sur un repère cartésien en associant à chaque point de coordonnées $(x_i, 0)$ un segment vertical appelé bâton, parallèle à l'axe des ordonnées dont la longueur est proportionnelle à la fréquence (f_i) ou à l'effectif (n_i). Cette

représentation permet de donner une idée générale sur la forme de la distribution et permet aussi de repérer les valeurs aberrantes.

b- Polygone des fréquences ou des effectifs. Il s'agit de la ligne brisée joignant les sommets des bâtons du diagramme précédent.



c- Fonction de répartition :

La fonction de répartition **F(X),** est la fonction qui à chaque valeur **x** de **R** associe la proportion d'individus pour lesquels la valeur de la variable **X** est inférieure ou égale à **x**.

Notation :F(x) = P(X≤ x)

Remarque :

- Si $\boldsymbol{x = x_i}$ alors $\boldsymbol{F(x_i) = f_1 + f_2 + \cdots + f_i = F_i}$
- Si $\boldsymbol{x_i \leq x < x_{i+1}}$ alors $\boldsymbol{F(x) = F_i + 0 = F_i}$

Conclusion : $F(x) = F_i$ pour tout $x$tel que $\boldsymbol{x_i \leq x < x_{i+1}}$

La représentation graphique de **F(x)** est appelée courbe cumulative, c'est une courbe **"en escalier"** dont les paliers sont horizontaux, puisque **F(x)** est constante sur chaque intervalle $\boldsymbol{[x_i , x_{i+1}[}$.

$F(x) = 0$ si $x < x_1$ et $F(x) = 1$ si $x \geq x_k$

$F(-\infty) = 0$ et $F(+\infty) = 1$

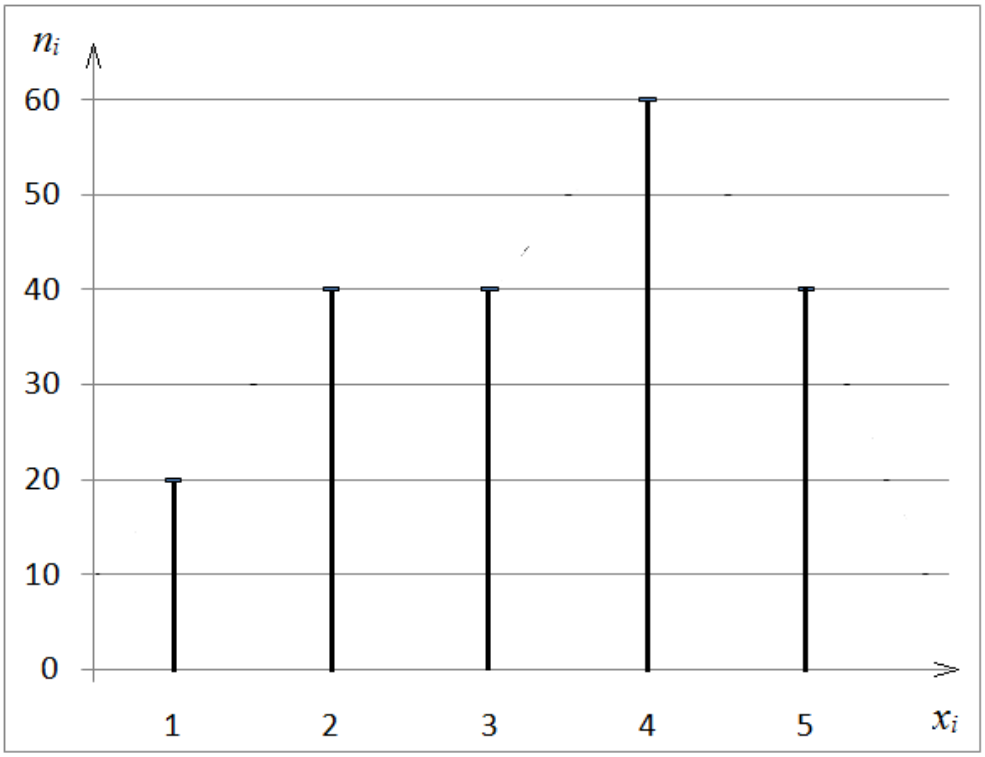
Remarque : On obtient la courbe cumulative des effectifs cumulés en remplaçant les fréquences cumulées $\boldsymbol{F_i}$ par les effectifs cumulés $\boldsymbol{N_i}$.

Exemple: Une enquête portant sur le nombre d'enfants à charge a été réalisée auprès des habitants d'une citée. Cette enquête a donné les résultats suivants :

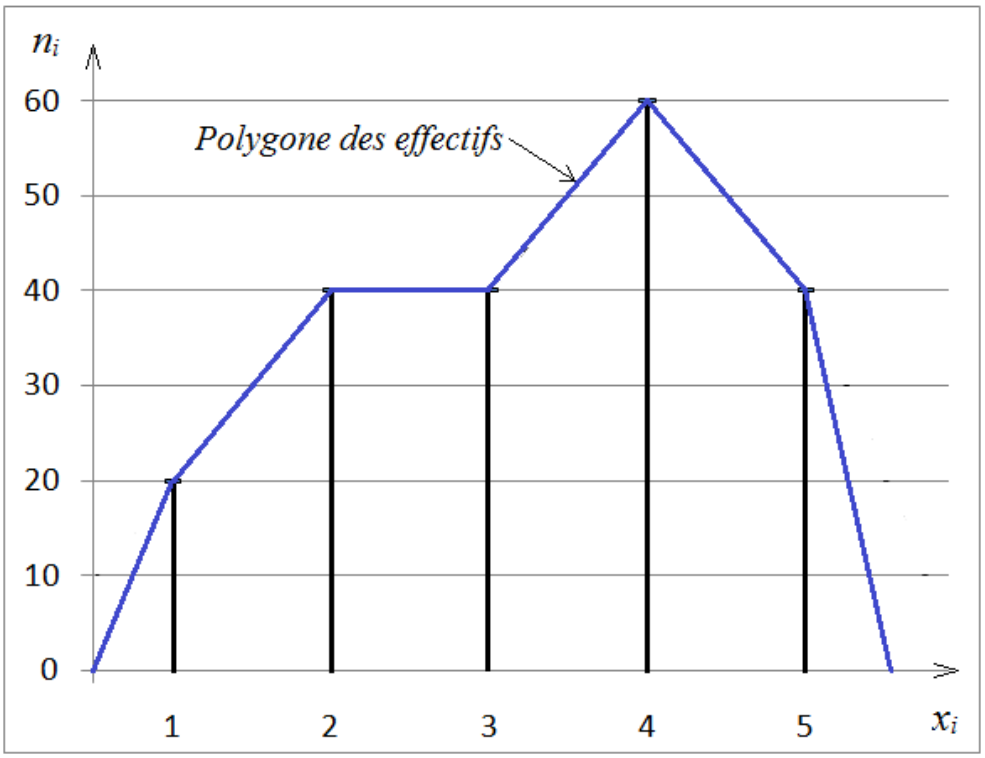
Nombre d'enfants x_i	Nombre de familles n_i
1	20
2	40
3	40
4	60
5	40

1- Représenter graphiquement la distribution.

a) Diagramme en bâtons :

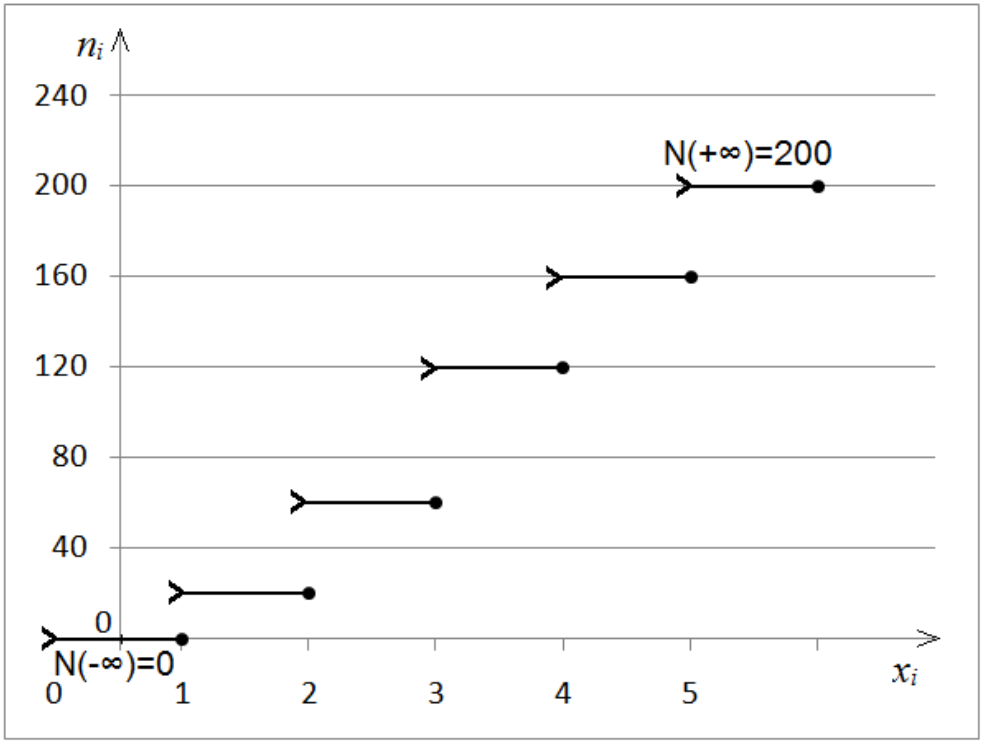


b) Polygone des effectifs (fréquences).



x_i	n_i	N_i	f_i	$F_i \uparrow$	$F_{i\downarrow}$
1	20	0	0,1	0	1
2	40	20	0,2	0,1	0,9
3	40	60	0,2	0,3	0,7
4	60	120	0,3	0,5	0,5
5	40	160	0,2	0,8	0,2
Total	200	200	1,0	1,0	0,0

c) Fonction de répartition :



1.3.2.2. Caractère quantitatif continu.

Dans le cas d'un caractère quantitatif continu, l'établissement du tableau des fréquences implique d'effectuer au préalable une répartition en classes des données. Cela nécessite de définir le nombre de classes attendu et donc l'amplitude associée à chaque classe ou intervalle de classe. En règle générale, on choisit des classes de même amplitude. Pour que la distribution en fréquence ait un sens il faut que chaque classe comprenne un nombre suffisant de valeurs.

Diverses formules empiriques permettent d'établir le nombre de classes pour une population (échantillon) de taille N, principalement :

a) La règle de ***Sturge*** : Nombre de classe $k = 1 + (3.3 \log N)$

b) La règle de ***Yule*** : Nombre de classe $k = 2.5\ (N)^{0.25}$

L'intervalle (l'amplitude) a_i de chaque classe est obtenu de la manière suivante :

$$a_i = \frac{X_{max} - X_{min}}{Nombre\ de\ classes}$$

où $X_{max}\ et\ X_{min}$ désignent la plus grande et la plus petite valeur de la variable dans la série statistique.

1.3.2.3. Représentations graphiques.

Les représentations graphiques appropriées sont :

a- L'histogramme.

L'histogramme des effectifs (fréquences) de la distribution statistique, est formé d'un ensemble de rectangles juxtaposés, où la hauteur du rectangle est proportionnelle à l'effectif n_i (fréquence f_i). Ceci n'est vrai que si les classes ont la même amplitude, en revanche si les classes ont des amplitudes inégales des modifications s'imposent. Pour **conserver la proportionnalité entre la hauteur et l'effectif (fréquence),** sur l'axe des ordonnées, au lieu de porter l'effectif ou la fréquence, on indique le rapport de l'effectif ou de la fréquence sur l'amplitude de classe, d'où l'utilisation des densités d'effectifs ou de fréquences. On définit la densité d'effectif (de fréquence) d'une classe par :

$$\boldsymbol{d_i = \frac{n_i}{a_i}} \ \left(densité d'effectif\right) \quad et \quad \boldsymbol{d_i = \frac{f_i}{a_i}} \ (densité\ de\ fréquence)$$

où $n_i,\ f_i\ et\ a_i$ sont respectivement l'effectif, la fréquence et l'amplitude des classes.

b- Le polygone des effectifs (fréquences)

- On subdivise l'histogramme en sous rectangles de même base égale à l'amplitude de référence notée $\boldsymbol{a_r}$. L'amplitude de référence étant choisie comme la **plus petite des amplitudes vérifiant l'égalité :** $a_i = ka_r$ où k désigne le nombre de classe.

Après avoir ajouté aux extrémités deux rectangles fictifs de hauteur nulle et de base a_r, on joint par des segments de droites, les milieux des bases supérieurs des rectangles de l'histogramme.

c- Courbe cumulative.

On construit la courbe des fréquences cumulées en joignant les points (e_i, F_i), où e_i est la borne supérieure de la $i^{ème}$ classe $[e_{i-1}, e_i[$ et F_i est la fréquence cumulée de cette même classe. On note $\boldsymbol{F_i = P(X \leq e_i)}$

Exemple : Dans le cadre de l'étude de la population gélinottes huppées, les valeurs de la longueur de la rectrice principale peuvent être réparties de la façon suivante :

158, 152, 171, 163, **140**, 157, 162, 171, 158, 164, 163, 159, 153,
158, 152, 165, 156, 162, 150, 154, 155, 162, 155, 164, 164, 157,
159, 153, 163, 158, **174**, 162, 156, 151, 160, 158, 162, 166, 162,
153, 165, 158, 150, 160, 160, 149, 159, 158, 164, 158.

- Détermination du nombre de classes :

Règle de ***Sturge*** : $k = 1 + (3.3 \log N) = 1 + (3.3 \log 50) = 6.60$

Règle de ***Yule*** : $k = 2.5\ (N)^{0.25} = 2.5\ (50)^{0.25} = 6.64$

Les valeurs sont peu différentes, la valeur choisie est celle calculée par la règle de *Sturge*.

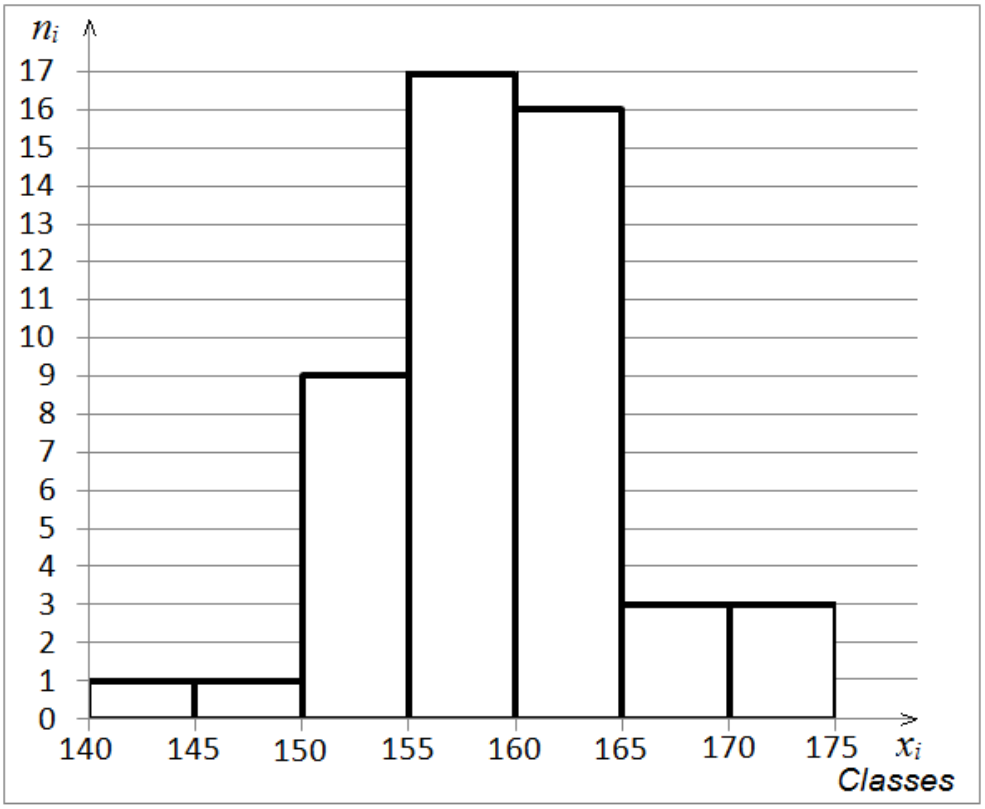
- Détermination de l'amplitude de classe a_i

$a_i = \frac{Etendue}{Nombre\ de\ classes} = \frac{174-140}{6{,}60} = 5{,}15$ mm que l'on arrondi, par commodité, à 5 mm.

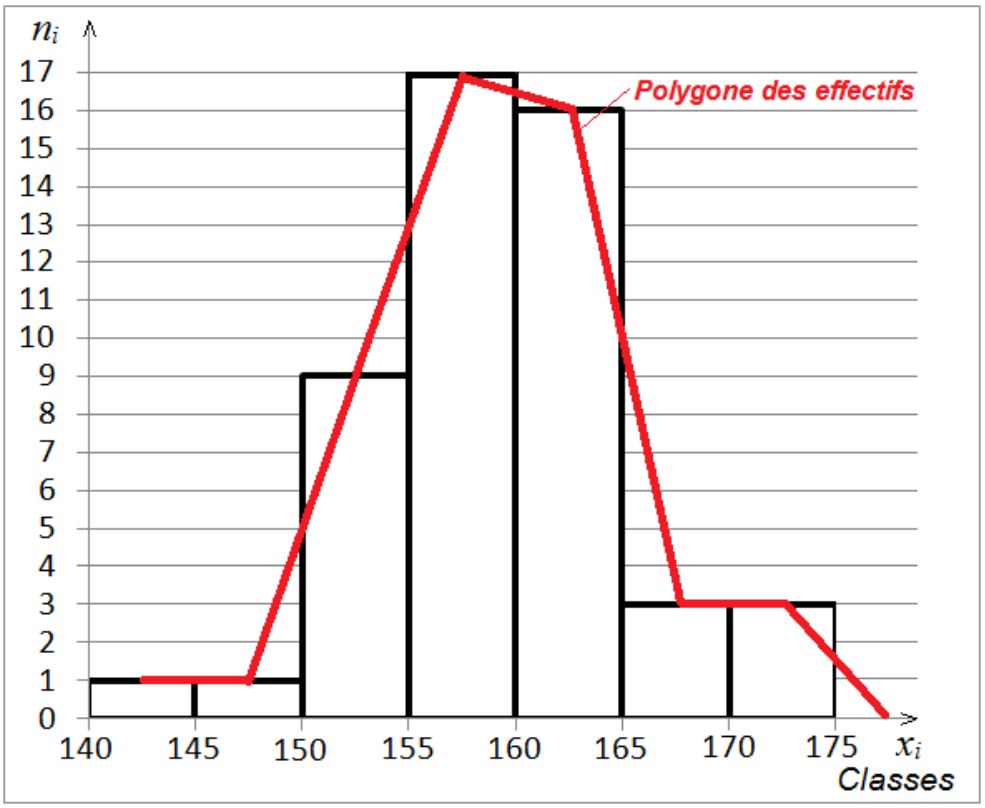
Classes	c_i	n_i	$f_i\,\%$	$F_{i\uparrow}\,\%$
[140 ; 145[	142.5	1	2	2
[145 ; 150[	147.5	1	2	4
[150 ; 155[	152.5	9	18	22
[155 ; 160[	157.5	17	34	56
[160 ; 165[	162.5	16	32	88
[165 ; 170[	167.5	3	6	94
[170 ; 175[	172.5	3	6	100

Représentations graphiques

a) Histogramme.



b) Le polygone des effectifs



1.4. Paramètres caractéristiques des distributions statistiques

1.4.1. Paramètres de position (tendance centrale).

On a vu dans la première partie comment condenser les informations pour les rendre plus lisibles et utilisables. On est ainsi passé d'une liste de plusieurs dizaines, centaines, éventuellement milliers de données à un tableau ou un graphique reposant sur un regroupement de celles-ci en quelques classes.

On souhaite maintenant synthétiser d'avantage l'information pour les caractères quantitatifs en mettant en évidence des nombres permettant de décrire au mieux la population observée.

La première idée concerne naturellement la " **tendance centrale**" de la population. Cela peut signifier, prendre la **classe** de plus grand effectif, trouver un **nombre** séparant la population en deux parties contenant chacune 50 % de l'effectif total ou calculer une **valeur moyenne**. Ces trois grandeurs présentent un grand intérêt et conduisent à définir les ***caractéristiques de position*** utilisées en statistique.

1.4.1.1. Le mode M_O (dominante).

On appelle *mode* d'une série statistique la ou les valeurs du caractère dont l'effectif est le plus élevé. Dans le cas d'une répartition en classes nous ne pouvons pas parler de mode, car

les modalités sont des classes et non des valeurs. La notion de mode correspond à une idée d'intensité plutôt qu'a une idée d'effectif. C'est vrai que dans le cas d'une variable discrète il y a superposition des deux notions, mais pour une variable continue l'amplitude entre en jeu. Les classes peuvent être d'amplitudes inégales et une classe qui a le plus grand effectif n'est pas nécessairement la classe où le caractère est le plus intense. La classe modale est définie comme suit : **C'est la classe qui correspond au plus grand rapport** $\frac{n_i}{a_i}$ (densité d'effectif) où $\frac{f_i}{a_i}$ (densité de fréquence)**.**

Propriétés :

- Le mode correspond à un sommet sur l'histogramme ou sur le diagramme en bâtons.
- Le mode, s'il existe n'est pas unique.
- Le mode est une caractéristique peut utiliser en pratique car il ne fait pas intervenir l'ensemble des valeurs.

Exemple : Dans le but de déterminer le temps de réaction (en secondes) de l'être humain au son, 50 personnes ont été soumise à l'expérience suivante :
On a enregistré le temps mis pour réagir après avoir entendu un signal sonore. Les résultats de l'expérience sont enregistrés dans le tableau ci-dessous :

Classes	n_i	f_i	$\frac{n_i}{a_i}$	$\frac{f_i}{a_i}$
[0.45 ; 0.51 [	2	0,04	33.33	0,66
[0.51 ; 0.57 [	8	0,16	133.33	2,66
[0.57 ; 0.63 [	18	0,36	300	6
[0.63 ; 0.69 [	16	0,32	266.66	5,55
[0.69 ; 0.75 [	4	0,08	66.66	1,33
[0.75 ; 0.81 [	2	0,04	33.33	0,66

La classe modale est [0.57 ; 0.63 [correspond à l'effectif maximum.

Remarque : Il faut faire attention, dans cet exemple, le plus grand effectif n_i coïncide avec la plus grande densité d'effectif $\frac{n_i}{a_i}$ ce qui n'est pas toujours le cas.

Lorsque les variables sont groupées par classes il est parfois utile de remplacer la notion de mode par la notion de classe modale, pour cela on effectue une interpolation linéaire à l'intérieur de cette classe, la détermination se fait de la façon suivante :

Identification de la classe à laquelle appartient le mode, soit $[b_{i-1} ; b_i [$ la dite classe. Pour obtenir la valeur exacte du mode en utilise l'expression suivante :

$$M_o = b_{i-1} + a_i \left[\frac{\Delta_1}{\Delta_1+\Delta_2}\right] \text{ où,}$$

b_{i-1} : Limite inférieur de la classe modale,

Δ_1 : Ecart des effectifs entre la classe modale et la classe **précédente**,

Δ_2 : Ecart des effectifs entre la classe modale et la classe **suivante**,

a_i : Amplitude de la classe modale.

Remarque :

1- Dans le cas où les classes de la série statistique ont des amplitudes différentes, l'utilisation des effectifs pour calculer le mode n'est pas toujours fiable, alors il faut utiliser la notion de densité des classes $d_i = \frac{n_i}{a_i}$ où $\frac{f_i}{a_i}$, dans ce cas la classe modale sera la classe ou la densité est maximale. Dans l'expression précédente du calcul du mode, il faut porter les modifications suivantes :

Δ_1 : Différence de **densité** entre la classe modale et la classe **précédente**,

Δ_2 : Différence de **densité** entre la classe modale et la classe **suivante**.

2- Lorsque les classes adjacentes à la classe modale ont des densités de fréquences égales, le mode coïncide avec le centre de la classe modale, nous signalons ici que le mode dépend beaucoup de la répartition en classes.

3- Une variable statistique peut présenter plusieurs modes locaux : on dit alors qu'elle est plurimodale. Cette situation est importante car elle met en évidence l'existence de plusieurs sous populations, donc l'hétérogénéité de la population étudiée.

a- Cas d'une variable quantitative discrète (non classée)

Par définition, le mode correspond à la valeur de la variable pour laquelle l'effectif (fréquence) est le plus élevé.

Exemple : Recensement de 12398 familles dans une population dont le nombre d'enfants de moins de 10 ans est le suivant :

Nombre d'enfants	0	1	2	3	4
Nombre de familles	2601	6290	2521	849	137

L'effectif le plus élevé est de 6290, alors le mode $M_O = 1$ enfant par famille.

b- Cas d'une variable quantitative continue ou discrète classée

La classe modale est la classe dont la fréquence par **unité d'amplitude** est la plus élevée ; cette classe correspond donc au rectangle le plus haut de l'histogramme des fréquences. Lorsqu'on veut être plus précis, on peut déterminer à l'intérieur de la classe modale la valeur exacte du mode.

Exemple 1 : Soit la distribution suivante.

Classes	c_i	n_i	$N_i \uparrow$	f_i	$F_i \uparrow$
[140 ; 145[	142.5	1	1	0.02	0.02
[145 ; 150[	147.5	1	2	0.02	0.04
[145 ; 155[	152.5	9	11	0.18	0.22
[155 ; 160[	**157.5**	**17**	**28**	**0.34**	**0.56**
[160 ; 165[	162.5	16	44	0.32	0.88
[165 ; 170[	167.5	3	47	0.06	0.94
[170 ; 175[	172.5	3	50	0.06	1.00
		50		1.00	

Les classes étant toutes de même amplitude (égale à 5), les hauteurs des rectangles de l'histogramme des effectifs sont donc égales aux effectifs.

La classe modale correspond à la classe [155 ; 160[dont l'effectif est le plus élevé.

b_{i-1} : Limite inférieure de la classe modale =155,

Δ_1 : Ecart des effectifs entre la classe modale et la classe précédente = (17- 9) = 8,

Δ_2 : Ecart des effectifs entre la classe modale et la classe suivante = (17-16) = 1

a_i : Amplitude de la classe modale = 5.

d'où le mode vaut : **159.44**

Exemple 2: Lors d'une étude concernant la résistance d'un métal, on a réalisé 100 expériences de rupture en charge d'un fil de même épaisseur et l'on a noté les poids limites dans chaque cas. Le tableau ci-dessous représente la répartition par classes des résultats.

Classes	n_i	c_i	$n_i c_i$	N_i	a_i	$di = \frac{ni}{a_i}$
[700 ; 750[	10	725	7250	10	50	0,2
[750 ; 800[	23	775	17825	33	50	0,46
[800 ; 840[	4	820	3280	37	40	0,1
[840 ; 880[	15	860	12900	52	40	0,375
[880 ; 920[	**32**	**900**	**28800**	**84**	**40**	**0,8**
[920 ; 960[	16	940	15040	100	40	0,4
Total	100		85095			

Du fait que les classes ont des amplitudes inégales, ceci nous conduit à utiliser la notion de densité d'effectif pour déterminer la classe modale et par suite déterminer la valeur du mode. La classe modale étant la classe [880 ; 920[elle correspond à la densité d'effectif la plus élevée (0,8).

b_{i-1} : Limite inférieur de la classe modale = 880,

Δ_1 : Ecart des effectifs entre la classe modale et la classe précédente $(32 - 15) = 17$,

Δ_2 : Ecart des effectifs entre la classe modale et la classe suivante $(32 - 16) = 16$,

a_i : Amplitude de la classe modale = 40.

d'où le mode vaut : **900,60**

1.4.1.2. La médiane *M*é.

La médiane est la valeur de la variable située au **milieu** d'une série statistique **ordonnée,** par valeurs croissantes ou décroissante, telle que la moitié des individus prennent une valeur qui lui soit inférieure et l'autre moitié prenant une valeur qui lui soit supérieure. On note que la médiane ne dépend que de l'ordre des modalités et par conséquent elle n'est pas influencée par les observations aberrantes.

Comme pour le mode, la détermination de la médiane dépend de la nature de la variable et ne peut être **calculée** que pour les caractères **quantitatifs.**

1- Cas d'une variable discrète

La détermination du rang de la médiane nécessite, avant tout, de ranger par ordre croissant ou décroissant les valeurs observées, il y a deux cas :

a) La série comporte un nombre **impair** de valeurs, soit N valeurs, la médiane sera la valeur de rang$\left(\frac{N+1}{2}\right)$.

b) La série comporte un nombre **pair** de valeurs, on parle d'intervalle médian dont les bornes sont définies par la $\left(\frac{N}{2}\right)$ i^{ème} valeur, et la $\left(\frac{N}{2}+1\right)$ i ^{ème} valeur.

Remarque:

1) Toute valeur appartenant à cet intervalle peut être considérée comme médiane.
2) Certains auteurs proposent de prendre comme médiane le centre de l'intervalle médian.
3) La médiane n'est pas forcément une valeur observée.

Exemple1: On considère la répartition de 9 familles selon le nombre d'enfants :

Nombres d'enfants par famille	0	0	1	1	2	3	3	3	4
Rang (ordre croissant)	1	2	3	4	5ème	6	7	8	9
	4 Observations				$Mé$	4 Observations			

La médiane dans ce cas correspond à la cinquième valeur c'est-à-dire la médiane = 2 enfants par famille. On dit qu'il y a autant de familles qui ont moins de deux enfants que de familles qui ont plus de deux enfants.

Exemple 2: On considère la répartition de 10 familles selon le nombre d'enfants :

Nombres d'enfants par famille	0	0	1	1	2	3	3	3	4	4
Rang (ordre croissant)	1	2	3	4	5ème	6ème	7	8	9	10
	4 Observations				Intervalle Médian		4 Observations			

Dans ce cas, on parle plutôt d'intervalle médian [2 ; 3[correspondant à la 5ème et la 6ème valeur ; si on considère que la médiane correspond au centre de cet intervalle alors elle est égale à la moyenne arithmétique $\frac{2+3}{2} = 2.5$, qui n'est pas une valeur observée.

Exemple 3: Une étude effectuée par un chercheur à propos du nombre de forages pétroliers en Afrique a conduit à la distribution suivante :

Nombres de forages x_i	0	1	2	3	4	5	6	7
Nombre de pays n_i	10	13	7	7	5	3	3	1

Détermination de la médiane et du mode de cette distribution.

- La médiane est la valeur de la variable qui partage la série ordonnée en deux parties de même effectif, d'où $\boldsymbol{M}$é **= 2**
- Le mode est la valeur de la variable qui se répète le plus, donc $\boldsymbol{M_o}$ **= 1**

2- Cas d'une variable continue

Dans ce cas, la méthode de détermination de la médiane est différente de la méthode du cas précédent ; on commence tout d'abord de repérer la classe qui contient la moitié de l'effectif $(N/2)$, celle ci représente la classe médiane. Cette classe peut être également repérée sur le diagramme des effectifs (fréquences) cumulés croissants. Ensuite, par interpolation linéaire à l'intérieur de l'intervalle médian, on peut calculer la valeur exacte de la médiane en utilisant l'expression suivante :

$$M\text{é} = b_{i-1} + a_i \left[\frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}}\right] \text{ où,}$$

- $[b_{i-1}, b_i[$: intervalle de la classe médiane,
- a_i : amplitude de la classe médiane,
- N_i : effectif cumulé croissant de la classe médiane,
- N : effectif total.
- N_{i-1}: effectif cumulé croissant de la classe juste avant la classe médiane.

La même démarche pourrait être utilisée en remplaçant les fréquences absolues par les fréquences relatives

$$M\text{é} = b_{i-1} + a_i \left[\frac{0.5 - Fi-1}{Fi - F_{i-1}}\right] \text{ où,}$$

$-$ F_i: Fréquence cumulée croissante de la classe médiane,

$-$ F_{i-1} : Fréquence cumulée croissante de la classe qui précède juste la classe médiane.

Exemple 1: Dans une région, on a relevé la superficie de 200 exploitations agricoles. Les résultats obtenus sont distribués comme suit :

Classes	n_i	N_i	f_i	$F_{i\uparrow}$
[0 ; 10[	30	30	0.15	0.15
[10 ; 30[	80	110	0.40	0.55
[30 ; 50[	60	170	0.30	0.85
[50 ; 100[	20	190	0.10	0.95
[100 ; 200[	10	200	0.05	1.00
Total	200			

Déterminer la médiane par interpolation linéaire.

Méthode *:* Sur la colonne des effectifs cumulés croissants, on remarque que la moitié de l'effectif de l'échantillon examiné, appartient à l'intervalle [10 ; 30[, alors la médiane est incluse dans cette intervalle, d'où :

$$Mé = b_{i-1} + a_i \left[\frac{\frac{n}{2} - N_{i-1}}{N_i - N_{i-1}}\right] =$$

$$= 10 + 20 \left[\frac{200/2 - 30}{110 - 30}\right] = 27.5 \text{ ha (utilisation des effectifs cumulés).}$$

ou bien,

$$Mé = b_{i-1} + a_i \left[\frac{0.5 - F_{i-1}}{Fi - F_{i-1}}\right] =$$

$$= 10 + 20 \left[\frac{0.5 - 0.15}{0.55 - 0.15}\right] = 27.5 \text{ ha (utilisation des fréquences cumulées).}$$

Exemple 2 : Soit la répartition de 100 individus selon l'âge :

Classes	n_i	N_i	f_i	$F_{i\uparrow}$
[05 ; 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
Total	100		1.00	

$$Mé = b_{i-1} + a_i \left[\frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}}\right] =$$

$= 20 + 10 \left[\frac{50-36}{56-36}\right] = 27$ ans (en utilisant les effectifs cumulés)

$$Mé = b_{i-1} + a_i \left[\frac{0.5 - F_{i-1}}{Fi - F_{i-1}}\right] =$$

$= 20 + 10 \left[\frac{0.5-0.36}{0.56-0.36}\right] = 27$ ans (en utilisant les fréquences cumulées).

1.4.1.3. La moyenne

a) La moyenne arithmétique.

La moyenne arithmétique, notée $\bar{x}$, est la mesure la plus commune de tendance centrale, elle se définit comme la somme des valeurs divisée par le nombre de valeurs.

La moyenne arithmétique est un indicateur de tendance centrale, fondé sur les valeurs, elle est la valeur unique que devraient avoir toutes les unités statistiques, considérées comme identiques, d'une population ou d'un échantillon pour que leur total demeure inchangé. En plus de cette moyenne il existe d'autres types de moyennes en particulier la moyenne géométrique, la moyenne quadratique et la moyenne harmonique.

1- Cas d'une variable discrète :

Exemple : Soit la série suivante concernant la détermination de l'âge moyen des élèves d'une classe d'examen d'un lycée.

18 ; 20 ; 17 ; 17 ; 17 ; 16 ; 20 ; 18 ; 18

18 ; 19 ; 19 ; 19 ; 18 ; 18 ; 18 ; 19 ; 18

18 ; 18 ; 17 ; 16 ; 17 ; 17 ; 20 ; 17 ; 16

La moyenne arithmétique est donnée par l'expression :

$$\bar{x} = \sum_{i=1}^{27} x_i = \frac{18+20+17+.....+17+16}{27} = 17{,}89$$

Si on avait pris comme échantillon la totalité des élèves des classes d'examen de ce lycée, le calcul aurait été très long et on aurait commis des erreurs. C'est pour cette raison qu'on calcul généralement la moyenne **pondérée** où chaque valeur de la variable est multipliée par l'effectif correspondant. Soit la distribution suivante :

Rang	x_i	n_i	f_i *en %*
1	16	3	11.11
2	17	7	25.92
3	18	10	37.03
4	19	4	14.81
5	20	3	11.11
Total		27	

La moyenne pondérée est donnée par l'expression :

$$\bar{x} = \sum_{i=1}^{5} \frac{nix_i}{N} = \frac{3\times16+7\times17+10\times18+4\times19+3\times20}{27} = \frac{483}{27} = 17{,}89$$

Elle est obtenue en multipliant chaque valeur de la variable par l'effectif correspondant. Le nombre obtenu est divisé par l'effectif total.

On peut aussi utiliser les fréquences pour calculer la moyenne, sachant que la fréquence f_i correspondant à la valeur x_i est égale à $\frac{ni}{N} = \frac{Effectifpartiel}{Effectiftotal}$, d'où la moyenne est donnée par l'expression $\bar{x} = \sum_{i=1}^{5} f_i x_i$

$$\bar{x} = 0.11\times16 + 0.26\times17 + 0.37\times18 + 0.15\times 19 + 0.11\times20 = 17.89$$

2- Cas d'une variable continue répartie en classes.

Dans cette situation, pour calculer la moyenne, on suppose que la répartition des valeurs de la variable est uniforme dans chaque classe. Ceci est équivalent de dire qu'il y a **concentration de toutes les modalités d'une classe à son centre.**

Exemple : Une enquête sur la répartition d'une population agricole, selon l'âge, a donné les résultats suivants :

Classes	n_i	f_i	c_i
[15 ; 25[	197	0.197	20
[25 ; 35[	207	0.207	30
[35 ; 45[	151	0.151	40
[45 ; 55[	189	0.189	50
[55 ; 65[	127	0.127	60
[65 ; 75[	108	0.108	70
[75 ; 85[	21	0.021	80
Total	1000	1.000	

Les valeurs de la variable sont représentées par les centres des classes c_i d'où les expressions permettant de calculer la moyenne sont:

$$\bar{x} = \sum_{i=1}^{7} \frac{nic_i}{N} = \frac{197\times20+207\times30+151\times40+189\times50+127\times60+108\times70+21\times80}{1000} = 42.5 \text{ ans}$$

$$\bar{x} = \sum_{i=1}^{7} f_i x_i = 0{,}197 \times 20 + 0{,}207 \times 30 + 0{,}151 \times 40 + 0{,}189 \times 50 +$$
$$+ 0{,}127 \times 60 + 0{,}108 \times 70 + 0{,}021 \times 80 = 42.5 \text{ ans.}$$

Remarque : On note ici qu'il est possible d'utiliser une autre expression du calcul de la moyenne, qui permet de réduire d'une manière considérable les calculs, appelée méthode de changement de variable.

b) Nous pouvons à titre d'information citer d'autres types de moyenne **(géométrique** G **, harmonique** H**, quadratique** Q**).**

Résultat comparatif :

Pour une même série statistique, on montre que les quatre moyennes vérifient toujours l'ordre suivant : $H < G < \bar{x} < Q$

Conclusion :

1- Un inconvénient de la moyenne arithmétique est qu'elle est très sensible aux valeurs extrêmes de la série.

2- La moyenne géométrique est peu sensible aux valeurs extrêmes de la série.

3- La moyenne harmonique est plus sensible aux plus petites valeurs de la série qu'aux plus grandes.

1.4.2. Paramètres de dispersion :

1.4.2.1. L'étendue: L'étendue, notée E, est la différence entre valeur maximale et la valeur minimale de la série étudiée : $E = V_{max} - V_{min}$

Ce paramètre est souvent utilisé dans les contrôles de fabrication, pour lesquels on donne, à priori des marges de construction. Son intérêt est limité par le fait qu'il dépend uniquement des valeurs extrêmes de la distribution qui peuvent être des valeurs aberrantes.

Remarque : L'étendue est exprimée avec la même unité que celle de la variable, elle n'est pas très informative, car elle ne tient pas compte de la répartition des données dans le segment $[V_{max}\ ;\ V_{min}]$.

1.4.2.2. Quartiles : Le quartile est une extension de la médiane puisqu'il s'agit de partager l'effectif en quatre parties égales. Nous distinguons les cas suivants :

- **Cas d'une variable statistique continue :** On appelle quartiles les nombres réels Q_1, Q_2 et Q_3 pour lesquels les fréquences cumulées de la variable sont respectivement 0.25, 0.5 et 0.75. Les quartiles partagent la série ordonnée en 4 parties de même effectif, le 2ème quartile Q_2 coïncide avec la médiane.

L'intervalle interquartile est l'intervalle $\mathbf{[}Q_1\mathbf{;}\ Q_3\mathbf{]}$ qui contient 50 % des valeurs de la série. L'écart interquartile est la différence entre le 3ème quartile et le 1er quartile qui vaut $\mathbf{[}Q_3 - Q_1\mathbf{]}$.

- **Cas d'une variable statistique discrète :** Nous savons que la courbe des fréquences cumulées est une courbe en escalier, s'il existe une valeur de la variable pour laquelle la fréquence cumulée est égale à 0.25, 0.5 et 0.75 alors la quantité correspondante est égale à cette valeur de la variable, sinon les quantités seront déterminées par interpolation linéaire.

Remarque : L'intervalle interquartile possède une unité de mesure qui est celle de la variable étudiée, il est plus précis que la médiane, car il ne tient compte que des valeurs proches de celle-ci.

1.4.2.3. Variance et écart type.

a- Variance $Var(X)$: La variance est un indicateur de dispersion d'une série statistique par rapport à sa moyenne. Par définition, la variance d'une série est la moyenne des carrés des écarts des valeurs de cette variable à sa moyenne arithmétique définie par:

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

Toute fois ce calcul n'est pas très commode car le calcul nécessite n opérations de soustractions, n mises au carré, n multiplications et $(n-1)$ additions. L'expression simplifiée de la variance dans laquelle la somme ne nécessite plus de soustraction, peut être mise sous la forme :

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

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

b- Ecart type $\sigma(X)$: L'écart type d'une série statistique est la racine carrée de la variance notée $\sigma(X)$exprimé par l'expression$\sigma(X) = \sqrt{Var(X)}$.

1.4.3. Coefficient de variation:

Le coefficient de variation noté *CV* est une grandeur sans dimension, égale au rapport de l'écart type à la moyenne, c'est un indice de dispersion relatif exprimé en pourcentage, il est indépendant du choix des unités de mesure et de l'ordre de grandeurs des valeurs observées, il permet d'apprécier la représentativité de la moyenne par rapport à l'ensemble des observations, il permet aussi la comparaison des distributions de valeurs dont les échelles de mesure ne sont pas comparables, il est donné par l'expression : $C_v = \frac{\sigma(X)}{\bar{x}}$

Remarque : Plus la valeur du coefficient de variation est élevée, plus la dispersion autour de la moyenne est grande.

1.4.4. Représentation graphique par le diagramme de Tukey (boite à moustaches).

La boite à moustaches, ou box plot en Anglais, est un graphique mettant en exergue certains paramètres de position et de dispersion. Il permet de visualiser la répartition des données de la série statistique et rend compte aussi du niveau d'asymétrie de la dispersion des valeurs extrêmes de la distribution.

Eléments constitutifs d'une boite à moustaches :

a) Un rectangle de bornes Q_1 et Q_3 coupé au niveau de la médiane par un trait plein et au niveau de la moyenne par un trait en pointillés.

b) Des moustaches : Ce sont les extrémités gauche et droite du rectangle (parfois appelés pattes).

- La moustache gauche définie par la valeur de la série immédiatement supérieure à la quantité:$\boldsymbol{Q_1 - 1.5\ (Q_3 - Q_1)}$. S'il ya des valeurs inférieures à cette quantité dans la série, elles sont dites atypiques et sont représentées par des marqueurs.
- La moustache droite définie par la valeur de la série immédiatement inférieure à la quantité:$\boldsymbol{Q_3 + 1.5\ (Q_3 - Q_1)}$. S'il y a des valeurs supérieures à cette quantité dans la série, elles sont dites atypiques et sont représentées par des marqueurs.

Remarque :

- En général pour la représentation d'une série statistique par une boite à moustache (Diagramme de **TUKEY**), il est nécessaire de disposer au minimum de cinq paramètres caractéristiques.
- La valeur 1.5 est selon TUKEY est une grandeur pragmatique, qui a une raison probabiliste.
- Dans la boite à moustaches définie par TUKEY, celle-ci a pour hauteur la distance interquartile $\boldsymbol{(Q_3 - Q_1)}$ et les moustaches sont basées généralement sur 1.5 fois la hauteur de la boite. Dans ce cas, une valeur est atypique si elle dépasse de 1.5 fois l'écart interquartile au-dessus du 1er quartile ou au-dessous du 3ème quartile.
- Il est à signaler que la médiane et l'écart interquartile ne sont jamais influencés par les valeurs extrêmes.

<u>Exemple</u> :

Soit la série suivante représentant le nombre d'absences d'un étudiant aux différentes séances de travaux pratiques :

$\{4, 0, 1, 1, 2, 2, 2, 3, 3, 4, 2, 3, 4, 5, 2, 1, 3, 3, 4, 5\}$. Série brute.

$\{0,1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5\}$. Série ordonnée.

Paramètres nécessaires pour la construction d'une boite à moustaches.

1- La moyenne arithmétique (moyenne pondérée):

$$\bar{x} = \sum_{i=1}^{20} \frac{nixi}{N} = \frac{1\times0+3\times1+5\times2+5\times3+4\times4+2\times5}{20} = 2.7 \text{ absences}$$

2- $V_{min} = 0$ et $V_{max} = 5$

3- Le premier quartile $Q_1 = 2$

4- Le deuxième quartile (médiane) $Q_2 = 3$

5- Le troisième quartile $Q_3 = 4$

6- Valeurs atypiques : $Q_1 - 1.5\ (Q_3 - Q_1) = 2 - 1{,}5(4-2) = -1$

$Q_3 + 1.5\ (Q_3 - Q_1) = 4 + 1{,}5(4-2) = 7$

Comme : $-1 < 0 \quad et \quad 7 > 5$ alors il n ya pas de valeurs atypiques.