Chapitre 4

Statistique descriptive

4.1 Généralités

4.1.1 Population

Toute étude statistique concerne un ensemble Ω appelé population dont les éléments sont appelés des individus.

Définition 4.1.1
Une population c'est l'ensemble d'individus ou d'objets qui possèdent un ou plusieurs caractères spécifiques en commun.
Une population statistique est dite finie si l'on peut déterminer avec précision le nombre d'individus qui la composent sinon elle est dite infinie.

Exemple 31
• *Dans une étude sur le sport, la population peut être l'ensemble des personnes qui pratiquent un sport.*
• *Dans une étude sur les revenus mensuels dans une entreprise, la population peut être l'ensemble des personnes qui travaillent dans cette entreprise.*

4.1.2 Echantillon

Pour obtenir un renseignement exact concernant une variable X, il faut étudier tous les individus de la population. Quand cela n'est pas possible, on restreint l'étude à une partie de la population appelée échantillon.

Définition 4.1.2
Un échantillon est une partie finie représentative de la population c'est donc un sous ensemble E de Ω.

4.1.3 Variables

L'étude statistique consiste en l'analyse d'une variable X appelé parfois caractère qui sert à décrire l'aspect d'une population objet de l'étude. On distingue deux types de variables : qualitatives et quantitatives.

Définition 4.1.3
• *Une variable X est dite qualitative si les valeurs prises sont des mots ou des lettres.*
• *Une variable X est dite quantitative si les valeurs prises sont des nombres réels.*

Exemple 32
• *La couleur des cheveux, état du temps constaté à Alhoceima pendant les six premiers mois de l'année* 2020 *(pluvieux, orageux, beau, venteux, brouillard, ...), mode de transport pour se rendre à l' ENSAH (voiture, taxi, moto, bicyclette, à pied) définissent des variables qualitatives.*
• *La taille, le poids, le salaire, l'âge, les notes sur* 20 *obtenues en statistique par les étudiants de AP*2*, la hauteur des précipitations tombées chaque mois à Al Hoceima sont des variables quantitatives.*

4.2 Statistique

4.2.1 Notions de probabilités et statistiques

La statistique descriptive permet de décrire les données à l'aide de graphiques et de paramètres d'une façon compréhensible et utilisable.
La statistique inférentielle permet de faire des prévisions ou des généralisations à une population à partir d'échantillons.

4.2.2 Effectifs - Fréquences - Fréquences cumulées

L'étude concrète d'une variable X donne N valeurs qui constituent la distribution statistique de X (aussi appelé série statistique).
Cette distribution est, en générale, présentée d'une façon groupée :
• Sous la forme $\{(x_i, n_i)/1 \leq i \leq p\}$ dans le cas d'une variable qualitative ou quantitative discrète (avec $x_1 < x_2 < ... < x_p$ dans le cas d'une variable quantitative discrète).
• Sous la forme d'intervalles ou de classes $\{(]x_i, x_{i+1}], n_i)/1 \leq i \leq p\}$ dans le cas d'une variable quantitative continue .

Définition 4.2.1
l'effectif n_i est le nombre d'individus de la population ou de l'échantillon pour lesquels X prend la valeur x_i (dans le cas d'une variable qualitative ou quantitative discrète) ou une valeur de l'intervalle $]x_i, x_{i+1}]$ (dans le cas d'une variable quantitative continue).
La somme des effectifs est appelée la taille de la population ou de l'échantillon et est notée N.
$N = n_1 + n_2 + ... + n_p$

Définition 4.2.2
On appelle fréquence de la valeur x_i ou de la classe $]x_i, x_{i+1}]$ le nombre réel $f_i = \frac{n_i}{N}$. On a évidemment $\sum_{i=1}^{p} f_i = 1$.
C'est la proportion de l'effectif d'une valeur de la variable par rapport à N la taille totale de la population ou de l'échantillon.

Définition 4.2.3
On appelle fréquence cumulée de la valeur x_i ou de la classe $]x_i, x_{i+1}]$ la somme des fréquences

de cette valeur ou classe et des fréquences des valeurs ou classes qui la précèdent $F_i = \sum_{k=1}^{i} f_k$.
C'est la proportion des unités statistiques de la population ou de l'échantillon qui possèdent une valeur inférieure ou égale à une valeur x *donnée d'une variable quantitative.*

Exemple 33
Variable qualitative : La répartition des adultes d'une résidence selon le niveau d'instruction.

Niveau d'instruction	*Effectifs* n_i	*Frequences* f_i	*Frequences cumulée* F_i
Sans	25	$0,072$	$0,072$
Primaire	36	$0,103$	$0,202$
Secondaire	81	$0,231$	$0,433$
Universitaire	208	$0,594$	$1,027$
Totale	$N = 350$	1	

4.3 Paramètres statistiques

4.3.1 Moments

4.3.1.1 Moments simple

Les moments simples d'ordre p correspondent à une moyenne des puissances p.

Définition 4.3.1
Le moment simple d'ordre p *d'une variable statistique* x *est la moyenne des puissances* p*-ièmes des valeurs observées.*

Si les données sont écrites sous forme exhaustive, la formule mathématique du moment simple d'ordre p est :

$$\boxed{M_p = \frac{1}{N}\sum_{i=1}^{N} x_i^p}$$

Si les données sont regoupées sous forme de tableau d'effectifs de la forme :

Valeurs	v_1	v_2	...	v_k
Effectifs	n_1	n_2	....	n_k

La formule s'écrit :

$$\boxed{M_p = \frac{1}{N}\sum_{i=1}^{k} n_i v_i^p}$$

où $N = n_1 + n_2 + ... + n_k$.
Avec un tableau de fréquences, la formule s'écrit :

$$\boxed{M_p = \sum_{i=1}^{k} f_i v_i^p}$$

Les moments d'ordre p sont exprimés dans l'unité des données élevées à la puissance p : par exemple, si les x sont des quantités en mètres, le moment d'ordre 3 sera en mètres cubes.

4.3.2 Moments centrées

Les moments centrés sont les moments simples appliqués aux écarts par rapport à la moyenne. Autrement dit, on remplace les valeurs x_i par $x_i - \overline{x}$ dans les formules précédentes. Les formules mathématiques sont donc (selon que les données sont exhaustives ou regroupées) :

$$\boxed{\mu_p = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^p}$$

$$\boxed{\mu_p = \frac{1}{N}\sum_{i=1}^{k} n_i(v_i - \overline{x})^p = \sum_{i=1}^{k} f_i(v_i - \overline{x})^p}$$

où $\overline{x} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{i=1}^{k} n_i v_i = \sum_{i=1}^{k} f_i v_i$

• Cas particulier où $p = 1$

On calcule

$$\mu_1 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x}) = 0.$$

Le moment centré d'ordre 1 est toujours nul.

On interprète ce résultat en disant que les écarts à gauche de la moyenne (écarts par défaut) compensent exactement les écarts à droite (écarts par excès).

Exemple 34

On considère les données suivantes concernant une variable discrète V pouvant prendre les valeurs $0, 1, 2, 3$.

Valeurs	0	1	2	3
Effectifs	16	19	28	22

Les moments simples et les moments centrés d'ordres $1, 2$ et 3 sont :

p	*Moments simples*	*Moments centrés*
1	$2,01$	0
2	$5,69$	$1,65$
3	$17,97$	$-0,10$

4.3.3 Le mode

4.3.3.1 Variable qualitative ou quantitative discrète

Définition 4.3.2

Le mode est une valeur de la variable pour laquelle l'effectif ou la fréquence est maximal(e). Le mode est noté m_d.

Une distribution peut être unimodale, bimodale ou multimodale.

Exemple 35
Considérons la distribution des notes d'un groupe d'étudiants.

x_i	8/20	9/20	10/20	11/20	12/20	13/20	14/20
n_i	2	7	12	17	11	6	3

L'effectif maximal est 17*, donc le mode est* $m_d = 11/20$

Exemple 36
Considérons la distribution des couleurs des voitures dans un parking

x_i	*Rouge*	*Blanche*	*Verte*	*Jaune*	*Noire*	*Grise*
n_i	2	7	5	7	5	7

L'effectif maximal est 7*.*
La variable est qualitative. Ici on a trois modes : Blanche, Jaune et Grise. Cette distribution est multimodale.

4.3.3.2 Variable quantitative continue

Dans le cas d'une variable quantitative continue, les données sont regroupées en classes. Si les classes sont toutes de même amplitude, une classe modale est celle dont la fréquence ou l'effectif est le plus élevé.

Exemple 37
Soit la distribution suivante

$[x_i, x_{i+1}[$	$[500, 700[$	$[700, 900[$	$[900, 1100[$	$[1100, 1300[$
f_i	$0,21$	$0,34$	$0,25$	$0,2$

la fréquence maximale est 0.34, donc la classe modale est $[700, 900[$*.*

Remarque 4.3.1
Si les classes ne sont pas de même amplitude, on doit obligatoirement corriger les effectifs et les fréquences (c'est à dire rendre les classes de même amplitude) avant de déterminer le mode m_d*.*

4.3.4 La médiane

La médiane est la valeur m_e de la variable qui partage les éléments de la série statistique, préalablement classés par ordre croissant, en deux groupes d'effectifs égaux : 50% des individus présentent une valeur inférieure ou égale à la médiane et 50% présentent une valeur supérieure ou égale à la médiane.

4.3.4.1 Variable quantitative discrète

Soient $x_1, x_2, ..., x_N$ les valeurs prises par la variable. On les ordonne de la plus petite à la plus grande et on note $x_{(1)}$ la plus petite valeur, $x_{(2)}$ la deuxième valeur, ..., $x_{(i)}$ la i^{eme} valeur, ..., $x_{(N)}$ la plus grande valeur. Alors on a

$$m_e = \begin{cases} x_{(\frac{N+1}{2})} & \text{si } N \text{ impair} \\ \\ \dfrac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} & \text{si } N \text{ pair} \end{cases}$$

Exemple 38
Considérons la distribution suivante

x_i	10	20	30	40	50	60
n_i	4	9	5	8	3	4
effectifs cumulés	4	13	18	26	29	33

On a $N = 33$ *donc* $\dfrac{N+1}{2} = 17$ *et* $m_e = x_{(17)} = 30$ *car le premier effectif cumulé supérieur ou égal à* 17 *est* 18 *et* $x_{(18)} = 30$.

Exemple 39
Considérons la distribution suivante

x_i	10	20	30	40	50	60
n_i	3	8	4	9	3	3
effectifs cumulés	3	11	15	24	27	30

On a $N = 30$ *donc* $\dfrac{N}{2} = 15$ *et* $m_e = \dfrac{x_{(\frac{N}{2})} + x_{(\frac{N}{2}+1)}}{2} = \dfrac{x_{(15)} + x_{(16)}}{2} = \dfrac{30+40}{2} = 35$.
$x_{(16)} = 40$ *car le premier effectif cumulé supérieur ou égal à* 16 *est* 24 *et* $x_{(24)} = 40$.

4.3.4.2 Variable quantitative continue

La médiane est la solution de l'équation $F(x) = 0,5$. Pour la déterminer, on commence par déterminer la classe médiane $]x_i, x_{i+1}]$ qui vérifie

$$F(x_i) < 0,5 \text{ et } F(x_{i+1}) \geq 0,5$$

La médiane m_e (qui appartient à la classe médiane) est ensuite déterminée à partir d'une interpolation linéaire.

Exemple 40
On considère la distribution des salaires mensuels (en milliers de dirhams) du personnel d'une entreprise :

Classe	*Effectifs* n_i	*Fréquence* f_i	*Fréquence cumulée* $F(x_{i+1})$
$]2,3]$	15	$0,19$	$0,19$
$]3,4]$	20	$0,25$	$0,44$
$]4,6]$	20	$0,25$	$0,69$
$]6,10]$	24	$0,31$	1
Total	79	1	

On a $F(4) = 0,44 < 0,5$ *et* $F(6) = 0,64 > 0,5$*, donc la classe médiane est* $]4,6]$.
En considérant les triangles ABD et AIC avec $A(x_A, y_A), B(x_B, y_B), I(x_I, y_I), C(x_I, y_A), D(x_B, y_A)$ *de la figure ci-dessous, on a*

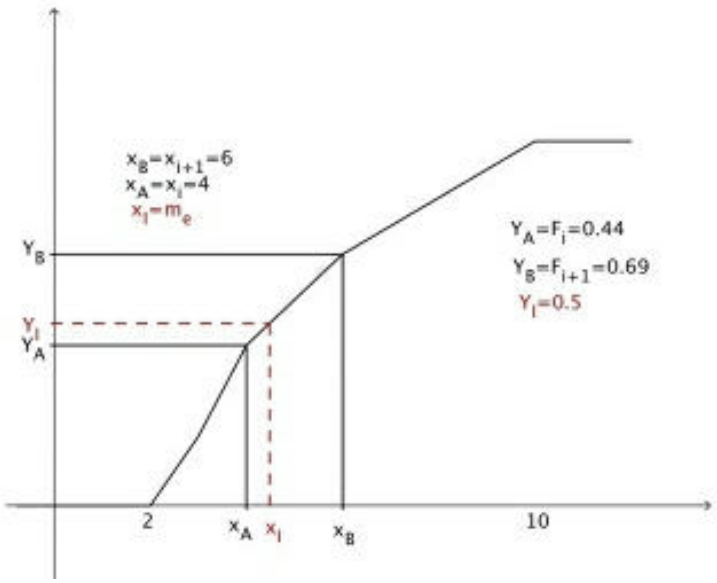

FIGURE 4.1 – Courbe des fréquences cumulées

$$tg(\alpha) = \frac{DB}{DA} = \frac{Y_B - Y_A}{x_B - x_A} = \frac{F(x_{i+1}) - F(x_i)}{x_{i+1} - x_i}$$
$$= \frac{CI}{AC} = \frac{Y_I - Y_A}{x_I - x_A} = \frac{0,5 - F(x_i)}{m_e - x_i}$$

d'où $$m_e = x_i + (x_{i+1} - x_i)\frac{0,5 - F(x_i)}{F(x_{i+1}) - F(x_i)}$$

Application numérique : $x_i = 4; x_{i+1} = 6; F_i = 0,44; F_{i+1} = 0,69$

donc $$\boxed{m_e = 4 + (6-4)\frac{0,5 - 0,44}{0,69 - 0,44} = 4,48}$$

4.3.5 Moyenne et variance empiriques

4.3.5.1 Moyenne empirique :

Définition 4.3.3
La moyenne empirique d'un échantillon est la somme de ses éléments divisée par leur nombre.

Si l'échantillon est noté $(x_1, \ldots, x_n)$, sa moyenne empirique est :

$$\overline{x} = \frac{1}{n}(x_1 + \ldots + x_n)$$

Si on réunit deux échantillons, de tailles respectives n_x et n_y, de moyennes respectives $\overline{x}$ et $\overline{y}$, alors la moyenne du nouvel échantillon sera

$$\frac{n_x\overline{x} + n_y\overline{y}}{n_x + n_y}$$

Si $(x_1, \ldots, x_n)$ est un échantillon et si on pose pour tout $i = 1, \ldots, n$, $y_i = ax_i + b$, où a et b sont deux constantes, alors la moyenne empirique de l'échantillon $(y_1, \ldots, y_n)$ est $\overline{y} = a\overline{x} + b$. En particulier, si $a = 1$ et $b = -\overline{x}$, le nouvel échantillon a une moyenne nulle. Centrer les données c'est leur retrancher la moyenne empirique de manière à la ramener à 0.

4.3.5.2 Variance empirique :

Les notions de variance et d'écart-type servent à quantifier la dispersion d'un échantillon autour de sa moyenne. La définition est la suivante :

Définition 4.3.4
Soit $(x_1, \ldots, x_n)$ un échantillon, et $\overline{x}$ sa moyenne empirique. On appelle variance de l'échantillon, la quantité, notée V, définie par :

$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \ .$$

On appelle écart-type de l'échantillon la racine carrée de la variance :

$$\sigma = \sqrt{V}.$$

L'avantage de l'écart-type sur la variance est qu'il s'exprime, comme la moyenne, dans la même unité que les données.
On utilise parfois le coefficient de variation, qui est le rapport de l'écart-type sur la moyenne :

$$CV = \frac{\sigma}{\overline{x}}$$

Proposition 4.3.1
Soit $(x_1, \ldots, x_n)$ un échantillon numérique. Considérons l'application EQ (erreur quadratique) qui à un nombre m associe :

$$EQ(m) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \ .$$

L'application EQ admet un minimum absolu pour $m = \overline{x}$. La valeur de ce minimum est la variance de l'échantillon.

Démonstration 9 *La fonction $EQ(m)$ est un polynôme de degré deux en m :*

$$EQ(m) = m^2 - 2m\overline{x} + \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

Elle est décroissante, puis croissante, et atteint son minimum au point où la dérivée s'annule, à savoir $m = \overline{x}$.□

Pour le calcul algorithmique, on calcule en général simultanément moyenne et variance grâce à la formule suivante.

Proposition 4.3.2
La variance empirique est :

$$V = \Big(\frac{1}{n}\sum_{i=1}^{n} x_i^2\Big) - \overline{x}^2 \ .$$

Démonstration 10 *Il suffit de développer les carrés dans la définition de V :*

$$\begin{aligned} V &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \Big(\frac{1}{n}\sum_{i=1}^{n} x_i^2\Big) - \frac{1}{n}\sum_{i=1}^{n} 2x_i\overline{x} - \frac{1}{n}\sum_{i=1}^{n}\overline{x}^2 \\ &= \Big(\frac{1}{n}\sum_{i=1}^{n} x_i^2\Big) - 2\overline{x}^2 + \overline{x}^2 \\ &= \Big(\frac{1}{n}\sum_{i=1}^{n} x_i^2\Big) - \overline{x}^2. \end{aligned}$$

4.4 Séries statistiques à une dimension

4.4.1 Tableau des distributions des fréquences

Définition 4.4.1
Une distribution statistique est une représentation des données collectées dans un tableau où figurent les valeurs que prenne la variable, les effectifs, les fréquences et les fréquences cumulées relatives à chaque valeur ou ensemble de valeurs prises par la variable.

Exemple 41 *Variable quantitative discrète :*
Les performances en saut en hauteur (en cm) de 10 athlètes sont : $191, 194, 197, 191, 200, 203, 200, 197, 203, 203$.

Hauteur en cm	*Effectifs* n_i	*Frequences* f_i	*Frequences cumulées* F_i
191	2	$0,2$	$0,2$
194	1	$0,1$	$0,3$
197	2	$0,2$	$0,5$
200	2	$0,2$	$0,7$
203	3	$0,3$	1
Totale	$N = 10$	1	

Exemple 42 *Variable quantitative continue :*
Etude de la consommation aux 100 km de 20 voitures d'un nouveau modèle : $5,56;\ 5,35;\ 5,98;\ 5,77;\ 5,18;\ 5,66;\ 5,28;\ 5,11;\ 5,58; 5,49;\ 5,59;\ 5,33;\ 5,55;\ 5,45;\ 5,76;\ 5,23;\ 5,57;\ 5,52;\ 5,8;\ 6,0.$

Consommation en litre	*Effectifs* n_i	*Frequences* f_i	*Frequences cumulées* F_i
$[5, 5.2]$	2	$0,1$	$0,1$
$]5.2, 5.4]$	4	$0,2$	$0,3$
$]5.4, 5.6]$	8	$0,4$	$0,7$
$]5.6, 5.8]$	4	$0,2$	$0,9$
$]5.8, 6]$	2	$0,1$	1
Totale	$N = 20$	1	

4.4.2 Représentations graphiques

4.4.2.1 Représentations graphiques d'une distribution de variables qualitative

i) Les tuyaux d'orgues :

Les tuyaux d'orgues des effectifs (respectivement des fréquences) de la distribution statistique, $\{(x_i, n_i)/1 \leq i \leq p\}$ (respectivement $\{(x_i, f_i)/1 \leq i \leq p\}$) s'obtient en traçant sur un repère orthonormé, pour tout $i = 1, \ldots, p$, un rectangle de base de centre x_i et de hauteur égale à l'effectif ou la fréquence de la valeur x_i.
Sur l'axe des abscisses on représente les modalités de la variable, alors que sur l'axe des ordonnées on représente les effectifs ou les fréquences selon que l'on désire tracer un diagramme des effectifs ou des fréquences.

Exemple 43 *Représentation du diagramme en tuyaux d'orgues des fréquences pour le niveau d'étude des adultes d'une résidence.*

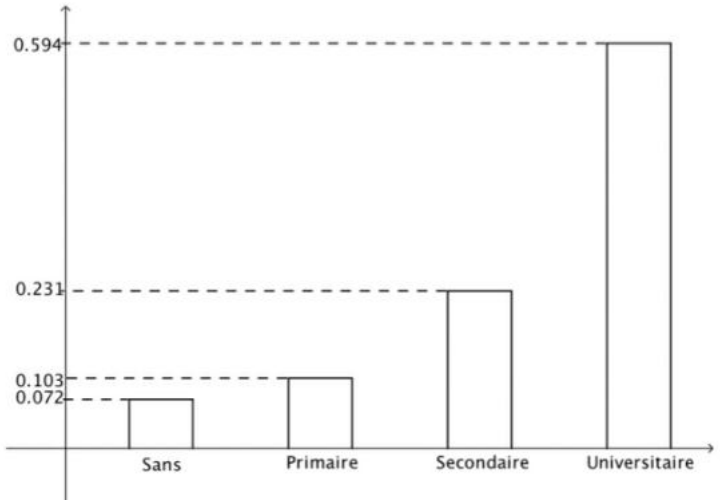

FIGURE 4.2 – Diagramme en tuyaux d'orgues

ii) Représentation circulaire :

C'est une représentation où chaque modalité est représentée par une portion du disque. Si S est l'aire du disque, l'aire d'une portion est égale à $f \times S$, où f est la fréquence de la modalité correspondante.
L'angle α de chaque portion s'obtient en multipliant la fréquence par 360°, l'angle du disque ($\alpha = f \times 360^\circ$).

Exemple 44 *Représentation du digramme circulaire des fréquences pour le niveau d'étude des adultes d'une résidence.*

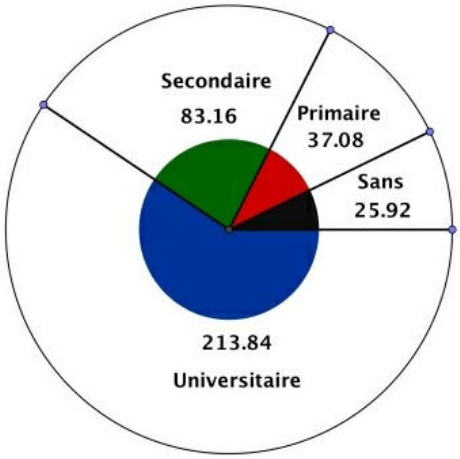

FIGURE 4.3 – Diagramme circulaire

4.4.3 Représentations graphiques d'une distribution de variables quantitatives discrètes

i) Diagramme en bâtons :

Le diagramme en bâtons des effectifs (respectivement des fréquences) de la distribution statistique $\{(x_i, n_i)/1 \leq i \leq p\}$ (respectivement $\{(x_i, f_i)/1 \leq i \leq p\}$) s'obtient en traçant sur un repère orthonormé les " bâtons " A_iB_i, c'est à dire les segments joignant les point $A_i(x_i, 0)$ et $B_i(x_i, n_i)$ (respectivement $B_i(x_i, f_i)$) pour $1 \leq i \leq p$.
Sur l'axe des abscisses on représente les valeurs de la variable, alors que sur l'axe des ordonnées on représente les effectifs ou les fréquences selon que l'on désire tracer un diagramme des effectifs ou des fréquences.

Exemple 45 *La distribution des performances en saut en hauteur de* 100 *athlètes sont représentées dans le tableau suivant :*

Hauteur en cm	*Effectifs* n_i	*Frequences* f_i	*Frequences cumulées* F_i
191	6	$0,06$	$0,06$
194	17	$0,17$	$0,23$
197	41	$0,41$	$0,64$
200	27	$0,27$	$0,91$
203	9	$0,09$	1
Totale	$N = 100$	1	

Représentation du diagramme en bâtons pour la distribution des performances en saut en hauteur de 100 *athlètes.*

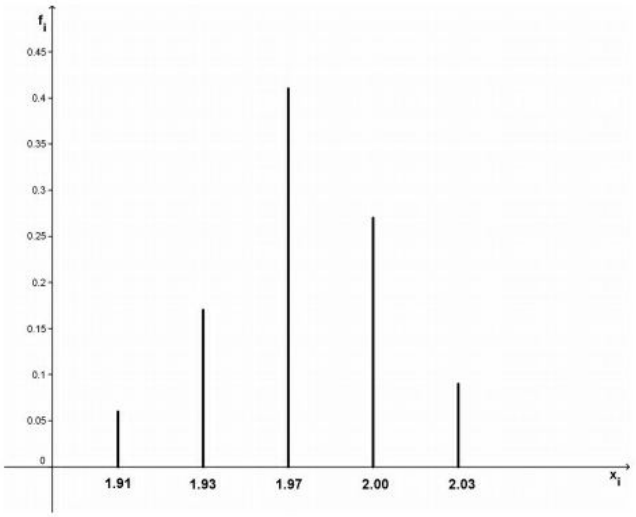

FIGURE 4.4 – Diagramme en bâtons

ii) Polygone des fréquences :
C'est une ligne brisée joignant les points de coordonnées (x_i, f_i). C'est aussi la ligne qui joint les sommets des bâtons du diagramme.

Exemple 46 *Représentation du polygone des fréquences pour la distribution des performances en saut en hauteur de* 100 *athlètes :*

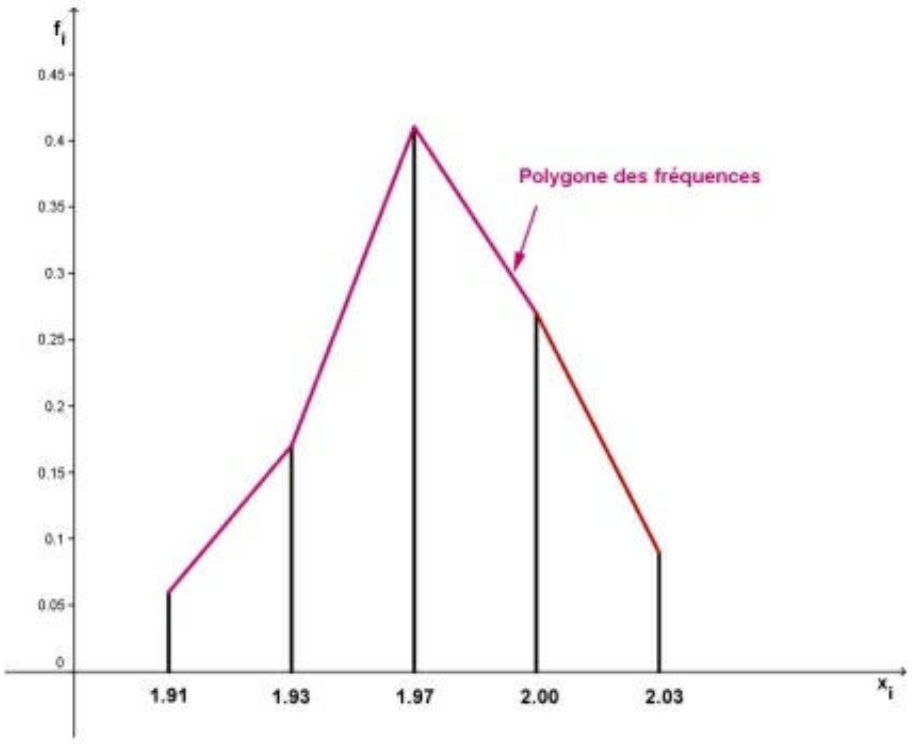

FIGURE 4.5 – Polygône des fréquences

iii) Courbe des fréquences cumulées :
C'est une courbe en escaliers qui représente la fonction :

$$F(x) = \begin{cases} 0 & \text{si } x < x_1 \\ \sum\limits_{\{i:x_i \leq x\}} f_i & \text{sinon} \end{cases}$$

Exemple 47 *Représentation de la courbe des fréquences cumulées pour la distribution des performances en saut en hauteur de* 100 *athlètes.*

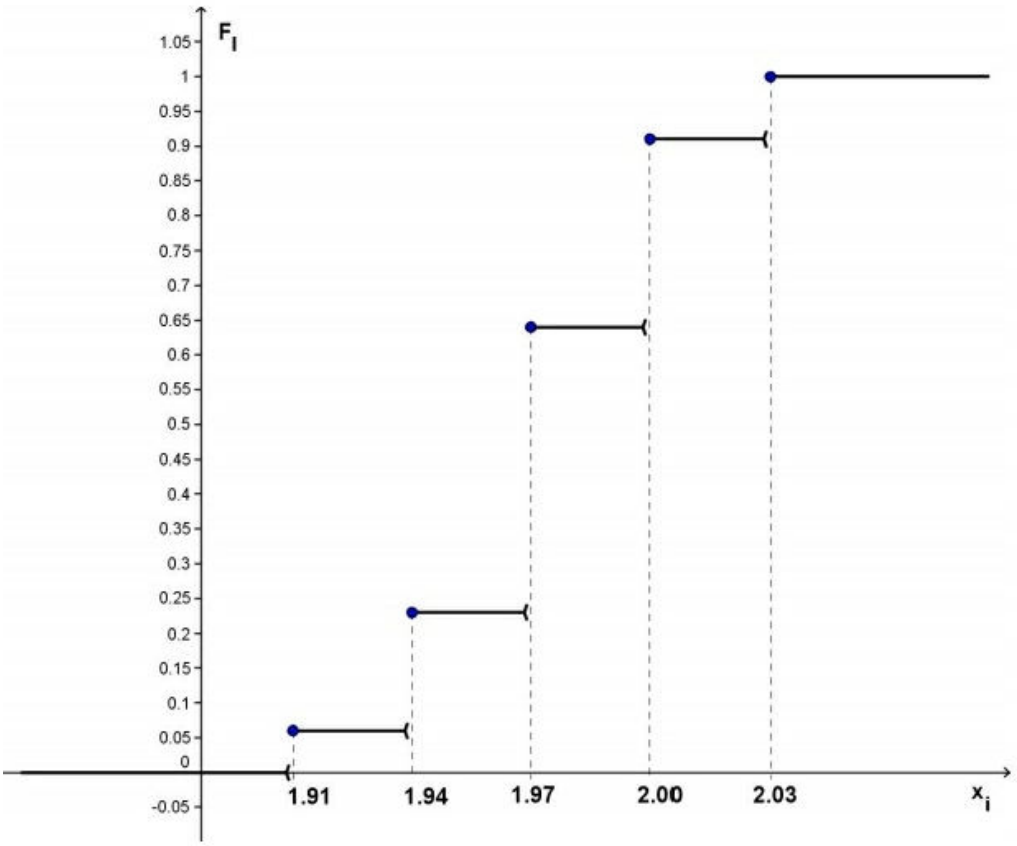

FIGURE 4.6 – Courbe des fréquences cumulées

4.4.3.1 Représentations graphiques d'une distribution de variables quantitatives continues

Considérons une variable continue X dont les valeurs se situent dans un intervalle I. On divise cet intervalle en k classes disjointes $]x_i, x_{i+1}]$, $i = 1, ..., p$.
On prendra toujours des classes de même amplitude ($x_{i+1} - x_i =$ constante).
Plus le nombre d'observations est grand plus le nombre de classes est élevé. On admet cependant, pour aider à la compréhension, que ce nombre devrait être entre 5 et 15.

Pour tout i, on note n_i le nombre de valeurs de X dans la classe $]x_i, x_{i+1}]$ qu'on appelle effectif de cette classe.
Pour dresser le tableau de distribution, on pourra suivre les étapes suivantes :
Etape 1 : Déterminer p le nombre de classes à considérer dans l'étude. Pour N l'effectif de la population ou de l'échantillon, on peut le calculer selon l'une des deux règles suivantes :
i) Règle de Sturge : $P = 1 + 3.3 \times log_{10}(N)$
ii) Règle de Yule : $P = 2.5 \times \sqrt[4]{N}$
Avec $p =$ l'entier naturel le plus proche de P .
Etape 2 : Calculer l'étendue $e = x_{max} - x_{min}$ où x_{min} est la valeur minimale de la variable X et x_{max} est la valeur maximale de la variable X.
Etape 3 : Diviser l'étendue e par p le nombre de classes, pour avoir une idée sur la valeur de l'amplitude des classes que l'on notera a. on a, $a = \dfrac{e}{p}$
Etape 4 : On construit alors les classes
$[x_{min}, x_{min} + a],]x_{min} + a, x_{min} + 2a], \ldots,]x_{min} + (p-1)a, x_{min} + pa]$
Etape 5 : S'assurer que chaque observation appartient à une et une seule classe.

Exemple 48 *Etude de la consommation aux* 100 *km de* 20 *voitures d'un nouveau modèle :*
$6,11;\ 6,05;\ 5,98;\ 5,77;\ 5,18;\ 5,66;\ 5,28;\ 5,11;\ 5,58;\ 5,49;\ 5,62;\ 5,33;\ 5,55;\ 5,45;$
$5,76;\ 5,23;\ 5,57;\ 5,52;\ 5,8;\ 6,0.$
Pour la méthode de Sturge $P = 1 + 3,3 \times log_{10}(20) = 5,293.$
Pour la méthode de Yule $P = 2.5 \times \sqrt[4]{20} = 5.287,$
D'où le nombre de classe est $p = 5.$
Nous avons $x_{min} = 5,11$ *et* $x_{max} = 6,11.$ *D'où* $e = 6,11 - 5,11 = 1$ *et* $a = \dfrac{e}{p} = \dfrac{1}{5} = 0,2.$

Consommation en litre	*Effectifs* n_i	*Frequences* f_i	*Frequences cumulées* $F(x)$
$[5,11; 5,31]$	4	$0,2$	$0,2$
$]5,31; 5,51]$	3	$0,15$	$0,35$
$]5,51; 5,71]$	6	$0,3$	$0,65$
$]5,71; 5,91]$	3	$0,15$	$0,8$
$]5,91; 6,11]$	4	$0,2$	1
Totale	$N = 100$	1	

Histogramme :
L'histogramme des effectifs (respectivement des fréquences) de la distribution statistique $\{(]x_i, x_{i+1}], n_i)/1 \leq i \leq p\}$ (respectivement $\{(]x_i, x_{i+1}], f_i)/1 \leq i \leq p\}$) s'obtient en traçant sur un repère orthonormé, pour tout $i = 1, \ldots, p$, un rectangle de base la longueur du segment $]x_i, x_{i+1}]$ et de hauteur égale à l'effectif ou la fréquence de cette classe.
Sur l'axe des abscisses on représente les bornes des classes $]x_i, x_{i+1}]$ de la variable c'est à dire les points $x_1, x_2, \ldots, x_p, x_{p+1}$, alors que sur l'axe des ordonnées on représente les effectifs ou les fréquences selon que l'on désire tracer un histogramme des effectifs ou des fréquences.

Exemple 49 *Représentation de l'histogramme des fréquences de la distribution de l'exemple* (48).

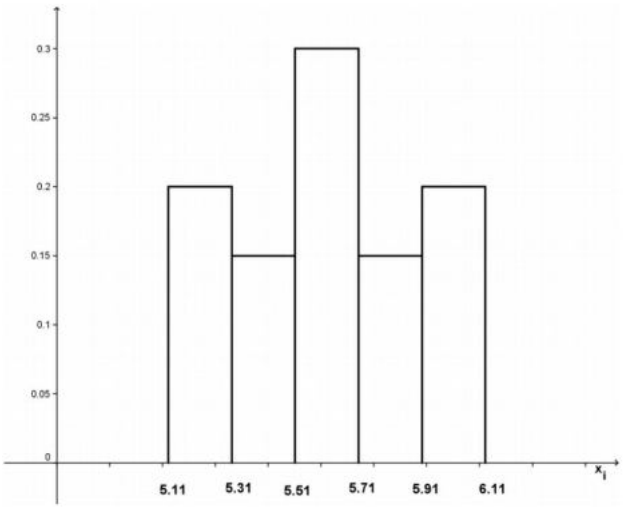

FIGURE 4.7 – Histogramme

4.4.3.2 Polygone des fréquences

Le polygone des fréquences de la distribution $\{(]x_i, x_{i+1}], f_i)/1 \leq i \leq p\}$ est la ligne brisée joignant les points de coordonnées (c_i, f_i) où $c_i = \frac{x_i + x_{i+1}}{2}$ le centre de la classe i, $i = 1, \ldots, p$.
Lorsque la borne inférieure de la première (resp. supérieure de la dernière) classe est observée c'est à dire l'intervalle est fermé en x_1 (resp. x_{p+1}) (comme c'est le cas dans l'exemple (48)), on complète la courbe en joignant les points $(c_0, 0)$ et (c_1, f_1) (resp. (c_p, f_p) et $(c_{p+1}, 0)$) où $c_0 = x_1 - \frac{a}{2}$ (resp $c_{p+1} = x_{p+1} + \frac{a}{2}$).
Lorsque la borne inférieure de la première (resp. la borne supérieure de la dernière) classe n'est pas observée c'est à dire l'intervalle est ouvert en x_1 (resp. en x_{p+1}) , on complète la courbe en joignant les points $(x_1, 0)$ et (c_1, f_1) (resp. (c_p, f_p) et $(x_{p+1}, 0)$).

Exemple 50 *Représentation du polygone des fréquences de la distribution de l'exemple* (48)

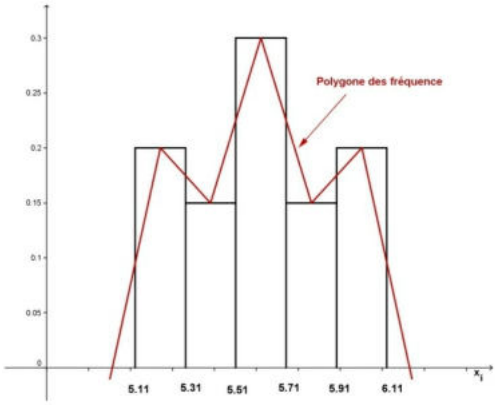

FIGURE 4.8 – Polygone des fréquences

4.4.3.3 Courbe des fréquences cumulées

La courbe des fréquences cumulées de la distribution $\{(]x_i, x_{i+1}], f_i)/1 \leq i \leq p\}$ s'obtient en joignant les points de coordonnées $(y, 0), (c_i, F_i)$ pour $i = 0, 1, \ldots, p$ et $(x, 1)$ pour $y \leq c_0$ et $x \geq c_p$ avec , $F_0 = 0, F_i = f_1 + \ldots + f_i$ et $c_i = x_{i+1}$ pour $i = 0, 1, \ldots, p$.
Lorsque la borne inférieure de la première classe est observée c'est à dire l'intervalle est fermé en x_1, $F(x_1) \neq 0$, (comme c'est le cas dans l'exemple (48)), on a $c_0 = x_1 - \frac{a}{2}$.
Lorsque la borne inférieure de la première classe n'est pas observée c'est à dire l'intervalle est ouvert en x_1, $F(x_1) = 0$, on a $c_0 = x_1$.

Exemple 51 *Représentation de la courbe des fréquences cumulées de la distribution de l'exemple* (48).

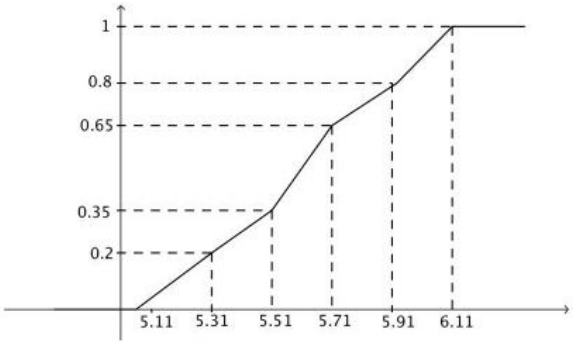

FIGURE 4.9 – Courbe des fréquences cumulées

4.4.4 Les mesures de dispersion

Les indicateurs de dispersion sont nombreux, les plus courants sont : L'étendue , l'écart interquartile, la variance, l'écart-type et le cœfficient de variation.

4.4.4.1 L'étendue

i) Variable quantitative discrète :
L'étendue mesure l'écart entre la plus petite valeur de la variable et la plus grande :

$$e = x_{max} - x_{min}$$

où x_{min} (resp. x_{max}) est la valeur minimale (resp. maximale) prises par la variable.

Exemple 52 *Soient les 4 séries statistiques suivantes :*
a) $10, 10, 10, 10, 20, 30, 30, 30, 30$ $\overline{x} = \dfrac{4 \times 20 + 1 \times 10 + 4 \times 30}{9} = \dfrac{180}{9} = 20$
b) $20, 22, 21, 20, 20, 19, 18, 20, 20$ $\overline{x} = \dfrac{1 \times 18 + 1 \times 19 + 5 \times 20 + 1 \times 21 + 1 \times 22}{9} = \dfrac{180}{9} = 20$
c)$1, 4, 6, 8, 20, 32, 34, 36, 39$ $\overline{x} = \dfrac{1 + 4 + 6 + 8 + 20 + 32 + 34 + 36 + 39}{9} = \dfrac{180}{9} = 20$
d) $10, 12, 14, 16, 20, 24, 26, 28, 30$; $\overline{x} = \dfrac{10 + 12 + 14 + 16 + 20 + 24 + 26 + 28 + 30}{9} = \dfrac{180}{9} = 20$
Ces quatre séries ont la même moyenne $\overline{x} = 20$ *et la même médiane* $m = 20$. *Pourtant ces*

séries sont très différentes. Cette différence provient de leur dispersion, en effet : $Etendue(a) = 30 - 10 = 20$, $Etendue(b) = 22 - 18 = 4$, $Etendue(c) = 39 - 1 = 38$ *et* $Etendue(d) = 30 - 10 = 20$.

ii) Variable quantitative continue :
Dans ce cas l'étendue est la différence entre la borne supérieure de la dernière classe et la borne inférieure de la première classe. $e = x_{max} - x_{min}$
où x_{min} (resp. x_{max}) est la borne inférieure (resp. supérieure) de la première (resp. dernière) classe.

4.4.4.2 Les quartiles

Nous savons que la médiane divise la distribution en deux parties égales. Il existe d'autres indicateurs utiles :
a) Les quartiles qui divise la distribution en quatre (4) parties égales.
b) Les déciles qui divise la distribution en dix (10) parties égales.
c) Les centiles qui divise la distribution en cent (100) parties égales.
Les quartiles sont notés Q_1, Q_2 et Q_3 et on a $F(Q_1) = 0.25, F(Q_2) = 0.5$ et $F(Q_3) = 0.75$. La médiane est le 2ème quartile, le 5ème décile et le 50ème centile.
i) Variable quantitative discrète :
On considère une série statistique dont les valeurs du caractère étudié, ont été rangés dans un ordre croissant :

$$x_1 \leq x_2 \leq \ldots \leq x_n.$$

La médiane m_e sépare la série en deux séries de même effectif. La série inférieure dont les valeurs du caractère sont inférieures ou égale à la médiane m_e, et la série supérieure dont les valeurs du caractère sont supérieures ou égale à la médiane m_e. On appelle premier (resp. troisième) quartile, la médiane de la série inférieure (resp. supérieure) on le note Q_1 (resp. Q_3).

Exemple 53 *Considérons la distribution suivante*

x_i	10	20	30	40	50	60
Effectifs n_i	3	8	4	9	3	3
Effectifs cumulées N_i	3	11	15	24	27	30

On a $N = 30$ *et* $m_e = 35$.
Série inférieure avec $N_1 = 15$ *:*

x_i	10	20	30
Eff. n_i	3	8	4
Eff. cum. N_i	3	11	15

Série supérieure avec $N_2 = 15$ *:*

x_i	40	50	60
Eff. n_i	9	3	3
Eff. cum. N_i	9	12	15

Donc N_1 est impair d'où $\frac{N_1+1}{2} = 8 \Rightarrow Q_1 = x_{(8)} = 20$ et $Q_3 = x_{\frac{N_2+1}{2}} = x_{(8)} = 40$

Exemple 54 *Considérons la distribution suivante*

x_i	10	20	30	40	50	60
Effectifs n_i	4	9	5	8	3	4
Effectifs cumulées N_i	4	13	18	26	29	33

On a $N = 33$ et $m_e = 30$.
Série inférieure avec $N_1 = 16$:

x_i	10	20	30
Eff. n_i	4	9	3
Eff. cum. N_i	4	13	16

Série supérieure avec $N_2 = 16$:

x_i	30	40	50	60
Eff. n_i	1	8	3	4
Eff. cum. N_i	1	9	12	16

donc N_1 est pair d'où $\frac{N_1}{2} = 8 \Rightarrow Q_1 = \frac{x_{(8)} + x_{(9)}}{2} = \frac{20+20}{2} = 20$ *et* $Q_3 = \frac{x_{(8)}+x_{(9)}}{2} = \frac{40+40}{2} = 40$.

ii) Variable quantitative continue :
Des techniques similaires à celles utilisées pour déterminer la médiane dans le cas continue permettent de déterminer ces indicateurs.
Pour le premier quartile

$$\begin{cases} x_i < Q_1 \leq x_{i+1} \\ \\ F(x_i) < 0.25 \leq F(x_{i+1}) \end{cases} \quad \text{et } Q_1 = x_i + (x_{i+1} - x_i)\frac{0,25 - F(x_i)}{F(x_{i+1}) - F(x_i)}$$

Pour le troisième quartile

$$\begin{cases} x_i < Q_1 \leq x_{i+1} \\ \\ F(x_i) < 0.75 \leq F(x_{i+1}) \end{cases} \quad \text{et } Q_3 = x_i + (x_{i+1} - x_i)\frac{0,75 - F(x_i)}{F(x_{i+1}) - F(x_i)}$$

Exemple 55 *Reprenons la distribution des salaires mensuels*

Classe	*Effectifs* n_i	*Fréquence* f_i	*Fréquence cumulée* $F(x_{i+1})$
$]2,3]$	15	$0,19$	$0,19$
$]3,4]$	20	$0,25$	$0,44$
$]4,6]$	20	$0,25$	$0,69$
$]6,10]$	24	$0,31$	1
Total	79	1	

$0,19 < F(Q_1) = 0,25 \leq 0,44 \Longrightarrow 3 < Q_1 \leq 4$ *d'où* $Q_1 = 3 + (4-3) \times \dfrac{0,25-0,19}{0,44-0,19} = 3,24$

et $0,69 < F(Q_3) = 0,75 \leq 1 \Longrightarrow 6 < Q_3 \leq 10$ *d'où* $Q_3 = 6 + (10-6) \times \dfrac{0,75-0,69}{1-0,69} = 6,19$

iii) L'écart interquartile :
Q_1 étant le premier quartile et Q_3 le troisième quartile, l'écart interquartile est la différence entre le troisième et le premier quartile, il est noté $R(Q) = Q_3 - Q_1$.
L'intervalle $[Q_1, Q_3]$ est appelé intervalle interquartile. Il contient 50% des observations, le reste se réparti avec 25% à gauche de Q_1 et 25% à droite de Q_3.
L'écart interquartile $R(Q)$ est la largeur de l'intervalle interquartile. C'est une mesure de dispersion des données autours de la médiane.
• Plus il est grand, plus les données sont dispersées autours de la médiane.
• Plus il est petit, plus les données sont proches de la médiane.

Exemple 56
Reprenons l'exemple de la distribution des salaires mensuels.
L'intervalle interquartile est $[3,24; 6,19]$ *et l'écart interquartile est* $R(Q) = 6,19 - 3,24 = 2,85$.

4.4.4.3 La variance et l'écart-type

La variance est un résumé statistique qui mesure la concentration ou la dispersion des observations autour de la moyenne. L'écart-type permet d'avoir une idée de la façon dont les valeurs de la série s'écartent par rapport à la moyenne, c'est donc une mesure de dispersion. Un écart-type faible correspond à une série concentrée autour de la moyenne.
i) Variable quantitative discrète :
La variance $V(x)$ est la moyenne arithmétique des carrés des écarts des valeurs de la variable à la moyenne arithmétique

$$V(x) = \frac{1}{N}\sum_i n_i \times (x_i - \overline{x})^2 = \sum_i f_i \times (x_i - \overline{x})^2 \text{ où } \sum_i n_i = N$$

La racine carrée de la variance est appelée l'écart-type

$$\sigma(x) = \sqrt{V(x)} = \sqrt{\frac{1}{N}\sum_i n_i \times (x_i - \overline{x})^2} = \sqrt{\sum_i f_i \times (x_i - \overline{x})^2}$$

Exemple 57 *Considérons la distribution suivante*

x_i	10	20	30	40	50	60
n_i	4	8	4	9	3	3

on a $N = 31$ *et* $\overline{x} = 32.58$

$$V(x) = \frac{4 \times (10-32,58)^2 + 8 \times (20-32,58)^2 + 4 \times (30-32,58)^2}{31}$$
$$+\frac{9 \times (40-32,58)^2 + 3 \times (50-32,58)^2 + 3 \times (60-32,58)^2}{31} = \frac{6993.5484}{31} = 225.598$$

donc $\boxed{V(x) = 225.598}$ *et* $\boxed{\sigma(x) = \sqrt{225.598} = 15,02}$.

ii) Variable quantitative continue :
La variance $V(x)$ est la moyenne arithmétique des carrés des écarts des centres des classes à la moyenne arithmétique

$V(x) = \dfrac{1}{N}\sum_i n_i \times (c_i - \overline{x})^2 = \sum_i f_i \times (c_i - \overline{x})^2$ où c_i est le centre de la classe $]x_i, x_{i+1}$ associée à l'effectif n_i.
La racine carrée de la variance est appelée l'écart-type

$$\sigma(x) = \sqrt{V(x)} = \sqrt{\frac{1}{N}\sum_i n_i \times (c_i - \overline{x})^2} = \sqrt{\sum_i f_i \times (c_i - \overline{x})^2}$$

Exemple 58 *Reprenons la distribution des salaires mensuels*

Classe	*Effectifs* n_i	*Fréquence* f_i	*Fréquence cumulée* $F(x_{i+1})$
$]2,3]$	15	$0,19$	$0,19$
$]3,4]$	20	$0,25$	$0,44$
$]4,6]$	20	$0,25$	$0,69$
$]6,10]$	24	$0,31$	1
Total	79	1	

On a $\overline{x} = 5,05$.

$$V(x) = \frac{15 \times (2,5 - 5,05)^2 + 20 \times (3,5 - 5,05)^2 + 20 \times (5 - 5,05)^2 + 24 \times (8 - 5,05)^2}{79}$$
$$= \frac{354,497}{79} = 4,487$$

donc $\boxed{V(x) = 4,487}$ *et* $\boxed{\sigma(x) = \sqrt{4,487} = 2,118}$

4.4.4.4 Cœfficient de variation

Tous les indicateurs de dispersion que nous avons vu jusqu'à présent dépendent des unités de mesure de la variable. Ils ne permettent pas de comparer des dispersions de distributions statistiques hétérogènes.
Le cœfficient de variation, qui est un nombre sans dimension, permet cette comparaison lorsque les valeurs de la variable sont positives. Il s'écrit

$$CV = \frac{\sigma(x)}{\overline{x}}$$

Si $CV < 0,5$ alors la dispersion n'est pas importante. Si $CV > 0,5$ alors la dispersion est importante.

Exemple 59 *Dans une maternité on a relevé le poids (en kg) à la naissance de* 47 *nouveau-nés. Les données collectées sont résumées dans le tableau suivant :*

Classe	n_i	c_i	$n_i c_i$	$c_i - \overline{x}$	$(c_i - \overline{x})^2$	$n_i(c_i - \overline{x})^2$
$]2,5;3,0]$	8	$2,75$	$22,00$	$-0,73$	$0,5329$	$22,00$
$]3,0;3,5]$	15	$3,25$	$48,75$	$-0,23$	$0,0529$	$0,7935$
$]3,5;4,0]$	20	$3,75$	$75,00$	$0,27$	$0,0729$	$1,4580$
$]4,0;4,5]$	4	$4,5$	$18,00$	$0,52$	$0,2704$	$1,0816$
Total	47		$163,75$			$7,5963$

$\overline{x} = \frac{163,75}{47} = 3,48$, $\sigma(x) = \sqrt{\frac{7,5963}{47}} = \sqrt{0,1616} = 0,4019$ *et* $CV = \frac{\sigma(x)}{\overline{x}} = \frac{0,4019}{3,48} = 0,1154$
Le cœfficient de variation étant faible, le poids à la naissance est concentré autour de la moyenne.

4.4.5 Mesure de forme

4.4.5.1 Symétrie et asymétrie

Une distribution est dite symétrique si le mode, la médiane et la moyenne sont confondus. Une distribution qui n'est pas symétrique est dite asymétrique.

Remarque 4.4.1 *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 cloche comme dans la figure ci-après*

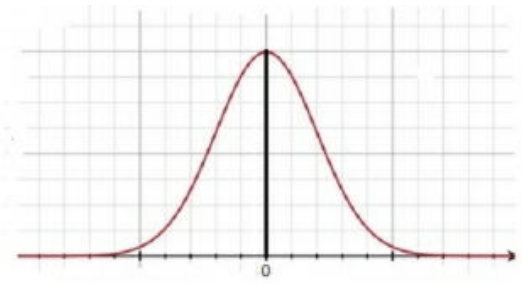

FIGURE 4.10 – Cloche

A la différence de la médiane et du mode, la moyenne arithmétique est fortement influencée par les valeurs extrêmes. Lorsque les valeurs sont distribuées de manière symétrique, la moyenne arithmétique coïncide avec la médiane et le mode.
Lorsque la distribution est asymétrique, la moyenne arithmétique dépasse la médiane si les valeurs extrêmes sont élevées et se situe en dessous de la médiane si les valeurs extrêmes sont basses.
Une distribution est dite asymétrique à droite, si la courbe du polygone des fréquences est étalée à droite, on a généralement : mode < médiane < moyenne.
Une distribution est dite asymétrique à gauche, si la courbe du polygone des fréquences est étalée à gauche, on a généralement : moyenne < médiane < mode.
La figure ci-dessous illustre ces différents cas lorsque la distribution ne présente qu'un seul mode.

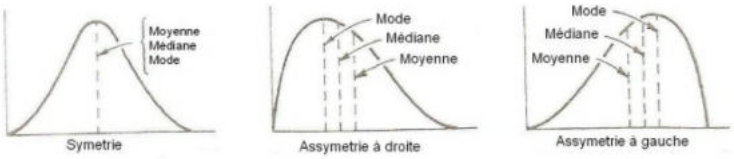

FIGURE 4.11 – Symétrie et asymétrie

4.4.5.2 Cœfficient d'asymétrie

le coefficient d'asymétrie a pour rôle de fournir une mesure de dissymétrie d'une distribution.

i) Cœfficient de d'asymétrie de Pearson :

Le premier cœfficient d'asymétrie de Pearson est basé sur une comparaison de la moyenne et de la médiane et est normalisé par l'écart-type. Il est calculé à partir de la formule suivante :

$$\boxed{A_{P1} = 3 \times \frac{\overline{x} - m_e}{\sigma}}$$ où m_e est la médiane.

Lorsque la distribution statistique est unimodale, on peut utiliser le second coefficient de Pearson basé sur une comparaison de la moyenne et du mode et est normalisé par l'écart-type. Il est calculé à partir de la formule suivante :

$$\boxed{A_{P2} = \frac{\overline{x} - m_d}{\sigma}}$$ où m_d est le mode.

ii) Cœfficient de d'asymétrie de Yule :

Le cœfficient d'asymétrie de Yule est basé sur les positions des trois quartiles et est normalisé par l'écart interquartile. Il est calculé à partir de la formule suivante :

$$\boxed{Y = \frac{Q_1 + Q_3 - 2Q_2}{R(Q)}}$$ où Q_1, Q_2, Q_3 les 3 quartiles , et $R(Q)$ l'écart interquartile.

iii) Cœfficient de d'asymétrie de Fisher :

Le coefficient d'asymétrie de Fisher est basé sur le moment d'ordre 3 et est normalisé par le cube de l'écart-type. Il est calculée à partir de la formule suivante :

$$\boxed{A_F = \frac{\mu_3}{\sigma^3}}$$

Tous les cœfficients d'asymétrie ont les mêmes propriétés.

• Si la distribution est symétrique, le coefficient est nul. On admettra que si le cœfficient de Fisher $A_F \in]-0.1, 0.1[$, la distribution est symétrique.

• Si la distribution est asymétrique à droite (resp. à gauche) c'est à dire la courbe est étalée à droite (resp. à gauche), le cœfficient est positif (resp. négatif).

Exemple 60

On considère la série statistique suivante (masse en grammes des œufs de poule d'un élevage).

masse x_i	40	45	50	55	60	65	70	75	80	85	90
Effectif n_i	16	20	75	141	270	210	165	63	21	12	7

$\overline{x}$	V	σ	μ_3	$m_e = Q_2$	m_d	Q_1	Q_3	$R(Q)$	A_{P1}	A_{P2}	A_Y	A_F
62.5	73.8	8.59	91.125	60	60	55	70	15	0.87	0.29	0.33	0.14

La distribution des masses est asymétrie à droite car les cœfficients d'asymétrie sont positifs.

4.4.5.3 Le cœfficient d'aplatissement

Le cœfficient d'aplatissement mesure le degré d'aplatissement d'une distribution. On l'obtient à partir du moment centré d'ordre 4.

• Cœfficient d'aplatissement de Pearson

$$\boxed{\beta_2 = \frac{\mu_4}{\sigma^4}}$$ où μ_4 est le moment d'ordre 4 et σ est l'ecart type.

• Cœfficient d'aplatissement de Ficher

$\boxed{F_2 = \beta_2 - 3}$ où β_2 est le coefficient d'aplatissement de Pearson

• Si $F_2 = 0$, le polygone statistique de la variable centrée réduite $\dfrac{X - \overline{x}}{\sigma(x)}$ à le même aplatissement qu'une courbe en cloche, on dit que la variable est mesokurtique.

• Si $F_2 > 0$, le polygone statistique de la variable centrée réduite est moins aplati qu'une courbe en cloche, la concentration des valeurs de la série autour de la moyenne est forte, on dit que la variable est leptokurtique.

• Si $F_2 < 0$, le polygone statistique de la variable centrée réduite est plus aplati qu'une courbe en cloche, la concentration des valeurs autour de la moyenne est faible, on dit que la variable est platykurtique.

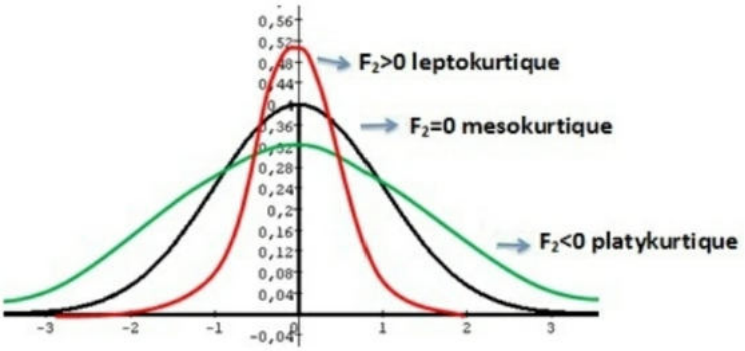

FIGURE 4.12 – Aplatissement

Exemple 61 *Reprenons la distribution des masse des oeufs de poule de l'exemple 60.*
$\mu_4 = 17523.91$, $V(x) = 73.8$, $\beta_2 = 3.22$ *et* $F_2 = 0.22 > 0 \Longrightarrow$ *la variable est leptokurtique et le polygone statistique de la variable centrée réduite est moins aplati qu'une courbe en cloche, la concentration des valeurs de la série autour de la moyenne est forte.*