Statistique Descriptive Univariée

La statistique a longtemps consisté en de simples dénombrements fournissant des renseignements sur la population ou l'économie d'un pays. De nos jours, nous pouvons dire que : la statistique est la branche des mathématiques qui consiste à la collecte, au classement, à l'analyse et à l'interprétation des données afin d'en tirer des conclusions et de faire des prévisions.

On ne doit pas confondre entre la statistique et les statistiques. Le mot "statistiques", au pluriel désigne l'ensemble des données chiffrées qui regroupe toutes les observations faites sur des faits relatifs à un même phénomène qui concerne un groupe d'objets.

La statistique descriptive ou encore appelée statistique exploratoire est un ensemble de méthodes permettant de synthétiser, décrire et résumer des données qui souvent très nombreuses sous des formes claires et compréhensibles.

1.1 Notions de base et terminologie

Nous allons commencer par définir les termes utilisés en statistiques pour désigner les observations chiffrées.

1.1.1 Population, individu, échantillon

Une population est l'ensemble des éléments auxquels se rapportent les données étudiées. En statistique, le terme "**population**" s'applique à des ensembles de toute nature : étudiants d'une académie, production d'une usine, poissons d'une rivière, entreprises d'un secteur donné...

Une population doit être bien définie. Sa définition est importante car elle conditionne l'homogénéité des unités observées et aussi la fiabilité des résultats

Dans une population donnée, chaque élément est appelé "**individu**" ou "unité statistique".

Lorsqu'on veut étudier les données relatives aux caractéristiques d'un ensemble d'individus ou d'objets il est difficile d'observer toutes les données lorsque leurs nombres sont élevés. Au lieu d'examiner l'ensemble qu'on appelle population on examine un nombre restreint qu'on appelle **échantillon**.

Les observations obtenues sur une population ou sur un échantillon constituent un ensemble de données auxquelles s'appliquent les méthodes de la statistique descriptive dont le but est de décrire le plus complètement et le plus simplement l'ensemble des observations qu'elles soient relatives à toute la population ou seulement à un sous-ensemble.

1.1.2 Caractère

Pour étudier une population, le statisticien ne retient que les caractères qui l'intéressent, un caractère étant une variable qui caractérise les individus de cette population. Les valeurs possibles d'un caractère sont appelées ses modalités.

Exemple 1.1.1 *Les modalités du caractère* ***sexe*** *sont masculin (codé M) et féminin (codé F).*

Exemple 1.1.2 *Les modalités du caractère* ***nombre d'enfants par famille*** *sont 0,1,2,3,4,5,...*

Il existe deux catégories de caractères : les caractères qualitatifs et les caractères quantitatifs.

Caractères qualitatifs et quantitatifs

Un caractère est dit quantitatif quand ses différentes modalités sont mesurables par des nombres qui en indiquent l'intensité, peut encore distinguer :

► Les caractères discrets sont ceux dont le nombre de modalités est fini ou dénombrable. Leurs valeurs peuvent être ou non des nombres entiers, par exemple le nombre de pages d'un livre, le nombre de personnes dans une famille. D'autre façon entre deux valeurs successives, aucune autre valeur n'est possible.

► Les caractères continus sont ceux qui ont une infinité de modalités, comme la taille, le poids des labradors de 3 ans, pression artérielle, taux de lipides, nombre de bactéries, durée de survie. De façon générale entre deux valeurs successives, il peut exister une infinité de valeurs.

Un caractère est dit qualitatif quand ses différentes modalités ne peuvent être désignées que par leurs qualités. Il existe deux types du caractère qualitatif, nominal et ordinal.

► Caractère qualitatif ordinal si nous pouvons comparer ces modalités entre elles et par conséquent ranger par ordre croissant ou décroissant, par exemple : le degré de sévérité d'une maladie (faible, moyenne, forte), état sanitaire des individus (Sain, Malade).

► Caractère qualitatif nominal s'il correspond à des noms et il n'y a aucun ordre précis, comme le sexe, la race , couleur de la robe de bovins, etc.

1.2 Représentations graphiques des caractères qualitatifs

On représente généralement caractère qualitatif par des graphiques qui utilisent des surfaces : représentation en cercle le diagramme à secteurs angulaires (dit camembert) ou tuyaux le diagramme en bandes (dit tuyaux d'orgue).

1.2.1 Tuyaux d'orgue

Les différentes modalités sont représentées par des rectangles dont la base est la même quelque soit la modalité, et la hauteur est proportionnelle à l'effectif ou à la fréquence, les distances entre les rectangles doivent être les mêmes.

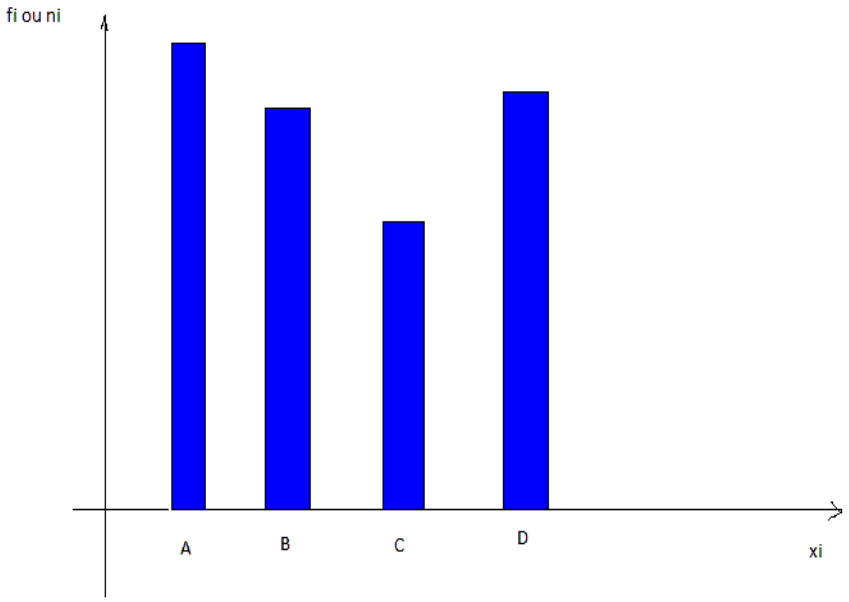


FIG. 1.1 – Tuyaux d'orgue.

1.2.2 Secteur angulaire

Secteur angulaire est un disque partagé en secteurs, chaque secteur[1] représentant une modalité et ayant une surface proportionnelle à la fréquence de cette modalité dans la série statistique.

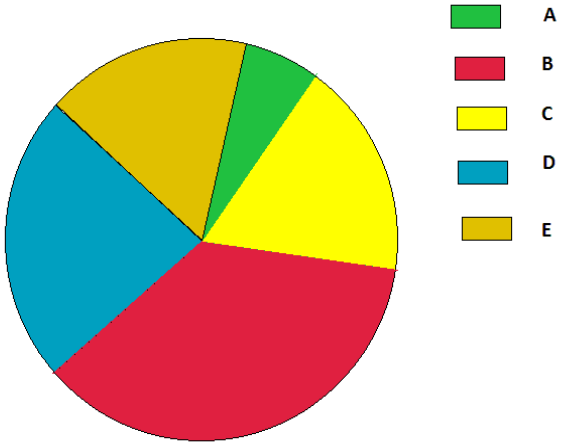


FIG. 1.2 – Secteur angulaire.

1.2.3 Exemple explicatif

On interroge 50 personnes sur leur dernier diplôme obtenu, la codification a été faite comme suit :

- Sans diplôme (Sd),
- Primaire(P),
- Secondaire (Se),
- Supérieur non universitaire (Su),
- Universitaire par U.

[1] Angle de chaque secteur est égale à

$$\theta_i^\circ = 360^\circ \times f_i$$

Pour cela on a obtenu la série suivante :

Sd	Sd	U	Sd	P	P	P	P	Se	U	P	P	P	P	P	P	Se
Se	Se	Se	Se	Se	Se	Se	Se	P	U	Se	Se	Su	Su	Su	Su	Su
Su	Su	Su	Su	U	U	U	U	P	Se	U	U	Sd	U	U	U	

Tableau statistique :

Modalité	Effectif n_i	Fréquence relative f_i	Angle θ_i
Se	13	$13/15 = 0.26$	93.6°
Su	9	$9/50 = 0.18$	64.8°
Sd	4	$4/50 = 0.08$	28.8°
U	12	$12/50 = 0.24$	86.4°
P	12	$12/50 = 0.24$	86.4°
Total	**50**	**1**	**360°**

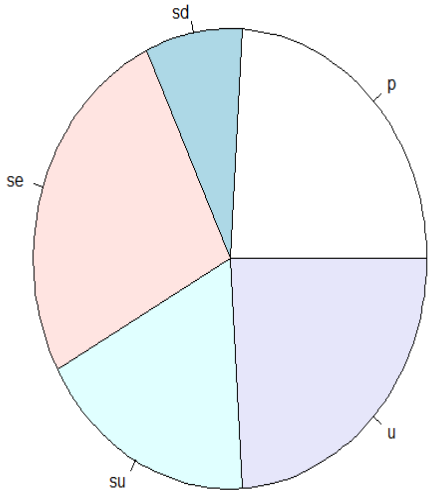


FIG. 1.3 – Secteur angulaire.

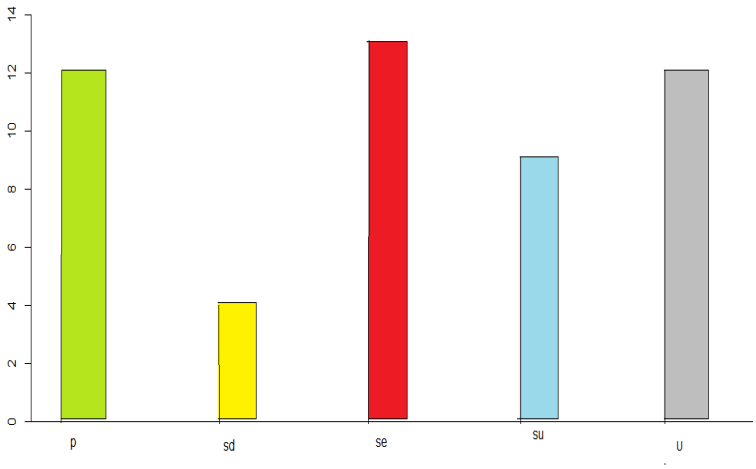


FIG. 1.4 – Tuyaux d'orgue.

Chapitre 2

Variable Statistique Discrète

On rappelle que une variable statistique est dite discrète lorsqu'elle ne peut prendre que des valeurs isolées dans son intervalle de variation.

2.1 Définitions

Définition 2.1.1 (Effectif) *L'effectif d'une modalité x_i d'un caractère x est le nombre d'individus présentant cette modalité. L'effectif correspondant à la $i^{ème}$ modalité du caractère x est noté n_i. L'effectif total est le nombre d'individus appartenant à la population statistique étudiée. L'effectif total sera noté N.*

$$\sum_{i=1}^{k} n_i = N$$

Remarque 1 : Le symbole sigma. Nous avons utilisé le symbole $\sum$. Il s'agit tout simplement d'une notation permettant de raccourcir certaines écritures. Ainsi, lorsqu'on fait la somme des valeurs indicées n_i (les effectifs de la série), au lien d'écrire $N = n_1 + ... + n_i + ... + n_k$, il est plus commode d'écrire $N = \sum_{i=1}^{k} n_i$

Définition 2.1.2 (Effectif cumulé croissant) *Les modalités d'un caractère variant de 1 à k, l'effectif cumulé croissant d'une modalité i est le nombre d'individus de la population*

présentant une modalité d'indice inférieur ou égal à i.

$$N_i = n_1 + n_2 + ...n_i = \sum_{j=1}^{i} n_j.$$

Exemple 2.1.1 *Considérons l'exemple du groupe de trente étudiants. On calcule les éffectifs cumulés correspondant à l'âge des étudiants dans le tableau suivant :*

Age	$Effectif$ n_i	$Effectif$ *cumulé* N_i
18	2	2
19	4	6
20	10	16
21	11	27
22	3	30
Total	30	

Définition 2.1.3 (Fréquence relative) *La fréquence relative d'une modalité est la proportion d'individus de la population totale qui présentent cette modalité : elle est obtenue en divisant l'effectif de cette modalité du caractère par l'effectif total et notée f_i, soit :*

$$f_i = \frac{n_i}{N}.$$

Définition 2.1.4 (Fréquence relative cumulée croissante) *La fréquence relative cumulée croissante de la valeur x_i de la distribution statistique X comme suit :*

$$F_i = f_1 + f_2 + ...f_i = \sum_{j=1}^{i} f_j.$$

Cette somme représente la proportion d'individus dans la population pour lesquels X prend une valeur inférieure ou égale à x_i.

Exemple 2.1.2 *On calcule les fréquences relatives cumulées de l'exemple 1.1.3 et les résultats obtenus dans le tableau suivant :*

Age	$Effectif$ n_i	$Fréquence$ $relative\ f_i$	$Fréquence\ relative$ $cumulée\ F_i$
18	2	$2/30 = 0.067$	0.067
19	4	$4/30 = 0.133$	0.200
20	10	$10/30 = 0.333$	0.533
21	11	$11/30 = 0.367$	0.900
22	3	$3/30 = 0.100$	1
Total	30	1	

2.2 Tableau statistique

Un tableau statistique est juste une liste de chiffres relative au caractère de la population que l'on souhaite étudier, présentée de façon la plus compréhensible possible. Les données peuvent être présentées individuellement, sous la forme suivante :

$Valeur\ observées$ x_i	$Effectifs$ n_i	$Effectifs$ $cumulés\ N_i$	$Fréquences$ f_i	$Fréquences$ $cumulées_i F$
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_k	n_k	N	f_k	$F_k = 1$
TOTAL	N		1	

TAB. 2.1 – Tableau statistique d'un caractère quantitatif discret.

2.2.1 Caractéristiques de tendance centrale (position)

Pour confirmer certaines impressions sur la série et pour en donner plus de précision, nous serons amenés à trouver une ou plusieurs valeurs centrales de la variable, capables de

résumer la série en caractérisant l'ordre de grandeur des observations. De telles valeurs centrales sont appelés paramètres de tendance centrale ou caractéristiques de position. Un indicateur de position doit être défini de manière rigoureuse et objective, doit tenir compte de l'ensemble des observations de la série et doit être exprimé dans la même unité que la variable.

Moyenne arithmétique

Soit la série statistique de données obsevées $x_1, \ldots, x_i, \ldots, x_N$, sa moyenne arithmétique est définie par :

Définition 2.2.1 (Moyenne arithmétique simple) *On appelle moyenne arithmétique la somme de toutes les données statistiques divisée par le nombre de ces données :*

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2.1}$$

Définition 2.2.2 (Moyenne arithmétique pondérée) *Si une valeur x_i de X est observée n_i fois, la formule 2.1 devient :*

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} n_i x_i. \tag{2.2}$$

on peut écrir la moyenne arithmétique à l'aide des fréquences f_i comme suit :

$$\overline{x} = \sum_{i=1}^{N} f_i x_i.$$

Exemple 2.2.1 *Voici le poids de 20 chiens de race Berger Allemand, tous sexes confondus, exprimés en kg :*

x_i	29	28	30	35	35	33	31	30	36	37	38	37	35	33	31	29	28	28	34	35

La moyenne arithmétique vaut : $\overline{x} = \frac{1}{20}\sum\limits_{i=1}^{20} x_i = \frac{652}{20} = 32.6$.

Exemple 2.2.2 *L'étude posologique d'un nouveau médicament faite sur N sujets a donné le tableau suivant :*

x_i	1	2	3	4	5	6	7	8	9
n_i	3	5	17	35	20	35	25	12	8

La moyenne arithmétique égale : $\overline{x} = \frac{1}{160}\sum_{i=1}^{9} n_i x_i = \frac{857}{160} = 5.356$.

Propriétés de la moyenne arithmétique

- La somme des écarts à la moyenne arithmétique est nulle : $\sum_{i=1}^{N} n_i\,(\overline{x} - x_i) = 0$.
- Elle est sensible aux valeurs extrêmes, il est parfois nécessaire de supprimer des valeurs extrêmes ou "aberrantes".
- Si l'on multiplie par un même nombre a chaque valeur de la série, la moyenne arithmétique est multipliée par ce nombre :

$$\frac{1}{N}\sum_{i=1}^{N} n_i a x_i = a\overline{x}.$$

- Si l'on ajoute (ou retranche) un même nombre à chaque valeur de la série, la moyenne arithmétique se trouve augmentée (diminuée) de ce nombre :

$$\frac{1}{N}\sum_{i=1}^{N} n_i (x_i + a) = a + \overline{x}.$$

- Propriété d'associativité : la moyenne arithmétique des moyennes arithmétiques calculées sur des sous-ensembles d'une série est égale à la moyenne arithmétique générale de la série.

Mode

On appelle mode ou valeur dominante d'une série statistique la valeur observée de la variable ayant le plus grand effectif (ou la fréquence la plus élevée). On note généralement le mode M*o*.

Exemple 2.2.3 *On considère les notes obtenues en biostatistique par un groupe de 20 étudiants : 7, 13, 5, 15, 12, 9, 7, 8, 14, 16, 13, 6, 13, 10, 13, 12, 10, 7, 12, 13.*

Le mode de cette série correspond à la note la plus fréquente, soit Mo =13, valeur qui apparaît cinq fois. L'interprétation en est que la note la plus fréquente est 13.

Médiane

La médiane est la valeur de la variable statistique telle qu'il y ait autant d'observations supérieures et d'observations inférieures à cette valeur. Elle partage la série statistique en deux parties d'égal effectif. En d'autres termes, la médiane[1] est la valeur de la variable située au **milieu** d'une série ordonnée telle que la moitié des individus prenne une valeur qui lui soit inférieure, l'autre moitié prenant par conséquent une valeur qui lui soit supérieure. La détermination de la médiane d'une série statistique nécessite d'abord de ranger par ordre croissant (ou décroissant) les valeurs observées.

- Si la série comporte un nombre impair de valeurs, soit N valeurs, la médiane sera la valeur de rang $\left(\frac{N+1}{2}\right)$.
- Si la série comporte un nombre pair de valeurs, on parle d'intervalle médian. Ce dernier est défini par :

$$\left[la\ \left(\frac{N}{2}\right)^{ième}\ valeur,\ la\ \left(\frac{N}{2}+1\right)^{ième}\ valeur\right[.$$

Toute valeur appartenant à cet intervalle fait fonction de médiane[2].

Exemple 2.2.4 *On considère la répartition de 9 ménages selon le nombre de chients par ménage.*

[1] Pour une répartition parfaitement symétrique on a : Moyenne = mode = médiane

[2] Certains ouvrages proposent de choisir comme médiane le centre de l'intervalle médian. La médiane, dans ce cas, n'est pas forcément une valeur observée.

$N°$ de chiens par ménage	0	0	1	1	**2**	3	3	3	4
range ordre (croissant)	1	2	3	4	**5**	6	7	8	9
					Me				

La médiane, dans ce cas, correspond à la cinquième valeur : $Me = 2$ chiens par ménage. On dit qu'il y a autant de ménage qui ont moins de deux chiens que de ménage qui ont plus de deux chiens.

Exemple 2.2.5 *On considère la répartition de 10 ménages selon le nombre de chiens par ménage.*

$N°$ de chiens par ménage	0	0	1	1	**2**	**3**	3	3	4	4
range ordre (croissant)	1	2	3	4	**5**	**6**	7	8	9	10
					intervalle	**médiane**				

Dans ce cas on parle plutôt d'intervalle médian [2, 3[, correspondant à [la cinquième valeur, sixième valeur[.

Quartiles

Les quantiles sont des caractéristiques de position partageant la série statistique ordonnée en quatre parties égaux.

Le premier quartile : ou $25^{ème}$ percentile, noté Q_1, est la valeur du caractère quantitatif telle que 25% des individus de l'échantillon ont une valeur inférieure à Q_1.

Le deuxième quartile Q_2 : est la médiane.

Le troisième quartile : ou $75^{ème}$ percentile, noté Q_3, est la valeur du caractère quantitatif telle que 75% des individus de l'échantillon ont une valeur inférieure à Q_3.

Exemple 2.2.6 *Dans l'exemple 1.2.2, nous avons le tableau des effectifs cumulés suivant :*

x_i	n_i	N_i	
1	3	3	
2	5	8	
3	17	25	
4	35	60	$\longleftarrow N/4 = 40$
5	20	80	
6	35	115	
7	25	140	$\longleftarrow 3N/4 = 120$
8	12	152	
9	8	160	

Alors les quartiles vaut :

$Q_1 = x_{[N/4]} = x_{[160/4]} = x_{[40]} = 4.$

$Q_3 = x_{[3N/4]} = x_{[3\times 160/4]} = x_{[120]} = 7.$

2.2.2 Caractéristiques de dispersion

Les caractéristiques de dispersion quantifient les fluctuations des valeurs observées et leur étalement. Il existe plusieurs mesures de la dispersion. Nous donnons, dans ce polycopie les principaux paramètres de dispersion qui sont l'étendue, l'écart interquartile, l'écart-type, la variance et le coefficient de variation.

Étendue

La différence entre la plus grande et la plus petite valeur observée, notée E

$$E = \max(x_i) - \min(x_i).$$

L'étendue est simple et facile à calculer. Toutefois, il est très sensible aux valeurs extrêmes

"aberrantes".

Exemple 2.2.7 *Reprenons les données de l'exemple 1.2.2, l'étendue est*

$$E = 9 - 1 = 8.$$

Ecart-interquartile

De par la définition des quartiles, l'intervalle interquartile $[Q_1,\ Q_3]$ contient 50% des observations. Sa longueur, est écart-interquartile notée $\mathbf{I}_Q$, alors la différence entre entre le troisième et le premier quartile :

$$I_Q = Q_3 - Q_1. \tag{2.3}$$

Variance

La variance est la moyenne arithmétique des carrés des écarts à la moyenne arithmétique

$$var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2. \tag{2.4}$$

Nous pouvons utiliser la formule développée de la variance pour simplifier le calcul de la formule 3.1. Cette formule est issue du théorème de König. La variance peut aussi s'écrire

$$var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2. \tag{2.5}$$

Démonstration. On a $var(X) = \frac{1}{N}\sum_{i=1}^{k} n_i \left(x_i - \overline{x}\right)^2$

D'après le développement du carré $\left(x_i - \overline{x}\right)^2 = x_i^2 - 2\overline{x}x_i + \overline{x}$.

On trouve

$$\begin{aligned} var(X) = &= \frac{1}{N}\sum_{i=1}^{k} n_i \left(x_i^2 - 2\overline{x}x_i + \overline{x}\right) \\ &= \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - 2\overline{x}\frac{1}{N}\sum_{i=1}^{k} n_i x_i + \overline{x}^2\frac{1}{N}\sum_{i=1}^{k} n_i \\ &= \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - 2\overline{x}\overline{x} + \overline{x}^2 \\ &= \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - 2\overline{x}^2 + \overline{x}^2 \\ &= \frac{1}{N}\sum_{i=1}^{k} n_i x_i^2 - \overline{x}^2. \end{aligned}$$

□

Exemple 2.2.8 *Prenons l'exemple 1.2.1, nous avons :*

$$\begin{aligned} var(X) &= \frac{1}{20}\sum_{i=1}^{20} x_i^2 - \overline{x}^2 \\ &= \frac{1}{20} \times (21468) - (32.6)^2 \\ &= 10.64. \end{aligned}$$

Ecart-type

L'écart-type[3] est défini comme la racine carrée positive de la variance

$$\sigma(X) = \sqrt{var(X)}. \tag{2.6}$$

Exemple 2.2.9 *Soit la série statistique 2, 3, 4, 4, 5, 6, 7, 9 de taille 8. On a*

$$\overline{x} = \frac{2+3+4+4+5+6+7+9}{8} = 5,$$

$$\begin{aligned} var(X) &= \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2 \\ &= \frac{1}{8}\left((2-5)^2 + (3-5)^2 + (4-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (7-5)^2 + (9-5)^2\right) \\ &= \frac{36}{8} = 4.5. \end{aligned}$$

On peut également utiliser la formule (2.1) de la variance, ce qui nécessite moins de calcul (surtout quand la moyenne n'est pas un nombre entier).

[3] L'unité de la variance est l'unité carré des valeurs observées. A cause de cette changement d'unité il est dificile d'utiliser la variance comme mesure de dispersion. mais l'écart-type qu'exprime dans les mêmes unités que les valeurs observées.

$$\begin{aligned} var(X) &= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{x}^2. \\ &= \frac{1}{8}(2^2+3^2+4^2+4^2+5^2+6^2+7^2+9^2) - 5^2 \\ &= \frac{1}{8}(236) - 25 = 4.5. \end{aligned}$$

Propriétés de la variance

1. La variance et l'écart-type sont toujours positifs.
2. Multiplication par une constante : $\forall\ a \in \mathbb{R} :\ var(aX) = a^2 var(X)$.
3. Addition d'une constante : $\forall\ b \in \mathbb{R} :\ var(X+b) = var(X)$.
4. Transformation linéaire : $\forall\ a,\ b \in\ \mathbb{R} : var(aX+b) = a^2 var(X)$.

Remarque 2.2.1 *La signification de l'écart-type et de la variance est simple : plus les valeurs observées sont peu dispersées (homogène), plus ces deux nombres sont petits et inversement, plus les valeurs observées sont hétérogènes, plus ces deux nombres sont grands.*

Représentation graphique des caractères discrets

Deux diagrammes permettent de représenter un caractère quantitatif discret : le diagramme en bâtons et le diagramme cumulatif.

Diagramme en bâtons Le diagramme en bâtons des effectifs (resp. des fréquences) d'une distribution statistique discrète est constitué d'une suite de segments verticaux d'abscisses x_i dont la longueur est proportionnelle à l'effectif (resp. la fréquence) de x_i.

Diagramme cumulatif Diagramme cumulative est un graphique en escalier dont les paliers horizontaux ont pour ordonnées respectivement F_i ou N_i. Les marches de l'escalier correspondent aux valeurs possibles de la variable statistique et sont à des hauteurs proportionnelles aux effectifs cumulés ou aux fréquences cumulées.

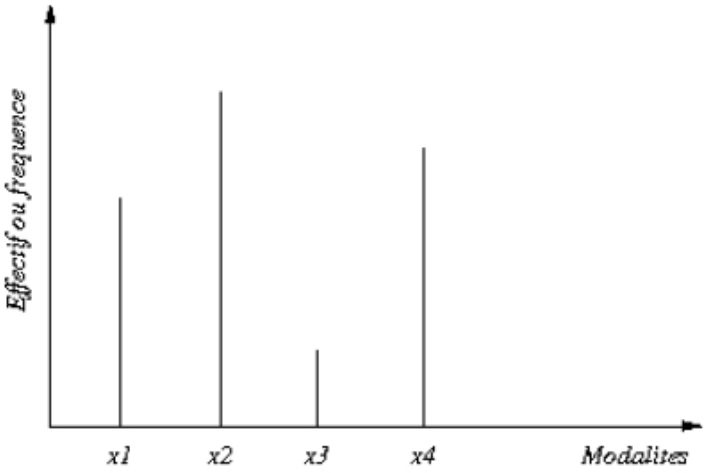


FIG. 2.1 – Diagramme en bâtons.

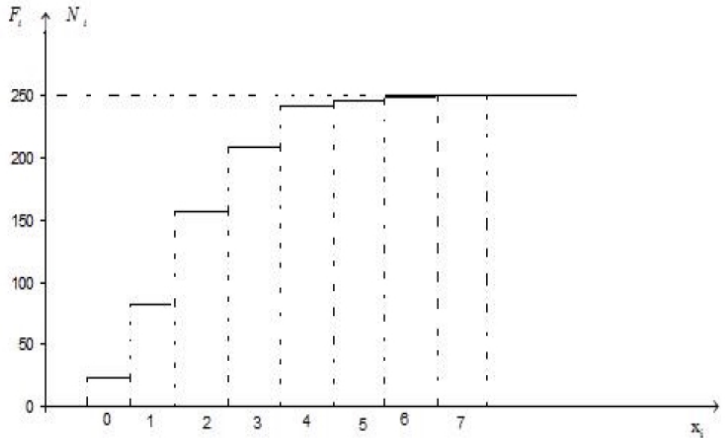


FIG. 2.2 – Diagramme cumulatif

Chapitre 3

Variable Statistique Continue

Lorsque le caractère quantitatif discret comprend un grand nombre de valeurs, nous préférons regrouper les valeurs en intervalles appelées classes pour rendre la statistique plus lisible. Nous partageons alors l'ensemble des valeurs du caractère en classes $[e_{i-1},\ e_i[$ avec $e_{i-1} < e_i$.

On choisit les classes pas trop nombreuses, mais suffisamment pour qu'il n'y ait pas de perte d'information.

On peut fixer le nombre de classes selon l'un des deux formules suivantes :

▶ **Règle de Sturge** :

$$nb.de\ classes = 1 + (3.3 \log N).$$

▶ **Règle de Yule :**

$$nb.de\ classes = 2.5\sqrt[4]{N}.$$

L'amplitude de classe est alors donnée par :

$$\frac{valeur \max - valeur \min}{nb.de\ classes}$$

Une classe est définie par ses extrémités e_{i-1}, e_i et son effectif n_i. Chaque classe est carac-

térisé par son centre et son amplitude :

► Le centre de la classe $[e_{i-1},\ e_i[$ noté C_i se définit de manière évidente par la valeur

$$C_i = \frac{e_{i-1} + e_i}{2}.$$

► La différence entre les deux extrémités est appelé amplitude de la classe. L'amplitude d'une classe i est

$$a_i = e_i - e_{i-1}.$$

3.1 Tableau statistique

Classes $[e_{i-1},\ e_i[$	*Centres* c_i	*Amplitude* a_i	*Effectifs* n_i	*Effectifs cumulés* N_i	*Fréquences* f_i
$[e_0,\ e_1[$	c_1	a_1	n_1	$N_1 = n_1$	f_1
$[e_1,\ e_2[$	c_2	a_2	n_2	$N_2 = n_1 + n_2$	f_2
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
$[e_{k-1},\ e_k[$	c_k	a_k	n_k	N	f_k
TOTAL			N		1

TAB. 3.1 – Tableau statistique d'un caractère quantitatif continue.

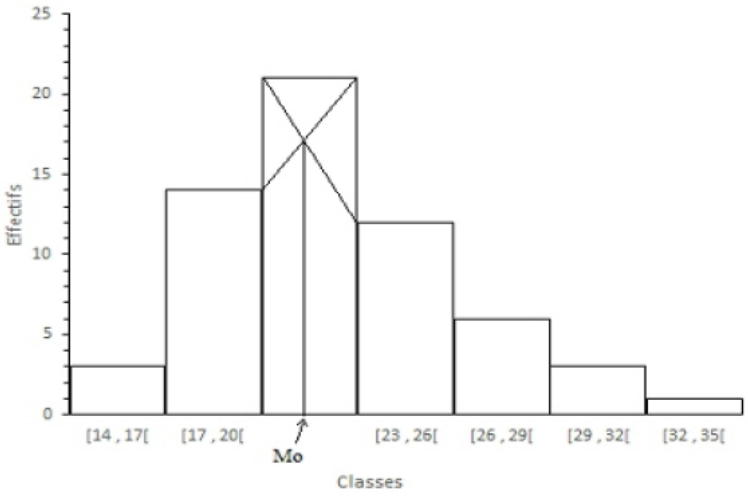
Dans le cas de la variable statistique continue, la moyenne arithmétique se calcul par la formule suivante :

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{k} n_i c_i.$$

3.2 Classe modale

C'est la classe ayant le plus grand éffectif par unité d'amplitude. Dans le cas d'une classe modale unique, on parle de distribution continue unimodale.

Graphiquement la classe modale est la base du rectangle ayant la hauteur la plus élevée. Cependant, on distingue deux cas selon que les amplitudes des classes sont :



3.2.1 Calcul du mode : effectifs groupés par classes d'amplitudes égales

Dans ce cas, la classe modale est la classe d'effectif n_i le plus élevé, soit $[e_i,\ e_{i-1}[$. L'effectif de la classe qui précède la classe modale est n_{i-1} et celui de la classe qui suit la classe modale est n_{i+1} alors

$$Mo = e_{i-1} + a_i \left(\frac{m_1}{m_1 + m_2} \right). \tag{3.1}$$

avec

$m_1 = n_i - n_{i-1}$.

$m_2 = n_i - n_{i+1}$.

FIG. 3.1 – Détermination graphique du mode pour une variable statistique continue. quand les classes sont d'amplitudes égales.

Exemple 3.2.1 *Dans le tableau ci-dessous, les valeurs d'une variable X ont été groupées par classes de valeurs d'amplitudes égales.*

$Classes$	n_i	N_i
[0, 5[	2	2
[5, 10[	7	9
[10, 15[	18	27
[15, 20[	3	30

TAB. 3.2 – Valeurs groupées par classes de valeurs d'amplitudes égales.

Appliquons la formule 3.1 en l'interprétant par rapport à la figure 3.2. Alors

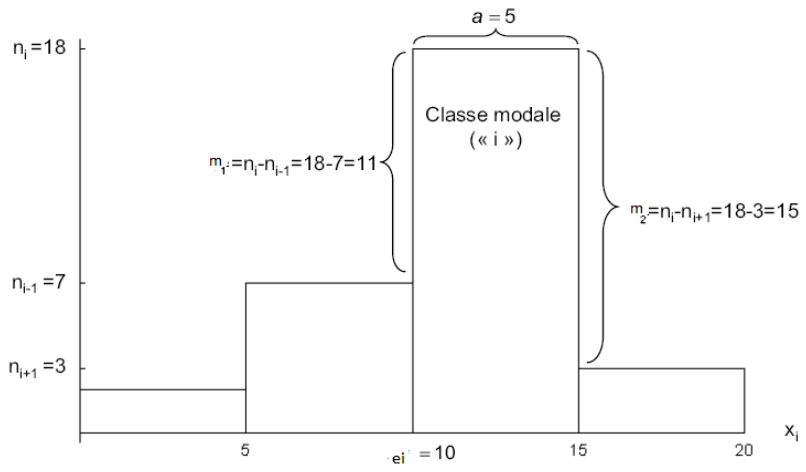


FIG. 3.2 – Calcul du mode quand les classes sont d'amplitudes égales.

$$Mo = e_{i-1} + a_i \frac{m_1}{m_1 \times m_2} = 10 + 5 \times \left(\frac{11}{11+15} \right) = 12.115.$$

3.2.2 Calcul du mode : effectifs groupés par classes d'amplitudes inégales

Dans ce cas, pour calculer le mode, il faut appliquer la formule 3.1, mais la définition de m_1 et de m_2 change, car il faut remplacer les effectifs n_i par les amplitudes corrigées $h_i = n_i/a_i$

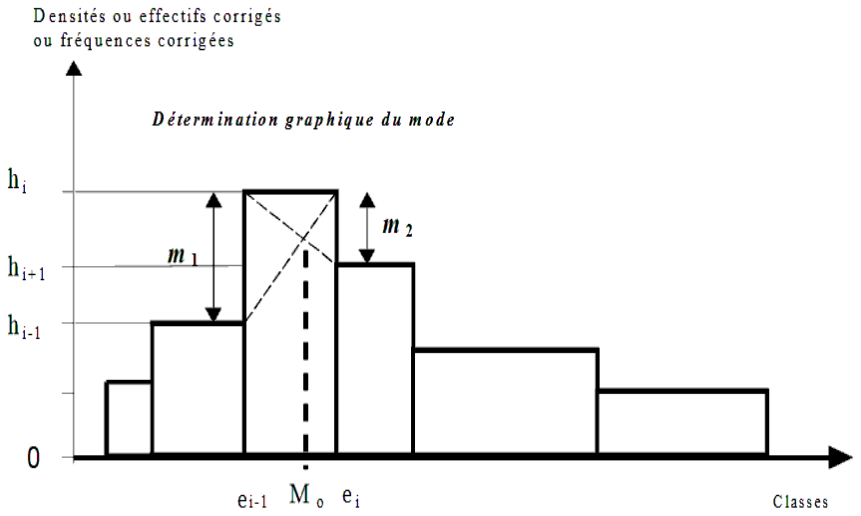


FIG. 3.3 – Détermination graphique du mode pour une variable statistique continue quand les classes sont d'inégales amplitudes.

Exemple 3.2.2 *Soit le tableau suivant où des données sont présentées par classes d'amplitudes d'amplitudes inégales.*

x_i	n_i	a_i	$h_i = \frac{n_i}{a_i}$
[0, 10[	9	10	0.9
[10, 12[	9	2	4.5
[12, 20[	12	8	1.5

TAB. 3.3 – Valeurs groupées par classes de valeurs d'inégales amplitudes.

On a donc

$$h_{i-1} = n_{i-1}/a_{i-1} = 9/10 = 0.9 \quad h_i = n_i/a_i = 9/2 = 4.5 \quad h_{i+1} = n_{i+1}/a_{i+1} = 12/8 = 1.5$$

$$m_1 = h_i - h_{i-1} = 4.5 - 0.9 = 3.6 \quad m_2 = h_i - h_{i+1} = 4.5 - 1.5 = 3$$

et $Mo = e_{i-1} + a_i \dfrac{m_1}{m_1 + m_2} = 10 + 2 \times \left(\dfrac{3.6}{3.6 + 3}\right) = 11.09$

Certains paramètres statistiques comme la médiane, les quartiles et, de manière générale les fractiles, peuvent être obtenus (de façon approchée) à partir des fréquences cumulées à l'aide d'une interpolation linéaire.

3.3 Médiane

Il n'y a aucune différence de calcul pour la médiane selon que les classes sont d 'amplitudes constantes ou variables. Le calcul de la médiane dans le cas de variable continue passe, d'abord, par la détermination de la classe médiane. Ensuite, par interpolation linéaire, on peut calculer la valeur précise de la médiane à l'intérieur de la classe médiane.

Pour déterminer la médiane, on recherche la classe qui la contient. C'est la classe $[e_{i-1}, e_i]$ telle que $F_{i-1} < 0,5 < F_i$ en notant F la fonction "fréquences relatives cumulées".

La médiane M est l'abscisse du point P d'ordonnée 0,5 (voir figure suivante). Son calcul est le suivant :

$$\frac{0.5 - F_{i-1}}{F_i - F_{i-1}} = \frac{Me - e_{i-1}}{e_i - e_{i-1}},$$

d'où

$$Me = e_{i-1} + (e_i - e_{i-1}) \left[\frac{0.5 - F_{i-1}}{F_i - F_{i-1}}\right]. \tag{3.2}$$

Remarque 3.3.1 *La même démarche pourrait être utilisée en remplaçant les fréquences*

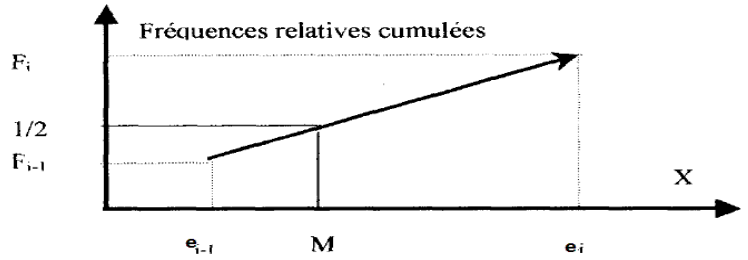


FIG. 3.4 – Détermination de la médiane par interpolation linéaire.

relatives par l'effectifs cumulés. L'expression de la médiane est donnée par :

$$Me = e_{i-1} + a_i \left[\frac{\frac{N}{2} - N_{i-1}}{N_i - N_{i-1}} \right].$$

Exemple 3.3.1 *En reprenant un exemple sur la répartition des 100 individus selon leur âge :*

Classes	$effectifs$ n_i	$effectifs\ cumulé$ $croissants\ N_i$	$fréquences$ f_i	$fréquences\ cumulées$ $croissants\ F_i$
[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
TOTAL	100		1	

Exemple 3.3.2 *Le calcul, par interpolation linéaire, de la médiane donne :*

$$\frac{Me - 20}{30 - 20} = \frac{0.50 - 0.36}{0.56 - 0.36}.$$

Ou encore, en utilisant les effectifs cumulés croissants :

$$\frac{Me - 20}{30 - 20} = \frac{50 - 36}{56 - 36}.$$

Dans notre exemple $\frac{N}{2} = 50$. La classe médiane est la classe à laquelle appartient la valeur médiane, c'est à dire la classe [20, 30[d'où

$$Me = 20 + 10\left[\frac{50 - 36}{56 - 36}\right] = 27.$$

C'est à dire que 50% des individus sont âgés de moins de 27 ans.

3.4 Quartiles

Déterminons le premier quartile Q_1, en examinant les fréquences relatives cumulées, rappelons que l'on doit atteindre 25% des valeurs les plus basses

$$Q_1 = e_{i-1} + a_i\left[\frac{0.25 - F_{i-1}}{F_i - F_{i-1}}\right].$$

Le calcul du troixième quartile Q_3 est du même type, dans laquelle on atteint 75% des valeurs les plus basses.

$$Q_3 = e_{i-1} + a_i\ \left[\frac{0.75 - F_{i-1}}{F_i - F_{i-1}}\right].$$

Exemple 3.4.1 *Le calcul, par interpolation linéaire du premier quartile Q_1 donne :*

$$\frac{Q_1 - 15}{20 - 15} = \frac{0.25 - 0.21}{0.36 - 0.21}.$$

Ou encore, en utilisant les effectifs cumulés croissants :

$$\frac{Q_1 - 15}{20 - 15} = \frac{25 - 21}{36 - 21}.$$

$$Q_1 = 15 + 5\left[\frac{25 - 21}{36 - 21}\right] = 16.33.$$

De même pour le calcul du troixième quartile Q_3 donne :

$$\frac{Q_3 - 40}{60 - 40} = \frac{75 - 74}{90 - 74}.$$

$$Q_3 = 40 + 20\left[\frac{75 - 74}{90 - 74}\right] = 41.25.$$

Pour les caractéristiques de dispersion la remarque qui nous a permis d'utiliser les centres de classes pour le calcul de la moyenne nous servira aussi pour développer un raisonnement analogue au sujet de la variance d'établir que

$$var(X) = \frac{1}{N}\sum_{i=1}^{N} n_i c_i^2 - \overline{x}^2 = \sum_{i=1}^{N} f_i c_i^2 - \overline{x}^2 \tag{3.3}$$

$$\sigma(X) = \sqrt{var(X)}. \tag{3.4}$$

3.5 Représentation graphique des caractères continus

Deux diagrammes permettent de représenter une variable quantitative continue : l'histogramme et la courbe cumulative.

3.5.1 Histogramme

Un histogramme est constitué d'un ensemble de rectangles dont la largeur est égale à l'amplitude a et la hauteur égale à n_i/a_i ou f_i/a_i, autrement dit la hauteur de ce rectangle sera le rapport de la fréquence observée de ces valeurs et de la différence entre ces bornes. Avant toute construction d'histogramme, il faut regarder si les classes sont d'amplitudes égales ou non. Si les classes sont d'amplitudes égales, alors la hauteur des rectangles est proportionnelle à l'effectif n_i ou à la fréquence relative f_i.

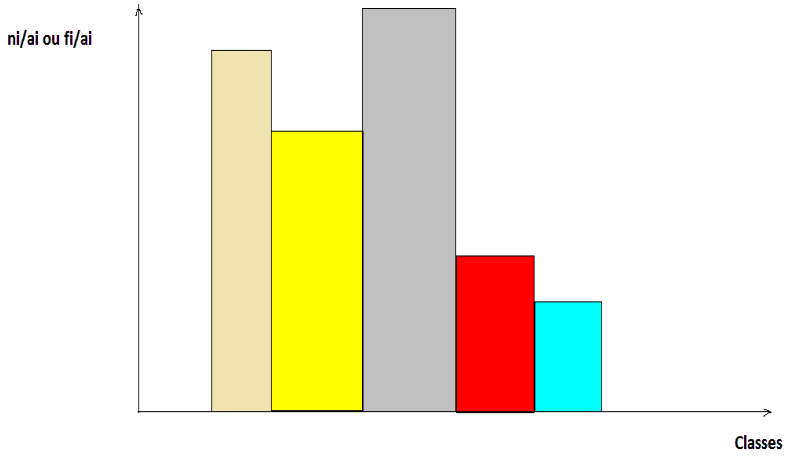


FIG. 3.5 – Histogramme

3.5.2 Courbe cumulative

La courbe cumulative ou courbe des fréquences cumulées est la représentation graphique des fréquences cumulées. Plus précisément, la courbe cumulative est la représentation graphique de la proportion $F(t)$ des individus de la population dont le caractère prend une valeur inférieure à t. Cette fonction, appelée fonction cumulative ou fonction de répartition[1]. La courbe cumulative est une succession de segments de droite reliant le point

[1] fonction de répartition, est :
1. définie pour tout $t \in \mathbb{R}$.
2. croissante (mais non strictement croissante)

$(a_{i-1};\ F(a_{i-1}))$ au point $(a_i;\ F(a_i))$.

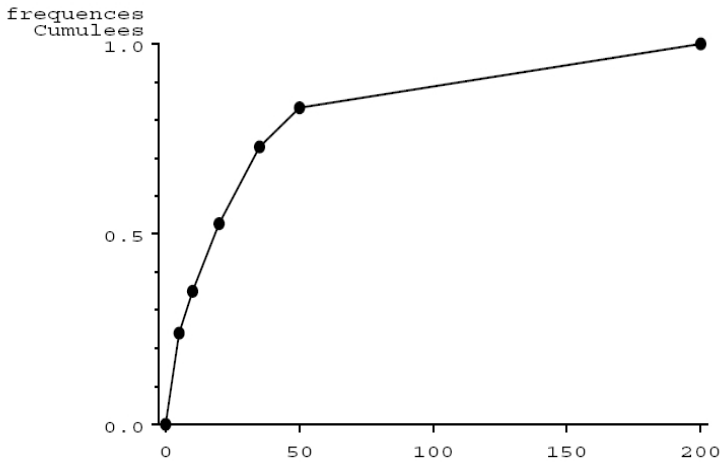


FIG. 3.6 – Courbe cumulative

3. nulle pour t inférieur à $\min_{1\leq i\leq n} x_i$.

4. égale à 1 pour t au moins égal à $\max_{1\leq i\leq n} x_i$.

3.6 Exercice

Sur une population bien portantes issues d'une certaine région, on a dosé le taux de cholestérol, exprimé en cg/l (xi) et on a obtenu les résultats suivants :

Classes	[84, 96[	[96, 108[	[108, 120[	[120, 132[	[132, 144[	[144, 156[	[156, 168[
Effectif n_i	2	8	16	19	2	2	1

1. Quel est le nombre d'individus ?
2. Quel est le caractère étudié ?
3. Déterminer le mode.
4. Déterminer les quartiles Q_1 et Q_3.
5. Déterminer la médiane.
6. Déterminer la moyenne arithmétique.
7. Déterminer l'écart-type.
8. Conclure.

Réponse :

1) La taille de l'échantillon vaut : N = 50

2) Le caractère : Le taux de cholestérol.

Tableau statistique

Classes	n_i	Centres	f_i	N_i	$f_i c_i$	$f_i c_i^2$
$[84, 96[$	02	90	0,040	2	3,600	324
$[96, 108[$	8	102	0,160	10	16,320	1664,640
$[108, 120[$	16	114	0,320	26	36,480	4158,720
$[120, 132[$	19	126	0,380	45	47,788	6032,880
$[132, 144[$	2	138	0,040	47	5,520	761,760
$[144, 156[$	2	150	0,040	49	6,000	900
$[156, 168[$	1	162	0,020	50	3,240	524,880
Total	50		1		119,040	14366,880

3) La classe modale : $Mo \in [120,\ 132[$

$m_1 = 19 - 16 = 3,\ a_i = 132 - 120 = 12,\ m_2 = 19 - 2 = 17$

$$Mo = 120 + 12\left(\frac{3}{3+17}\right) = 121.800.$$

4) La médiane :

$N/2 = 25$ alors la classe qui contient la médiane est $[108,\ 120[$

par interpolation linéaire on trouve

$$M = 108 + (120\ - 108) \times \left[\frac{25-10}{26-10}\right] = 119.25.$$

5) Les quartiles :

$Q_1 = X_{[N/4]} = 12.5 \implies Q_1 \in [108,\ 120[\,,\ alors$

$$Q_1 = 108 + (120\ -108) \times \left[\frac{12.5-10}{26-10}\right] = 109.88.$$

$Q_3 = X_{[3N/4]} = 37.5 \implies Q_3 \in [120,\ 132[\,,\ donc$

$$Q_1 = 120 + (132\ -120) \times \left[\frac{37.5-26}{45-26}\right] = 127.26.$$

6)La moyenne arithmétique :

$$\overline{X} = \sum_{i=1}^{7} f_i c_i = 119.040.$$

7) L'écart-type :

$var(x) = \sum_{i=1}^{7} f_i c_i^2 - (\overline{x})^2 = 14366.880 - (119.040)^2 = 196.358.$

donc

$\sigma(x) = \sqrt{var(x)} = \sqrt{196.358} = 14.012.$

8) Conclusion :

Comme $\overline{x} = 119.040$ et $\sigma(x) = 14.012$, alors l'écart-type est assez petit par rapport a la moyenne, donc on conclure que l'échantillon est bien homogène.

Notation et abréviations

Les différentes abréviations et notations utilisées tout au long de ce polycopié sont expliquées ci-dessous.

Notations	**Explication**
$[.]$	: partie entière
$\sum_{i=1}^{N}$	: somme pour i variant de 1 à N
Mo	: mode
Me	: médiane
$Cov(X,Y)$	: covariance entre X et Y
$\overline{x}$	: moyenne d'une série statistique X
$Var(X)$	: variance de X
$\sigma(X)$	: ecart-type
$\|.\|$	: valeur absolue
$\mathbb{R}$	: ensemble des réels
$\mathbb{N}$	: ensemble des entiers naturels
$\mathbb{N}^*$	: ensemble des entiers naturels non nuls
n_i	: éffectif ou fréquence absolue
N_i	: effectif cumulé croissant
f_i	: fréquence relative
F_i	: fréquence relative cumulée croissante
i.e.	: c'est-a-dire