

Le cours -1- de Biostatistique

1 Statistique descriptive univariée (a une variable)

1.1 Définitions et vocabulaires

Definition 1.1 *La statistique est une méthode scientifique qui consiste à réunir des données chiffrées sur des ensembles nombreux, puis à analyser, à commenter et à critiquer ces données.*

****** L'analyse statistique se subdivise en deux parties :

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

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

Definition 1.2 *La biostatistique c'est la statistique appliquée à la biologie.*

Exemple 1.1 *Etude descriptive des poids des étudiants inscrits en Master 1 de biologie à l'université de Mila.*

* Comme toute science, la statistique fait appel à un vocabulaire spécialisé :

- POPULATION : La collection d'objets ou de personnes étudiées (élèves, habitants, voitures ...).
- INDIVIDU : élément de la population étudiée. (un élève, un habitant, une voiture ...).
- ECHANTILLON : partie de la population étudiée. Nombre d'individus dans un échantillon noté n est appelé taille de l'échantillon
- LA TAILLE : représente le nombre d'individus d'un échantillon ou d'une population. Elle est symbolisée par " n " dans le cas d'un échantillon et par " N " dans le cas d'une population.
- VARIABLE (CARACTERE) : propriété commune aux individus de la population, que l'on veut étudier.

****** Un caractère peut être :

a)- qualitatif : on ne peut associer ni valeur numérique ni un ordre naturel (type de voiture, couleur des cheveux, ...).

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

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

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

- LES MODALITES : les différentes manières d'être que peut présenter un caractère.

Exemple 1.2 *Nous résumons les différents concepts dans cet exemple :*

- Population : l'ensemble des tous les employés d'une usine.

- Individu : chaque employé de l'usine.

- Caractère : le salaire, l'état matrimonial, le nombre d'enfants, ... etc.

- Les modalités du caractère : marié, célibataire, divorcé et veuf sont les modalités de l'état matrimonial, par exemple.

Nombre d'enfant	0	1	2	3	4	5	total
Nombre de famille effectif : n_i	16	18	14	11	3	2	64
Fréquence relative : f_i	0.250	0.281	0.218	0.172	0.047	0.031	1

Table 1

Notations :

- Effectif ou fréquence absolue : (noté n_i) nombre d'apparitions de la valeur associé à un caractère dans un échantillon.
- Fréquence relative : (noté f_i) $f_i = \frac{n_i}{n}$
- Série statistique : l'ensemble des valeurs du caractère avec en regard, les fréquences absolues ou relative correspondante.
- Les effectifs (la fréquence) cumulés (es) croissant (es) : la fréquence (effectif) cumulé croissante d'une valeur X_i est égale à la somme des effectifs ou fréquences des valeurs inférieure ou égale X_i . On note $n_i^\uparrow$ ($f_i^\uparrow$).
- Les effectifs (la fréquence) cumulés (es) décroissant (es) $n_i^\downarrow$ ($f_i^\downarrow$): la fréquence (effectif) cumulé décroissante d'une valeur X_i est égale à la somme des effectifs ou fréquences des valeurs supérieures ou égale X_i .

Remark 1.1 On a toujours

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

$$\sum_{i=1}^k f_i = 1, 0 \leq f_i \leq 1$$

* Représentation des données :

Après la réalisation d'une étude statistique, les données recueillies peuvent être résumées et représentées sous forme de

- Tableaux statistiques
- Graphes
- Paramètres statistiques

1. Représentation sous forme des tableaux statistiques

1)- Série statistique dans le cas d'une variable quantitative discrète (discontinu) (Table 1) :

Exemple 1.3 • Population étudiée : les familles

- L'échantillon sur lequel porte l'étude : familles d'un immeuble ; $n=64$.
- Le caractère étudié est le nombre d'enfants par famille. C'est un caractère quantitatif discret.

2)- Série statistique dans le cas d'une variable quantitative continue (Table 2) :

Exemple 1.4 poids de 122 nouveau-nés.

Les classes	Centre de classe c_i	Effectif n_i	Fréquence relative f_i	$n_i^\uparrow$	$f_i^\uparrow$
2.2-2.5	2.35	5	0.040	5	0.040
2.5-2.8	2.65	11	0.090	16	0.13
2.8-3.1	2.95	24	0.196	40	0.366
3.1-3.4	3.25	40	0.327	80	0.693
3.4-3.7	3.55	42	0.344	122	1

Table 2

Groupe SANGUIN	Effectif n_i	Fréquence relative f_i
O	40	0.40
A	43	0.43
B	12	0.12
AB	5	0.05

Table 3

3) Série statistique dans le cas d'une variable qualitative (Table 3) :

Exemple 1.5 *L'analyse du sang de 100 individus à donner les résultats suivants*

2. Représentation graphique d'une série statistique :

1) Caractère discret (discontinu):

- Diagramme en Bâtons (Figure 1): C'est un ensemble de bâtons ayant pour abscisses les valeurs $X_1, X_2, \dots, X_n$ du caractère et en chacun des points d'abscisses x_i une ordonnée proportionnelle à l'effectif n_i de X_i

- Polygone des Fréquences (Figure 2): On l'obtient en joignant par des segments de droite les extrémités des bâtons

2) Caractère continue :

Histogramme (Figure 3): c'est un ensemble de rectangles ayant pour largeur l'amplitude de la classe et pour hauteur l'effectif de la classe

3) Dans le cas d'une variable qualitatives

Les représentations graphiques que l'on rencontre avec les variables qualitatives sont assez nombreuses. Les deux plus courantes, qui sont aussi les plus appropriées (Figure 4), sont :

- Le diagramme en colonnes (en rectangles).
- Le diagramme en secteurs (diagramme en camembert).

Figure 1: diagramme en bâtons

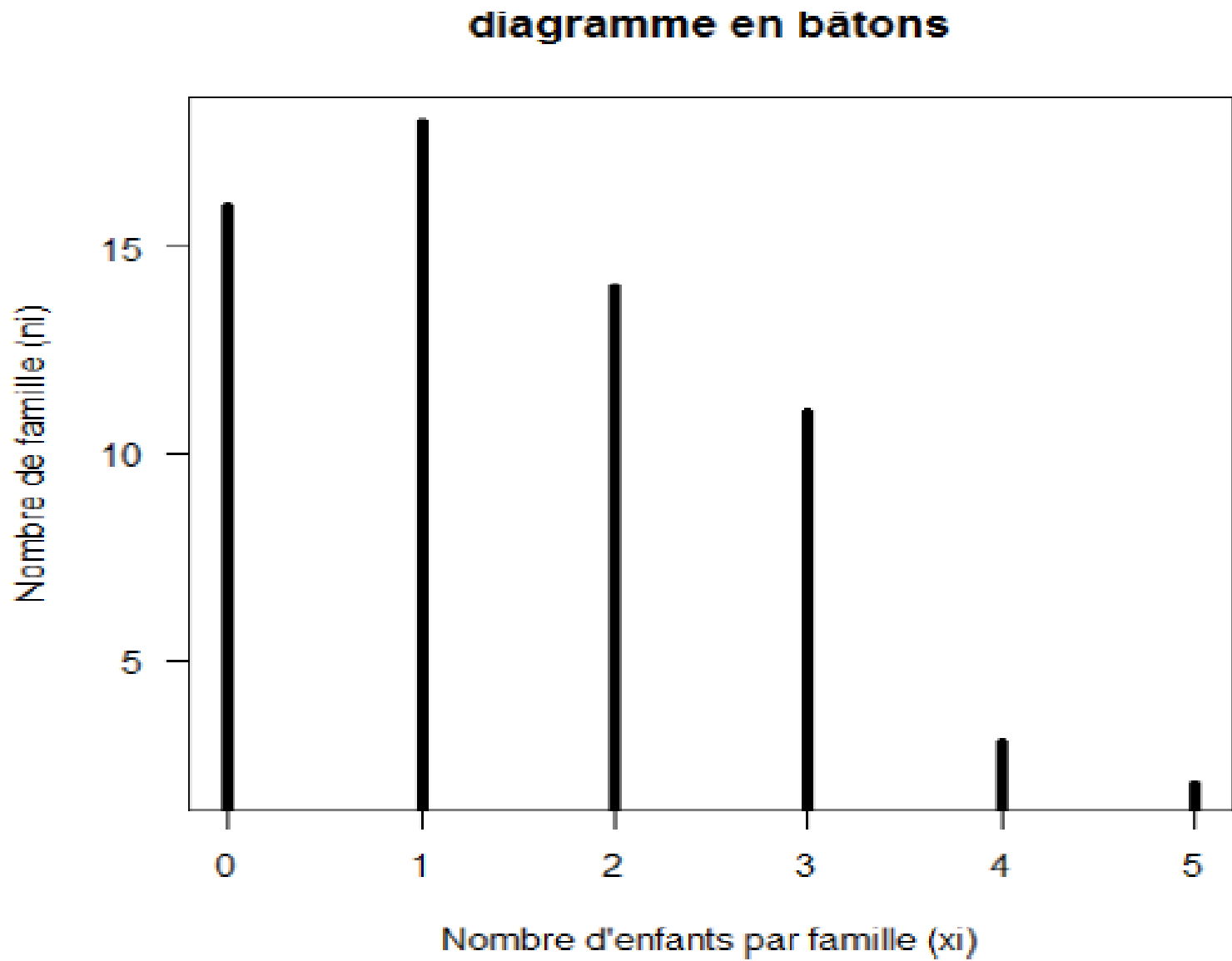


Figure 2: polygone des fréquences

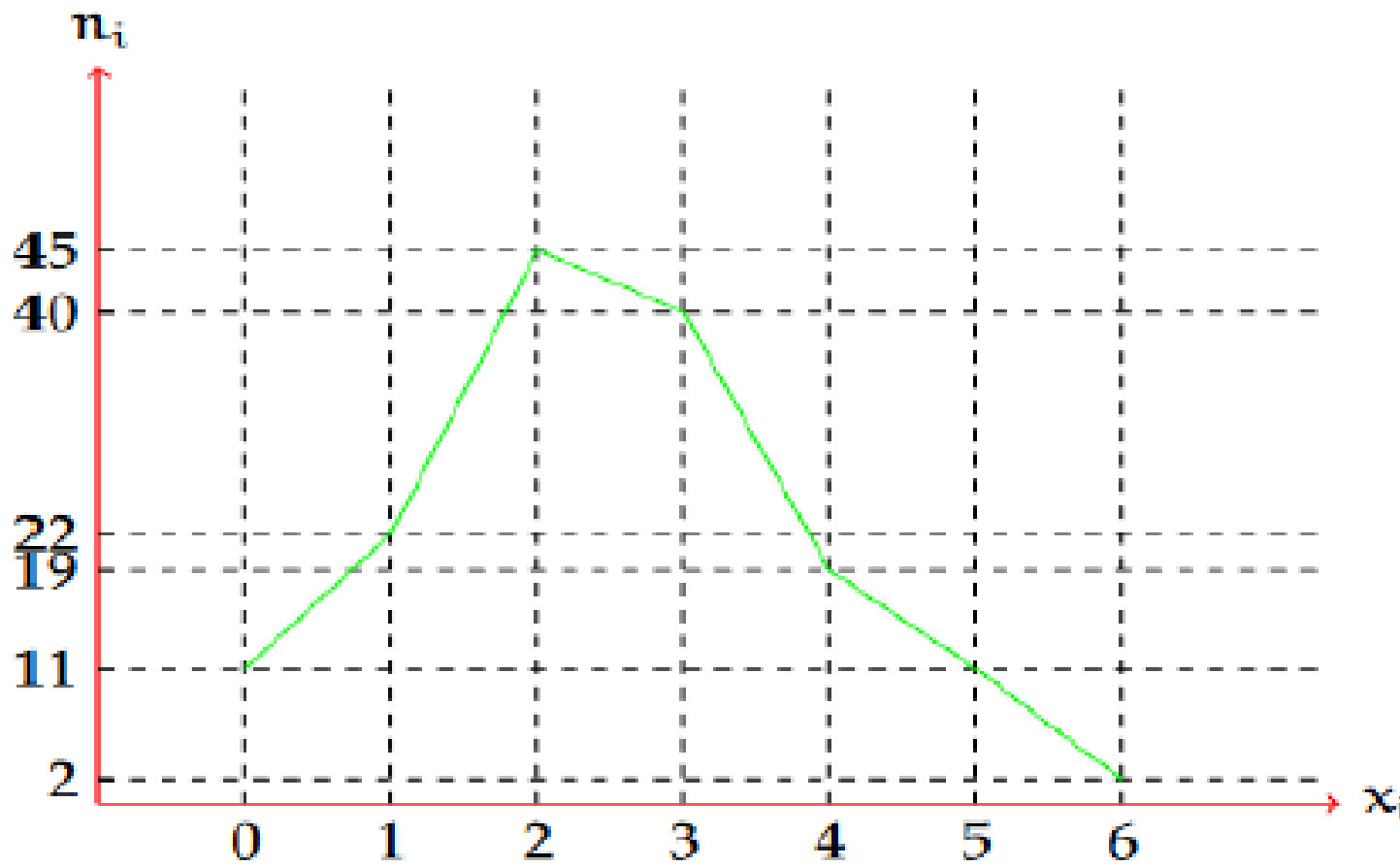


Figure 3: Histogramme

Tracer l'histogramme de l'exemple « poids des nouveaux nés

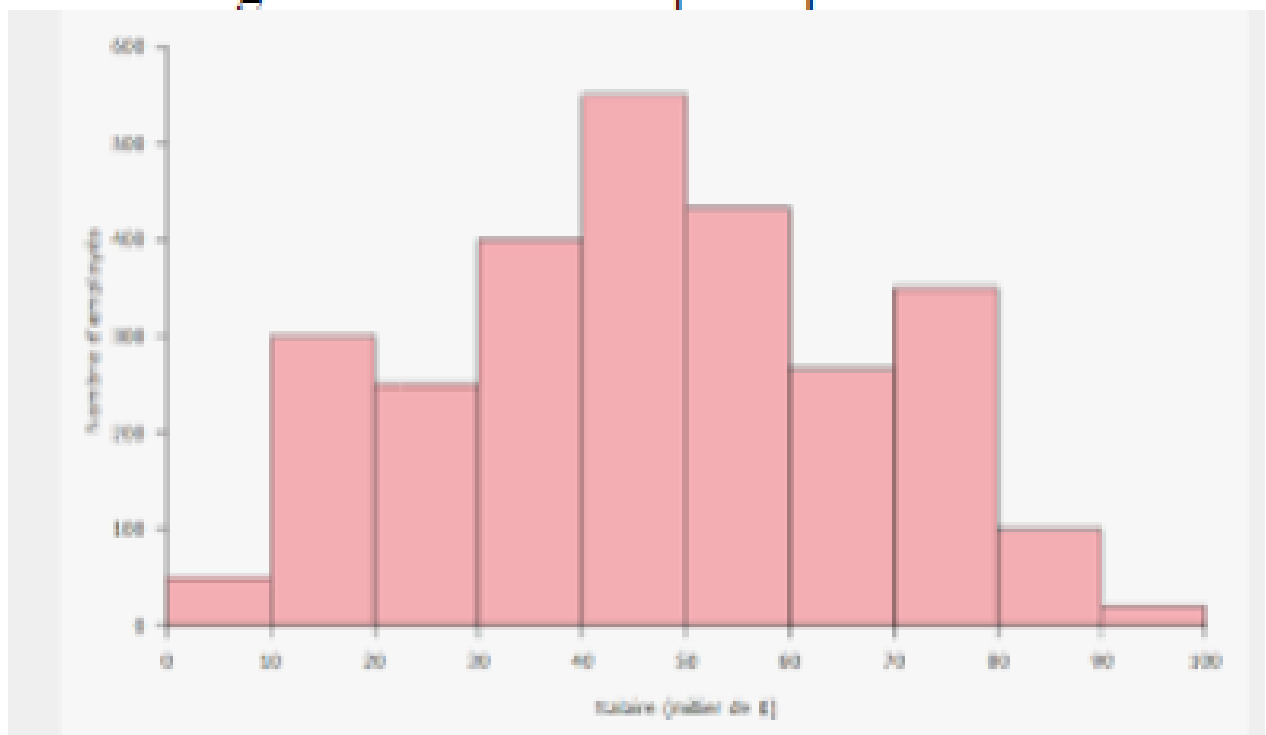
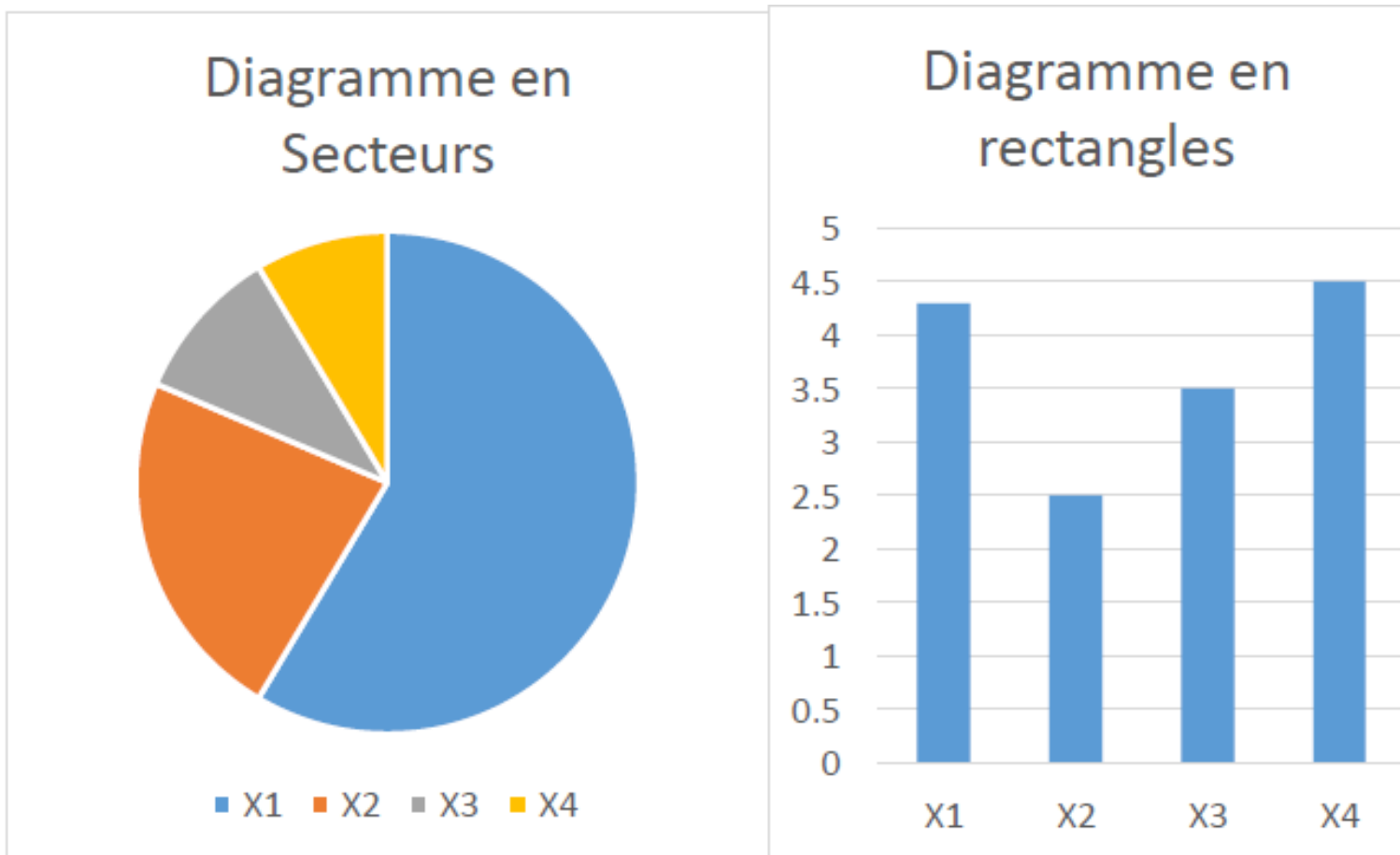


Figure 4: Diagramme en secteurs et en rectangles



2 Les Paramètres (Description Numérique)

La réduction des données permet de condenser les données sous forme des paramètres typiques qui sont les suivants:

- Paramètre de position
- Paramètre de dispersion

1) Paramètre de position

Ce sont des valeurs moyennes qui servent à caractériser l'ordre de grandeur des observations. Ce sont principalement: -La moyenne arithmétique - La médiane - Le mode.

2.1 Moyenne arithmétique :

Soit l'ensemble de mesure d'une même variable $X : X_1, X_2, \dots, X_n$ la moyenne arithmétique notée $\bar{X}$ définie par :

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Exemple 2.1 La moyenne arithmétique des valeurs 8,5,3,6,2

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{8 + 5 + 3 + 6 + 5 + 2}{5} = 4.8$$

- Lorsque les valeurs $X_1, X_2, \dots, X_n$ se répètent respectivement 1, 2, ..., n fois, on obtient "la moyenne arithmétique pondérée" en comptant chaque valeur X_i autant de fois qu'elle se présente: ceci revient à pondérer la valeur X_i

Classe	c_i	n_i
$[1, 2[$	1.5	3
$[2, 3[$	2.5	1
$[3, 4[$	3.5	2

Table 4

par l'effectif n_i qui lui correspond. On aura

$$\bar{X} = \frac{\sum_{i=1}^n x_i n_i}{n}$$

Exemple 2.2 Si les valeurs 8,5,3,6, et 2 se reproduisent respectivement 1,4,2,2,1 fois, la moyenne arithmétique est :

$$\bar{X} = \frac{8 \times 1 + 5 \times 4 + 3 \times 2 + 6 \times 2 + 2 \times 1}{10} = 4.8$$

- Si les valeurs X_i sont groupé dans des classes dont les centres c_i , la moyenne arithmétique est donnée par :

$$\bar{X} = \frac{\sum_{i=1}^n c_i n_i}{n}$$

Exemple 2.3 Table 4

$$\bar{X} = \frac{\sum_{i=1}^n c_i n_i}{3} = \frac{1}{6}(1.5 \times 3 + 2.5 \times 1 + 3.5 \times 2) = \frac{14}{6} = 2.33.$$

2.2 La médiane

a- Cas d'une variable statistique discrète

La médiane est le quantile d'ordre $\frac{1}{2}$. Elle partage donc la série des observations en deux ensembles d'effectifs égaux. Si la série possède un nombre impair de valeurs, la médiane sera $\left(\frac{n+1}{2}\right)^{ieme}$ valeur

Si la série compte un nombre pair de valeurs, la médiane sera la demi somme de la $\left(\frac{n}{2}\right)^{ieme}$ et la $\left(\frac{n}{2} + 1\right)^{ieme}$ valeurs.

Exemple 2.4 On a la série de 15 valeurs suivantes : 12,10,11,4,5,4,2,1,6,8,7,8,9,9,4.

On ordonne d'abord les valeurs : 1,2,4.

On a $n = 15$ impair donc $me = \left(\frac{7+1}{2}\right)^{ieme}$

Exemple 2.5 4, 5, 8, 8, 9, 11, 12, 14, 17, 19. La médiane $me = \frac{9+11}{2} = 10$.

b- Cas d'une variable statistique continue (médiane par interpolation)

$$me = l_1 + \frac{\left(\frac{n}{2} - n_{l_1}^{\uparrow}\right)}{n_0} a.$$

Avec:

- l_1 : la borne inférieure de la classe médiane (classe médiane : la classe correspondante à $n_i^{\uparrow}$ égale ou supérieure à $\frac{n}{2}$).

Classe	n_i
$[1.60 - 1.65[$	3
$[1.65 - 1.70[$ (la classe modale)	8
$[1.70 - 1.75[$	2

Table 5

- a : l'amplitude de la classe médiane.
- n_0 : l'effectif de la classe médiane.
- $n_{l_1}^\uparrow$: l'effectif cumulé jusqu'à l_1 (avant $\frac{n}{2}$).

Remark 2.1 *La médiane Graphiquement (le cas continue)*

La médiane est le point d'intersection de deux graphiques cumulés croissant et décroissant

2.3 Le mode

a- Cas d'une variable statistique discrète

Le mode d'une série est la valeur la plus fréquente de la série

Exemple 2.6 *le mode de la série $\{4, 2, 4, 3, 2, 2\}$ est 2.*

b- Cas d'une variable statistique continue

Dans ce cas le mode se calcule par la forme

$$M_0 = L_i + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) a.$$

où :

- L_i : la borne inférieure de la classe modale (la classe qui a le plus grand effectif).
- Δ_1 : l'effectif de la classe modale - l'effectif de la classe précédente (n_i, n_{i-1}) .
- Δ_2 : l'effectif de la classe modale - l'effectif de la classe suivante (n_{i+1}, n_i) .
- a : l'amplitude de la classe modale.

Exemple 2.7 *Table 5*

$L_i = 1.65, \Delta_1 = 8 - 3 = 5, \Delta_2 = 8 - 2 = 6, a = 1.70 - 1.65 = 0.05$. Donc :

$$M_0 = 1.65 + \left(\frac{5}{5 + 6} \right) \times 0.05 = 1.67.$$

Remark 2.2 *Une variable peut avoir plusieurs modes.*

Exemple 2.8 *le mode de la série $\{4, 2, 4, 3, 2, 2\}$ est 2.*

2) Paramètre de dispersion

2.4 Étendu

On appelle étendu, notée "e", la différence entre la plus grande valeur et la plus petite valeur observée ($e = X_{max} - X_{min}$)

Exemple 2.9 $X = \{6, 6, 7, 7, 8, 9, 9, 10, 10\}$.

$$e = 10 - 6 = 4.$$

2 - Variance et écart type

a) Cas d'une variable discrète

• Variance :

La variance σ^2 (ou $var(x)$) d'une série est la moyenne arithmétique des carrés des écarts des valeurs de la série à leur moyenne

$$var(x) = \sigma^2 = \frac{\sum_{i=1}^n n_i (X_i - \bar{X})^2}{n}.$$

La variance est donnée aussi par

$$var(x) = \left(\frac{1}{n} \sum_{i=1}^n n_i x_i^2 \right) - \bar{X}^2.$$

2.5 Ecart type

L'écart type σ est la racine carrée de la variance i.e, $\sigma = \sqrt{\sigma^2}$.

b) Cas d'une variable continue

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

$\sigma = \sqrt{\sigma^2}$ est appelé l'écart type de la série.