Chap 5 : Statistique descritive

HDHIRI I.
LFSI2

- On appelle Statistique l'ensemble des méthodes et techniques permettant d'analyser des ensembles d'observations (nous parlerons de données).
- Les méthodes en question relèvent essentiellement des mathématiques et font largement appel à l'outil informatique pour leur mise en oeuvre.

Population : ensemble concerné par une étude statistique. Si l'on s'intéresse aux notes d'un groupe d'étudiants, ce groupe constitue la population.

Individu (ou unité statistique) : on désigne ainsi tout élément de la population considérée. Dans l' exemple indiqué ci-dessus, un individu est tout étudiant du groupe

échantillon : Sous-ensemble de la population sur lequel sont réalisées les observations.

Caractère (ou variable statistique): c'est la caractéristique observée (age, salaire, sexe, couleur des yeux...).

Modalités Les formes que prend le caractère exple: masculin, féminin

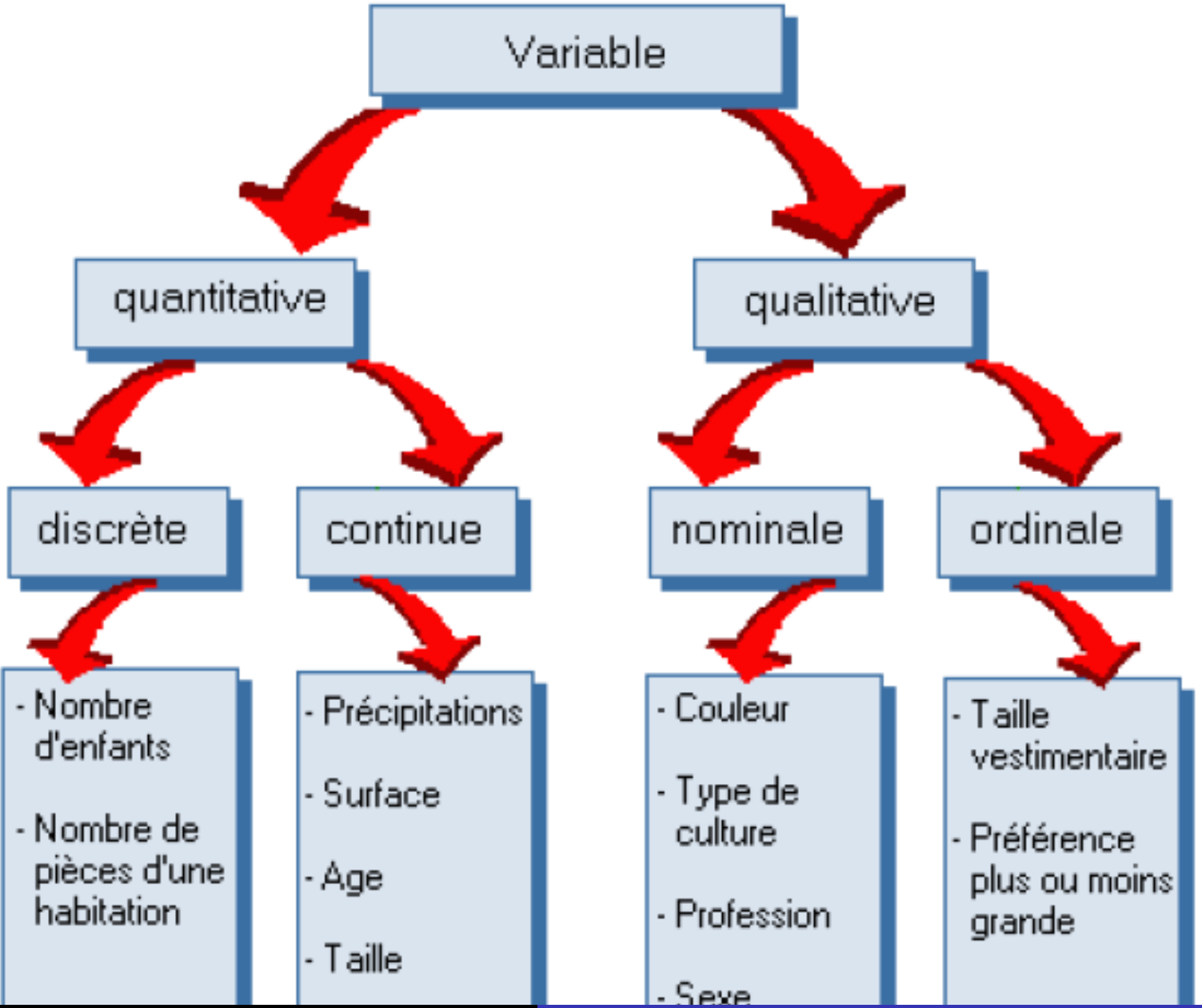
Données (statistiques) l'ensemble des individus observés, des variables considérées et les observations de ces variables sur ces individus.

Caractères qualitatifs Ne résultent ni d'une mesure par un instrument ni d'un comptage.

- Nominal : modalités exprimables par des noms et non hiérarchisées (Couleur des yeux, nationalité...)
- Ordinale : traduit le degré d'un état sans que ce degré ne puisse être défini par un nombre. Modalités hiérarchisées(Mention : Passable, Assez bien, Bien, Très Bien)

Caractère quantitatif : mesurables ; résultent d'une mesure ou d'un comptage.

- discret : il peut prendre seulement des valeurs isolées : résulte d'un comptage (Nombre de frères et soeurs, nombre d'années d'étude ...)
- continu : peut prendre n'importe quelle valeur dans un intervalle donné (le poids, l'age ...)


Variable
quantitative
qualitative
discrète
continue
nominale
ordinale
- Nombre d'enfants
- Nombre de pièces d'une habitation
- Précipitations
- Surface
- Age
- Taille
- Couleur
- Type de culture
- Profession
- Sexe
- Taille vestimentaire
- Préférence plus ou moins grande

Exemple introductif On a noté l'age de 10 étudiants de la FSG. les données sont listées comme suit :
22 21 19 21 20 19 20 20 21 20

Ce n'est très commode à lire

Tableaux statistiques: On appelle tableau statistique un tableau dont la première colonne comporte l'ensemble des r observations distinctes de la variable X (modalités) rangées par ordre croissant et non répétées ; nous les noterons $x_i; i = 1; ...; r$. Dans une seconde colonne, on dispose, en face de chaque valeur x_i, le nombre de réplications qui lui sont associées. Ces réplications sont appelées effectifs et notées n_i. On utilise également les fréquences $f_i = \frac{n_i}{n}$.

Dans l'exemple précédant, les modalités sont 19, 20, 21 et 22 ($n = 10$ et $r = 4$)

Modalité : x_i	19	20	21	22	Total
Effectif : n_i	2	4	3	1	$n = n_1 + ... + n_4 = 10$
Fréquence $f_i = \frac{n_i}{n}$	0.2	0.4	0.3	0.1	$f_1 + ... + f_4 = 1$

- Dans le cas des caractères continus, les valeurs sont mises en classes $[x_i, x_{i+1}[$ Leurs valeurs extrêmes sont appelées bornes.
- L'amplitude de la classe Δ = borne supérieure - la borne inférieure.
- Le point central de la classe est situé à mi chemin entre les bornes. $c_i = x_i + \frac{\Delta_i}{2}$
- Dans certains cas, la limite inférieure de la première classe ou supérieure de la dernière classe n'est pas précisée. On parle de classes ouvertes
- En cas de classes d'amplitudes différentes, **la densité de fréquence** $\frac{f_i}{\Delta_i}$ permet de comparer les effectifs ou les fréquences d'une classe à l'autre.

Exemple : les tranches de revenus dans un groupe de 100 salariés

Classe $[x_i, x_{i+1}[$	[0,500[	[500,1000[	[1000,2000[	[2000,3000[
Amlitude Δ_i	500	500	1000	1000
Centre $c_i = x_i + \frac{\Delta_i}{2}$	250	750	1500	2500
Effectif : n_i	14	38	37	11
Densité d'effectif $d_i = \frac{n_i}{\Delta_i}$	0.28	0.076	0.037	0.011
Fréquence $f_i = \frac{n_i}{n}$	0.2	0.4	0.3	0.1
Densité de fréquence $d_i = \frac{f_i}{\Delta_i}$	0.28%	0.076 %	0.037 %	0.011 %

Definition

L'effectif cumulé d'une modalité x_i est le nombre d'individus de la population présentant une modalité d'indice inférieur ou égal à i.

La fréquence cumulée d'une modalité x_i est la propotion d'individus de la population présentant une modalité d'indice inférieure ou égal à i.

Exemple :

Modalité : x_i	19	20	21	22	Total
Effectif : n_i	2	4	3	1	$n = n_1 + ...n_4 = 10$
Effectif cumulé $N_i = \frac{n_i}{n}$	2	6	9	10	$N_1 + ...N_4 = 10$
Fréquence cumulée $F_i = \frac{n_i}{n}$	0.2	0.6	0.9	1	$F_1 + ...F_4 = 1$

Exemple de caractère qualitatif: Répartition des groupes sanguins dans un groupe de 100 personnes ($n = 100$)

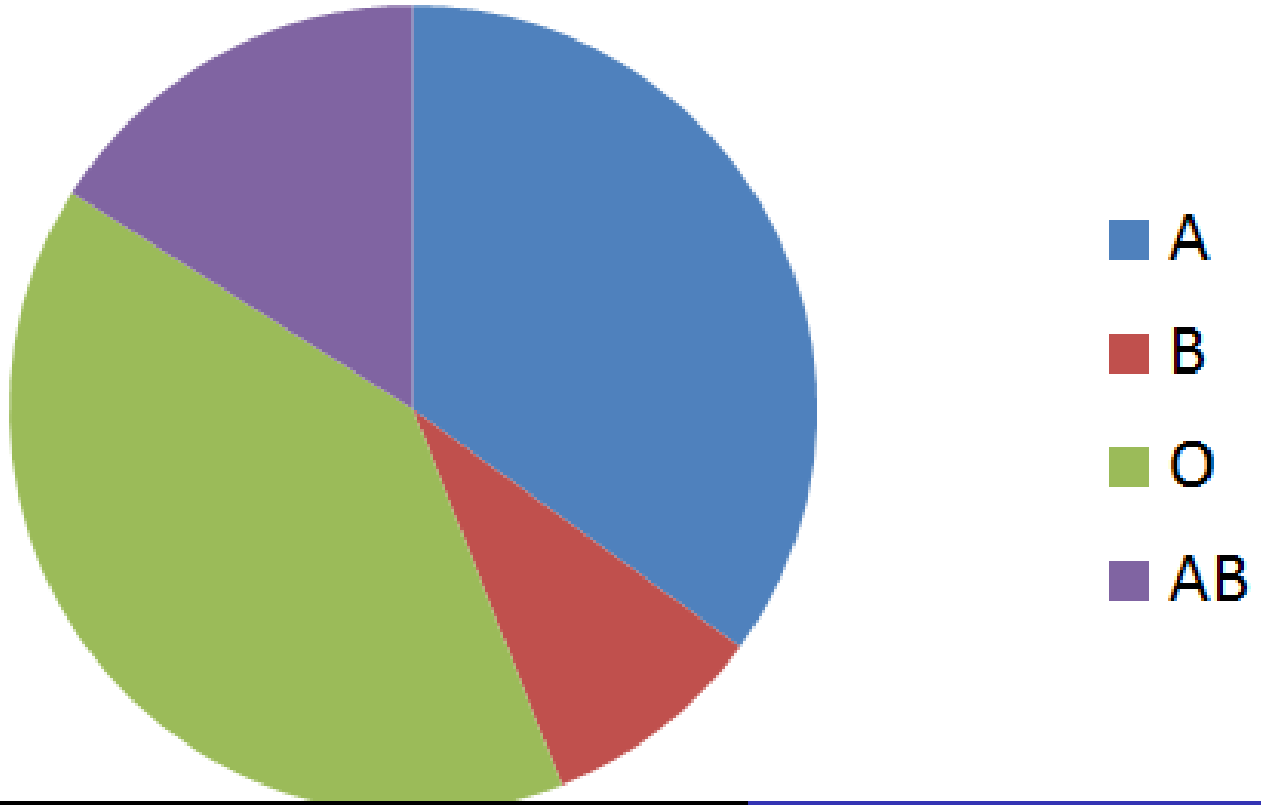
Cathégorie	A	B	AB	o	Total
Effectif : n_i	35	9	16	40	$n = n_1 + ... + n_4$
Fréquence $f_i = \frac{n_i}{n}$	35%	9%	16%	40%	$f_1 + ... + f_4 = 1$
Angle $\theta_i = 360.f_i$	126	32.4	57.6	144	$\theta_1 + ... + \theta_4 = 360$

Représentations graphiques des caractères qualitatifs :
Les modalités d'un caractère qualitatif ne sont pas ordonnées, on les représente généralement par des graphiques utilisant des surfaces.

Diagramme sectoriel : la surface attribuée à chaque catégorie est proportionnelle à l'importance de la catégorie dans l'ensemble de la population étudiée. on associe à chaque cathégorie l'angle $\theta_i = 360 \times f_i$.

Exemple : Répartition des groupes sanguins

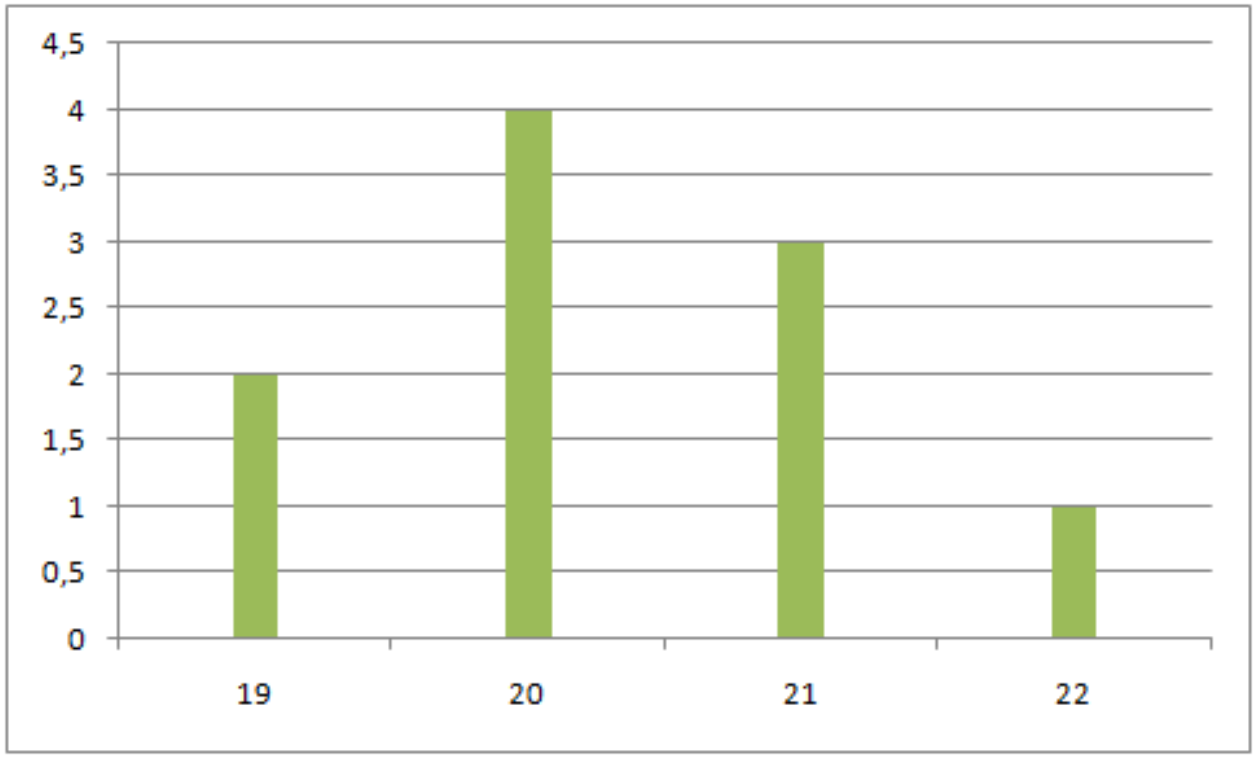
Cathégorie	A	B	AB	o	Total
Fréquence $f_i = \frac{n_i}{n}$	35%	9%	16%	40%	$f_1 + ... + f_4 = 1$
Angle $\theta_i = 360.f_i$	126	32.4	57.6	144	$\theta_1 + ... + \theta_4 = 360$



Représentation graphique des caractères quantitatifs discrets Un caractère quantitatif discret est représenté par un diagramme en bâtons. On trace parallèlement à l'axe des ordonnées, en regard des x_i qui sont portés en abscisse, un segment de longueur proportionnel à fi ou ni

Exemple : Age des étudiants

Modalité : x_i	19	20	21	22	Total
Effectif : n_i	2	4	3	1	$n = n_1 + ... + n_4 = 10$
Fréquence $f_i = \frac{n_i}{n}$	0.2	0.4	0.3	0.1	$f_1 + ... + f_4 = 1$

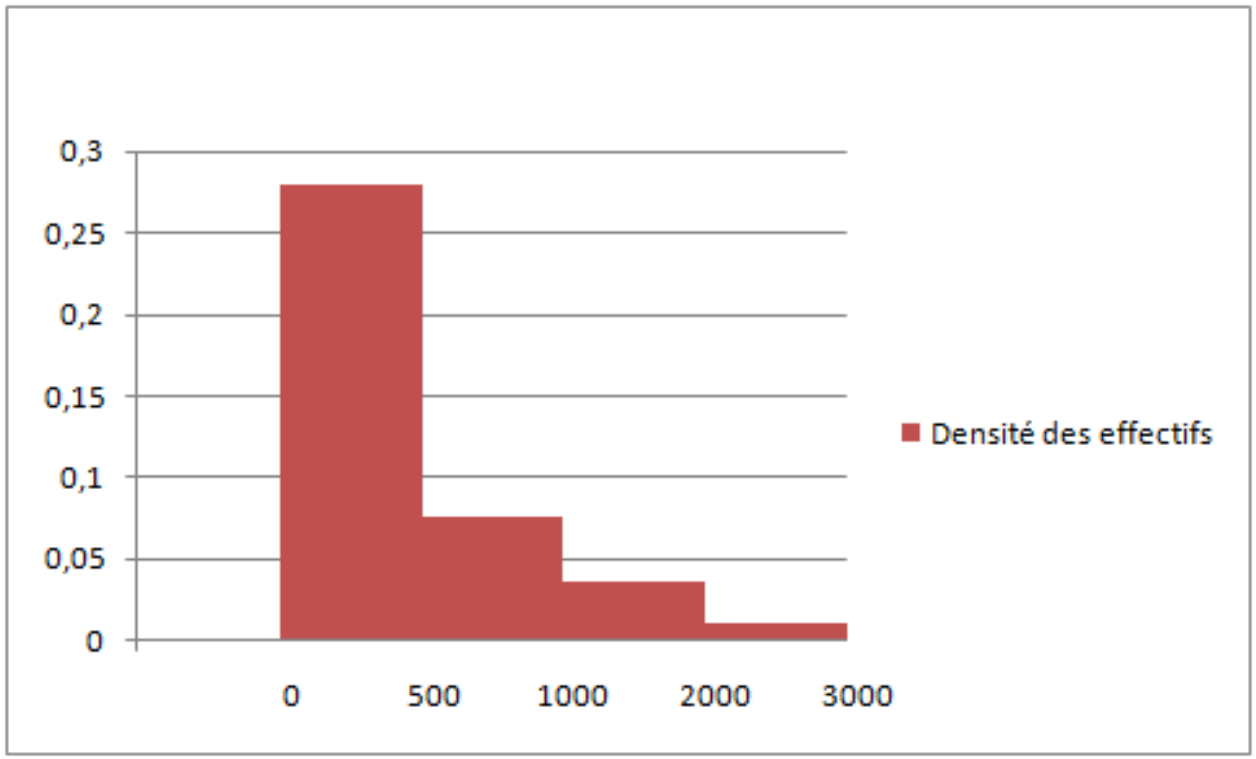


Représentation graphique des caractères quantitatifs continus :

Ce type de caractère est réprésenté à l'aide d'un **histogramme** Composé de rectangles ayant comme base l'intervalle de classe et comme hauteur la densité d'effectif ou de fréquence

Exemple : Salaires de 100 employés

Classe $[x_i, x_{i+1}[$	[0,500[	[500,1000[	[1000,2000[	[2000,3000[
$d_i = \frac{n_i}{\Delta_i}$	0.28	0.076	0.037	0.011



Indicateurs statistiques

1. La médiane Me: C'est la valeur de la variable qui partage la population statistique étudiée en deux effectifs égaux, les individus étant ordonnés selon les valeurs de la variable.
   - Ce sera donc la valeur de la variable telle que 50% de la population se situe au-dessus et 50% se situe en dessous.
   - la médiane est alors la valeur de la variable à laquelle est associée une fréquence cumulée de 50%

Données brutes : On considère la série **ordonnée**

$$x_1 \quad x_2 \quad \dots \quad x_n$$

$(x_1 \leq x_2 \dots \leq x_n)$.

- n impair : $Me = x_{\frac{n+1}{2}}$
- n pair : $Me = \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n+2}{2}})$.

Exemple : Donner la médiane des séries statistiques suivantes :

1. $$7 \quad 6 \quad 5 \quad 6 \quad 8 \quad 5 \quad 4 \ (n = 7)$$
2. $$11 \quad 13 \quad 6 \quad 9 \quad 6 \quad 8 \quad 7 \quad 9 \ (n = 8)$$

On ordonne les données

1. $4 \quad 5 \quad 5 \quad 6 \quad 6 \quad 7 \quad 8$ Donc $Me = x_4 = 6$
2. $6 \quad 6 \quad 7 \quad 8 \quad 9 \quad 9 \quad 11 \quad 13$ Donc
   $Me = \frac{1}{2}(x_4 + x_5) = \frac{1}{2}(8 + 9) = 8.5$

Données classées dans un tableau :

- Pour une variable discrète, on prend la valeur à laquelle est associée une fréquence cumulée de 0.5 Quand la médiane ne tombe pas sur une valeur exacte de la variable, par convention on retient la valeur de la variable immédiatement supérieure.
- Quand la variable est continue, le calcul se fait par approximation : on traite les variables par interpolation linéaire comme si les effectifs étaient uniformément répartis à l'intérieur d'une classe.

Exemple : Age des étudiants

Modalité : x_i	19	**20**	21	22	Total
Fréquence cumulée F_i	0.2	**0.6**	0.9	1	$F_1 + ...F_4 = 1$

Me $= 20$.

2. Le Mode C'est la modalité qui admet l'effectif le plus élévé. Une série peut etre unimodale (un seul mode) ou plurimodale (plusqu'un mode).

3. Les quantiles
   - Les quartiles sont les valeurs de la variable qui partagent la population en 4 groupes de même effectif. $Q_3 - Q_1$ est dit intervalle inter-quartile.
   - Les déciles sont les valeurs de la variable qui partagent la population en 10 groupes de même effectif.
   - Les déciles sont les valeurs de la variable qui partagent la population en 100 groupes de même effectif.

Exemple : on considère la serie suivante représentant la note de 19 étudiants :

11 – 14 – 08 – 17 – 13 – 14 – 11 – 16 – 11 – 10 – 02 – 09 – 12 – 15 – 13 – 16 – 12 – 10 – 09

On ordonne les données :

02 – 08 – 09 – 09 – **10** – 10 – 11 – 11 – 11 – **12** – 12 – 13 – 13 – 14 – **14** – 15 – 16 – 16 – 17

- Le mode est 11
- Les quartiles sont : $Q_1 = 10$, $Q_2 = Me = 12$ et $Q_3 = 14$.
- Les déciles sont $D_1 = 08$, $D_2 = 09$,...,$D_5 = Me$, ... $D_9 = 16$.

4. La moyenne arithmétique :

- Données brutes

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

- Données groupées pour caractère discret

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

- Données classées pour caractère continus : Il suffit de remplacer x_i par c_i

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i$$

Exemple : À partir de la distribution statistique des âges du groupe de dix étudiants, on peut se demander quel est l'âge moyen du groupe.

$$22\ 21\ 19\ 21\ 20\ 19\ 20\ 20\ 21\ 20$$

On a

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i = \frac{22+21+19+21+20+19+20+20+21+20}{10} = 20.3$$

Autrement,

Modalité : x_i	19	20	21	22	Total
Effectif : n_i	2	4	3	1	$n = n_1 + ... + n_4 = 10$
$n_i x_i$	38	80	63	22	$\sum_{i=1}^{4} n_i x_i = 203$

Ainsi, $\overline{x} = \frac{203}{10} = 20.3$.

- La médiane n'est pas influencée par les valeurs extrêmes de la variable mais elle se prête mal aux calculs statistiques
- La moyenne est facile à calculer mais elle est fortement influencée par les valeurs extrêmes
- La somme des écarts à la moyenne est nulle: $\sum_{i=1}^{k}(x_i - \overline{x}) = 0$.

4. L'étendue On appelle étendue, la différence entre la plus grande et la plus petite modalité du caractère.
5. L'écart type : s Il mesure l'écart entre les données et leur moyenne
6. La variance : s^2 : Le carré de l'écart type.

- Données brutes

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

- Données groupées pour caractère discret

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i{x_i}^2 - \overline{x}^2$$

- Données classées pour caractère continus :

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(c_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i{c_i}^2 - \overline{x}^2.$$

Exemple : Age des étudiants

Modalité : x_i	19	20	21	22	Total
Effectif : n_i	2	4	3	1	$n = n_1 + ... + n_4 = 10$
$n_i x_i^2$	722	1600	1323	484	$\sum_{i=1}^{4} n_i x_i^2 = 4129$

On a

$$s^2 = \frac{1}{10}\sum_{i=1}^{4} n_i {x_i}^2 - \overline{x}^2 = 4129 - 20.3^2 = 3716.91$$

D'ou

$$s = \sqrt{3716.91} = 60.97$$

Exemple : Salaire des employés

Classe	[0,500[	[500,1000[	[1000,2000[	[2000,3000[	Total
Centre c_i	250	750	1500	2500	-
Effectif : n_i	14	38	37	11	-
$n_i c_i$	3500	28500	55500	27500	115000
$(n_i c_i^2).10^3$	875	21375	83250	68750	144250

- Le salaire moyen est $\overline{x} = \frac{1}{100}\sum_{i=1}^{4} n_i c_i = \frac{144250\,10^3}{100} = 1150$
- La variance est donnée par :
  $s^2 = \frac{1}{100}\sum_{i=1}^{4} n_i {c_i}^2 - \overline{x}^2 = 1442500.$
- L'écart type est $s = \sqrt{1442500} = 1201.04$