Rappels de statistiques descriptives

MEBREK N.

L'analyse des données recueillies est l'un des objectifs de la statistique.

Le vocabulaire de la statistique vient des premières études qui portaient sur la démographie

On effectue généralement des mesures sur les individus qui composent une population.

Ces mesures sont rattachées à des variables ou caractères.

L'ensemble sur lequel porte l'étude s'appelle Population

Un élément de cet ensemble est un individu

Répartition des étudiants selon leur note

[0;4]	20
[4;6]	60
[6;8]	90
[8;10]	100
[10;12]	70
[12;14]	70
[14;16]	40
[16;20]	20

Ici,la variable d'observation (la note) est découpée en classes

À chaque modalité correspond un effectif observé

Deux types de caractères ou variables sont observées

- Les variables continues

  Variables de type quantitatif,
  On peut calculer une moyenne (loyer, taille, âge…)

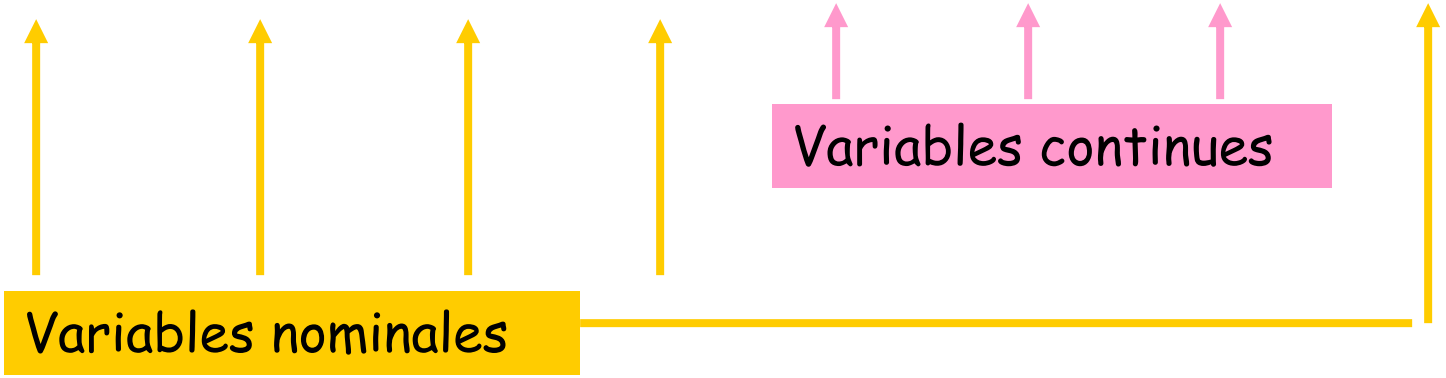
- Les variables nominales

  Variables de type qualitatif, (profession, statuts matrimoniaux, catégories…)

variables

individus

arrdt	pièces	parking	étage	surface	loyer	charges	bailleur
7	1	0	1	37	4100	390	AGF
9	5	0	1	150	13500	1200	AGF
10	3	1	4	78	6120	590	AGF
12	5	1	3	87	7305	958	LOC
14	3	1	1	69	5865	822	SGI
14	3	1	1	68	6644	500	LOC
14	2	1	15	56	5345	788	SOLVEG

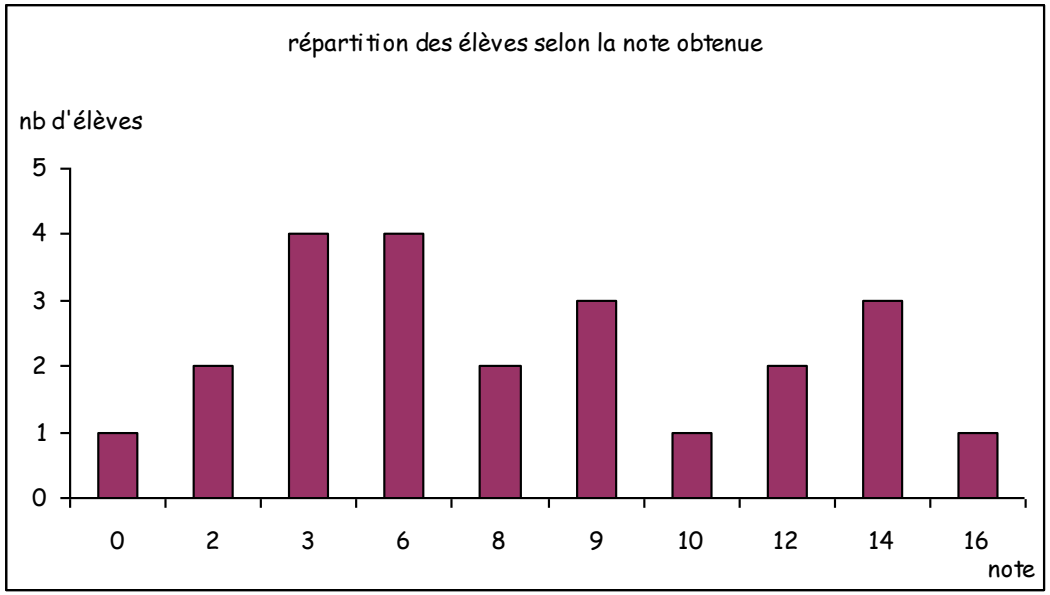


Une série statistique

Variable

note	effectif
0	1
2	2
3	4
6	4
8	2
9	3
10	1
12	2
14	3
16	1

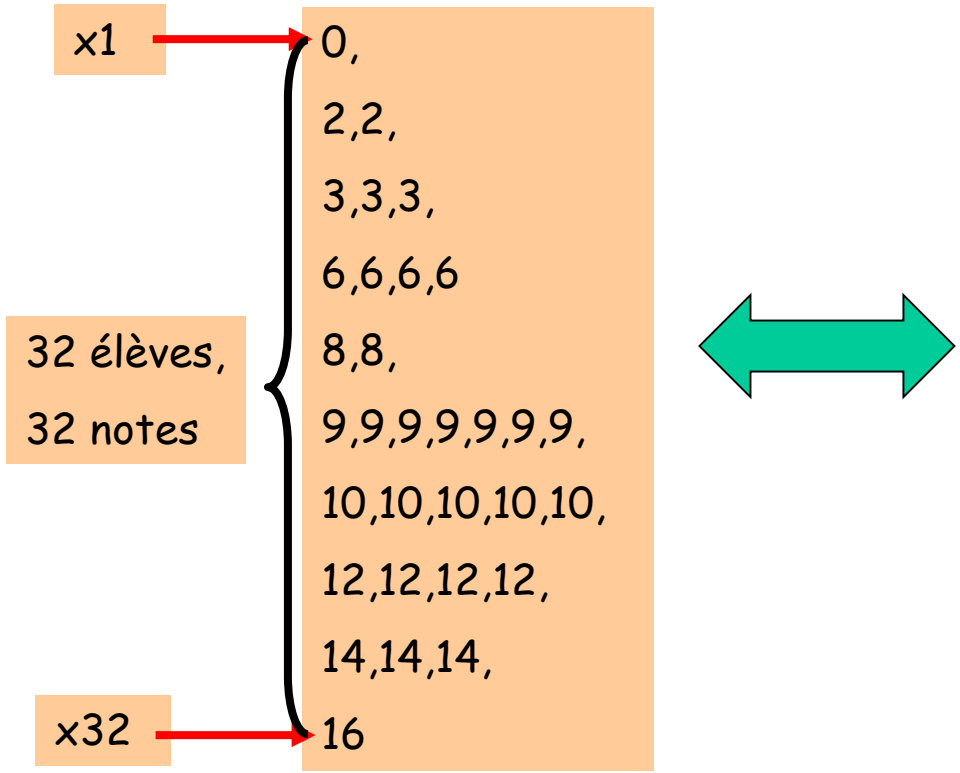
fréquence 3/23 = 0,131



Effectif total : 23

Indicateurs statistiques

Nous considérons une population de n individus pour lesquels une variable xest observée.
On notera (x1) la valeur de x pour l'individu 1,
(xi) la valeur de x pour l'individu i et
(xn) la valeur de x pour l'individu n.



note	effectif
0	1
2	2
3	3
6	4
8	2
9	7
10	5
12	4
14	3
16	1

La médiane :

La médiane est la valeur de la série qui partage la distribution en deux sous-ensembles d'égal effectif.

50% des valeurs sont supérieures à la médiane.
50% des valeurs sont inférieures à la médiane.

Si on range les valeurs selon un ordre croissant, la médiane se trouve au rang

$$i = \frac{n+1}{2}$$

Si n est impair i est un entier et la valeur médiane existe dans la série statistique.

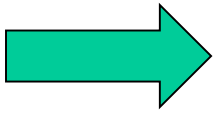
Par contre si n est pair le rang i tombe entre deux valeurs, la médiane correspondra à la moyenne des 2 valeurs qui encadrent ce rang.

Moyenne :

.

La moyenne de X est égale à la somme des Xi divisée par l'effectif N.

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$



$$\frac{x_1 + x_2 + \ldots\ldots + x_n}{N}$$

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i$$

Propriété de la moyenne : $\sum_{i=1}^{n}(x_i - \overline{X}) = 0$

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

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

$$\sum_{i=1}^{n} x_i - n\overline{X} = 0$$

4	-7
7	-4
10	-1
13	2
14	3
18	7
11	

66

Propriété de la moyenne : $\sum_{i=1}^{n}(x_i - \overline{X}) = 0$

La **variance** de x est égale à la somme des carrés des écarts à la moyenne divisée par l'effectif N.

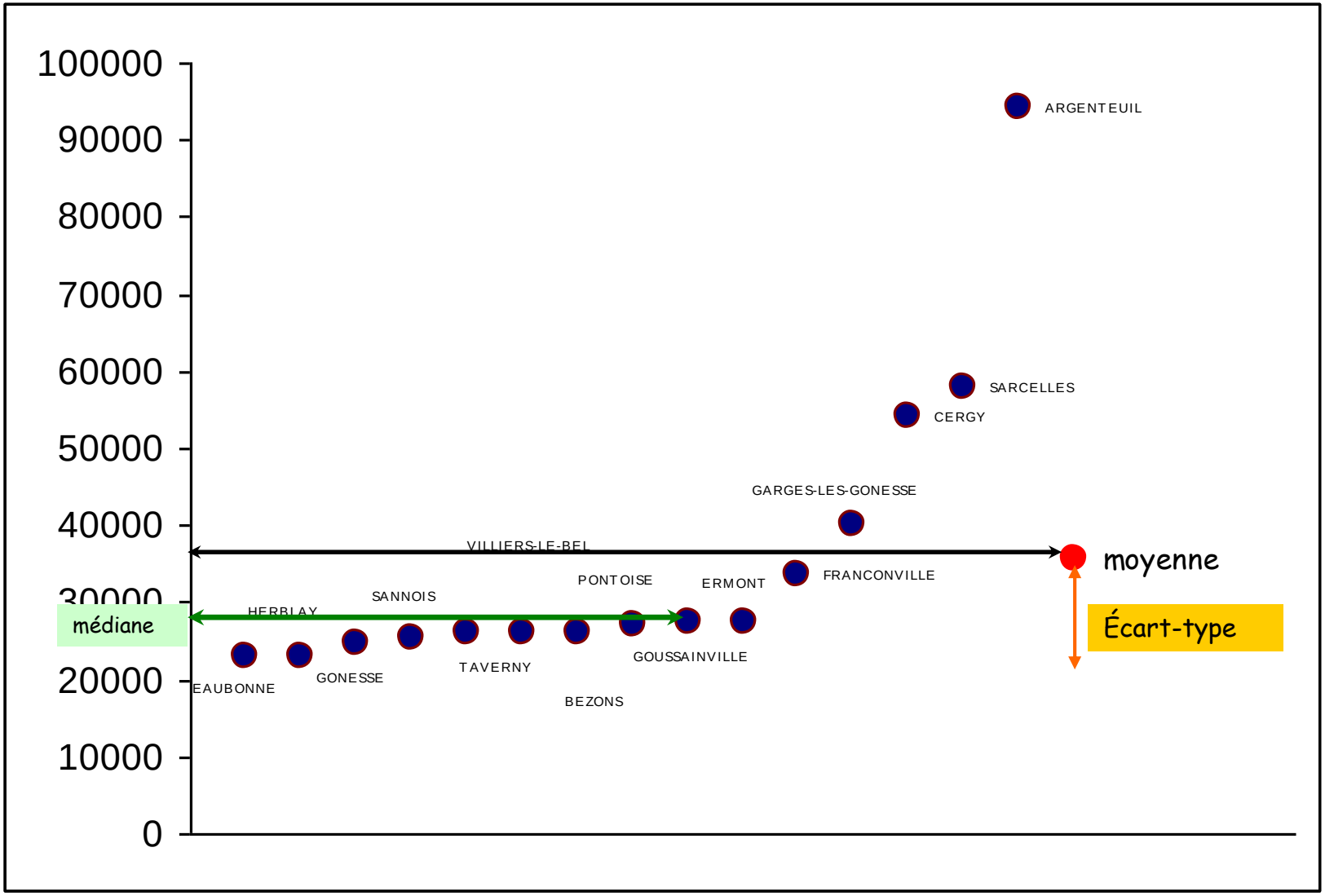
$$\frac{1}{N}\sum_{i=1}^{N}(Xi-\overline{X})^2 = S^2$$

L'écart type d'une variable est égal à la racine carrée de sa variance.

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Xi-\overline{X})^2}=S$$

L'écart type s'exprime dans la même unité que les observations

Moyenne et écart-type d'une série statistique



Code	NOM	PSDC99
95 203	EAUBONNE	22870
95 306	HERBLAY	23081
95 277	GONESSE	24721
95 582	SANNOIS	25331
95 607	TAVERNY	25905
95 063	BEZONS	26087
95 680	VILLIERS-LE-BEL	26089
95 280	GOUSSAINVILLE	27224
95 500	PONTOISE	27418
95 219	ERMONT	27542
95 252	FRANCONVILLE	33494
95 268	GARGES-LES-GONESSE	39963
95 127	CERGY	53995
95 585	SARCELLES	57940
95 018	ARGENTEUIL	94019

Médiane 27 224

Moyenne 35 711,9

Écart-type 18 784,9

coefficient de variation 0,53

écart-type/moyenne

Exércice : Afin de tester le taux de mortalité des nouveaux nés, N femmes ont été questionnées,

pour chacune d'elles on a noté le nombre d'enfants vivants, Les résultats sont dans le tableau suivant:

1	3	0	0	3	2	1	2	3	2
1	1	0	2	1	2	0	3	1	1
2	2	3	2	1	3	2	3	2	0
1	2	4	1	4	4	0	4	2	4
0	2	1	1	0	2	0	1	2	3

1) draisser le tableau des fréquences absolus , fréquences relatives et fréquences cumulées,

2) donner les représentations graphiques des deux fréquences,

3) Calculer la moyenne, la médiane, la variance et l'écart type.

Les quartiles :

Le terme vient de quart. On distingue 3 quartiles Q1, Q2 et Q3. Ce sont des quantiles d'ordre 25%, 50% et 75%. Partageant la série statistique en 4 parties de même taille.

25 % des observations sont inférieures à Q1
50% des observations sont inférieures à Q2
75% des observations sont inférieures à Q3

La médiane est le deuxième quartile ou le quantile d'ordre 50%.

Les déciles

Le terme vient de dix. Les déciles notés D1, D2,….,D9 partagent la série en 10 parties d'égal effectif.

Les centiles

Le terme vient de cent. Les paramètres C1, C2……C99 (99 centiles) d'ordre 1, 2,…99% partagent la série en 100 parties de même taille.

Intervalle interquartile :

C'est un indicateur de dispersion.
Il mesure l'écart entre le $3^{ème}$ et 1^{er} quartile.
Cet intervalle correspond à 50% des observations.
Cette mesure n'est pas sensible aux valeurs éloignées.
On calcule la longueur de l'intervalle : Q3-Q1.

L'intervalle de Kelley :

D9-D1, mesure l'écart pour 80% des observations situées autour de la médiane.

Série en classes

Classes	**Effectifs**
M_1	n_1
M_2	n_2
M_2	n_2
.	.
.	.
$\mathbf{M_i}$	$\mathbf{n_i}$
.	.
.	.
M_k	n_k

Ce tableau renseigne sur l'effectif ou fréquence absolue pour chacune des classes M_i.

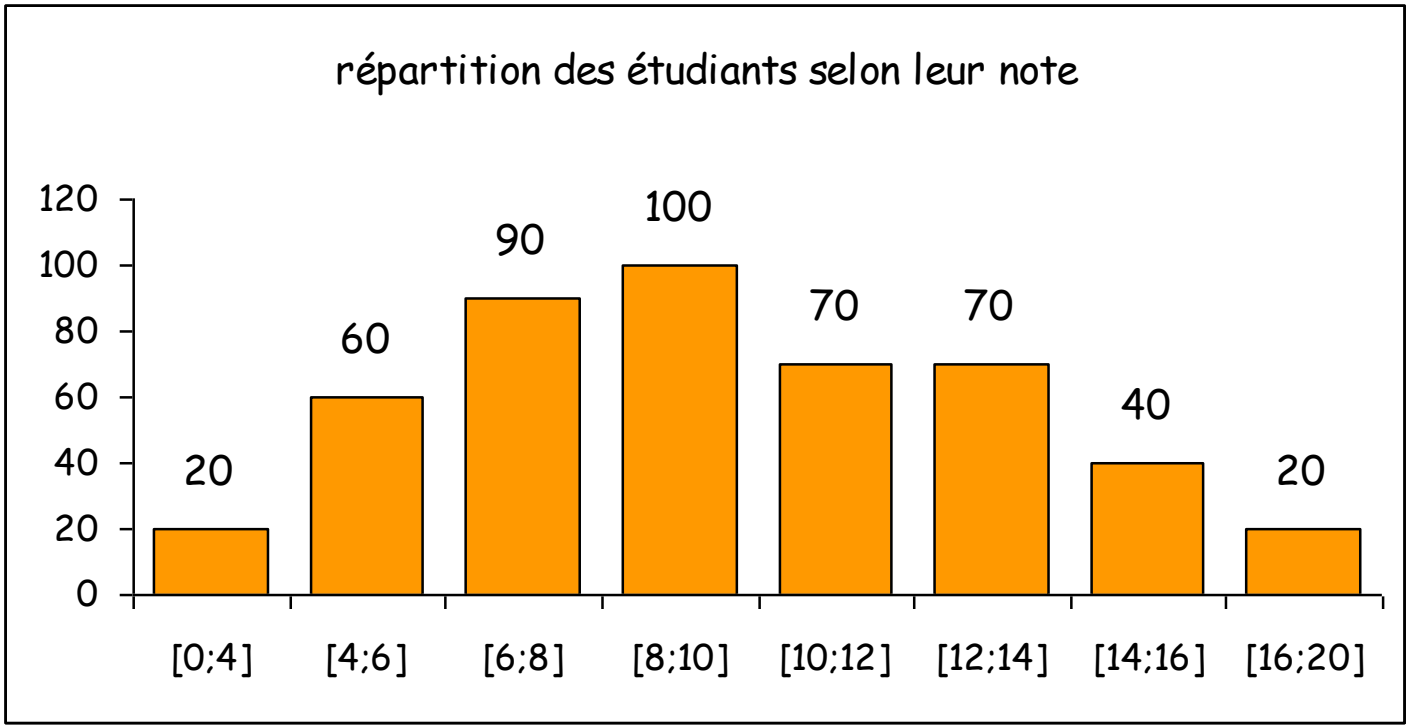
N est l'effectif total
k étant le nombre de classes.

Les n_i représentent les effectifs des différentes classes.

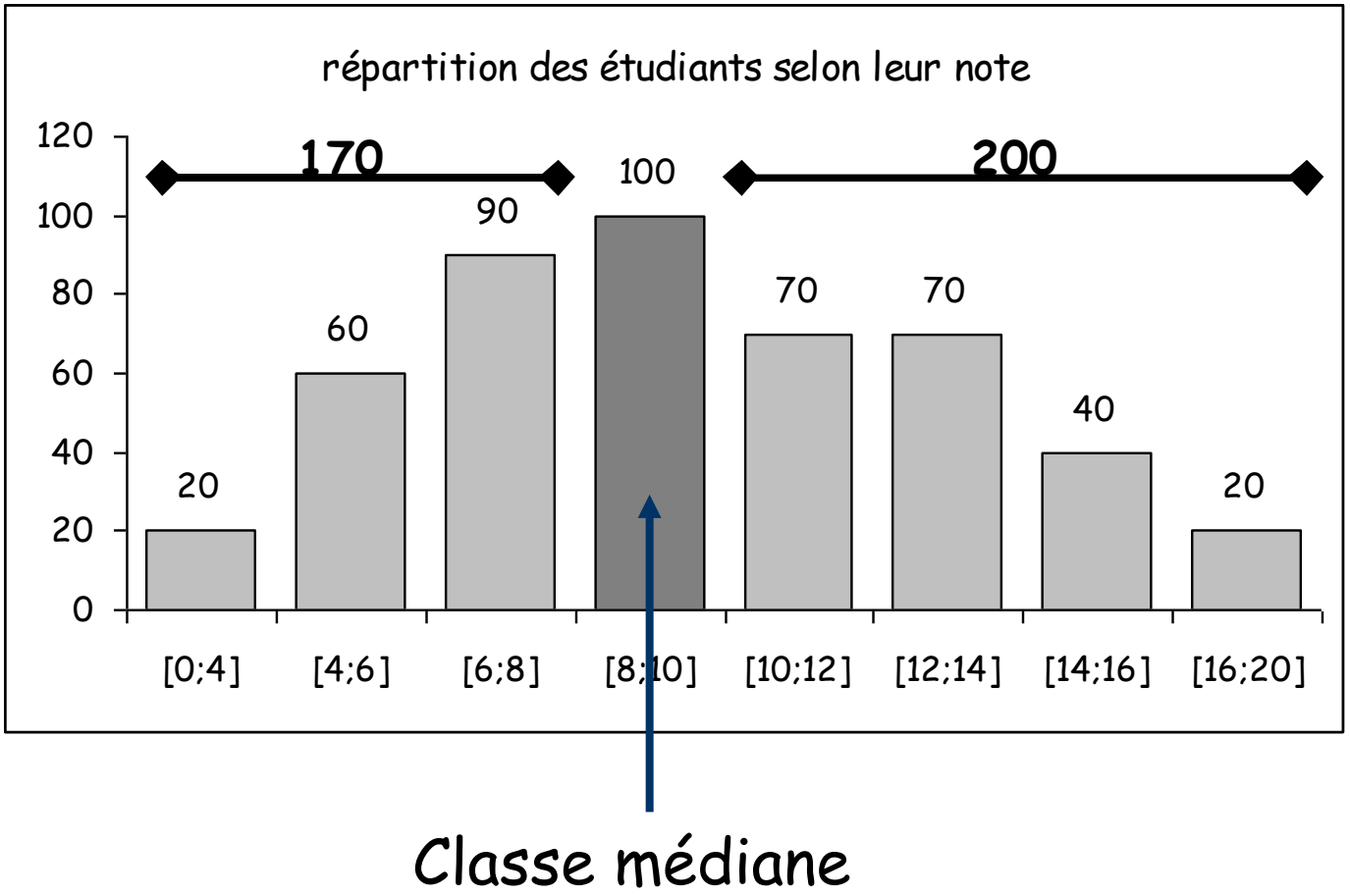
$$N = \sum_{i=1}^{k} n_i$$

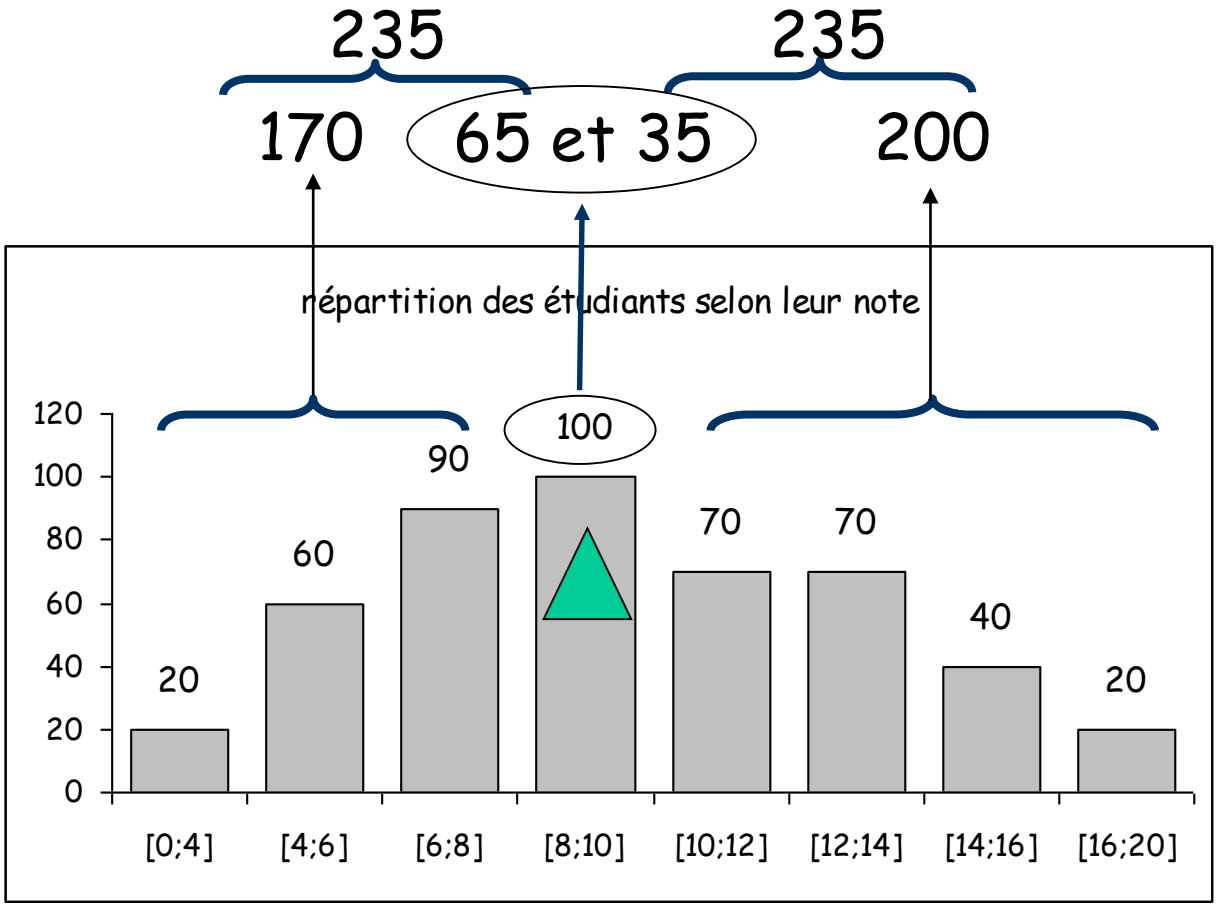
sous Excel deux méthodes sont possibles pour découper une série en classes :
-Utiliser la fonction fréquence
-Le tableau croisé dynamique et la fonction grouper.

Analyse statistique d'une série en classes



Estimation de la médiane sur une série en classes :





On a 65/100 des effectifs de la classe modale qui sont inférieurs au mode estimé,

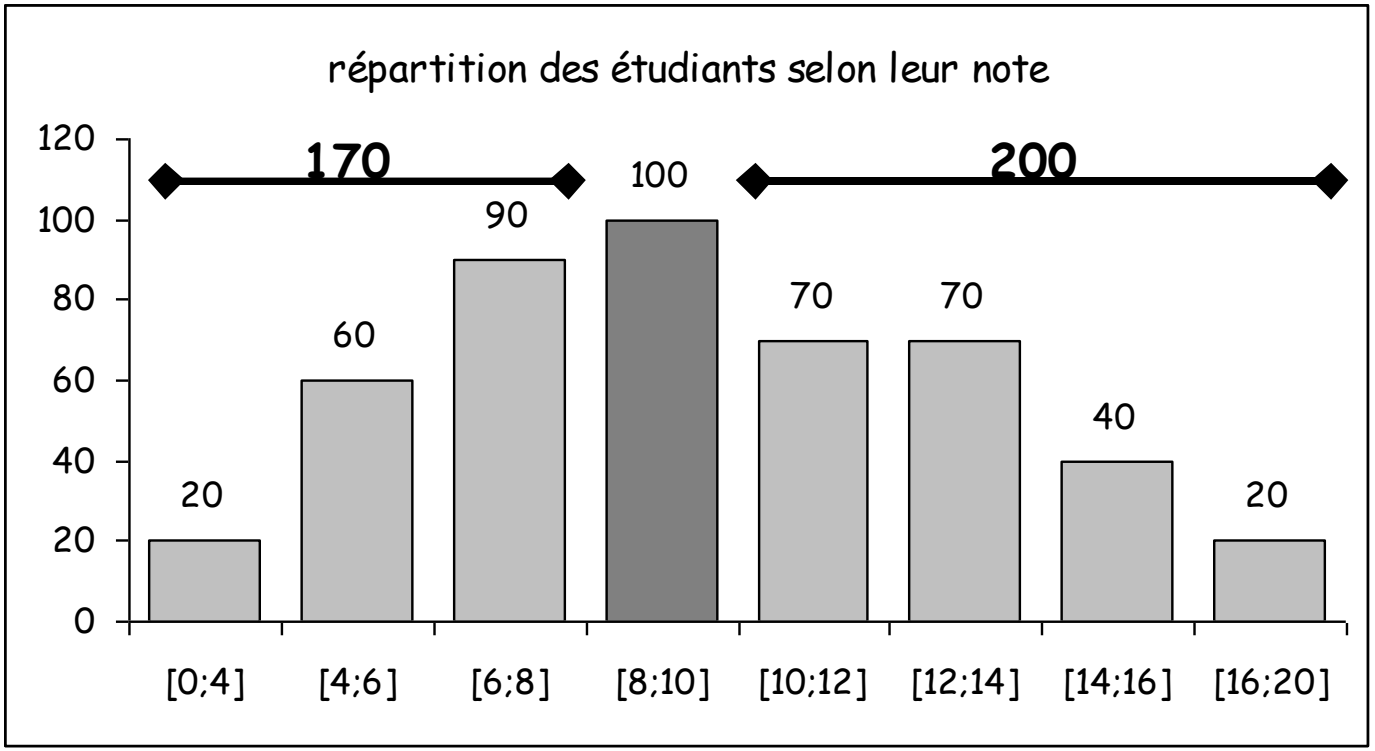
On a 35/100 des effectifs de la classe modale qui sont inférieurs au mode estimé

On divise la classe modale en deux parties (65% et 35% de l'amplitude de la classe) pour trouver le mode estimé.

Ici pour une amplitude de 2, la classe modale est divisée en une sous classe (8 ; 9,3) et une autre (9,3 ; 10). Le mode estimé est donc 9,3.

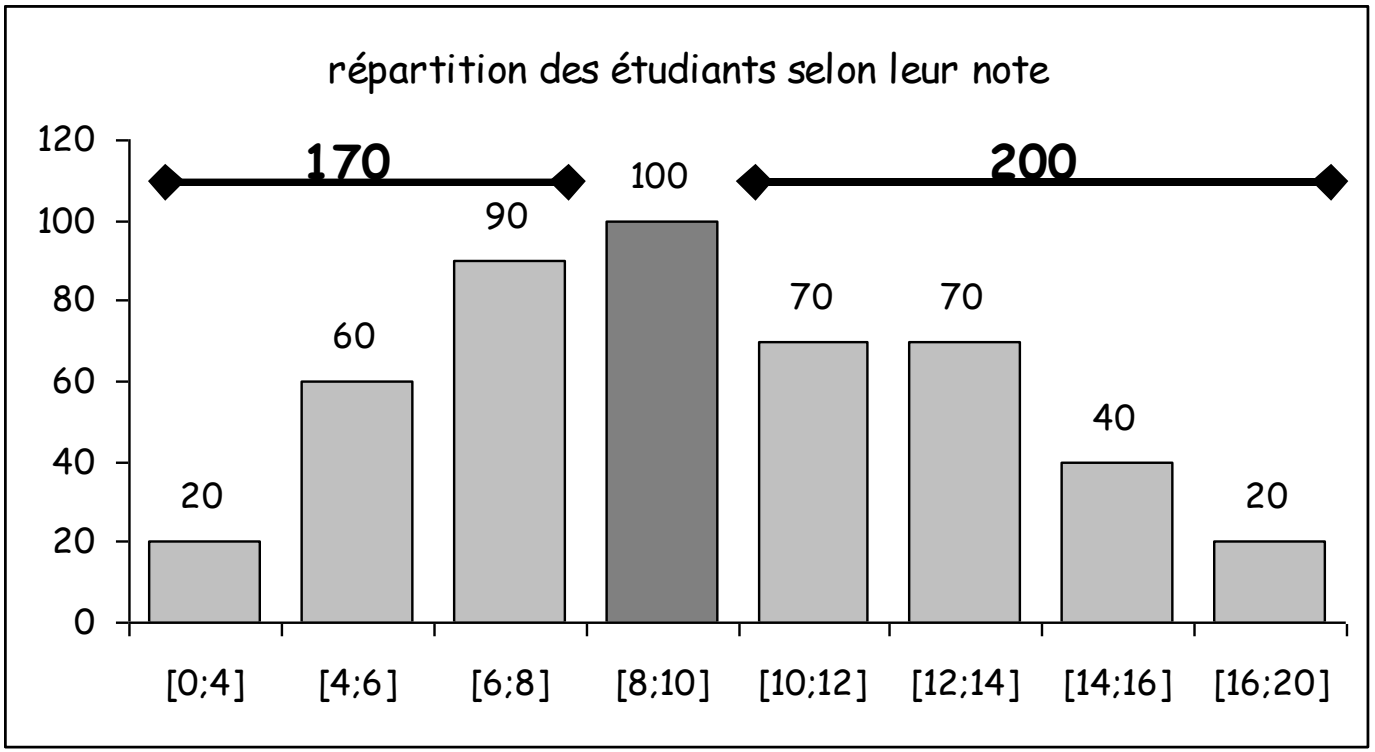
L'estimation de la médiane utilise les paramètres suivants :

Bi : borne inférieure de la classe médiane (8)
N : effectif total (470)
Em : effectif de la classe médiane (100)
Ec : effectif cumulé des classes inférieures à la classe médiane (170)
A : l'amplitude de la classe médiane (10-8=2)



$$Médiane = Bi + \left(\frac{\frac{N}{2} - Ec}{Em} \right) \times A$$

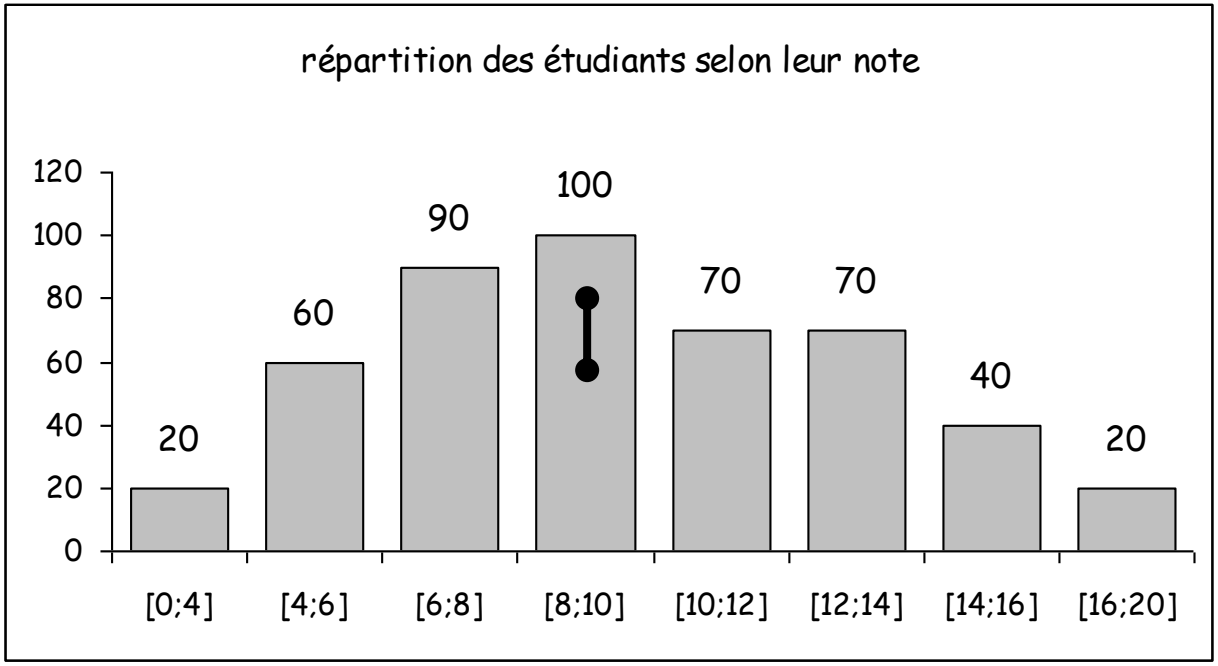
Dans notre exemple :
8 + [(235-170) / 100] x 2 = 8 + (0,65) x 2 = 9,3



Le mode :

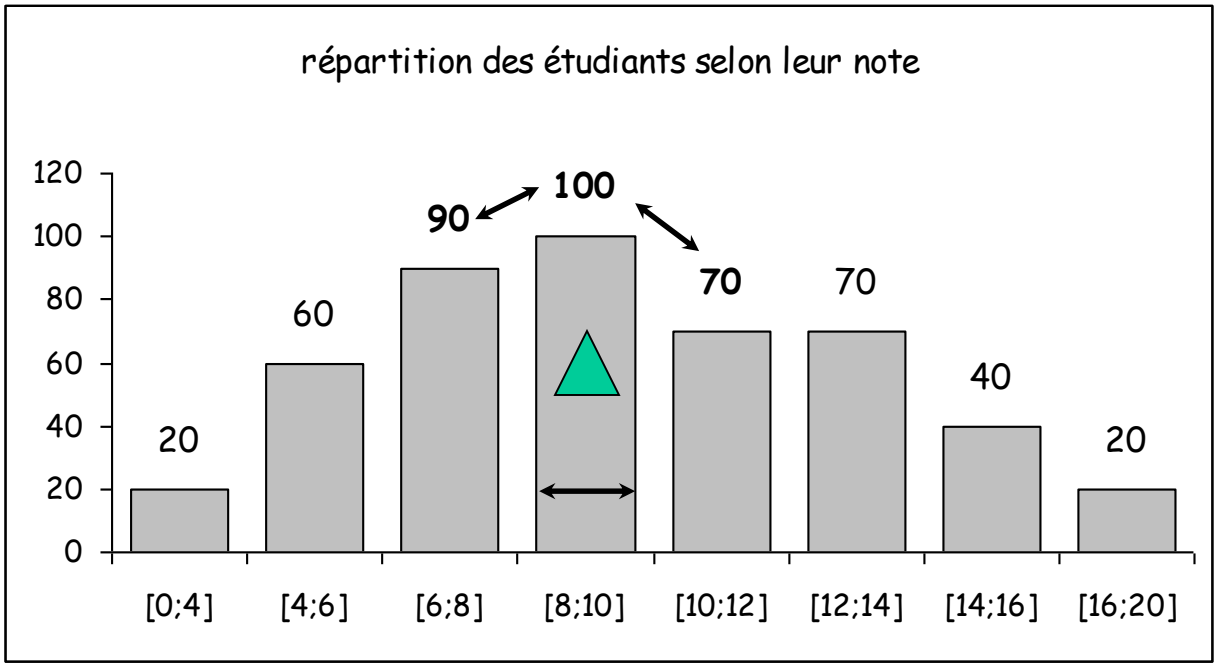
C'est la valeur la plus fréquente de la série statistique, c'est à dire celle qui a l'effectif le plus important.

Dans une série en classes, on détermine d'abord la classe modale,



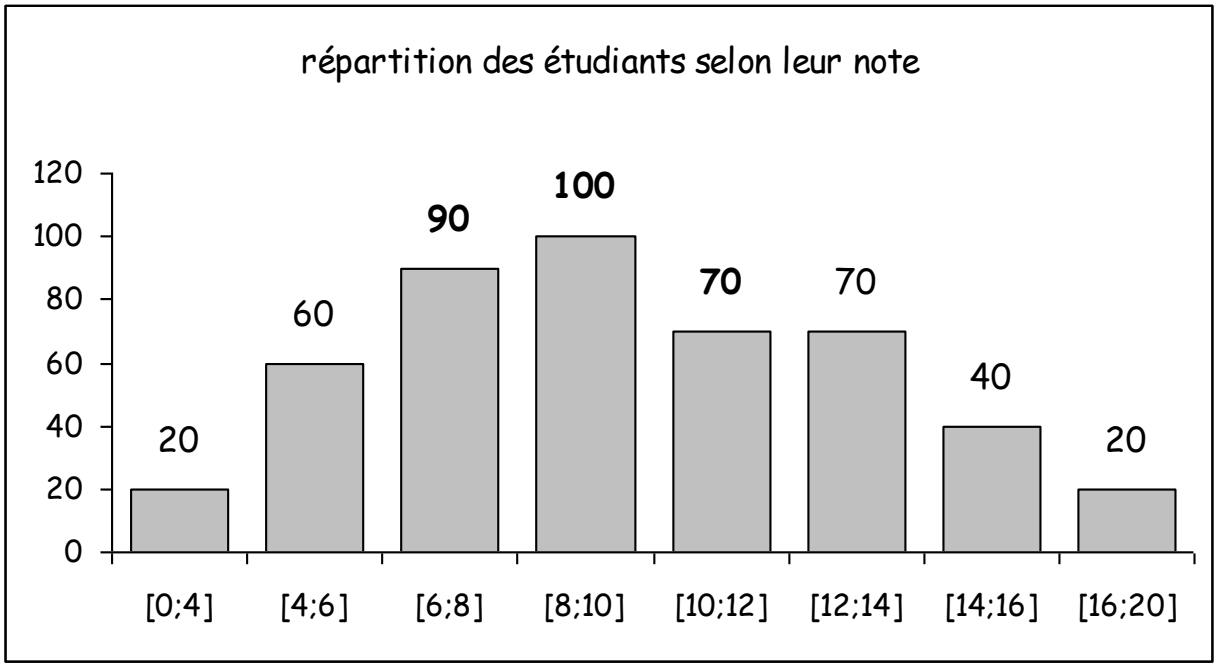
L'estimation du mode utilise les paramètres suivants :

Bi : borne inférieure de la classe modale (8)
Ep : l'écart entre l'effectif de la classe modale et l'effectif de la classe précédente (100-90=10)
Es : l'écart entre l'effectif de la classe modale et l'effectif de la classe suivante (100-70=30)
A : l'amplitude de la classe modale (10-8=2)



Estimation du mode : $$Bi + \left(\frac{Ep}{Ep + Es} \right) \times A$$

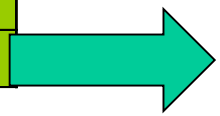
Dans notre exemple : 8 + (10/10+30)x2 = 8+(0,25x2) = 8,5



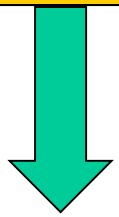
fréquences et concentration

On observe la répartition des 543 communes des Pyrénées selon leur population

classes de population	effectifs	population
moins de 500	366	78 981
de 500 à 1 000	92	62 906
de 1 000 à 2 000	43	57 609
de 2 000 à 5 000	25	82 419
de 5 000 à 10 000	8	54 205
de 10 000 à 50 000	8	160 239
plus de 50 000	1	82 157
ensemble	543	578 516



Population du département



Nombre de communes

$$fi = \frac{ni}{N}$$

classes	effectifs	fréquence relative
moins de 500	366	0,674
de 500 à 1 000	92	0,169
de 1 000 à 2 000	43	0,079
de 2 000 à 5 000	25	0,046
de 5 000 à 10 000	8	0,015
de 10 000 à 50 000	8	0,015
plus de 50 000	1	0,002
ensemble	543	1,00

La somme des fréquences relatives est égale à 1