❄ Chapitre 9 ❄

Statistiques descriptives

Objectif du chapitre :

- Analyser les données brutes d'une série : caractère, type, modalité
- Déterminer et interpréter une fréquence, une moyenne, une médiane et des quartiles.
- Trier des données brutes d'une série en dressant un tableau d'effectifs et en déterminant des classes

I. Définitions et vocabulaire des statistiques

La **population** est l'ensemble des individus sur lesquels portent l'étude statistique. (Par exemple classe de seconde, hommes, habitants de la France ...)

Le **caractère** (ou **variable**) d'une série statistique est une propriété étudiée sur chaque individu :

- Lorsque le caractère ne prend que des valeurs (ou **modalités**) numériques, il est **quantitatif** :
  - **discret** s'il ne peut prendre que des valeurs isolées (notes, âge ...)
  - **continu** dans le cas contraire (poids, taille ...). Dans ce cas on effectue souvent un regroupement des valeurs par **classes**.
- Sinon, on dit qu'il est **qualitatif** (couleur des yeux, sport pratiqué ...) : les modalités ne sont pas des nombres.

A chaque valeur (ou classe) est associée un **effectif** n : c'est le nombre d'individus associés à cette valeur.

Faire des **statistiques**, c'est recueillir, organiser, synthétiser, représenter et exploiter des données, numériques ou non, dans un but de comparaison, de prévision, de constat...

Les plus gros "consommateurs" de statistiques sont les **assureurs** (risques d'accidents, de maladie des assurés), les **médecins** (épidémiologie), les **démographes** (populations et leur dynamique), les **économistes** (emploi, conjoncture économique), les **météorologues** ...

❄ Définition 1:

On considère une série statistique à caractère quantitatif, dont les p valeurs sont données par :
$x_1, x_2, \ldots, x_p$ d'effectifs associés $n_1, n_2, \ldots, n_p$ avec $n_1 + n_2 + \ldots + n_p = N$.

- A chaque valeur (ou classe) est associée une **fréquence** f_i : c'est la proportion d'individus associés à cette valeur.
- $f_i = \dfrac{n_i}{N}$ est un nombre compris entre 0 et 1, que l'on peut écrire sous forme de pourcentage.
- L'ensemble des fréquences de toutes les valeurs du caractère s'appelle la **distribution des fréquences** de la série statistique.

Exemple 1:

Voici les notes obtenues à un contrôle dans une classe de 30 élèves : (**Série A :**)
2 – 3 – 3 – 4 – 5 – 6 – 6 – 7 – 7 – 7 – 8 – 8 – 8 – 8 – 8 – 9 – 9 – 9 – 9 – 9 – 9 – 10 – 10 – 11 – 11 – 11 – 13 – 13 – 15 – 16
On peut représenter cette série par un tableau d'effectifs, et le compléter par la distribution des fréquences :

Notes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Eff.	0	1	2	1	1	2	3	5	6	2	3	0	2	0	1	1	0	0	0
Fréq. en %	0	3	7	3	3	7	10	17	20	7	10	0	7	0	3	3	0	0	0

Remarque :
On peut vérifier que la somme des fréquences est égale à 1 (ou à 100 si on les exprime en pourcentages).

On peut aussi faire un regroupement par classe, ce qui rend l'étude moins précise, mais qui permet d'avoir une vision plus globale.

Exemple 2:
Toujours pour la **série A**, si on regroupe les données par classes d'amplitude 5 points, on obtient :

Notes	[0 ; 5 [	[5 ; 10 [	[10 ; 15 [	[15 ; 20 [	total
Effectif	4	17	7	2	30
Fréquence	0,13	0,57	0,23	0,07	1

II. Caractéristiques de position

1. Médiane (Vu en classe de 3^e)

Définition 2:
Soit une série statistique ordonnée dont les n valeurs sont $x_1 \leqslant x_2 \leqslant x_3 \leqslant \cdots \leqslant x_n$.
Une **médiane** est un nombre Me qui permet de diviser cette série en deux sous-groupes de même effectif.

- Si n est **impair**, M est la valeur de cette série qui est située au milieu, à savoir la valeur dont le rang est $\dfrac{n+1}{2}$, notée $x_{\frac{n+1}{2}}$.
- Si n est **pair**, M est le centre de l'intervalle médian, qui est l'intervalle formé par les deux nombres situés « au milieu » de la série, à savoir $x_{\frac{n}{2}}$ et $x_{\frac{n}{2}+1}$.

Exemple 3:
1. Une médiane de la série « 2 – 5 – 6 – 8 – 9 – 9 – 10 » est 8.
2. Une médiane de la série « 2 – 5 – 6 – 8 – 9 – 9 » est 7.
3. Une médiane de la série « 2 – 5 – 6 – 6 – 9 – 10 » est 6.

2. Quartiles

Définition 3:
Soit une série statistique, on appelle **quartiles** de la série un triplet de réels (Q_1 ; Q_2 ; Q_3) qui sépare la série en quatre groupes de même effectif.

Remarque :
Par définition, si X est une série statistique, $Q_2 = Me(X)$.

La calculatrice donne les valeurs de Q_1, Me et Q_3.
Nous verrons dans la partie VI. d'autres façons de trouver la médiane et les quartiles.

Exemple 4:
Pour la **série A**, la calculatrice nous donne $Q_1 = 7$, $Me = 8,5$ et $Q_3 = 10$.

3. Moyenne pondéré

Définition 4:

Soit une série statistique à caractère quantitatif, dont les p valeurs sont données par x_1, x_2, …, x_p d'effectifs associés n_1, n_2, …, n_p avec $n_1 + n_2 + ... + n_p = N$.

La **moyenne pondérée** de cette série est le nombre noté $\overline{x}$ qui vaut

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + ... + n_p x_p}{n_1 + n_2 + ... + n_p} = \frac{1}{N} \sum_{i=1}^{p} n_i x_i.$$

Remarque :

Lorsque la série est regroupée en classes, on calcule la moyenne en prenant pour valeurs x_i le **centre de chaque classe**; ce centre est obtenu en faisant la moyenne des deux extrémités de la classe.

Exemple 5:

- Dans la **série A**, la moyenne du contrôle est égale à

$$\overline{x} = \frac{2 \times 1 + 3 \times 2 + \cdots + 16 \times 1}{30} = \frac{254}{30} \approx 8,47$$

- si on regroupe par classe d'amplitude 5 points, une estimation de la moyenne est :

$$\overline{x} = \frac{2,5 \times 4 + 7.5 \times 17 + \cdots + 17,5 \times 2}{30} = \frac{260}{30} \approx 8,67$$

Remarque :

On peut aussi calculer une moyenne à partir de la distribution de fréquences :

$$\overline{x} = f_1 x_1 + f_2 x_2 + \cdots + f_p x_p = \sum_{i=1}^{p} f_i x_i.$$

Propriété 1 : *Linéarité de la moyenne*

- Si on ajoute (ou soustrait) un même nombre k à toutes les valeurs d'une série, alors la moyenne de cette série se trouve augmentée (resp. diminuée) de k.
- Si on multiplie (ou divise) par un même nombre non nul k toutes les valeurs d'une série, alors la moyenne de cette série se trouve multipliée (resp. divisée) par k.

Exemple 6:

On considère la **série A** :

- Si on ajoute $1,5$ points à chaque note du contrôle, alors la moyenne de classe devient

$$\overline{x} = 8,47 + 1,5 = 9,97$$

- Si on augmente chaque note de 10%, cela revient à multiplier chaque note par $1,1$, ce qui donne

$$\overline{x} = 8,47 \times 1,1 = 9,32$$

Propriété 2 : *Moyenne par sous-groupes*

Soit une série statistique, d'effectif total N, de moyenne $\overline{x}$.

Si on divise cette série en deux sous-groupes **disjoints** d'effectifs respectifs p et q (avec $p + q = N$) de moyennes respectives $\overline{x}_1$ et $\overline{x}_2$, alors on a :

$$\overline{x} = \frac{p}{N} \times \overline{x}_1 + \frac{q}{N} \times \overline{x}_2.$$

Exemple 7:
On suppose par exemple que les 12 garçons de la classe de la **série A** ont obtenu une moyenne globale de 8 sur 20.

- La moyenne du groupe formé par les filles de la classe vérifie : $9,47 = \frac{12}{30} \times 8 + \frac{18}{30} \times \overline{m}_f$.
- Soit $\overline{m}_f = \frac{30}{18}\left(9,47 - \frac{12}{30} \times 8\right) = 10,45$.

III. Caractéristiques de dispersion

Définition 5:
On appelle **étendue** d'une série discrète X le réel défini par $e(X) = \max(X) - \min(X)$.

Il s'agit de la première mesure de la dispersion d'une série statistique. Son principal mérite a longtemps été d'exister, et de fournir une information sur la dispersion très simple à obtenir.

Exemple 8:
L'étendue de la **série A** est de $e(A) = 16 - 2 = 14$.

1. L'écart interquartile

Définition 6:
On appelle **intervalle inter-quartiles** l'intervalle $[\ Q_1\ ;\ Q_3\ ]$.
L'amplitude de cet intervalle est appelée **écart inter-quartiles**. Cette valeur est l'indicateur de dispersion associé à la médiane.

Exemple 9:
- Dans la **série A**, l'intervalle inter-quartile est l'intervalle $[\ 7\ ;\ 10\ ]$ dont l'écart vaut $10 - 7 = 3$.
- Cet intervalle comprend donc la moitié des notes de la série située au centre de celle-ci.

2. L'écart type

Définition 7:
- L'écart type est un indicateur de dispersion associé à la moyenne : il mesure la dispersion des valeurs autour de la moyenne. Plus l'écart-type est grand, plus les valeurs sont dispersées et moins la moyenne représente de façon significative la série.
- L'écart type, noté σ, de la série statistique est obtenu à l'aide de la calculatrice.

Voir **Fiche méthode** : "Calcul de paramètres statistiques"

Remarque :
La connaissance des formules qui suivent sont facultatives
Il est possible de définir de manière rigoureuse l'écart type :

- L'écart type σ de la série statistique est défini comme la racine carrée de la variance V.

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

- La variance de la série statistique est :

$$V = \frac{n_1(x_1 - \bar{x})^2 + \ldots + n_k(x_k - \bar{x})^2}{N}$$

La variance est aussi appelée la moyenne quadratique de la série centrée.

IV. Effectifs et fréquences cumulés

Définition 8:

Quand les valeurs d'un caractère quantitatif sont rangées dans l'ordre croissant,

- **L'effectif cumulé croissant [respectivement décroissant]** d'une valeur est la somme des effectifs des valeurs inférieures [respectivement supérieures] ou égales à cette valeur,
- **la fréquence cumulée croissante [respectivement décroissante]** d'une valeur est la somme des fréquences des valeurs inférieures [respectivement supérieures] ou égales à cette valeur.

Exemple 10:

On reprend l'exemple de la **série A** de notes du chapitre précédent, on obtient :

Notes	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
Eff.	0	1	2	1	1	2	3	5	6	2	3	0	2	0	1	1	0	0	0
E.C.C.	0	1	3	4	5	7	10	15	21	23	26	26	28	28	29	30	30	30	30
E.C.D.	30	30	29	27	26	25	23	20	15	9	7	4	4	2	2	1	0	0	0

Ce tableau peut par exemple nous permettre de calculer une médiane de la série :

l'effectif étant de 30, on choisit la moyenne entre la $15^{ième}$ et la $16^{ième}$ note, lues dans la ligne des E.C.C.

On obtient $Me = \dfrac{8+9}{2} = 8,5$.

Exemple 11:

Toujours pour l'exemple de la **série A** par classes, on s'intéresse cette fois-ci à la fréquence :

Notes	[0 ; 5 [	[5 ; 10 [	[10 ; 15 [	[15 ; 20 [
Effectif	4	17	7	2
Fréquence en %	13	57	23	7
F.c.c.	13	70	93	100
F.c.d.	100	87	3	7

V. Représentation graphique d'une série statistique

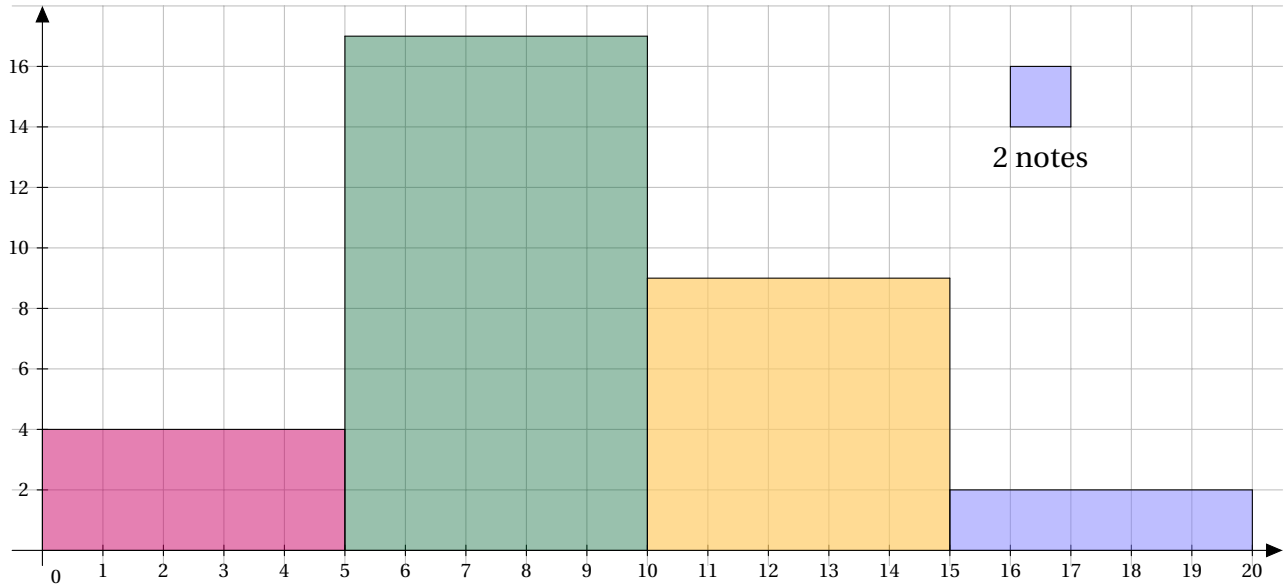
1. Histogramme

Lorsque le caractère étudié est **quantitatif** et lorsque les modalités sont regroupées en **classes**, on peut représenter la série par un **histogramme** : l'aire de chaque rectangle est alors proportionnelle à l'effectif (ou à la fréquence) associée à chaque classe.

Lorsque les classes ont la même **amplitude**, c'est la hauteur qui est proportionnelle à l'effectif.

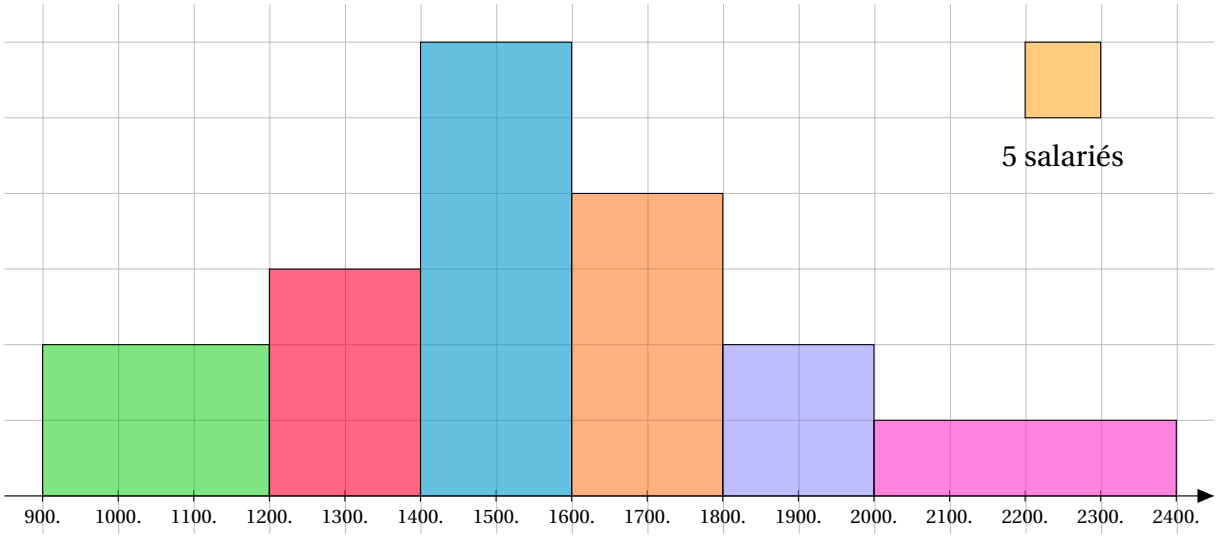
Exemple 12:

Histogramme de la **série A** pour laquelle les amplitudes sont toute égales à 5 :



Exemple 13:

Exemple d'un histogramme représentant la répartition des salaires dans une entreprise, l'amplitude des classes n'étant pas régulières :



On obtient le tableau suivant :

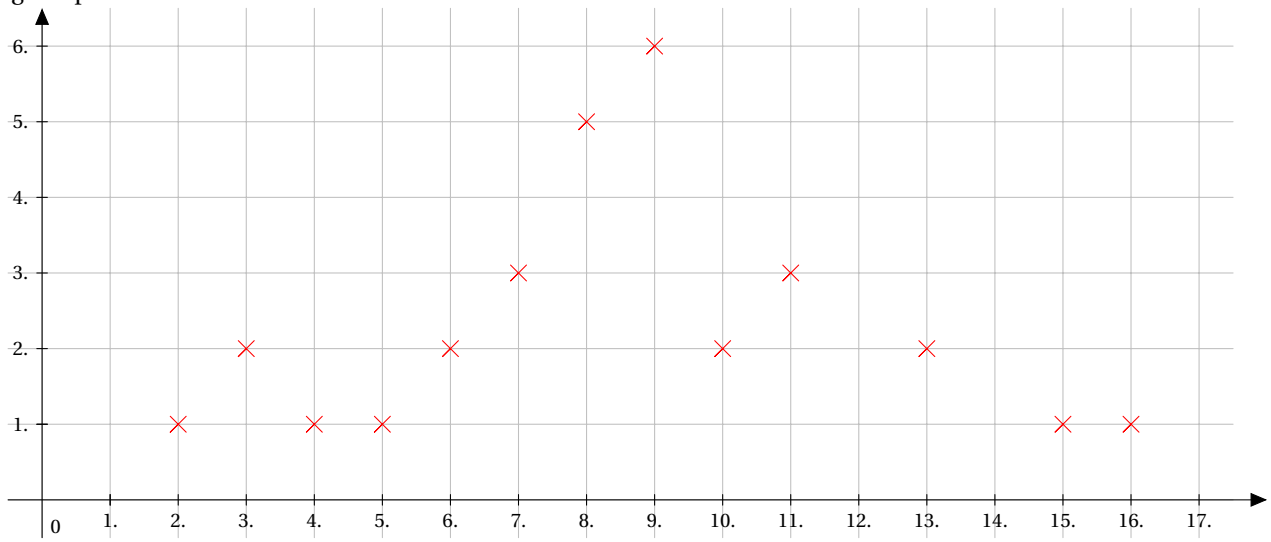
Salaires	[900;1200]	[1200;1400]	[1400;1600]	[1600;1800]	[1800;2000]	[2000;2400]
Effectif	30	30	60	40	20	10

2. Nuage de points

Lorsque le caractère étudié est **quantitatif et discret**, on peut représenter la série par un **nuage de points** : chaque couple de valeurs est représenté par un point dans un repère orthogonal.

Exemple 14:

Nuage de points de la **série A** :

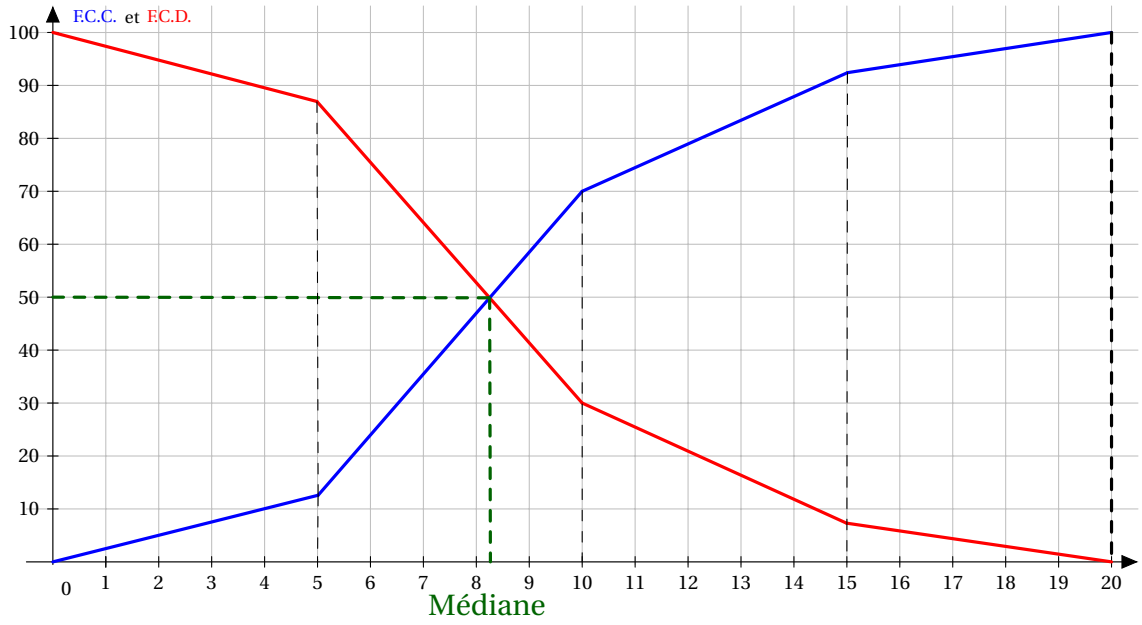


3. Courbe des fréquences cumulées

Enfin, Lorsque le caractère étudié est **quantitatif** et lorsque les modalités sont regroupées en **classes**, on peut effectuer la **courbe des fréquences cumulées** (croissantes ou décroissantes) appelée aussi **polygone** des fréquences cumulées.

Exemple 15:

Polygone des fréquences cumulées croissantes et décroissantes de la **série A** :



On peut grâce à ces polygones déterminer une médiane de la série de deux manières :

- Soit en déterminant le point du polygone d'ordonnée 50% : on trouve environ $M = 8,2$,
- soit en lisant l'abscisse du point d'intersection des deux courbes.

VI. Médiane, quartile et représentations graphiques associées

Dans la partie précédente, on a réussi a déterminer graphiquement une médiane d'une série statistique à caractère quantitatif continu.

En effet, il n'est pas possible pour ce type de série de la faire analytiquement. Cependant pour les série à caractère quantitatif discret, il est tout a fait possible de le faire "à la main".

1. Calcul de médiane et de quartiles

Exemple 16:

Les résultats d'une enquête sur le prix d'une baguette de pain dans 30 boulangeries d'une ville sont regroupés dans le tableau suivant.

Prix *(en centimes d'euros)*	70	75	80	85	90	95	100	105	110	115	120	Total
Nombres de boulangerie	1	1	2	4	7	3	3	4	2	2	1	30

Méthode 1 : *Calcul de la médiane (avec N pair)*

L'effectif total de la série statistiques est de 30 donc N est pair.
Une fois les valeurs de la série statistique triées par ordre croissant, une médiane est donc la moyenne des deux valeurs du milieu.

- La première valeur du milieu est la valeur située à la position $\frac{30}{2} = 15$
- La deuxième valeur du milieu est la valeur située à la position $\frac{30}{2} + 1 = 16$

Pour pouvoir trouver facilement quelle valeur se situe en position 15 et 16, il faut dresser le tableau des Effectifs cumulés croissants (E.c.c.).

Prix *(en centimes d'euros)*	70	75	80	85	90	95	100	105	110	115	120	Total
Nombres de boulangerie	1	1	2	4	7	3	3	4	2	2	1	30
E.c.c.	1	2	4	8	15	18	21	25	27	29	30	X

La 15^e valeur de la série vaut donc 90 et la 16^e valeur vaut 95.
On fait la moyenne de la 15^e valeur et la 16^e valeur pour obtenir la médiane :

$$\begin{aligned} Me &= \frac{90+95}{2} \\ &= 92.5 \end{aligned}$$

La médiane de cette série statistique est de 92.5 centimes.
On peut en conclure que la moitié des baguettes de pain de la ville coute 92.5 centimes ou moins.

Méthode 2 : *Calcul des quartiles*

L'effectif total de la série statistique est de 30.

- Le premier quartile est la première valeur située au delà de la position $\frac{1}{4} \times 30 = 7.5$ donc la 8^e valeur.
- Le troisième quartile est la première valeur située au delà de la position $\frac{3}{4} \times 30 = 22.5$ donc la 23^e valeur.

Pour pouvoir trouver facilement quelles valeurs se situent en position 8 et 23, on reprend le tableau des Effectifs cumulés croissants (E.c.c.).

Prix *(en centimes d'euros)*	70	75	80	85	90	95	100	105	110	115	120	Total
Nombres de boulangerie	1	1	2	4	7	3	3	4	2	2	1	30
E.c.c.	1	2	4	8	15	18	21	25	27	29	30	X

Donc le premier quartile Q_1 vaut 85 et Q_3 vaut 105.
On peut en conclure qu'au moins un quart des baguettes de pain ont un prix inférieur à 85 centimes d'euros et au moins 75% des baguettes ont un prix inférieur à un euros et cinq centimes

On peut aussi calculer la valeur de l'écart interquartile :

$$Q_3 - Q_1 = 105 - 85 = 20$$

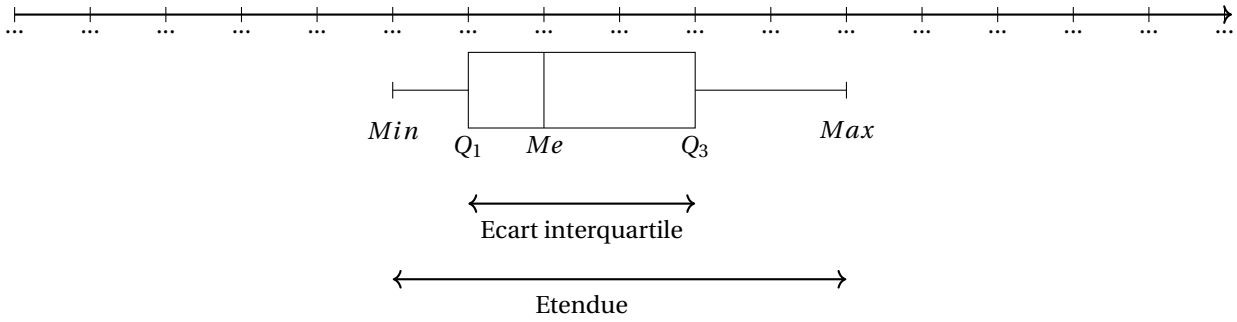
On peut en conclure que la moitié des baguettes de pain sont dans une fourchette de prix de 20 centimes d'euros

2. Représentation graphique associée à la médiane et aux quartiles

Les valeurs extrêmes (minimum et maximum), la médiane et les deux quartiles permettent de partager une série en quatre parties contenant chacune environ un quart de l'effectif total. On a ainsi une idée de la répartition des valeurs de cette série.

Méthode 3 : *Construction du diagramme en boite*

- Tracer un axe qui **DOIT** être gradué en fonction de la série statistique étudiée.
- La « boite » est limitée par Q_1 et Q_3. Sa longueur est $Q_3 - Q_1$
- Elle montre la médiane Me.
- Les « moustaches » sont limitées par les valeurs extrêmes.

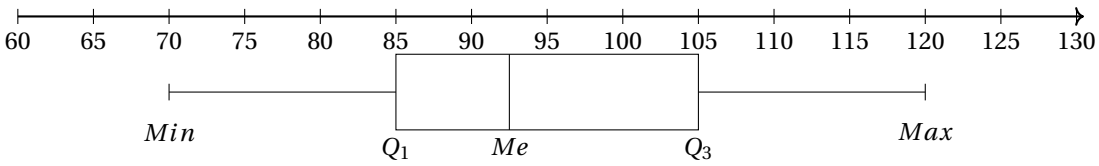


Exemple 17:

En reprenant l'étude fait sur nos boulangeries ci-dessus, on a :

- $Q_1 = 85$ et $Q_3 = 105$
- $Me = 92.5$
- $Min = 70$ et $Max = 120$

Le diagramme en boîte de cette série statistique est donc :



Remarque :

Il est souvent intéressant de tracer le diagramme en boîte de deux séries statistiques différentes sur le **même** axe gradué, afin de pouvoir mieux les comparer.