LES STATISTIQUES

I Vocabulaire et définitions

Définition n°1. ***Population, individus***

La population, c'est l'ensemble des individus sur lesquels portent l'étude statistique.

Exemple n°1.

On étudie le nombre d'arbres malades dans une forêt.
Les individus sont les arbres, la population est la forêt.

Définition n°2. ***Caractère (ou variable)***

C'est une propriété étudiée sur chaque individu.

Exemple n°2.

Pour notre notre forêt, le caractère est le fait d'être malade ou non.

Définition n°3. ***Nature du caractère, valeur ou modalité***

Le caractère est **quantitatif** quand il prend des valeurs (ou modalités) numériques. Il peut alors être **quantitatif discret** si les valeurs sont isolées ou **quantitatif continu** dans le cas contraire.
Le caractère est **qualitatif** quand les modalités qu'il prend **ne sont pas numériques**.

Remarque n°1.

On a tendance à utiliser le terme « valeur » pour un caractère quantitatif et plutôt le terme « modalité » pour un caractère qualitatif.

Exemple n°3. ***Qualitatif***

Pour notre forêt, le caractère est qualitatif.

Exemple n°4. ***Quantitatif discret***

On étudie le nombre de stylos par élève dans notre classe.
La population est la classe, les individus sont les élèves, le caractère est le nombre de stylos. C'est un caractère quantitatif discret.

Exemple n°5. ***Quantitatif continu***

On étudie la hauteur en mètres des bâtiments d'une ville.
La population est l'ensemble des bâtiments de la ville, les individus sont les bâtiments, le caractère est la hauteur en mètres. C'est un caractère quantitatif continu.

Remarque n°2. ***Regroupement par classe***

Dans le cas d'un caractère quantitatif continu, on regroupe les valeurs par **classes** (qui sont des **intervalles**). On peut également le faire quand le caractère est quantitatif discret.

Définition n°4. ***Centre de classe***

Le centre de la classe, c'est la moyenne des extrémités de la classe

Définition n°5. ***Effectif***

C'est le nombre d'individus qui possèdent le caractère étudié.

Définition n°6. ***Mode, Valeur modale, classe modale***

On appelle mode ou valeur (*resp* classe) modale, la valeur (*resp* classe) qui possède le plus grand effectif.

II Fréquences, distribution des fréquences

Soit p un nombre entier.
On considère une série statistique, dont le caractère étudié peut prendre les valeurs (ou modalités) x_1 , x_2 , x_3 , ... , x_p .
On note n_1 , n_2 , n_3 , ... , n_p les effectifs correspondants et
on pose $N=n_1+n_2+n_3+...+n_p$
(N est l'effectif total, c'est à dire le nombre d'individus qui composent la population)

▪ Pour $i \in [\![1 ; p]\!]$, on pose $f_i=\frac{n_i}{N}$,
f_i est alors la fréquence associée à x_i
▪ L'ensemble de ces fréquences est appelé la distribution des fréquences.

Exemple n°6. Série A

Voici les notes obtenues à un contrôle dans une classe de 30 élèves :
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

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

➔ Le mode ou la valeur modale est 9 (6 élèves ont eu 9 : c'est le plus grand effectif)

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

Notes	[0 ; 5 [	[5 ; 10 [	[10 ; 15 [	[15 ; 20 [	total
Centre	$\frac{0+5}{2}=2,5$	$\frac{5+10}{2}=7,5$	$\frac{10+15}{2}=12,5$	$\frac{15+20}{2}=17,5$	
Effectif	4	17	7	2	30
Fréquence	0, 13	0, 57	0, 23	0, 07	1

➔ La classe modale est [5 ; 10 [(17 élèves ont eu entre 5 inclus et 10 exclu : c'est le plus grand effectif)

III Indicateurs de tendance centrale

Les indicateurs de tendance centrale sont le mode, la médiane et la moyenne.

Définition n°7. Moyenne pondérée

Soit une série statistique à caractère quantitatif, dont les p valeurs sont données par x_1 , x_2 , x_3 , ... , x_p d'effectifs associés n_1 , n_2 , n_3 , ... , n_p avec $N = n_1 + n_2 + n_3 + ... + n_p$

La moyenne pondérée de cette série est le nombre $\overline{x}$ tel que :

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + n_3 x_3 + ... + n_p x_p}{n_1 + n_2 + n_3 + ... + n_p} = \frac{1}{N} \sum_{i=1}^{p} n_i x_i$$

Remarque n°3.

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 n°7. (avec la série A)

▪ Si on note $\overline{x}$ la moyenne du contrôle alors

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

▪ Si on regroupe par classe d'amplitude 5 points, une estimation de la moyenne est : $\overline{x} = \frac{2{,}5\times 4 + 7{,}5\times 17 + ... + 17{,}5\times 2}{30} = \frac{260}{30} \approx 8{,}67$

Remarque n°4.

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

$$\overline{x} = f_1 x_1 + f_2 x_2 + f_3 x_3 + ... + f_p x_p = \sum_{i=1}^{p} f_i x_i$$

Propriété n°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 n°8. (toujours avec la série A)

▪ Si on ajoute 1, 5 points à chaque note du contrôle, alors la moyenne de classe devient $\overline{x} \approx 8{,}47 + 1{,}5 = 9{,}97$

▪ Si on augmente chaque note de 10%, cela revient à multiplier chaque note par 1, 1, ce qui donne $\overline{x} \approx 8{,}47\times 1{,}1 = 9{,}317$

Propriété n°2. Moyenne par sous groupe

Soit une série statistique, d'effectif total N et 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 n°9. ***(toujours avec la série A)***

Énoncé

On suppose que les 12 garçons de la classe de la série A ont obtenu une moyenne globale de 8 sur 20. Déterminer la moyenne des filles.

Réponse :

Notons $\overline{x}_f$ la moyenne des filles. $\overline{x}_f$ vérifie l'égalité suivante :

$9{,}47 = \frac{12}{30}\times 8 + \frac{18}{30}\times \overline{x}_f$.

Après résolution : $\overline{x}_f = 10{,}45$

IV Indicateurs de dispersion

Les principaux indicateurs de dispersion sont l'étendue, l'écart inter-quartile, la variance et l'écart-type.

Définition n°8. ***Étendue***

On appelle étendue d'une série X le réel défini par $e(X) = \max(X) - \min(X)$

Exemple n°10. ***(toujours avec la série A)***

La plus grande valeur est 16, la plus petite est 2.

Donc en notant e l'étendue de la série, on obtient : $e = 16 - 2 = 14$

Définition n°9. ***Quartiles***

Soit une série statistique ordonnée, on appelle :

- **premier quartile** et on note Q_1 la valeur de la série telle qu'au moins 25% des valeurs soient inférieures (ou égales) à Q_1
- **troisième quartile** et on note Q_3 la valeur de la série telle qu'au moins 75% des valeurs soient inférieures (ou égales) à Q_3

Remarque n°5.

On ne parle pas de Q_2 , on lui préfère la médiane.

Exemple n°11. ***(toujours avec la série A)***

Énoncé :

Déterminer les premier et troisième quartiles de la 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

Réponse :

On pense à vérifier que les valeurs sont bien rangées dans l'ordre croissant. (si ce n'est pas le cas, on le fait)

La série comporte 30 valeurs :

- $\frac{1}{4}\times 30 = 7{,}5$, le premier quartile est donc la 8eme valeur de la série et vaut : 7. $\boldsymbol{Q_1 = 7}$
- $\frac{3}{4}\times 30 = 22{,}5$, le troisième quartile est donc la 23eme valeur de la série et vaut : 10. $\boldsymbol{Q_3 = 10}$

Remarque n°6. ***Mais pourquoi je n'obtiens pas le « bon résultat » ?***

Pour Q1	cours	calculatrice	tableur
série n°1	270	270	275,5
série n°2	90	90	144
série n°3	47	59	65
série n°4	196	185	196

Pour Q3	cours	calculatrice	tableur
série n°1	630	630	568
série n°2	547	547	467
série n°3	622	644,5	633,25
série n°4	456	546	456

Nos connaissances, à ce niveau, nous oblige à simplifier la définition des quartiles. C'est le choix fait dans l'immense majorité des manuels scolaires et dans ce cours. Cela n'est bien sûr pas sans conséquence, car parfois la calculatrice ou le tableur ne donneront pas les mêmes réponses que nous. Rassurez-vous (ou pas) la calculatrice et le tableur ne sont pas toujours d'accord entre eux non plus… La preuve avec les quatre séries suivantes :
Série n°1 : 46, 270, 293, 382, 630, 952
Série n°2 : 49, 90, 198, 302, 387, 547, 763
Série n°3 : 34, 47, 71, 263, 282, 622, 667, 968
Série n°4 : 39, 174, 196, 252, 331, 401, 456, 637, 944

Définition n°10. ***Intervalle interquartile, écart interquartile.***

On appelle intervalle inter-quartiles, l'intervalle $[Q_1 \ ; \ Q_3]$ et l'amplitude de cet intervalle : $Q_3 - Q1$ est appelée écart inter-quartiles.

Exemple n°12. ***(toujours avec la série A)***

L'intervalle inter-quartile est l'intervalle [7 ; 10]
L'écart inter-quartile vaut 10 – 7 = 3.

Définition n°11. ***La variance***

La variance d'une série statistique est la moyenne des carrés des écarts à la moyenne.

Remarque n°7.

Nous n'utiliserons pas la variance cette année, mais la définition suivante en dépend.

Définition n°12. ***L'écart-type***

L'écart-type d'une série statistique est la racine carrée de la moyenne des carrés des écarts à la moyenne.
Il est en général noté : σ qui se lit : « sigma »

Exemple n°13. ***(toujours avec la série A)***

Notes	[0 ; 5 [	[5 ; 10 [	[10 ; 15 [	[15 ; 20 [	total
Centre	2,5	7,5	12,5	17,5	
Effectif	4	17	7	2	30
Fréquence	0, 13	0, 57	0, 23	0, 07	1

▪ La moyenne est :

$$\overline{x} = \frac{2{,}5\times 4+7{,}5\times 17+\ldots+17{,}5\times 2}{30} = \frac{260}{30} = \frac{26}{3} \approx 8{,}67$$

▪ L'écart-type vaut alors :

$$\sigma = \sqrt{\frac{4\times\left(2{,}5-\frac{26}{3}\right)^2+17\times\left(7{,}5-\frac{26}{3}\right)^2+7\times\left(12{,}5-\frac{26}{3}\right)^2+2\times\left(17{,}5-\frac{26}{3}\right)^2}{4+17+7+2}} \approx 3{,}8$$

V Résumé du cours

Définitions

Population, individus	La population, c'est l'ensemble des individus sur lesquels portent l'étude statistique.
Caractère	C'est une propriété étudiée sur chaque individu.
Nature de caractère, valeur ou modalité.	Le caractère est **quantitatif** quand il prend des valeurs (ou modalités) numériques. Il peut alors être **quantitatif discret** si les valeurs sont isolées ou **quantitatif continu** dans le cas contraire. Le caractère est **qualitatif** quand les modalités qu'il prend **ne sont pas numériques**.
Classe	Dans le cas d'un caractère quantitatif continu, on regroupe les valeurs par **classes** (qui sont des **intervalles**). On peut également le faire quand le caractère est quantitatif discret.
Centre de classe	Le centre de la classe, c'est la moyenne des extrémités de la classe
Effectif	C'est le nombre d'individus qui possèdent le caractère étudié.
Mode, valeur modale, classe modale	On appelle mode ou valeur (*resp* classe) modale, la valeur (*resp* classe) qui possède le plus grand effectif.

Indicateurs de tendance centrale

Les indicateurs de tendance centrale sont le mode, la médiane et la moyenne.

Moyenne pondérée (avec les effectifs)

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + n_3 x_3 + ... + n_p x_p}{n_1 + n_2 + n_3 + ... + n_p} = \frac{1}{N} \sum_{i=1}^{p} n_i x_i$$

Moyenne pondérée (avec les fréquences)

$$\overline{x} = f_1 x_1 + f_2 x_2 + f_3 x_3 + ... + f_p x_p = \sum_{i=1}^{p} f_i x_i$$

Moyenne pondérée (par groupe)

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

Indicateurs de dispersion

Les principaux indicateurs de dispersion sont l'étendue, l'écart inter-quartile, la variance et l'écart-type.

Étendue

Étendue = plus grande valeur du caractère – plus petite valeur du caractère

Quartiles

Soit une série statistique ordonnée, on appelle :
- **premier quartile** et on note Q_1 la valeur de la série telle qu'au moins 25% des valeurs soient inférieures (ou égales) à Q_1
- **troisième quartile** et on note Q_3 la valeur de la série telle qu'au moins 75% des valeurs soient inférieures (ou égales) à Q_3

Intervalle inter-quartile, écart inter-quartile

On appelle intervalle inter-quartiles, l'intervalle $[Q_1 ; Q_3]$ et l'amplitude de cet intervalle : $Q_3 - Q1$ est appelée écart inter-quartiles.

Écart-type

L'écart-type d'une série statistique est la racine carrée de la moyenne des carrés des écarts à la moyenne.
Il est en général noté : σ qui se lit : « sigma »

Voir l'exemple n°13

COMPLÉMENT DE COURS

VI Effectifs et fréquences cumulés

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 n°14. (toujours avec la série A)

Notes	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Eff	1	2	1	1	2	3	5	6	2	3	0	2	0	1	1
E.C.C	1	3	4	5	7	10	15	21	23	26	26	28	28	29	30
E.C.D	30	29	27	26	25	23	20	15	9	7	4	4	2	2	1

Ce tableau peut par exemple nous permettre de calculer la médiane de la série :
l'effectif étant de 30, on choisit la moyenne entre la 15e et 16e note, lues dans la ligne des E.C.C. :

$$M = \frac{8+9}{2} = 8{,}5$$

Exemple n°15. (toujours avec la série A)

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	70
F.C.D	100	87	30	7

VII Représentation graphique d'une série statistique

VII.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.

Méthode n°1. Construire un histogramme

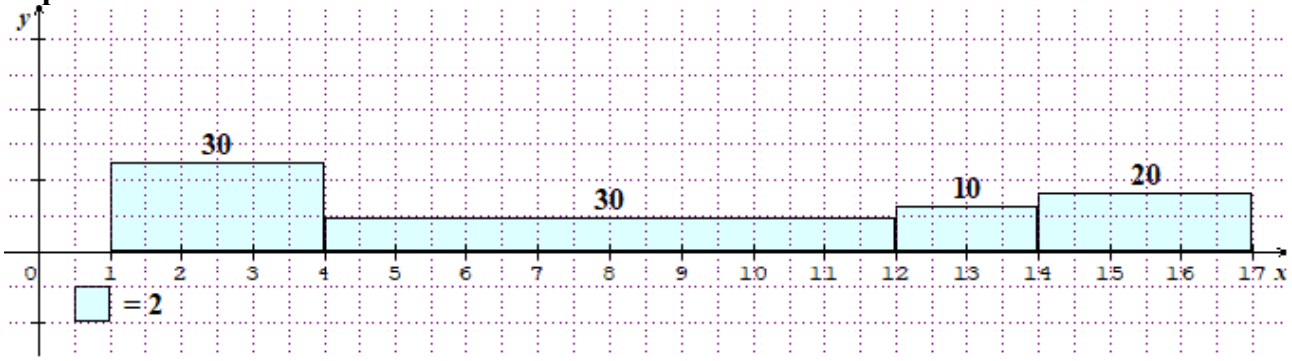
Énoncé :

On donne la série statistique suivante :

Classes	[1 ; 4 [	[4 ; 12 [	[12 ; 14 [	[14 ; 17 [
Effectif	30	30	10	20
Fréquence en %				

Construire un histogramme représentant cette série.

Réponse :



On gradue l'axe des abscisses et on y place les classes.
Ensuite on choisit « la valeur d'un carreau ». Ici on a choisi : « un carreau représente 2 unités ».
Il nous reste à déterminer la hauteur de chaque rectangle afin que son aire soit proportionnelle à l'effectif qu'il représente.

Classe	[1 ; 4 [	[4 ; 12 [	[12 ; 14 [	[14 ; 17 [
Effectif n_i	30	30	10	20
Amplitude en carreaux = largeur du rectangle : l_i	6	16	4	6
Hauteur du rectangle h_i	$\frac{30}{2\times 6}=2{,}5$	$\frac{30}{2\times 16}=0{,}9375$	$\frac{10}{2\times 4}=1{,}25$	$\frac{20}{2\times 6}=\frac{5}{3}$

Si on note « la valeur d'un carreau » alors $\boxed{h_i=\frac{n_i}{t\times l_i}}$

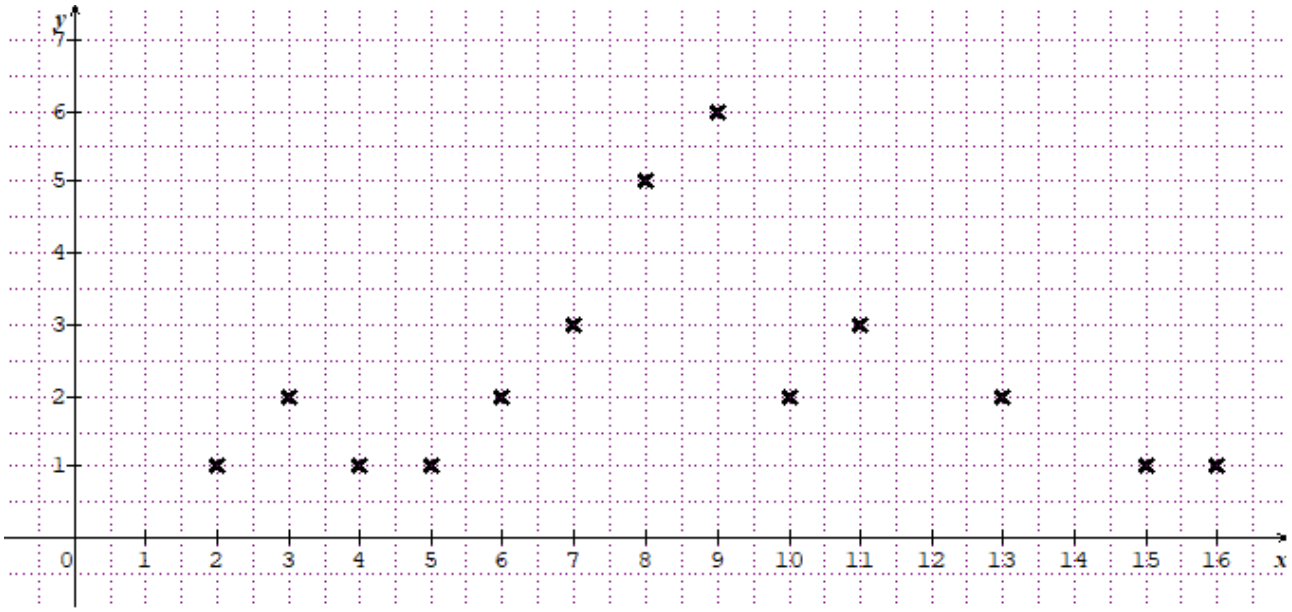
Remarque n°8.

Lorsque les classes ont la même amplitude, la hauteur est aussi proportionnelle à l'effectif. On rappelle que ce n'est pas le cas en général.

VII.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.

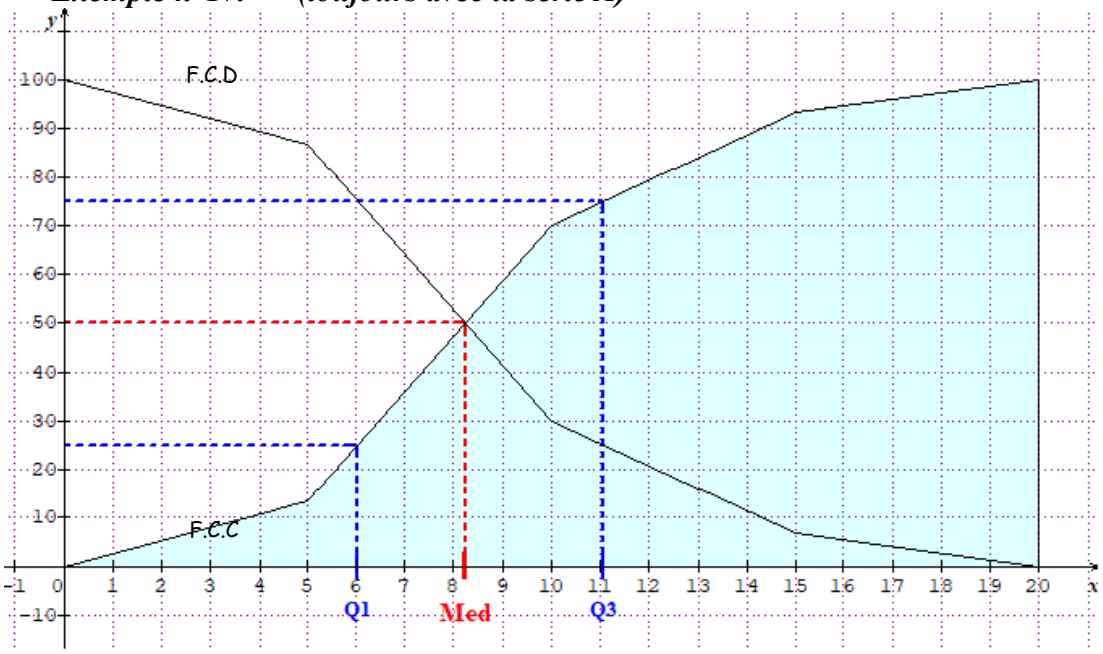
Exemple n°16. ***(toujours avec la série A)***



VII.3 Courbe des fréquences cumulées

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 n°17. ***(toujours avec la série A)***



On peut grâce à ces polygones déterminer la 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.