Fiche méthode : Statistiques

Dans une étude statistique, on étudie une **population**, c'est-à-dire un ensemble **d'individus** (personnes ou objets), et on classe ces individus selon **les valeurs d'un caractère** en donnant pour chaque caractère un **effectif** (nombre d'individus possédant cette valeur du caractère). Une **série statistique** est l'ensemble des résultats d'une étude : valeurs du caractère et effectifs correspondants, qui sont en général donnés dans un tableau.

Dans toute la suite, on considère la série statistique définie par le tableau suivant :

où les valeurs x_i sont ordonnées par ordre croissant (c'est-à-dire $x_1 < x_2 < \cdots < x_p$)

Valeurs du caractère x_i	x_1	x_2	x_3	…	x_p	Total
Effectifs n_i	n_1	n_2	n_3	…	n_p	N

I. Distribution de fréquences

Application 1 : Caractère quantitatif discret

Une société de prêt-à-porter fait une étude sur le nombre de jeans achetés en 2009 par les 16-25 ans.

Nombre de jeans achetés	0	1	2	3	4	5	6	Total
Nombre d'individus	**30**	**80**	45	25	45	20	5	250
Fréquence	$\mathbf{0,12}$	$\mathbf{0,32}$	0,18	0,1	0,18	0,08	0,02	1
FCC	0,12	0,44	0,62	0,72	0,9	0,98	1	

1. Quelle est la population étudiée et le caractère étudié ?

   La population est constituée des jeans achetés en 2009 par les 16-25 ans.
   Le caractère étudié est le nombre d'achats.
2. Calculer les fréquences.

   Exemples : $\frac{30}{250} = \mathbf{0,12}$ $\quad \frac{80}{250} = \mathbf{0,32}$
3. Calculer les fréquences cumulées croissantes (FCC).
4. En déduire le pourcentage d'individus qui ont acheté quatre jeans ou moins en 2009.

   D'après les FCC, on a 90% d'individus qui ont acheté quatre jeans ou moins en 2009.
5. Quelle est la fréquence des individus ayant acheté au maximum cinq jeans en 2009?

   D'après les FCC, la fréquence des individus ayant acheté au maximum cinq jeans en 2009 est 0,98.
6. Compléter la phrase «62% de jeunes entre 16 et 25 ans ont acheté au plus **2** jeans en 2009. »
7. Quel est le pourcentage de jeunes entre 16 et 25 ans qui ont acheté au moins 3 jeans en 2009 ?

   $100\% - 62\% = 38\%$

Application 2 : Caractère qualitatif

Voici un sondage concernant le transport des élèves entre le domicile et le lycée pour les élèves d'une classe :

Moyen de transport	À pied	En moto	En bus	Voiture parents	Total
Effectif	10	2	12	**8**	32
Fréquence	31,25%	6,25%	37,5%	**25%**	100%

Donner la population étudiée, le caractère étudié ainsi que son type.

La population étudiée est la classe, le caractère est qualitatif, il s'agit du mode de transport des élèves.
Exemple fréquence : $\frac{8}{32} = \mathbf{0,25} = \mathbf{25\%}$

Caractère :

- Un caractère est **qualitatif** lorsqu'il n'est pas chiffré, on le qualifie de qualité (Opinion, couleurs, oui/non d'un référendum)
- Un caractère est **quantitatif** lorsqu'il est chiffré, il en existe deux type :
  - Discret : prend que des valeurs isolées que l'on notera x_i (âge, nombre d'enfants)
  - Continu : prend ses valeurs dans des intervalles appelées classes (taille, poids).

Effectif :

- **L'effectif** d'une valeur n_i est le nombre de fois où apparait la valeur x_i dans la série statistique.
- **L'effectif total** N est la **somme** de tous les effectifs.
  $$N = n_1 + n_2 + \cdots n_p$$
- **L'effectif cumulé croissant** de x_i est la somme des effectifs des valeurs qui lui sont inférieures.

Fréquence :

- **La fréquence** f_i d'une valeur d'effectif n_i est le quotient :
  $$f_i = \frac{\boldsymbol{n_i}}{\boldsymbol{N}} = \frac{\boldsymbol{effectif}}{\boldsymbol{effectif\ total}}$$
  On calcule la proportion d'individus vérifiant un critère donné (le caractère) parmi la population totale.
- Une fréquence f_i est toujours comprise entre 0 et 1.
- La somme des fréquences est égale à 1.
- **La fréquence cumulée croissante** de f_i est la somme des fréquences qui lui sont inférieurs.

Remarque : Une fréquence s'exprime souvent sous forme de pourcentage.

II. Couple Moyenne– écart-type.

Application 3 : Caractère quantitatif continu

Deux types de machines automatiques *M*1 et *M*2 permettent de remplir des paquets de sucre en poudre. Les fabricants assurent que leurs machines respectives *M*1 et *M*2 remplissent en moyenne des paquets de sucre de 1 kg.
On prélève deux échantillons de 100 paquets de sucre en poudre.
Les résultats obtenus après vérification du contenu exact de chaque paquet ont fourni les renseignements suivants.

	Machine M_1	
Centre des classes	Contenance (en kg)	Effectif
0,9825	[0,980 ; 0,985[	1
0,9875	[0,985 ; 0,990[	14
0,9925	[0,990 ; 0,995[	15
0,9975	[0,995 ; 1,000[	22
1,0025	[1,000 ; 1,005[	18
1,0105	[1,005 ; 1,016[	30
		100

	Machine M_2	
Centre des classes	Contenance (en kg)	Effectif
0,9825	[0,980 ; 0,985[	5
0,9875	[0,985 ; 0,990[	10
0,9925	[0,990 ; 0,995[	6
0,9975	[0,995 ; 1,000[	35
1,0025	[1,000 ; 1,005[	40
1,06125	[1,005 ; 1,117 5[	4
		100

1. Calculer la moyenne pour chacune des machines *M*1 et *M*2 après avoir rajouté le centre de chaque classe aux deux tableaux.

Méthode :
Si les valeurs de la série ont été regroupées en classe, les formules s'appliquent en prenant pour valeur de x_i le centre de chaque classe.

Pour la machine M1, on trouve :
$$\overline{x_1} = \frac{0{,}9825 \times 1 + 0{,}9875 \times 14 + \cdots + 1{,}0105 \times 30}{100} = 1$$

Pour la machine M2, on trouve :
$$\overline{x_2} = \frac{0{,}9825 \times 5 + 0{,}9875 \times 10 + \cdots + 1{,}06125 \times 4}{100} = 1$$

Les machines M1 et M2 remplissent en moyenne des paquets de sucre de 1kg.
Les affirmations des deux fabricants sont exactes.

2. Calculer l'écart-type pour chacune des machines *M*1 et *M*2 (Arrondir les résultats à 0,001 près).

Avec la calculatrice :
Pour la machine M1, on trouve à la calculatrice :
$$\sigma_1 \approx 0{,}008$$
Pour la machine M2, on trouve à la calculatrice :
$$\sigma_2 \approx 0{,}014$$

Avec la variance :
$$V_1 = \frac{1 \times (0{,}9825 - 1)^2 + \cdots + 30(1{,}0105 - 1)^2}{100}$$
$$= 0{,}00006895$$
$$\sigma_1 = \sqrt{V_1}$$
$$\approx 0{,}008$$

3. Comparer les deux séries à l'aide du couple moyenne-écart-type.

La machine M1 semble plus performante car son écart-type est plus petit, c'est-à-dire on aura le minimum d'écart autour de la moyenne.

Moyenne sans effectif :
La moyenne $\bar{x}$ de la série statistique dont les valeurs du caractère sont $x_1 ; \ldots ; x_N$ est le nombre :
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N} \times \sum_{i=1}^{N} x_i$$

Moyenne avec effectif :
La moyenne $\bar{x}$ de la série statistique dont les valeurs du caractère sont $x_1 ; \ldots ; x_p$ et les effectifs correspondants : $n_1 ; \ldots ; n_p$ est le nombre :
$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_p n_p}{N} = \frac{1}{N} \times \sum_{i=1}^{p} x_i n_i$$

Moyenne avec fréquence :
Si on note f_i la fréquence de la valeur x_i, on a :
$$\bar{x} = x_1 f_1 + x_2 f_2 + \cdots + x_p f_p = \sum_{i=1}^{p} x_i f_i$$

Variance :
La variance V est la moyenne des carrés des écarts entre chaque valeur et la moyenne $\bar{x}$.
$$V = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \cdots + n_p(x_p - \bar{x})^2}{N}$$
$$= \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{i=1}^{p} \boldsymbol{n_i}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^2$$

Ecart-type :
On appelle l'écart type de la série le réel $\boldsymbol{\sigma} = \sqrt{\boldsymbol{V}}$

Couple Moyenne - Ecart-type :
Le couple $(\bar{x} ; \sigma)$ prend en compte toutes les valeurs de la série, et sont, par conséquent, **influencées par les valeurs extrêmes.**
L'écart-type mesure la dispersion des valeurs autour de la moyenne : plus il est grand, plus les valeurs sont dispersées et moins la moyenne représente de façon significative la série.

Remarque : La moyenne est le nombre qui, substitué à chaque valeur de la série, redonne le même nombre.

III. Couple Médiane – écart interquartile.

Application 4 : Caractère quantitatif discret

Deux classes A et B de 1re S souhaitent comparer leurs moyennes trimestrielles de mathématiques. Les résultats obtenus sont les suivants.

CLASSE A

Notes sur 20	6	7	8	9	10	11	12	13	14	15	16	17	*Total*
Nombre d'élèves	0	1	3	3	6	4	5	4	3	1	0	0	30
ECC	0	1	4	7	13	17	22	26	29	30	30	30	
Fréquence (%)	0	3	10	10	20	13	17	13	10	3	0	0	100 (Arrondis)
FCC(%)	0	3	13	23	43	56	73	86	96	100	100	100	

CLASSE B

Notes sur 20	6	7	8	9	10	11	12	13	14	15	16	17	*Total*
Nombre d'élèves	1	1	2	3	2	5	3	3	2	3	1	1	27
ECC	1	2	4	7	9	14	17	20	22	25	26	27	
FCC(%)	4	7	15	26	33	52	63	74	81	93	96	100	

1. Donner la médiane et les quartiles de la classe A puis celle de la classe B. Interpréter les résultats.

Méthode 1 : Avec les FCC :

Classe A :

- $23\% < 25\% < \mathbf{43\%}$
  $\boldsymbol{Q_1 = 10}$
  Au moins 25% des notes sont ≤ à 10.
- $43\% < 50\% < \mathbf{56\%}$
  $\boldsymbol{M_e = 11}$
  Au moins 50% des notes sont ≤ à 11.
- $73\% < 75\% < \mathbf{86\%}$
  $\boldsymbol{Q_3 = 13}$
  Au moins 75% des notes sont ≤ à 11.

Classe B :

- $15\% < 25\% < \mathbf{26\%}$
  $\boldsymbol{Q_1 = 9}$
  Au moins 25% des notes sont ≤ à 9.
- $33\% < 50\% < \mathbf{52\%}$
  $\boldsymbol{M_e = 11}$
  Au moins 50% des notes sont ≤ à 11.
- $74\% < 75\% < \mathbf{81\%}$
  $\boldsymbol{Q_3 = 14}$
  Au moins 75% des notes sont ≤ à 14.

Méthode 2 : Avec les effectifs :

Classe A :

$N = 30$ est pair.

- $\frac{N}{4} = 7{,}5$
  Le premier quartile est la 8ème valeur.
  $Q_1 = 10$
- $\frac{N}{2} = 15$
  La médiane est la moyenne entre la 15ème et la 16ème valeur.
  $M_e = \frac{11+11}{2} = 11$
- $\frac{3N}{4} = 22{,}5$
  Le troisième quartile est la 23ème valeur.
  $Q_3 = 13$

Classe B :

$N = 27$ est impair.

- $\frac{N}{4} = \frac{27}{4} = 6{,}75$
  Le premier quartile est la 7ème valeur.
  $Q_1 = 9$
- $\frac{N+1}{2} = \frac{27+1}{2} = 14$
  La médiane est la 14ème valeur.
  $M_e = 11$
- $\frac{3N}{4} = 20{,}25$
  Le troisième quartile est la 21ème valeur.
  $Q_3 = 14$

Médiane :

La médiane M_e est la valeur qui partage la série en deux séries statistiques de mêmes effectifs.

En classant les valeurs de la série statistique dans **l'ordre croissant**, on a :

- Si N (effectif total) est impair M_e est la $\frac{N+1}{2}$ème valeur (valeur centrale de la série ordonnée).
- Si N est paire, la médiane est la demi-somme (moyenne) des valeurs $\frac{N}{2}$ et $\frac{N}{2}+1$

Quartile :

Les quartiles Q_i sont les valeurs qui partagent la série en quatre séries statistiques de même effectif.

- $\boldsymbol{Q_1}$ est la plus petite des valeurs de la série tel que $\mathbf{25\%}$ au moins de ces valeurs lui soient inférieures ou égales.
- $\boldsymbol{Q_3}$ est la plus petite des valeurs de la série tel que $\mathbf{75\%}$ au moins de ces valeurs lui soient inférieures ou égales.
- Le deuxième quartile peut être vu comme la **médiane** de la série. C'est la plus petite des valeurs de la série telle qu'au moins $\mathbf{50\%}$ de ces valeurs lui soient inférieures ou égales.

Intervalle interquartile :

L'intervalle $[Q_1\,;Q_3]$ est appelé **l'intervalle interquartile.**

Ecart interquartile :

L'écart interquartile est la différence entre Q_3 et Q_1 :

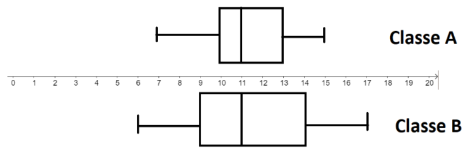
$$E_Q = Q_3 - Q_1$$

Etendue :

L'étendue est la **différence** entre la plus **grande valeur** et la plus **petite valeur** prise par le caractère :

$$e = x_{max} - x_{min}$$

2. Représenter le diagramme en boite de chaque classe au-dessus de l'axe gradué.



3. Déterminer l'écart interquartile des notes de chaque classe et interpréter les résultats.
   Comparer les deux séries à l'aide du couple médiane-écart interquartile.

Pour la classe A, on a $E_{Q_A} = Q_3 - Q_1 = 13 - 10 = 3$
Pour la classe B, on a $E_{Q_B} = Q_3 - Q_1 = 14 - 9 = 5$
La classe A semble avoir un niveau plus homogène que la classe B car les notes sont plus resserrées autour de la médiane et l'étendue des notes est plus petite.

Diagramme en boite :

Un diagramme en boîte est une représentation graphique qui résume une série statistique ordonnée à l'aide de cinq paramètres :

Les valeurs extrêmes (minimum et maximum), la médiane $M_e$et les quartiles Q_1et Q_3.

L'étendue de la série est : $e = x_{max} - x\_min$.

C'est la longueur totale du diagramme.

L'intervalle $[Q_1\,;Q_3]$est l'intervalle interquartile, il contient au moins **50%** des valeurs de la série.

L'écart interquartile est la longueur de la boîte.

Couple médiane – écart interquartile

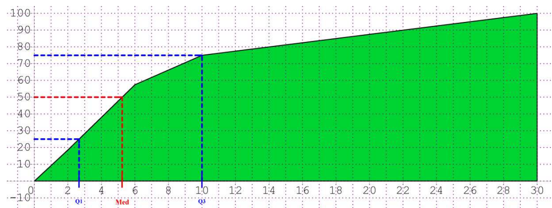
Le couple $(M_e\,;Q_3 - Q_1)$ permet de caractériser la série par une valeur centrale M_e et un réel qui donne une mesure de son étalement autour de la médiane. Il ne prend pas en compte les valeurs extrêmes ; il caractérise surtout les valeurs centrales.

Remarque : On préfèrera la médiane à la moyenne pour des séries fortement « asymétrique ».

Application 5 : Caractère quantitatif continu

On a relevé les temps d'attente des skieurs sur une remontée mécanique pendant 1 heure :

Temps d'attente (en min)	$[0;2[$	$[2;6[$	$[6;10[$	$[10;30]$	Total
Nombre de skieurs	20	42	19	27	108
ECC	20	62	81	108	
FCC	19%	57%	75%	100%	



Donner la médiane et les quartiles de cette série.

Interpréter les résultats dans le contexte de l'exercice.

- $M_e \approx 5{,}2$.
  Au moins 50% des skieurs attendent moins de 5 minutes 12 secondes (environ).
- $Q_1 \approx 2{,}7$.
  Au moins 25% des skieurs attendent moins de 2 minutes 42 secondes (environ).
- $Q_3 = 10$.
  Au moins 75% des skieurs attendent moins de 10 minutes.

Courbe des fréquences cumulées croissantes :

La courbe des fréquences cumulées croissantes associée à une série statistique dont le caractère est quantitatif continu est la ligne brisée passant par les points ayant pour abscisse la borne supérieur des classes (deuxième nombre dans l'intervalle) donnant les valeurs du caractère et pour ordonnée la fréquence cumulée croissante correspondant à ces valeurs du caractère.

Cas particulier : Caractère quantitatif continu

Médiane et quartiles dans le cas où les valeurs de la série sont regroupées par classe.

On réalise la courbe de fréquences cumulées croissantes.

- Pour déterminer la médiane, on cherche l'abscisse du point de cette courbe, dont l'ordonnée est égale 50%.
  Cette abscisse est alors la médiane de la série.
- Pour déterminer le premier quartile, on cherche l'abscisse du point de cette courbe, dont l'ordonnée est égale à 25%.
  Cette abscisse est alors le premier quartile de la série.
- Pour déterminer le troisième quartile, on cherche l'abscisse du point de cette courbe, dont l'ordonnée est égale à 75%.
  Cette abscisse est alors le troisième quartile de la série.