Statistique descriptive

Ce que dit le programme :

L'étude et la comparaison de séries statistiques menées en classe de seconde se poursuivent avec la mise en place de nouveaux outils dans l'analyse de données.
L'objectif est de faire réfléchir les élèves sur des données réelles, riches et variées issues, par exemple, de fichiers mis à disposition par l'INSEE (Institut National de la Statistique et des Études économiques).

CONTENUS	CAPACITÉS ATTENDUES	COMMENTAIRES
Statistique descriptive, Analyse de données Caractéristiques de dispersion :	Utiliser de façon appropriée les deux couples usuels qui permettent de résumer une série statistique : (moyenne, écart-type) et (médiane, écart interquartile).	On utilise la calculatrice ou un logiciel pour déterminer la variance et l'écart-type d'une série statistique.
Variance, écart-type. Diagramme en boîte.	Étudier une série statistique ou mener une comparaison pertinente de deux séries statistiques à l'aide d'un logiciel ou d'une calculatrice.	Des travaux réalisés à l'aide d'un logiciel permettent de faire observer des exemples d'effets de structure lors du calcul de moyennes.

I. Paramètres de position d'une série statistique

1.1) Moyenne

1.1.a) Moyenne arithmétique

On considère ***une série statistique à une variable*** quantitative (***caractère*** quantitatif), observé(e) sur N individus d'une population E. Cette série statistique peut être représentée dans un tableau de données :

Individus i	1	2	...	N
Valeurs x_i	x_1	x_2	...	x_N

N est l'***effectif total*** de la population ;
x_i représente la valeur du caractère pour l'individu i.
Le nombre i peut aussi être interprété comme un « *indice* » qui indique le rang de l'individu i.

Définition 1.
Nous savons déjà calculer ***la moyenne*** notée $\overline{x}$ de N valeurs $x_1, x_2, \cdots, x_N$ (par exemple la moyenne des notes dans une matière).

Pour calculer la moyenne des x_i , il suffit de calculer la somme de toutes les valeurs x_i puis diviser par l'effectif total, ici N. On a alors :

$$\boxed{\overline{x}=\frac{x_1+x_2+\cdots+x_N}{N}}$$

$\overline{x}$ s'appelle aussi la ***moyenne arithmétique*** des N valeurs x_i.

Ici, nous n'avons affecté aucun coefficient à ces notes (dont toutes les valeurs ont un coefficient = 1). On dit aussi que $\overline{x}$ représente la ***moyenne brute*** de la série.

Exemple 1.

On considère la série statistique suivante : 9 ; 10 ; 11 ; 12 ; 16 ; 20. Par exemple, les notes de mathématiques d'un élève au 1[er] trimestre. Calculer la moyenne brute de cette série statistique.

Ici, il y a 6 notes, donc $N = 6$.
La moyenne de ces 6 notes est :

$$\overline{x}=\frac{9+10+11+12+16+20}{6}=\frac{78}{6}=13$$

La ***moyenne brute*** (sans coefficients) de ces notes est donc égale à 13.

1.1.b) Moyenne *pondérée* (avec coefficients ou effectifs partiels)

On considère ***une série statistique à une variable*** quantitative, observée sur N individus d'une population E. On relève k valeurs possibles $x_1, x_2, \cdots, x_k$ du caractère dans cette population.

On note n_1 l'***effectif partiel*** de x_1, donc x_1 se répète n_1 fois.
n_2 l'effectif partiel de x_2; ... et n_k l'effectif partiel de x_k.
On obtient alors la formule :

Effectif total = Somme des effectifs partiels

ou encore : $$\boxed{N=n_1+n_2+\cdots+n_k}$$

On obtient alors une série statistique à une variable que l'on peut présenter dans un tableau de données :

Valeurs x_i	x_1	x_2	...	x_k	Total
Effectifs partiels	n_1	n_2	...	n_k	N

Ici, x_i représente la i-ème valeur du caractère et note n_i l'effectif partiel de x_i.

Définition 2.
On considère ***une série statistique à une variable*** quantitative, observée sur N individus d'une population E et prenant $\boldsymbol{k}$ valeurs $x_1, x_2, \cdots, x_k$ affectées des effectifs partiels $n_1, n_2, \cdots, n_k$ respectivement.
Alors, ***la moyenne*** notée $\bar{x}$ des k valeurs $x_1, x_2, \cdots, x_k$ se calcule comme suit :

$$\boxed{\bar{x}=\frac{n_1x_1+n_2x_2+\cdots+n_kx_k}{n_1+n_2+\cdots+n_k}} \quad \text{ou encore :} \quad \boxed{\bar{x}=\frac{n_1x_1+n_2x_2+\cdots+n_kx_k}{N}}$$

$\bar{x}$ s'appelle aussi la ***moyenne*** des N valeurs x_i affectés des effectifs patiels n_i.
Ce qui revient à affecter un coefficient n_i à chaque valeurs x_i.
On dit que $\bar{x}$ représente la ***moyenne pondérée*** ou simplement ***moyenne*** de la série.

Exemple 2.

On considère les deux séries statistique suivante :

A : 8 ; 8 ; 12 ; 12 ; 14 ; 12 et B : 9 ; 10 ; 11 ; 12 ; 16 ; 20.

Par exemple, les notes de mathématiques d'un élève au 1er trimestre.

1. Calculer la moyenne de la série A de deux manières.
2. Calculer la moyenne de la série B, sachant que 16 et 20 sont des notes de DM – devoir maison – donc de coefficient 1 ; et que les autres notes correspondent à des DS – devoirs surveillés – donc de coefficient 2.

1°) Ici, il y a 6 notes, donc $N=6$.

1ère *manière* : On calcule la moyenne brute de ces 6 notes. On obtient :

$$\overline{x_A}=\frac{8+8+12+12+14+12}{6}=\frac{66}{6}=11 \quad \text{donc} \quad \overline{x_A}=11.$$

2ème *manière* : On remarque que les notes se répètent. Sur les six notes il n'y a que trois notes différentes. 8 se répète 2 fois ; 12 se répète 3 fois et 14 apparaît 1 fois.
On calcule alors une moyenne pondérée :

$$\overline{x_A}=\frac{2\times8+3\times12+1\times14}{6}=\frac{66}{6}=11 \quad \text{donc} \quad \overline{x_A}=11.$$

On obtient bien le même résultat.

2°) Dans cette deuxième question. Les notes sont affectées de différents coefficients.

$$\overline{x_B}=\frac{2\times9+2\times10+2\times11+2\times12+1\times16+1\times20}{2+2+2+2+1+1}=\frac{120}{10}=12.$$

Cet élève avait obtenu une *moyenne brute* $\bar{x}=13$ et si on tient compte des coefficients, sa *moyenne* baisse à $\bar{x}=12$.

1.1.c) Moyenne d'une série statistique dont les valeurs sont groupées en classes

Définition 3.
On considère une série statistique à une variable quantitative, dont les valeurs sont groupées en k classes $[x_0 ; x_1[$; $[x_1 ; x_2[; \cdots ;$ $[x_{k-1} ; x_k[$; affectées des effectifs partiels $n_1, n_2, \cdots, n_k$ respectivement.

On appelle c_i le centre de la i-ème classe, c'est-à-dire la moyenne des deux bornes de chaque classe. Alors ***la moyenne*** de la série statistique dont les ***valeurs sont groupées en classes***, est égale à la moyenne des k centres $c_1, c_2, \cdots, c_k$ dont les effectifs partiels correspondants sont $n_1, n_2, \cdots, n_k$ respectivement. Ce qui donne :

$$\boxed{\bar{x}=\frac{n_1 c_1+n_2 c_2+\cdots+n_k c_k}{n_1+n_2+\cdots+n_k}}$$ ou encore : $$\boxed{\bar{x}=\frac{n_1 c_1+n_2 c_2+\cdots+n_k c_k}{N}}$$

Exemple 3.
On considère la série statistique suivante, représentant la répartition des temps mis pour aller à l'école des élèves dans une classe de Seconde de 35 élèves :

Temps t_i (en min)	[0;5[	[5;10[	[10;15[	[15;20[	[20;25[	[25;30[	Total
Effectifs n_i	3	7	8	12	4	1	35

Calculer le temps moyen que met un élève ce cette classe pour aller à l'école.

Les valeurs de cette série sont groupées en classes. Autrement dit, on ne connaît pas avec précision les valeurs de la série.
Pour calculer la moyenne pondérée d'une telle série, on doit calcule les centres des classes : c_i = moyenne des valeurs extrêmes de chaque classe $[a ; b[$: $c_i=\frac{a+b}{2}$.
On obtient ainsi le tableaux des effectifs avec les centres des classes :

Temps t_i (en min)	[0;5[	[5;10[	[10;15[	[15;20[	[20;25[	[25;30[	Total
Effectifs n_i	3	7	8	12	4	1	35
Centres c_i	2,5	7,5	12,5	17,5	22,5	27,5	X

La moyenne est alors égale à :

$$\bar{x}=\frac{n_1 c_1+n_2 c_2+\cdots+n_k c_k}{n_1+n_2+\cdots+n_k}$$

$$\bar{x}=\frac{3\times 2,5+7\times 7,5+8\times 12,5+12\times 17,5+4\times 22,5+1\times 27,5}{35}$$

$$\bar{x}=\frac{487,5}{35}=13,928\ldots$$

Conclusion. Le temps moyen que met un élève de cette classe, entre son domicile et le lycée est d'environ 14 minutes.

1.1.d) Utilisation de la calculatrice

À la calculatrice, on rentre toutes les valeurs de la série (ou les centres des classes) dans la liste **L1** et les effectifs dans **L2** (pour une moyenne avec coefficients ou effectifs partiels), puis on calcule les différents éléments de la série dont la moyenne, la médiane,... avec l'instruction **1-Var** ou **1-Var-Stats** ou l'équivalent.

Casio 35+ ou supérieur	TI 82 ou supérieur	Numworks (la nouvelle !)
MENU STATS On obtient 4 colonnes **L1** pour les valeurs x_i **L2** pour les effectifs partiels n_i (Taper chaque valeur puis Entrer) Puis Touche F2 Calc puis Touche F1 1-Var Vous obtenez la liste des éléments caractéristiques de la série : $\overline{x}$ = moyenne de la série ; $\sum x$ = Somme des valeurs $\sum x^2$ =Somme des carrés S x = pas au pgm σ x = écart-type n =effectif total min X = Valeur minimale Q_1 = 1er quartile Med = Médiane de la série Q_3 = 3ème quartile max X = Valeur maximale	Touche STAT Puis Edit On obtient 3 colonnes **L1** pour les valeurs x_i **L2** pour les effectifs partiels n_i (Taper chaque valeur puis Entrer) Touche STAT Puis Calc On obtient 1 liste : Cliquez sur 1 : **1-Var Stats** On obtient une nouvelle fenêtre : 1-Var Stats **List : L1** (avec 2nde **1**) **FreqList : L2** (avec 2nde **2**) **Calculate :** Taper Entrer Vous obtenez une liste des éléments caractéristiques de la série ... Voir colonne de gauche	Cliquer sur Statistiques puis OK. Vous obtenez quatre onglets : Données. Histog. Boîte. Stats. Cliquez sur Données vous obtenez un tableau pour 4 variables : Valeurs V1. Effectifs N1. pour la variable V1...etc. Saisir les données dans V1 et les effectifs dans N1, puis avec les flèches de direction remonter à l'un des onglets Données. Histog. Boîte. Stats. vous obtenez ce que vous voulez. Choisissez Stats pour obtenir la liste des éléments caractéristi-ques de la série ... Voir colonne de gauche

1.1.e) Calcul de la moyenne en utilisant les fréquences

On considère *une série statistique à une variable* quantitative, observée sur N individus d'une population E et prenant k valeurs $x_1, x_2, \cdots, x_k$ affectées des effectifs partiels $n_1, n_2, \cdots, n_k$ respectivement.

Les fréquences correspondantes sont $f_1, f_2, \cdots, f_k$ avec : $\boxed{f_i = \frac{n_i}{N}}$

Définition 4.
La moyenne notée $\overline{x}$ des k valeurs $x_1, x_2, \cdots, x_k$ affectées des fréquences respectives $f_1, f_2, \cdots, f_k$ se calcule comme suit :

$$\boxed{\overline{x} = f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}$$

En effet : $\bar{x}=\dfrac{n_1x_1+n_2x_2+\cdots+n_kx_k}{N}$

$$\bar{x}=\frac{n_1x_1}{N}+\frac{n_2x_2}{N}+\cdots+\frac{n_kx_k}{N}$$

$$\bar{x}=\frac{n_1}{N}x_1+\frac{n_2}{N}x_2+\cdots+\frac{n_k}{N}x_k$$

D'où : $\bar{x}=f_1x_1+f_2x_2+\cdots+f_kx_k$

1.2) Médiane d'une série statistique

1.2.a) Définition

On considère *une série statistique à une variable* quantitative, observée sur N individus d'une population E et prenant k valeurs $x_1, x_2, \cdots, x_k$ affectées des effectifs partiels $n_1, n_2, \cdots, n_k$ respectivement.

Définition 5.

On appelle ***médiane de la série statistique***, toute valeur ***m*** du caractère qui partage la série en deux parties de même effectif. Il y a donc autant de valeurs inférieures que de valeurs supérieures à la médiane.

Remarque

Dans cette définition, une médiane pourrait prendre *plusieurs valeurs possibles* dans certaines situations. Nous allons voir que nous allons donner ***une méthode*** qui permet à tous de trouver ***la même médiane***.

1.2.b) Méthodes pour calculer la médiane d'une série

On considère *une série statistique à une variable* quantitative, observée sur N individus d'une population E et prenant k valeurs $x_1, x_2, \cdots, x_k$ affectées des effectifs partiels $n_1, n_2, \cdots, n_k$ respectivement.
On suppose que les valeurs de la série sont rangés par ordre croissant.

1ère méthode. *Par un calcul direct :*

Propriété 1. (*Très importante*).

On procède en plusieurs étapes :

- Tout d'abord, ne devons commencer par ranger les valeurs la série statistique dans l'ordre croissant (avec répétition si nécessaire) ;
- Déterminer l'***effectif total***. Ici ***N*** ;

- Déterminer ***le rang de la médiane*** dans la série ;
- La médiane est la valeur correspondante à ce rang trouvé.

On distingue deux cas possibles :

1. Si l'effectif total ***N est impair***, alors la médiane est égale à ***la valeur centrale*** de la série ; son rang est $\frac{N+1}{2}$
2. Si l'effectif total ***N est pair***, alors *toute valeur comprise entre les deux valeurs centrales* est une médiane de la série. En général, on prend pour médiane ***la moyenne des deux valeurs centrales***, de rangs $\frac{N}{2}$ et $\frac{N}{2}+1$.

Notation.

La médiane est notée généralement : Me.

Exemple

On considère les deux séries statistiques suivantes. Calculer la médiane de chaqye série

1°) Série A : 8 ; 9 ; 9 ; 10 ; 11 ; 12 ; 13 ; 14 ; 15.
2°) Série B : 8 ; 9 ; 9 ; 10 ; 11 ; 12 ; 13 ; 14 ; 15 et 16

1°) La série A est déjà rangée par ordre croissant et contient 9 valeurs.
L'effectif total est égal à $N = 9$. C'est un nombre impair. Donc, la médiane est égale à ***la valeur centrale*** de la série.

Le rang le la médiane est : $\frac{N+1}{2}=\frac{9+1}{2}=\frac{10}{2}=5$

Me est donc la cinquième valeur de la série ordonnée, rangée par ordre croissant. On obtient : 8 ; 9 ; 9 ; 10 ; **11** ; 12 ; 13 ; 14 ; 15.

Conclusion. La médiane de la série A est donc : $\boxed{Me=11}$

1°) La série B est (aussi) déjà rangée par ordre croissant et contient 10 valeurs.
L'effectif total est égal à $N = 10$. C'est un nombre pair. Donc, la médiane est égale à ***la moyenne des deux valeurs centrales*** de la série.

Pour trouver le rang des deux valeurs centrales, je calcule : $\frac{N}{2}=\frac{10}{2}=5$

Me est donc égale à la moyenne de la 5ème et la 6ème valeurs de la série ordonnée, rangée par ordre croissant. On obtient : 8 ; 9 ; 9 ; 10 ; **11** ; **12** ; 13 ; 14 ; 15 ; 16

Conclusion. La médiane de la série B est donc : $Me=\frac{11+12}{2}=11{,}5$. $\boxed{Me=11{,}5.}$

2ème méthode. Sur un graphique :

Propriété 2. (*Très importante*).

On considère *une série statistique à une variable quantitative*, dont les valeurs sont groupées en k classes $[x_0 ; x_1[$; $[x_1 ; x_2[; \cdots ;$ $[x_{k-1} ; x_k[$; affectées des effectifs partiels $n_1, n_2, \cdots, n_k$ ou des fréquences $f_1, f_2, \cdots, f_k$ respectivement

1°) Si on construit le polygone des effectifs cumulés croissants (ECC), alors la médiane est la valeur qui correspond à la moitié de l'effectif total $\frac{N}{2}$.

2°) Si on construit le polygone des fréquences cumulées croissantes (FCC), alors la médiane est la valeur qui correspond à une **FCC = 0,5**.

Exemple 4

L'accueil téléphonique d'une entreprise a reçu 120 appels entre 9h et 13h, répartis comme suit :

Heures	9h-10h	10h-11h	11h-12h	12h13h	Total
Nombre d'appels	25	45	30	20	120

1°) Construire le polygone des effectifs cumulés croissants (ECC) de cette série.

2°) Déterminer la médiane de la série par lecture graphique. Donner votre résultat en heures et minutes.

3°) Peut-on faire un calcul direct, en supposant que la répartition des appels est « *uniforme* ».

4°) Reprendre l'exercice en utilisant les fréquences cumulées croissantes.

1°) On calcule les ECC dans un tableau :

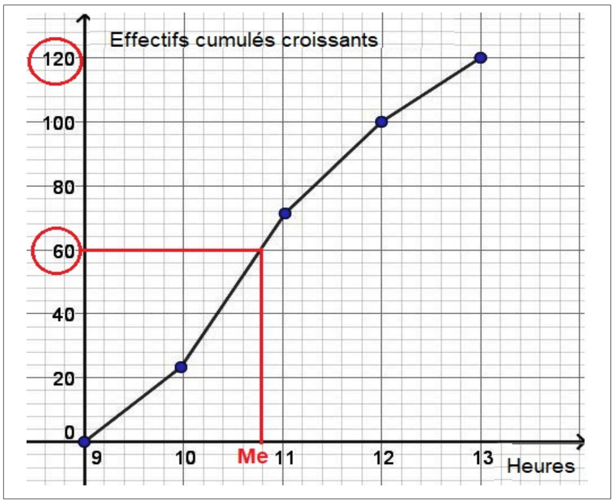
Heures ***jusqu'à***	9h	10h	11h	12h	13h
Nombre d'appels	0	25	70	100	120

A 10 heures, on n'a pas encore atteint la moitié de l'effectif total et à 11 heures, on vient de le dépasser.

Donc $Me \in [10,11]$.

On dit que l'intervalle [10 ; 11] est la ***classe médiane de la série***.

On obtient la courbe ci-contre :



2°) L'éffectif total est égal à 120. Donc la moitié de l'effectif est égale à 60.

Par lecture graphique. La médiane est l'antécédent de 60. Ce qui donne environ :

$$\boxed{\mathrm{Me}=10{,}8}$$

Pour convertir ce résultat en heures et minutes, on convertit les décimales en minutes en faisant un tableau de proportionnalité.

1 heure correspond à 60 min

0,8 heure correspond à x min.

Ce qui donne : $x\times 1=0{,}8\times 60$. On obtient $x=48$ minutes.

Conclusion. La médiane de cette série est égale à $\boxed{\text{Me = 10h 48 min}}$.

Remarque

On aurait pu faire autrement, en posant : 1h = 60 min. Donc : 0,1h = $\frac{1}{10}\mathrm{h}=6\,\mathrm{min}$.

Et par suite : $0{,}8\,\mathrm{h}=8\times 0{,}1\,\mathrm{h}=8\times 6=48\,\mathrm{min}$. D'où le résultat.

3°) Pour déterminer la médiane par un calcul direct, on utilise le théorème de Thalès.

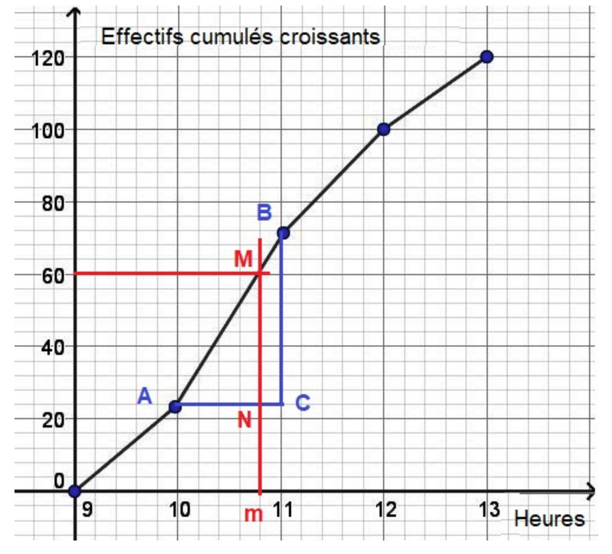
On pose *Me = m*. Dans le repère orthogonal (O ; I ; J), on considère le triangle *ABC* défini par les points *A*, *B* et C correspondant au ECC comme suit :

A(10 ; 25), B(11 ; 70) et C(11 ; 25).

Le point M qui correspond à la médiane a pour coordonnées :

M(m ; 60) et N(m ; 25).

Dans le triangle ABC, on a : $M\in[AB]$, $N\in[AC]$ et les droites (MN) et (BC) sont parallèles à l'axe des ordonnées.



Donc, d'après le théorème de Thalès, on a égalité des rapports :

$$\frac{AM}{AB}=\frac{AN}{AC}=\frac{MN}{BC}$$

Je garde les deux derniers rapports : $\frac{AN}{AC}=\frac{MN}{BC}$

Ce qui donne : $\frac{m-10}{11-10}=\frac{60-25}{70-25}$ donc : $\frac{m-10}{1}=\frac{35}{45}$.

Donc : $m=10+\frac{35}{45}$ ou encore : $m=\frac{485}{45}\simeq 10{,}77778\simeq 10{,}8$. Donc : $\boxed{\mathrm{Me}\simeq 10{,}8}$

On retrouve une valeur exacte de la médiane qui, arrondi au dixième donne la même valeur que le résultat obtenu par lecture graphique !

4°) Reprendre l'exercice en utilisant les fréquences cumulées croissantes. C'est facile.

1.3) Mode d'une série statistique

Définition 6.

Dans une série statistique à une variable, la valeur la plus fréquente s'appelle ***le mode de la série statistique***. C'est la (ou les) valeur(s) du caractère dont l'effectif est le plus grand.

Dans une série statistique à une variable où les valeurs sont groupées en classes, la classe la plus fréquente s'appelle ***le mode*** ou ***la classe modale*** de la série statistique.

1.4) Résumer une série statistique

La moyenne, la médiane et le mode sont les valeurs principales de ***tendance centrale d'une série statistique***. Elles permettent de synthétiser la série statistique étudiée à l'aide d'un petit nombre de « ***valeurs caractéristiques*** ».

On pourra faire un autre résumé en utilisant les indicateurs de dispersion d'une série statistique.

II. Paramètres de dispersion

2.1) Étendue d'une série statistique

Définition 7.

L'étendue d'une série statistique est la différence entre la plus grande et la plus petite valeur de la série. Si x_{min} et x_{max} désignent les valeurs minimale et maximale de la série, alors l'étendue, notée **e**, de la série statistique est : $\boxed{e=x_{max}-x_{min}.}$

Exemple 5 de la notion de dispersion

On considère les deux séries statistiques formées des 5 notes (rangées dans l'ordre croissant) de deux élèves au 1er trimestre.

- Série A : 8 ; 9 ; 9 ; 10 ; 11 ; 11 ; 12. Série B : 2 ; 5 ; 7 ; 10 ; 13 ; 15 ; 18.

Dans les deux séries, il y a 7 valeurs, la médiane est égale à la valeur centrale : Me = M'e = 10. Elles ont donc la même médiane, la même moyenne $\overline{x_A}=\overline{x_B}=10$, mais pas la même étendue : $e_A=12-8=4$ et $e_B=18-2=16$.

Ces deux séries ne se ressemblent pas ! Les valeurs de la série A sont très *resserrées*, l'élève A, malgré les difficultés, n'a pas de très bonnes notes, mais n'a pas de très mauvaises notes non plus ! ***L'élève A est régulier***.

Les valeurs de la série B sont très *dispersées*, l'élève B est très lunatique, il est capable d'avoir de très bonnes notes, mais également de très mauvaises notes ! ***L'élève B est très irrégulier***.

Remarque

Nous verrons en classe de première un outil de calcul de cette dispersion, appelé ***l'écart-type***, noté s ou σ (*sigma*). Cela revient à calculer, d'une certaine façon, la moyenne des écarts absolus à la moyenne. (Classes de 1ère S, ES et STMG).

Dans une ***répartition (ou distribution) « normale »*** des valeurs, si l'effectif total est assez grand, ***68% des valeurs sont comprises entre les valeurs*** $\bar{x}-\sigma$ ***et*** $\bar{x}+\sigma$ **.**
Ces deux séries n'ont donc pas le même écart-type. De plus $\sigma_A < \sigma_B$.
L'écart-type de la série A est plus petit (valeurs resserrés) que l'écart-type de la série B (valeurs dispersées). Tout un programme !!

2.2) Les quartiles

Définition 8.

On considère une variable statistique quantitative dont les valeurs sont rangées par ordre croissant.
Le premier quartile est égal à la plus petite valeur Q_1 des termes de la série pour laquelle ***au moins 25% des données sont inférieures ou égales à*** $\mathbf{Q_1}$.

Le troisième quartile est égal à la plus petite valeur Q_3.
des termes de la série pour laquelle ***au moins 75% des données sont inférieures ou égales à*** $\mathbf{Q_3}$.

Remarque

Le deuxième quartile n'est autre que la médiane, qui correspond à 50% des effectifs de la série statistique : $\mathbf{Q_2 = Me}$.

En pratique

Etant donné une série statistique à une variable quantitative, d'effectif total N et dont les valeurs sont rangées par ordre croissant. (Sinon, on commence par les ranger par ordre croissant.
Pour déterminer les deux quartiles, on partage la série en 4 groupes de même effectif.
On calcule $\frac{N}{4}$ qui correspond à 25% des valeurs. ***Le rang de*** Q_1 est le premier entier supérieur ou égal à $\frac{N}{4}$. De même, pour déterminer Q_3, on calcule $\frac{N\times 3}{4}$ qui correspond à 75% des valeurs. ***Le rang de*** Q_3 est le premier entier supérieur ou égal à $\frac{N\times 3}{4}$.

Exemple 6

Déterminer l'étendue, la médiane, le premier et le troisième quartiles de la série statistique suivante : 20 ; 52 ; 31 ; 4 ; 78 ; 5 ; 62 ; 34 ; 4 ; 9 ; 10 ; 45 ; 12.

a) Calcul de l'étendue

Les valeurs minimale et maximale de la série sont : x_{min}= 4 et x_{max} = 78.
Donc e = x_{max}−x_{min} = 78 − 4. Donc **e = 74**.

b) Recherche de la médiane

- Je commence par ranger les valeurs de la série par ordre croissant :
  4 ; 4 ; 5 ; 7 ; 10 ; 12 ; 20 ; 31 ; 34 ; 49 ; 52 ; 62 ; 78.
- L'effectif total de la série est $N = 13$
- N est ***impair***. Donc, la médiane est égale à la valeur centrale de la série.
- Donc le rang de la médiane est $\frac{N+1}{2}=\frac{13+1}{2}=7.$ Donc la médiane est la 7ème valeur de la série. Par conséquent : **Me = 20**.

c) Recherche des quartiles

- L'effectif total de la série est $N = 13$.
- Je divise N par 4 et j'obtiens le rang du 1er quartile : 13 ÷ 4 = 3,25. Donc Q_1 est la 4ème valeur de la série. Soit **$Q_1 = 7$**.
- De même, je multiplie (13÷4) par 3 et j'obtiens le rang du 3ème quartile : (13÷4)×3=9,75. Donc Q_3 est la 10-ème valeur de la série. Soit **Q3=49** .

Remarque

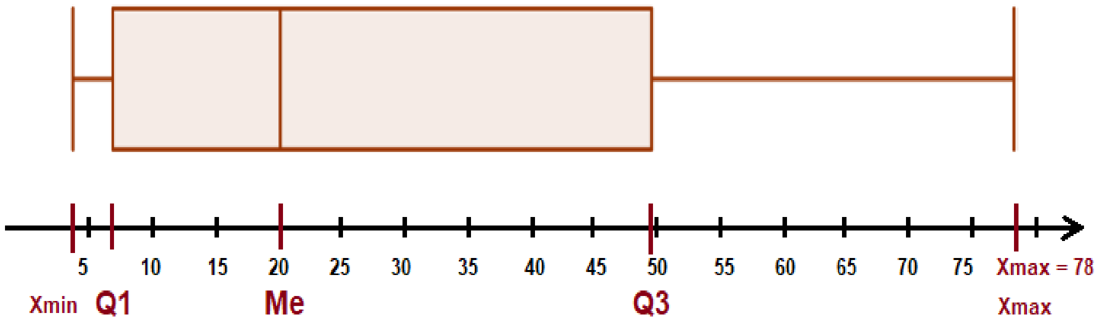
D'une manière analogue, on pourrait définir ***les déciles*** ***d'une série statistique*** correspondant : D_1 à 10%, D_2 à 20%,... et D_9 à 90% de l'effectif total de la série.
Les deux déciles les plus importants sont : D_1 et D_9. Par exmple, on s'intéresse aux 10% les plus riches ou les plus pauvres d'une population ...

2.3) Le diagramme en boîte

Définition 9.

Un ***diagramme en boîte*** — appelé aussi ***boîte à moustaches*** — est une illstration de cinq des paramètres d'une série statistique : le minimum x_{min}, le premier quartile, la médiane, le troisième quartile et le maximum x_{max} sur une graduation couvrant toute l'étendue de la série. [*Bien choisir l'unité* !]

En reprenant, l'exemple précédent, comme e = 74, nous devons créer une graduation de 74 unités. En prenant le mm, on peut placer x_{min}= 4 ; Q1=7 ; Me=20 ; Q3=49 ; et x_{max}= 78.
Au-dessus de la graduation, à 1 cm environ (ou un grand carreau), on place les deux moustaches aux extrêmités puis un rectangle joignant les deux quartiles et une barre dans le rectangle représentant la médiane.



En généram, on utilise des diagrammes en boîte pour comparer deux séries statistiques. On construira deux diagrammes en boîte sur une même graduation allant du minimum de x_{min} et x'_{min} jusqu'au maximum de x_{max} et x'_{max}.

Exemple 7 non résolu

Un médecin effectue des recherches sur l'efficacité d'un nouveau médicament bêta-bloquant [famille de médicaments, destinés à diminuer le rythme cardiaque des malades atteints de tachycardie (pouls supérieur à 60 battements par minute].
Il a donc séparé les malades en deux groupes : le groupe A reçoit le traitement du nouveau médicament et le groupe B reçoit un « placebo » (médicament sans principe actif).

Groupe A : 74 – 91 – 91 – 84 – 95 – 93 – 95 – 95 – 102 – 81 – 116 – 88 – 95 – 74 – 88 – 95 – 109 – 83 – 114 – 88 – 89 – 95 – 88 – 89 – 95 – 96

Groupe B : 94 – 95 – 113 – 95 – 104 – 113 – 94 – 144 – 105 – 153 – 79 – 153 – 123 – 108 – 114 – 92 – 110 – 123 – 84 – 93 – 83 – 123 – 123 – 114 – 96 – 104 – 94 – 97 – 93 – 82 – 98 – 82 – 83 – 105 – 83 – 105 – 93 – 94 – 84 – 93.

1°) Déterminer, à la calculatrice, les moyennes, les médiane et les deux quartiles des deux séries.
2°) Construire les diagrammes en boîte de chacune des deux séries dans un même graphique, en utilisant la même graduation.
3°) L'effet du médicament semble-t-il satisfaisant ? Expliquez.

2.4) Variance et écart-type d'une série statistique

2.4.a) Variance et écart-type d'une série statistique simple

Définition 10.
On considère ***une série statistique à une variable*** quantitative, observée sur N individus d'une population E et prenant N valeurs $x_1, x_2, \cdots, x_N$.
Alors, ***la variance de la série statistique***, notée V, désigne la moyenne des carrés des écarts à la moyenne $\bar{x}$. Autrement dit :

$$\boxed{V=\frac{1}{N}\left[(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_N-\bar{x})^2\right]}$$

Avec le signe Σ = "Somme" et i = "indice" variant de 1 à N :

$$\boxed{V=\frac{1}{N}\sum_{i=1}^{i=N}\left(x_i-\overline{x}\right)^2}$$

L'écart-type (lire "*sigma*") de la série est défini par : $\boxed{\sigma=\sqrt{V}}$ ou $\boxed{V=\sigma^2}$.

En général, la variance sert à calculer l'écart-type. C'est ce dernier qui permet de faire des comparaisons entre deux séries statistiques. Comme la variance, l'écart-type permet de caractériser la dispersion des valeurs x_k par rapport à la moyenne $\overline{x}$.
Une différence d'utilisation entre σ et $V=\sigma^2$, est que σ est *de même dimension que les valeurs* x_k, donc les valeurs x_k peuvent être directement comparées à σ.

Exemple 8.

On considère deux séries statistiques donnant les notes de deux élèves à 5 contrôles de mathématiques. On cherche à les comparer :

Elève 1 : 9 ; 11 ; 10 ; 8 ; 12
Elève 2 : 3 ; 17 ; 10 : 5 : 15

- Tout d'abord, on commence par calculer la moyenne et la médiane de chacune des deux séries :
  Ces deux élèves ont la même moyenne $\overline{x_1}=10=\overline{x_2}$ et la même médiane $Me_1 = 10 = Me_2$. Donc ces paramètres ne permettent pas de les comparer.
- On calcule l'étendue des deux séries : $e_1 = 12 - 9 = 3$ et $e_2 = 17 - 3 = 14$.
  Déjà, $e_1 < e_2$, donc l'étendue permet de voir que les notes de l'élève 1 sont « plus ramassées » et les notes de l'élève 2 sont « plus dispersées ». On pourrait en déduire que l'élève 1 est « régulier » et l'élève 2 est « irrégulier ».
- Calculons maintenant l'écart-type « *à la main* ». Pour cela, nous avons besoin de construire des tableaux de valeurs. Pour l'élève 1, on a :

Valeurs x_i	8	9	10	11	12
$x_i-\overline{x}$	– 2	– 1	0	1	2
$\left(x_i-\overline{x}\right)^2$	4	1	0	1	4

Donc : $V=\frac{1}{N}\left[(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_5-\overline{x})^2\right]$

$$V_1=\frac{1}{5}\left[(-2)^2+(-1)^2+0^2+(+1)^2+(+2)^2\right]$$

$$V_1=\frac{1}{5}[4+1+0+1+4]$$

$V_1=\frac{1}{5}\times10.$ D'où : $V_1=2.$

Et, par conséquent : $\sigma_1=\sqrt{V_1}$. Donc : $\boxed{\sigma_1=\sqrt{2}}$ et $\sigma_1\simeq1{,}4$.

Un calcul analogue montre que : $V_2 = \dfrac{148}{5} = 29{,}5$. $\boxed{\sigma_2=\sqrt{29{,}6}}$ donc $\sigma_2 \simeq 5{,}4$.

On remarque que $\sigma_1 < \sigma_2$. Comme pour l'étendue, on pourrait en déduire que l'élève 1 est plus « régulier » alors que l'élève 2 est « irrégulier ».

Remarque

« ***Plus l'écart-type est grand, plus les valeurs de la série sont dispersées*** ».

En général, (nous le verrons en classe de Terminale), dans une répartition (ou distribution) « normale », environ **68% des valeurs sont comprises entre les valeurs** $\bar{x}-\sigma$ **et** $\bar{x}+\sigma$.

2.4.b) Variance et écart-type d'une série statistique avec effectifs partiels

Définition 11.

On considère ***une série statistique à une variable*** quantitative, observée sur ***N*** individus d'une population E et prenant ***k*** valeurs $x_1, x_2, \cdots, x_k$ affectées des effectifs partiels $n_1, n_2, \cdots, n_k$ respectivement.

Pour calculer ***la variance de la série statistique***, notée ***V***, il suffit de multiplier par les effectifs partiels. Autrement dit :

$$\boxed{V=\frac{1}{N}\left[n_1(x_1-\bar{x})^2+n_2(x_2-\bar{x})^2+\cdots+n_k(x_k-\bar{x})^2\right]}$$

$$\boxed{V(X)=\frac{1}{N}\sum_{i=1}^{i=k} n_i(x_i-\bar{x})^2}$$

D'une manière analogue, l'écart-type est défini par : $\boxed{\sigma=\sqrt{V}}$ ou $\boxed{V=\sigma^2}$.

Voir exercices en classe.