

Statistiques descriptives

1. Vocabulaire statistique

Pour pratiquer les sciences statistiques, on doit collecter un grand nombre de données. Pour cela, on effectue une **enquête statistique** sur un ensemble de personnes ou d'objets.

■ DÉFINITION : Population

L'ensemble sur lequel cette enquête est effectuée est appelé la **population**.

■ DÉFINITION : Individu

Chaque élément de cette population est appelé un **individu**.

Il est rare de pouvoir interroger tous les individus d'une population, donc on en interroge un sous-ensemble, qui permet une estimation des résultats de l'enquête.

■ DÉFINITION : Échantillon

Un tel sous-ensemble est appelé un **échantillon**.

On s'intéresse à l'ensemble des résultats d'une enquête statistique pour en tirer des informations.

■ DÉFINITION : Caractère

La question posée lors d'une enquête statistique porte sur une particularité de l'ensemble étudié, appelée le **caractère**.

■ DÉFINITION : Valeurs du caractère

Les différentes réponses possibles sont les **valeurs du caractère**. On les note x_i .



Attention

Ces « valeurs » ne sont pas forcément numériques !

On peut regrouper le vocabulaire dans un tableau :

Vocabulaire ensembliste	Vocabulaire statistique	Vocabulaire probabiliste
Ensemble	Population	Ensemble des issues
Sous-ensemble	Échantillon	Événement
Élément	Individu	Issue
Propriété	Caractère	×

Exemple :

On interroge tous les élèves de premier du lycée franco-allemand de Buc.

- Population : l'ensemble des élèves de première.
- Individu : un élève de première.
- Échantillon : la classe de première ES.
- Caractère : le nombre de frères et sœurs.
- Valeurs du caractère : 0, 1, 2, 3, ...

Il existe différents types de caractères :

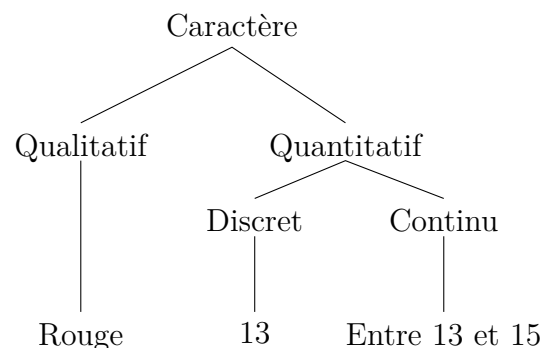
■ DÉFINITION : Caractères quantitatifs

Le caractère est dit **quantitatif** si on peut le mesurer, en associant un nombre à chaque valeur du caractère de la population. Il y a alors deux types de caractères quantitatifs possibles :

- On dit qu'il est quantitatif **discret** s'il peut prendre un nombre fini de valeurs (nombre de frères et sœurs...)
- On dit qu'il est quantitatif **continu** s'il peut prendre toutes les valeurs d'un intervalle de façon continue (taille, poids...)

■ DÉFINITION : Caractères qualitatifs

Le caractère est dit **qualitatif** si on ne peut pas le mesurer (couleur des yeux...).



L'intérêt des statistiques descriptives est de tirer des informations de l'ensemble des valeurs collectées.

■ DÉFINITION : Série statistique

On appelle **série statistique** l'ensemble des valeurs collectées.

■ DÉFINITION : Effectif

Le nombre de fois où apparaît une réponse est appelé l'**effectif** de la valeur du caractère. Les effectifs sont notés n_i .

■ DÉFINITION : Fréquence

La **fréquence** d'une valeur du caractère est le quotient de l'effectif de la valeur par l'effectif total. Elle peut s'exprimer en pourcentage comme l'exprime la propriété suivante.

■ PROPRIÉTÉ : (admise)

On a

$$f_i = \frac{\text{effectif de la valeur}}{\text{effectif total}}$$

avec $0 \leq f_i \leq 1$

En pourcentages,

$$p_i = \frac{\text{effectif de la valeur}}{\text{effectif total}} \times 100$$

avec $0 \leq p_i \leq 100$

Remarque : La somme des fréquences de toutes les valeurs de la série statistique est 1.

2. Caractères quantitatifs

A. Caractères quantitatifs discrets

C'est dans le cas des caractères quantitatifs discrets que l'on peut obtenir le plus d'informations. Nous allons définir les caractéristiques que vous devez savoir donner.

- Population : Élèves du Lycée Franco-Allemand
- Échantillon : Classe de première SMP
- Caractère : Nombre de frères et sœurs

Par habitude, on écrit la série statistique sous forme d'un tableau :

Nombre de frères et sœurs, noté x_i	0	1	2	3	4	5	Total
Nombre d'élèves, noté n_i	2	11	11	3	1	2	30

a. Étendue

■ DÉFINITION : Étendue

Soit $(x_i ; n_i)_{1 \leq i \leq p}$ une série statistique. L'**étendue** de la série est l'écart entre la plus grande la plus petite valeur du caractère, c'est $x_{\max} - x_{\min}$.

Exemple :

Dans l'exemple précédent, l'étendue est $5 - 0 = 5$.

b. Mode

■ DÉFINITION : Mode

On appelle **mode** d'une série statistique une valeur du caractère qui a le plus grand effectif.

Exemple :

Dans l'exemple précédent, les modes sont 1 et 2.

c. Moyenne

■ DÉFINITION :

La **moyenne** d'une série statistique est égale au quotient de la somme de ses valeurs par l'effectif total.

Exemple :

Mélanie et Julie lancent deux dés et notent le total des points obtenus : 10 ; 12 ; 5 ; 2 ; 3 ; 6 ; 7 ; 11

Calculer la moyenne des points obtenus. $\frac{10 + 12 + 5 + 2 + 3 + 6 + 7 + 11}{8} = \frac{56}{8} = 7$

■ PROPRIÉTÉ :

La moyenne est la note qui, substituée à chaque valeur du caractère de la série, donne la même somme.

■ DÉFINITION :

Soit $(x_i ; n_i)_{1 \leq i \leq p}$ une série statistique. La **moyenne pondérée** d'une série statistique est égale à la somme des produits de chaque valeur du caractère par son effectif, le tout divisé par l'effectif total. Ainsi

$$\bar{x} = \frac{\sum_{i=1}^p x_i n_i}{\sum_{i=1}^p n_i}$$

Exemple :

Dans l'exemple des frères et sœurs, calculer le nombre moyen de frères et sœurs des élèves de la classe de première SMP interrogée. On divise 56 par 30 pour obtenir 1,87.

Remarque : Une moyenne n'est pas nécessairement égale à l'une des valeurs de la série. Elle prend en compte les valeurs extrêmes d'une série.

■ PROPRIÉTÉ :

Une moyenne est toujours comprise entre la plus petite et la plus grande valeur du caractère.

Dans la suite, on considère les notes obtenues en mathématiques par Manon, Julien et Séverine au dernier trimestre. Les DM sont coefficientés 1, les interrogations écrites coefficientées 3 et les contrôles coefficientés 5. Cela signifie qu'une note du contrôle compte cinq fois plus qu'une note de DM, ou qu'une note de contrôle équivaut à cinq fois la même note en DM ! Pour comprendre l'importance des coefficients, observons l'exemple suivant :

Exemple :

Manon a obtenu

9 au contrôle, 20 à l'interrogation écrite, 7 et 8 aux DM.

Julien a obtenu

12 au contrôle, 11 à l'interrogation écrite et 13 et 14 aux DM.

et Séverine a obtenu

15 au contrôle, 4 à l'interrogation écrite, 16 et 17 aux DM.

Compléter les tableaux de la feuille suivante avec les valeurs de chacune des séries.

Pour chaque élève, répondre aux questions suivantes :

- Pour Manon :
 - Étendue : $20 - 7 = 13$
 - Moyenne : $\frac{120}{10} = 12$
- Pour Julien :
 - Étendue : $14 - 11 = 3$
 - Moyenne : $\frac{120}{10} = 12$
- Pour Séverine :
 - Étendue : $17 - 4 = 13$
 - Moyenne : $\frac{120}{10} = 12$

Quelle appréciation écrire sur le bulletin de chacun ?

Vous constatez que les élèves n'ont pas le même profil. Leur professeur ne leur mettra donc pas la même appréciation. Comment se rendre compte de cela grâce aux statistiques ? La *médiane* est une réponse...

d. Médiane

■ DÉFINITION :

Une **médiane** est un nombre qui sépare une série **ordonnée** en deux séries de même effectif. On peut donc dire que

- la moitié des valeurs de la série sont inférieures ou égales à la médiane.
- la moitié des valeurs de la série sont supérieures ou égales à la médiane.

■ PROPRIÉTÉ :

- Si l'effectif total de la série est impair, une médiane est la valeur de la série qui, une fois ordonnée, est au milieu de cette série.
Soit N l'effectif total de la série. Une médiane est la valeur du terme de rang $\frac{N+1}{2}$ de la série.
- Si l'effectif total de la série est pair, une médiane est un nombre qui se trouve entre les deux valeurs du milieu de la série. C'est, par exemple, la moyenne de ces deux valeurs de la série, c'est-à-dire la moyenne des valeurs $x_{\frac{N}{2}}$ et $x_{\frac{N}{2}+1}$.

Exemple :

Un élève a obtenu les résultats suivants pendant le trimestre : 11 – 16 – 18 – 8 – 12

- Quelle est la moyenne de la série ? $\frac{11 + 16 + 18 + 8 + 12}{5} = \frac{65}{5} = 13$.
- Quelle est une médiane de la série ? Comment interpréter ce nombre ? On classe d'abord la série dans l'ordre croissant, et on regarde la valeur centrale. 8 – 11 – **12** – 16 – 18. Une médiane est donc 12. Il y a donc 50 % des notes obtenues par l'élève qui sont inférieures à 12 et 50 % des notes obtenues par l'élève qui sont supérieures à 12.

Exemple :

Un élève a obtenu les résultats suivants pendant le trimestre : 11 – 16 – 18 – 8 – 12 – 13

- Quelle est la moyenne de la série ? $\frac{11 + 16 + 18 + 8 + 12 + 13}{6} = \frac{78}{6} = 13$.
- Quelle est une médiane de la série ? Comment interpréter ce nombre ? On classe d'abord la série dans l'ordre croissant, et on regarde la moyenne des valeurs centrales. 8 – 11 – **12** – **13** – 16 – 18. Une médiane est donc $\frac{12 + 13}{2} = 12,5$. Il y a donc 50 % des notes obtenues par l'élève qui sont inférieures à 12,5 et 50 % des notes obtenues par l'élève qui sont supérieures à 12,5.

Exemple :

Voilà les séries des notes de Séverine, Julien et Manon. Ordonner chacune des séries de notes, et déterminer une médiane de chaque série.

- Manon : 7 - 8 - 9 - 9 - 9 - 9 - 9 - 20 - 20 - 20 ; une médiane est $\frac{9 + 9}{2} = 9$.
- Julien : 11 - 11 - 11 - 12 - 12 - 12 - 12 - 12 - 13 - 14 ; une médiane est $\frac{12 + 12}{2} = 12$
- Séverine : 4 - 4 - 4 - 15 - 15 - 15 - 15 - 15 - 16 - 17 ; une médiane est $\frac{15 + 15}{2} = 15$.

Exemple :

Dans l'exemple des frères et sœurs, quelle est une médiane du nombre de frères et sœurs dans la classe interrogée ? Comment interpréter ce résultat ? Une médiane de la série est 2. Il y a donc 50 % des élèves qui ont moins de deux frères et sœurs et 50 % des élèves qui ont plus de deux frères et sœurs.

Remarque : Pour déterminer une médiane d'une série, il faut l'ordonner. Une médiane est un indicateur de position qui ne dépend pas des valeurs extrêmes.

e. Quartiles

■ DÉFINITION :

Soit $x_1 \leq x_2 \leq \dots \leq x_N$ une série statistique d'effectif total N dont les valeurs sont classées par ordre croissant.

- Le **premier quartile** Q_1 de la série est la plus petite valeur de la série telle qu'au moins 25 % des valeurs de la série lui soient inférieures ou égales.
- Le **troisième quartile** Q_3 de la série est la plus petite valeur de la série telle qu'au moins 75 % des valeurs de la série lui soient inférieures ou égales.

Remarque : Le deuxième quartile Q_2 correspond à la médiane.

Remarque : Pour tout calcul de médiane ou quartile, il convient d'abord d'ordonner les valeurs de la série par ordre croissant.

MÉTHODE :

En pratique, pour déterminer Q_1 , on calcule $\frac{1}{4}N$ puis on considère la première valeur d'indice supérieur ou égal à $\frac{1}{4}N$: cette valeur est le premier quartile. De même pour le troisième quartile avec $\frac{3}{4}N$

DÉFINITION :

On appelle **intervalle interquartile** l'intervalle $[Q_1 ; Q_3]$. Il regroupe environ 50% des valeurs de la série autour de la médiane. L'**écart interquartile** est $Q_3 - Q_1$, c'est-à-dire l'amplitude de l'intervalle précédent.

Exemple :

Voici la série des salaires mensuels dans une certaine entreprise :

Salaire x_i (en euros)	1500	2000	3000	10 000
Effectif n_i	500	296	201	3

Calculer la moyenne, la médiane, les quartiles et l'écart interquartile de cette série. Interpréter.

$$\bar{x} = \frac{1\,500 \times 500 + 2\,000 \times 296 + 3\,000 \times 201 + 10\,000 \times 3}{500 + 296 + 201 + 3} = \frac{1\,975\,000}{1\,000} = 1\,975 \text{ euros.}$$

$$\frac{N}{4} = 250; \quad \frac{N}{2} = 500; \quad \frac{N}{2} + 1 = 501 \text{ et } \frac{3N}{4} = 750.$$

Une fois les valeurs rangées par ordre croissant, la valeur du caractère d'indice 250 est 1 500, celle d'indice 500 est 1 500, celle d'indice 501 est 2 000, celle d'indice 750 est 2 000.

Donc le premier quartile est 1 500 euros, la médiane est $\frac{1500 + 2000}{2} = \frac{3500}{2} = 1\,750$ euros et le troisième quartile est 2 000 euros.

Ainsi, 25 % des salariés gagnent au plus 1500 euros, 50 % des salariés gagnent au plus 1750 euros et 75 % des salariés gagnent au plus 2 000 euros.

L'intervalle interquartile est $[1500 ; 2000]$. Ainsi 50 % des salaires sont compris entre 1500 et 2000 euros.

f. Boîte à moustaches

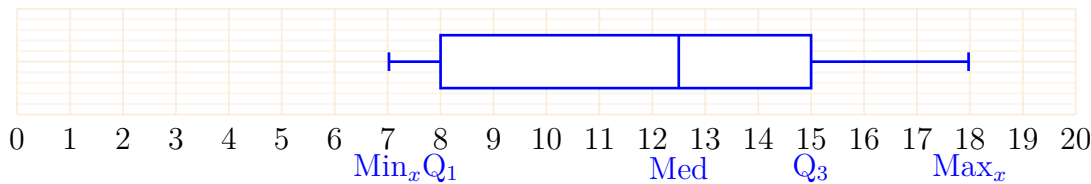
On appelle **boîte à moustaches** un diagramme qui permet de résumer toutes les informations obtenues jusqu'à maintenant :

Méthodologie :



Exemple :

On considère une série statistique correspond à des notes d'élèves (comprises entre 0 et 20). Si $Q_1 = 8$, $med = 12,5$ et $Q_3 = 15$, pour une série dont le minimum est 7 et le maximum est 18, alors



L'écart interquartile est $Q_3 - Q_1 = 15 - 8 = 7$.

C'est la longueur de la boîte.

L'étendue est donc la longueur de la boîte avec ses moustaches.

f. Écart-type

On observe les résultats en mathématiques de deux élèves, Alice et Bastien, au cours d'un trimestre.

- Alice : 11 / 9 / 13
- Bastien : 12 / 3 / 18
- On constate que $\overline{x_A} = \overline{x_B} = 11$.

Notes	11 / 9 / 13	12 / 3 / 18
Moyenne	11	11
Écarts par rapport à la moyenne	0 / -2 / +2	+1 / -8 / +7
Moyenne des écarts par rapport à la moyenne	0	0
Valeurs absolues des écarts par rapport à la moyenne	0 / 2 / 1	1 / 8 / 7
Moyenne des valeurs absolues des écarts par rapport à la moyenne	$\frac{4}{3} \approx 1,33$	$\frac{16}{3} \approx 5,33$
Carrés des écarts par rapports à la moyenne	0 / 4 / 4	1 / 64 / 49
Moyenne des carrés des écarts par rapport à la moyenne	$\frac{8}{3} \approx 2,67$	38
Racine carrée de la moyenne des carrés des écarts par rapport à la moyenne	$\sqrt{\frac{8}{3}} \approx 1,63$	$\sqrt{38} \approx 6,16$

Soit $(x_i ; n_i)_{1 \leq i \leq p}$ une série statistique.

■ DÉFINITION : Variance

On appelle **variance** le nombre $V(x) = \frac{\sum_{i=1}^p n_i (x_i - \overline{x})^2}{\sum_{i=1}^p n_i}$.

■ DÉFINITION : Écart-type

On appelle **écart-type** le nombre $\sigma(x) = \sqrt{V(x)} = \sqrt{\frac{\sum_{i=1}^p n_i (x_i - \overline{x})^2}{\sum_{i=1}^p n_i}}$.

Exemple :

On considère une série statistique des notes de 28 élèves d'une classe, que l'on regroupe dans le tableau suivant :

x_i	2	6	10	14	18	Σ
n_i	2	7	9	9	1	28

La moyenne de la série statistique ci-contre :

$$\bar{x} = \frac{\sum_{i=1}^N n_i x_i}{\sum_{i=1}^N n_i} = \frac{280}{28} = 10$$

La variance de cette série statistique est :

$$V(x) = \frac{448}{28} = 16$$

L'écart-type de cette série statistique est :

$$\sigma(x) = \sqrt{\frac{448}{28}} = 4$$

■ PROPRIÉTÉ : Lemme de König¹-Huygens²

L'écart-type est la racine carrée de la différence entre moyenne des carrés et le carré de la moyenne, c'est-à-dire

$$\sigma(x) = \sqrt{\overline{x^2} - \bar{x}^2}$$

Démonstration.

$$\begin{aligned}\sigma(x)^2 &= \frac{\sum_{i=1}^p n_i (x_i - \bar{x})^2}{\sum_{i=1}^p n_i} \\&= \frac{\sum_{i=1}^p n_i (x_i^2 - 2x_i \bar{x} + \bar{x}^2)}{\sum_{i=1}^p n_i} \\&= \frac{\sum_{i=1}^p (n_i x_i^2 - 2n_i x_i \bar{x} + n_i \bar{x}^2)}{\sum_{i=1}^p n_i} \\&= \frac{\sum_{i=1}^p n_i x_i^2 - 2\bar{x} \sum_{i=1}^p n_i x_i + \bar{x}^2 \sum_{i=1}^p n_i}{\sum_{i=1}^p n_i} \\&= \frac{\sum_{i=1}^p n_i x_i^2}{\sum_{i=1}^p n_i} - 2\bar{x} \frac{\sum_{i=1}^p n_i x_i}{\sum_{i=1}^p n_i} + \bar{x}^2 \frac{\sum_{i=1}^p n_i}{\sum_{i=1}^p n_i} \\&= \overline{x^2} - 2\bar{x} \times \bar{x} + \bar{x}^2 \times 1 \\&= \overline{x^2} - 2\bar{x}^2 + \bar{x}^2 \\&= \overline{x^2} - \bar{x}^2\end{aligned}$$

Donc $\sigma(x) = \sqrt{\overline{x^2} - \bar{x}^2}$ car $\sigma(x) > 0$. □

1. Johann Samuel König (ou Koenig), mathématicien, philosophe et juriste allemand (1712-1757). Il fut élève de Jean Bernoulli et de Leibniz, ami de Voltaire.

2. Christiaan Huygens (1er avril 1629 - 8 juillet 1695) est un grand mathématicien et physicien hollandais du XVII^{ème} siècle.

Exemple :

Dans l'exemple précédent, on rappelle que $\bar{x} = 10$. On calcule $\overline{x^2}$, qui vaut $\frac{\sum_{i=1}^p n_i x_i^2}{\sum_{i=1}^p n_i} = \frac{3248}{28}$

$$\text{Alors } \sigma(x) = \sqrt{\overline{x^2} - \bar{x}^2} = \sqrt{\frac{3248}{28} - \left(\frac{280}{28}\right)^2} = 4$$

On peut vérifier grâce à la calculatrice :

rad STATISTICS			rad STATISTICS		
Data	Graph	Stats	Data	Graph	Stats
Number of data points	n	28	Mean	μ	10
Minimum	Min	2	Standard deviation	σ	4
First quartile	Q1	6	Variance	σ^2	16
Median	Med	10	Sample mean	$\bar{x}$	10
Third quartile	Q3	14	Sample std deviation	s	4.073401
Maximum	Max	18	Sample variance	s ²	16.59259
Range	R	16	Sum of values	$\sum x$	280
Interquartile range	IQR	8	Sum of squared values	$\sum x^2$	3248
			Mode 1	Mod1	10



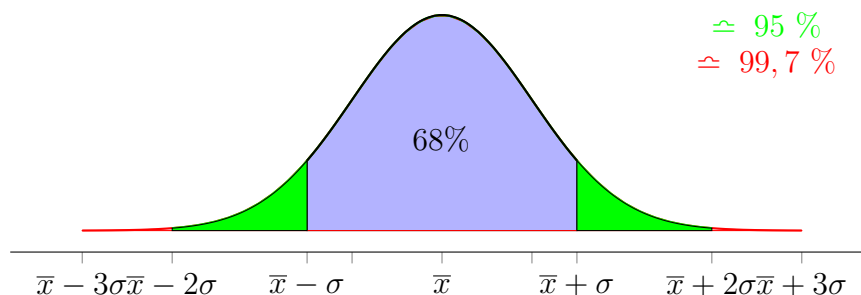
g. Répartition normale

Soit (x_i, n_i) une série statistique. Soient $\bar{x}$ sa moyenne et σ son écart-type.

■ DÉFINITION :

On dit que la répartition est **normale** si au moins l'une des trois conditions suivantes est vérifiée :

- si les valeurs du caractère comprises entre $\bar{x} - \sigma$ et $\bar{x} + \sigma$ représentent au moins 68 % de l'effectif total.
- si les valeurs du caractère comprises entre $\bar{x} - 2\sigma$ et $\bar{x} + 2\sigma$ représentent au moins 95 % de l'effectif total.
- si les valeurs du caractère comprises entre $\bar{x} - 3\sigma$ et $\bar{x} + 3\sigma$ représentent au moins 99 % de l'effectif total.



■ DÉFINITION :

Cette courbe est appelée **courbe en cloche** ou encore **courbe de Gauss**³.

Exemple :

Si $\bar{x} = 10$ et $\sigma = 4$, alors :

- $\bar{x} - \sigma = 6$ et $\bar{x} + \sigma = 14$. On a donc $7 + 9 + 9 = 25$ élèves, soit $89,3\%$.
- $\bar{x} - 2\sigma = 2$ et $\bar{x} + 2\sigma = 18$. On a donc 28 élèves, soit 100% .

3. Johann Carl Friedrich Gauß, né le 30 avril 1777 à Brunswick et mort le 23 février 1855 à Göttingen, est un mathématicien, astronome et physicien allemand. Surnommé « le prince des mathématiciens », il est considéré comme l'un des plus grands mathématiciens de tous les temps.

h. Changement de variable affine

■ PROPRIÉTÉ :

Soit $(x_i ; n_i)_{1 \leq i \leq p}$ une série statistique d'effectif total N de variance $V(x)$ et d'écart-type $\sigma(x)$.

Soit $(y_i ; m_i)_{1 \leq i \leq p}$ la série statistique définie pour tout $i \in \llbracket 1 ; p \rrbracket$, $y_i = ax_i + b$ où $a \in \mathbb{R}^*$ et $b \in \mathbb{R}$. On note $V(y)$ la variance de cette série et $\sigma(y)$ son écart-type. Alors

$$\bar{y} = a\bar{x} + b ; \quad V(y) = a^2 V(x) \quad \text{et} \quad \sigma(y) = |a| \sigma(x)$$

Démonstration. On pose $S = \sum_{i=1}^p n_i$

$$\begin{aligned} \bar{y} &= \frac{1}{S} \sum_{i=1}^p n_i y_i \\ &= \frac{1}{S} \sum_{i=1}^p n_i (ax_i + b) \\ &= \frac{1}{S} \sum_{i=1}^p an_i x_i + bn_i \\ &= \frac{1}{S} \left(a \sum_{i=1}^p n_i x_i + b \sum_{i=1}^p n_i \right) \\ &= a \times \frac{1}{S} \sum_{i=1}^p n_i x_i + b \times \underbrace{\frac{1}{S} \sum_{i=1}^p n_i}_{=1} \\ &= a\bar{x} + b \end{aligned}$$

De plus,

$$\begin{aligned} V(y) &= \overline{y^2} - (\bar{y})^2 \\ &= \left(\frac{1}{S} \sum_{i=1}^p n_i y_i^2 \right) - (a\bar{x} + b)^2 \\ &= \left(\frac{1}{S} \sum_{i=1}^p n_i (ax_i + b)^2 \right) - a^2 \bar{x}^2 - 2ab\bar{x} - b^2 \\ &= \frac{1}{S} \sum_{i=1}^p (a^2 n_i x_i^2 + 2ab n_i x_i + b^2 n_i) - a^2 \bar{x}^2 - 2ab\bar{x} - b^2 \\ &= \frac{1}{S} \left(\sum_{i=1}^p (a^2 n_i x_i^2) + 2ab \sum_{i=1}^p (n_i x_i) + b^2 \underbrace{\sum_{i=1}^p n_i}_{=S} \right) - a^2 \bar{x}^2 - 2ab\bar{x} - b^2 \\ &= \frac{1}{S} \sum_{i=1}^p (a^2 n_i x_i^2) + 2ab \times \frac{1}{S} \sum_{i=1}^p (n_i x_i) + b^2 - a^2 \bar{x}^2 - 2ab\bar{x} - b^2 \\ &= a^2 \times \frac{1}{S} \sum_{i=1}^p (n_i x_i^2) + 2ab \times \frac{1}{S} \sum_{i=1}^p (n_i x_i) - a^2 \bar{x}^2 - 2ab\bar{x} \\ &= a^2 \overline{x^2} + 2ab\bar{x} - a^2 \bar{x}^2 - 2ab\bar{x} \\ &= a^2 \overline{x^2} - a^2 \bar{x}^2 \\ &= a^2 V(x) \end{aligned}$$

Enfin

$$\begin{aligned}\sigma(y) &= \sqrt{V(y)} \\ &= \sqrt{a^2 V(x)} \\ &= |a| \sqrt{V(x)} \\ &= |a| \sigma(x)\end{aligned}$$

Autre preuve de la formule de la variance, en se ramenant à la définition

$$\begin{aligned}V(y) &= \frac{1}{S} \sum_{i=1}^p n_i (y_i - \bar{y})^2 \\ &= \frac{1}{S} \sum_{i=1}^p n_i (y_i^2 - 2y_i \bar{y} + \bar{y}^2) \\ &= \frac{1}{S} \sum_{i=1}^p (n_i y_i^2 - 2n_i y_i \bar{y} + n_i \bar{y}^2) \\ &= \frac{1}{S} \left(\sum_{i=1}^p n_i y_i^2 - \sum_{i=1}^p 2\bar{y} n_i y_i + \sum_{i=1}^p \bar{y}^2 n_i \right) \\ &= \frac{1}{S} \left(\sum_{i=1}^p n_i y_i^2 - 2\bar{y} \sum_{i=1}^p n_i y_i + \bar{y}^2 \underbrace{\sum_{i=1}^p n_i}_{=S} \right) \\ &= \frac{1}{S} \left(\sum_{i=1}^p n_i (ax_i + b)^2 - 2(a\bar{x} + b) \sum_{i=1}^p (n_i (ax_i + b)) + (a\bar{x} + b)^2 S \right) \\ &= \frac{1}{S} \left(\sum_{i=1}^p (n_i a^2 x_i^2 + 2abn_i x_i + b^2 n_i) - 2a\bar{x} \sum_{i=1}^p (an_i x_i + bn_i) - 2b \sum_{i=1}^p (an_i x_i + bn_i) + (a^2 \bar{x}^2 + 2ab\bar{x} + b^2) S \right) \\ &= \frac{1}{S} \left(a^2 \sum_{i=1}^p (n_i x_i^2) + 2ab \sum_{i=1}^p n_i x_i + b^2 \sum_{i=1}^p n_i - 2a^2 \bar{x} \sum_{i=1}^p n_i x_i - 2ab\bar{x} \sum_{i=1}^p n_i - 2ab \sum_{i=1}^p (n_i x_i) - 2b^2 \sum_{i=1}^p n_i \right. \\ &\quad \left. + a^2 \bar{x}^2 S + 2ab\bar{x} S + b^2 S \right) \\ &= a^2 \bar{x}^2 + 2ab\bar{x} + b^2 - 2a^2 \bar{x} \times \bar{x} - 2ab\bar{x} - 2ab\bar{x} - 2b^2 + a^2 \bar{x}^2 + 2ab\bar{x} + b^2 \\ &= a^2 \bar{x}^2 - 2a^2 \bar{x}^2 + a^2 \bar{x}^2 \\ &= a^2 \bar{x}^2 - a^2 \bar{x}^2 \\ &= a^2 (\bar{x}^2 - \bar{x}^2) \\ &= a^2 V(x)\end{aligned}$$

□

B. Caractères quantitatifs continus

Pour les caractères quantitatifs continus, on sépare les valeurs du caractère en intervalles de même taille.

■ DÉFINITION :

Les intervalles dans lesquels on regroupe les valeurs sont appelés les *classes*.
La taille de chaque classe est appelée l'**amplitude de la classe**.

On considère l'enquête suivante :

- Population : Élèves du Lycée Franco-Allemand
- Échantillon : Classe de première SMP
- Caractère : Taille, en cm

On a obtenu les résultats suivants :

156 - 159 - 168 - 161 - 156 - 176 - 174 - 162 - 158 - 161 - 174 - 161 - 151 - 166 - 167 - 176 - 153 - 173 - 166 - 175 -
166 - 161 - 175 - 160 - 174 - 166 - 161 - 175 - 172 - 160

Regrouper cette série statistique par classes d'amplitude 5 cm et calculer les fréquences en % arrondies à l'unité :

Tailles, x_i	$150 \leq t < 155$	$155 \leq t < 160$	$160 \leq t < 165$	$165 \leq t < 170$	$170 \leq t < 175$	$175 \leq t < 180$	Total
Effectifs, n_i	??	??	??	??	??	??	??
Fréquences, f_i	??	??	??	??	??	??	??
Fréquences, en %	??	??	??	??	??	??	??

Histogramme des effectifs

On peut associer aux classes un **histogramme**, dans lequel la légende est **indispensable**.

Si les classes ont toutes la même amplitude, alors la hauteur des rectangles est proportionnelle à l'effectif de chaque classe, comme dans la figure de gauche ci-dessous. Sinon, dans cet exemple, en regroupant les deux premières classes, on obtient la représentation de droite :

?? PythonTeX ??