STATISTIQUES A UNE VARIABLE

1) VOCABULAIRE

A) GÉNÉRALITÉS

Définition :

L'ensemble sur lequel on travaille en statistique est appelé **population.**

Si cet ensemble est trop vaste, on en restreint l'étude à une partie appelée **échantillon.**

Un élément de cet ensemble est appelé **individu**.

La particularité commune que l'on étudie est appelée **caractère ou variable.**

Les valeurs prises par le caractère sont aussi appelées les **modalités**.

B) CARACTÈRE QUALITATIF ET CARACTÈRE QUANTITATIF

Définition :

Si la particularité étudiée ne s'exprime pas par un nombre, il s'agit d'un **caractère qualitatif.**

Exemple : Dans une population, être marié(e) est un caractère qualitatif à deux valeurs : oui ou non.

Définition :

Si cette particularité s'exprime par un nombre *(et que l'on peut ordonner ces nombres)* , il s'agit d'un **caractère quantitatif.**

- Si les valeurs du nombre exprimé sont isolées, il s'agit d'un **caractère discret.**
- Par contre, si ces valeurs sont prises dans tout un intervalle de $\mathbb{R}$, il s'agit d'un **caractère continu.**

Dans ce cas, le nombre désignant la modalité se note en général x_i.

Exemple : Caractère discret

Le nombre de frères et sœurs d'un élève est un caractère quantitatif discret car il ne peut prendre que les valeurs 0, 1, 2, 3, 4 …

Exemple : Caractère continu

Le temps de révision pour un contrôle pourrait être n'importe quel nombre t, tel que $2 \leqslant t \leqslant 3$ par exemple.
Les valeurs de ce caractère sont regroupées en **classes** ($[0\,;1[$; $[1\,;2[$ ...)

Remarques :

- L'amplitude des classes n'est pas forcément la même.
- En général, on fait l'hypothèse d'une répartition uniforme à l'intérieur de chaque classe …

C) EFFECTIFS ET FRÉQUENCES

Définition :

Le nombre d'individus, noté n_i, d'une modalité (ou valeur) est appelé **effectif.**

Le nombre total d'individus, noté N, de la population est appelé **effectif total.**

$$N = \sum_{i=1}^{k} n_i = n_1 + n_2 + \ldots + n_k \text{ s'il y a } k \text{ modalités.}$$

Le rapport $f_i = \frac{n_i}{N}$ est appelé **fréquence.**

Remarques :

- f_i est un nombre toujours compris entre 0 et 1. Souvent, les nombres f_i s'expriment par un pourcentage.
- La somme des nombres f_i est toujours égale à 1. $\sum_{i=1}^{k} f_i = 1$

Définition :

Une **série statistique** est l'ensemble des résultats d'une étude : valeurs du caractère et effectifs correspondants.

On représente souvent une série statistique sous forme d'un tableau.

Dans le cas d'une variable quantitative, on peut ordonner les différentes valeurs de la plus petite à la plus grande (ou de la plus grande à la plus petite) puis additionner les effectifs successifs : on obtient ainsi **les effectifs cumulés croissants** (ou décroissants).

On obtient de la même façon **les fréquences cumulées croissantes** (ou décroissantes).

2) REPRÉSENTATION GRAPHIQUE

Pour le cas général, on considère une série statistique (x_i) quantitative discrète :

Valeurs (x_i)	x_1	x_2	…	x_k	Total
Effectifs (n_i)	n_1	n_2	…	n_k	N
Fréquences (f_i)	f_1	f_2	...	f_k	1

On représente généralement une série quantitative discrète par **un diagramme en bâtons** ou en barres.

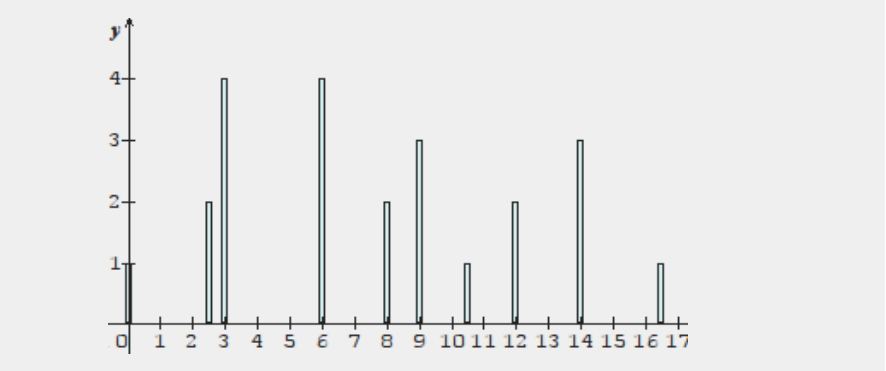
On peut aussi parfois utiliser un diagramme circulaire ou semi-circulaire, même si ces derniers sont plutôt réservés aux séries qualitatives.

Exemple: Voici les notes obtenues à un devoir dans une classe de seconde de 23 élèves.

0 – 12 – 9 – 10,5 – 2,5 – 8 – 3 – 8 – 3 – 14 – 6 – 2,5 – 6 – 16,5 – 14 – 6 – 9 – 3 – 6 – 14 – 12 – 3 – 9

On se propose de ranger ces valeurs dans un tableau:

Valeurs (x_i)	0	2,5	3	6	8	9	10,5	12	14	16,5
Effectifs (n_i)	1	2	4	4	2	3	1	2	3	1
Fréquences (f_i)	0,043	0,087	0,174	0,174	0,087	0,131	0,043	0,087	0,131	0,043
Effectifs cumulés croissants	1	3	7	11	13	16	17	19	22	23
Fréquences cumulées croissantes	0,043	0,130	0,304	0,478	0,565	0,696	0,739	0,826	0,957	1,000



3) LES PARAMÈTRES DE TENDANCE CENTRALE

Définition :

On appelle **mode** d'une série statistique une valeur du caractère dont l'effectif associé est le plus grand.

Exemple : La série de notes de la seconde admet deux modes : 3 et 6.

Définition :

La moyenne de la série (x_i) est le nombre réel, noté $\overline{x}$, tel que :

$$\overline{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{N} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} x_i f_i$$

Exemple : La moyenne des notes du devoir est :

- à partir de la distribution des effectifs : $\overline{x} = \frac{1\times 0 + 2\times 2{,}5 + 4\times 3 + \ldots + 1\times 16{,}5}{23} \approx 7{,}7$
- à partir de la distribution des fréquences : $\overline{x} = \sum_{i=1}^{10} x_i f_i \approx 7{,}7$

Remarque :

Si dans une série de notes, une note apparaît de manière exceptionnelle (0 par exemple) , on peut calculer la moyenne de la série privée de cette valeur . On dit qu'il s'agit d'une **moyenne élaguée**.

Propriété : Linéarité de la moyenne

Soit a et b deux nombres réels.

Si la série (x_i) a pour moyenne $\bar{x}$, alors la série (ax_i+b) a pour moyenne $\bar{x}\,'=a\bar{x}+b$

Preuve :

$$\bar{x}\,'=\frac{(a\,x_1+b)n_1+(a\,x_2+b)n_2+\ldots+(a\,x_k+b)n_k}{N}=\frac{a\,(x_1n_1+x_2n_2+\ldots+x_kn_k)+b(n_1+n_2+\ldots+n_k)}{N}$$

Ce qui donne $\bar{x}\,' = a\frac{(x_1n_1+x_2n_2+\ldots+x_kn_k)}{N}+b\ \frac{n_1+n_2+\ldots+n_k}{N}=a\,\bar{x}+b$

Définition :

La médiane est une valeur Me du caractère qui partage la population en deux sous-ensembles de même effectif. Les éléments du premier sous-ensemble correspondent à des valeurs du caractère inférieures ou égales à Me, ceux du second correspondent à des valeurs du caractère supérieures ou égales à Me.

Dans la pratique :

- Si l'effectif total N est impair, la médiane est la valeur du caractère située au rang $\frac{N+1}{2}$
- Si l'effectif total N est pair, la médiane est tout nombre situé entre la valeur du caractère occupant le rang $\frac{N}{2}$ et la valeur du caractère occupant le rang $\frac{N}{2}+1$ (On choisit souvent la demi-somme)

Exemple : Dans la série de notes de la classe de seconde, on a 23 valeurs . $\frac{23+1}{2}=12$

La médiane de cette série est donc la $12^{ème}$ valeur , c'est à dire Me = 8.

4) LES PARAMÈTRES DE DISPERSION

A) ÉTENDUE ET QUARTILES

Définition :

On appelle **étendue** , notée e d'une série statistique la différence entre la plus grande valeur, notée Max du caractère et la plus petite, notée Min .

$$e = Max - Min$$

Exemple : L'étendue de la série de notes de la seconde est $e = 16{,}5 - 0 = 16{,}5$.

Définition :

Le premier Quartile Q_1 d'une série statistique est la plus petite valeur de la série telle qu'au moins 25% des valeurs de celle-ci lui soient inférieures ou égales.
Le troisième Quartile Q_3 d'une série statistique est la plus petite valeur de la série telle qu'au moins 75% des valeurs de celle-ci lui soient inférieures ou égales.

Dans la pratique :

- Si $\frac{N}{4}$ est un entier, le premier quartile Q_1 est la valeur qui dans cette liste occupe le rang $\frac{N}{4}$ et le troisième quartile Q_3 est la valeur qui dans cette liste occupe le rang $\frac{3\,N}{4}$.
- Si $\frac{N}{4}$ n'est pas un entier, le premier quartile Q_1 est la valeur qui dans cette liste occupe le rang immédiatement supérieur à $\frac{N}{4}$ et le troisième quartile Q_3 est la valeur qui dans cette liste occupe le rang immédiatement supérieur à $\frac{3\,N}{4}$.

Exemple :

Pour la série des notes de seconde, il y a 23 valeurs.

$\frac{23}{4} = 5{,}75$, donc le premier quartile est la sixième valeur de la série, donc $Q_1 = 3$.

Au moins 25 % des élèves ont obtenu une note inférieure ou égale à 3

$\frac{3 \times 23}{4} = 17{,}25$, donc le troisième quartile est la dix-huitième valeur de la série, donc $Q_3 = 12$

Au moins 75 % des élèves ont obtenu une note inférieure ou égale à 12.

Remarques :

- Une série admet trois quartiles ; le deuxième, dont on ne fait pas usage au lycée, est associé à la valeur 50% .
- De nombreuses calculatrices considèrent les quartiles comme les médianes des deux séries obtenues après avoir partagé la série initiale par sa médiane … ce qui explique les différences constatées.
  Dans la pratique, ces différences ont peu d'importance vu la taille des séries.
- De la même façon, on peut définir les déciles d'une série statistique.

Définition :

L'intervalle interquartile d'une série statistique est l'intervalle [Q_1 ; Q_3] . Il contient au moins 50 % des valeurs.

L'écart interquartile d''une série statistique est le nombre $Q_3 - Q_1$

Exemple: L'écart interquartile de la série de notes de la classe de seconde est $12 - 3 = 9$

Remarques :

- L'écart interquartile mesure la dispersion des valeurs autour de la médiane ; plus l'écart est petit, plus les valeurs de la série appartenant à l'intervalle interquartile sont concentrées autour de la médiane.
- Contrairement à l'étendue qui mesure l'écart entre la plus grande et la plus petite valeur, l'écart interquartile élimine les valeurs extrêmes qui peuvent être douteuses, cependant il ne tient compte que de 50% de l'effectif …
- On peut correctement résumer une série statistique par le couple : **(médiane ; intervalle interquartile)**

B) ÉCART TYPE

Pour chaque valeur x_i de la série, son « éloignement » de la moyenne peut se mesurer par la distance $|x_i - \overline{x}|$. La moyenne pondérée de toutes ces distances fournit un très bon paramètre de dispersion. On l'appelle écart absolu moyen. Mais puisque la valeur absolue ne se prête pas trop aux calculs, il n'y a pas d'application dans les résultats obtenus en statistique.

Définitions :

La variance V est la moyenne des carrées des écarts des valeurs x_i à la moyenne $\overline{x}$, c'est à dire:

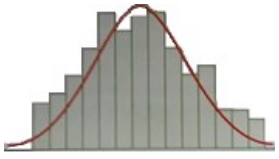
$$V = \frac{\sum_{i=1}^{p} n_i (x_i - \overline{x})^2}{N} = \sum_{i=1}^{p} f_i (x_i - \overline{x})^2$$

L'écart type σ est la racine carrée de la variance: $\sigma = \sqrt{V}$

Remarques :

- L'écart type est un paramètre plus fin que l'étendue, car il tient compte de la répartition des valeurs.
- L'écart type à la même unité que les valeurs de la série étudiée.
- L'écart type mesure la dispersion des valeurs de la série autour de la moyenne . Plus l'écart type est petit, plus les valeurs de la série sont concentrées autour de la moyenne.

  L'usage de l'écart type est particulièrement intéressant, lorsque le diagramme qui représente la série est assez symétrique et à la forme d'une courbe en cloche.
  On a alors environ 68 % des valeurs comprises dans l'intervalle $[\overline{x} - \sigma ; \overline{x} + \sigma]$



- On peut correctement résumer une série statistique par le couple : **(moyenne ; écart type)**