CHAP 10 : LES STATISTIQUES

I. VOCABULAIRE

1. **Population** : C'est l'ensemble des individus sur les quels porte une étude statistique.
2. **Échantillon** : C'est une partie de la population.
3. **Caractère** : Le caractère est l'information sur laquelle l'étude statistique est réalisée. Il peut être quantitatif s'il est mesurable, exemple *la taille, la masse*, etc., ou qualitatif dans le cas contraire, exemple *la couleur, la nationalité*, etc.
4. **Effectif** : C'est le nombre d'individus d'une population ou d'une partie de cette population.
5. **Modalité** : C'est l'une des différentes valeurs ou qualités de la variable d'une série statistique.
6. **Fréquence** : C'est le nombre de fois qu'une modalité est représentée par rapport à l'effectif total. Elle est donc toujours inférieure à 1 et la somme totale de toutes les fréquences donne 1.

Exemple illustratif du vocabulaire statistique

Situation

Un professeur interroge les élèves d'une classe de 4e (soit 25 élèves) pour connaître leur moyen de transport principal pour venir à l'école. Il note les réponses suivantes :

Bus, Bus, Voiture, Marche, Bus, Vélo, Voiture, Bus, Vélo, Bus, Voiture, Bus, Marche, Vélo, Bus, Bus, Bus, Voiture, Marche, Bus, Vélo, Voiture, Bus, Vélo, Bus

Analyse :

- **Population** : L'ensemble des 25 élèves de la classe.
- **Individu** : Chaque élève interrogé.
- **Échantillon** : Si on avait choisi uniquement 10 élèves sur les 25, cela aurait constitué un échantillon.
- **Caractère** : Le moyen de transport utilisé pour venir à l'école. C'est un caractère **qualitatif**.
- **Modalités** : Les différentes réponses possibles : *Bus, Voiture, Marche, Vélo.*
- **Effectifs** :
  - Bus : 12 élèves
  - Voiture : 5 élèves
  - Marche : 3 élèves
  - Vélo : 5 élèves
- **Fréquences** :
  - Bus : $\frac{12}{25} = 0{,}48$
  - Voiture : $\frac{5}{25} = 0{,}20$
  - Marche : $\frac{3}{25} = 0{,}12$
  - Vélo : $\frac{5}{25} = 0{,}20$

II. SÉRIE STATISTIQUE D'UNE VARIABLE

1. Définition

On appelle série statistique d'une variable x ou série statistique simple, la série obtenue si l'étude est réalisée sur un seul caractère x. Elle peut être *groupée* ou *non groupée* en classes. On la note (x_i, n_i). Avec $x = \{x_1, x_2, \ldots, x_p\}$ et n_i est l'effectif de la modalité x_i.

2. Effectif total, Moyenne, Variance, Écart-type, Fréquence partielle, Espéran

Effectif total : $\boxed{N = \sum_{i=1}^{p} n_i}$ **Moyenne** : $\boxed{\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_i x_i}$

Si la série est *groupée en classes*, alors : $\boxed{\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_i c_i}$ où $c_i = \frac{a+b}{2}$, le centre de la classe i de la forme $[a\ ;\ b[$.

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

Si la série est groupée en classes alors :

Variance : $\boxed{V(x) = \frac{1}{N}\sum_{i=1}^{p} n_i c_i^2 - \bar{x}^2}$ *La variance est toujours positive.*

Écart-type : $\boxed{\sigma(x) = \sqrt{V(x)}}$ **Fréquence partielle** : $\boxed{f_i = \frac{n_i}{N}}$

$$\boxed{\sum_{i=1}^{p} f_i = 1}$$

Exemple : On représente au tableau N°1 les tailles en cm de 10 jeunes garçons et au tableau N°2, les notes en maths de 11 élèves d'une classe de TS2.

Tableau N°1

Tailles x_i en cm	170	172	175	180	185
Effectifs n_i	1	2	3	3	1

Tableau N°2

Notes en maths x_i	[8 ;10[	[10 ;12[	[12 ;14[	[14 ;16[	[16 ;18[
Effectifs n_i	3	4	2	1	1

Pour chacun des tableaux ci-dessus, déterminer l'effectif total, la moyenne, la variance, l'écart-type et les fréquences partielles.

Solution

— **Tableau N°1 :** Série non groupée en classes.
$N = 1 + 2 + 3 + 3 + 1 = 10$.

$\bar{x} = \frac{1}{10}(1 \times 170 + 2 \times 172 + 3 \times 175 + 3 \times 180 + 1 \times 185) = 176{,}4$ cm.

$V(x) = \frac{1}{10}(1 \times 170^2 + 2 \times 172^2 + 3 \times 175^2 + 3 \times 180^2 + 1 \times 185^2) - (176{,}4)^2 = 19{,}84$.

$\sigma(x) = \sqrt{19{,}84} = 4{,}45$.

$f_1 = \frac{1}{10} \quad ; \quad f_2 = \frac{2}{10} \quad ; \quad f_3 = \frac{3}{10} \quad ; \quad f_4 = \frac{3}{10} \quad \text{et} \quad f_5 = \frac{1}{10}$.

— **Tableau N°2 :** Série groupée en classes.
$N = 3 + 4 + 2 + 1 + 1 = 11$.

$$\bar{x} = \frac{1}{11}\left[3 \times \left(\frac{8+10}{2}\right) + 4 \times \left(\frac{10+12}{2}\right) + 2 \times \left(\frac{12+14}{2}\right) + 1 \times \left(\frac{14+16}{2}\right) + 1 \times \left(\frac{16+18}{2}\right)\right] = 11,$$

$$V(x) = \frac{1}{11}\left(3 \times 9^2 + 4 \times 11^2 + 2 \times 13^2 + 1 \times 15^2 + 1 \times 17^2\right) - (11{,}73)^2 = 5{,}95.$$

$\sigma(x) = \sqrt{5{,}95} = 2{,}44$.

$f_1 = \frac{3}{11} \quad ; \quad f_2 = \frac{4}{11} \quad ; \quad f_3 = \frac{2}{11} \quad ; \quad f_4 = \frac{1}{11} \quad \text{et} \quad f_5 = \frac{1}{11}$.

III. SÉRIE STATISTIQUE DE DEUX VARIABLES

1. Définitions : Cas général, série non injective

On appelle série statistique de deux variables x et y, ou série statistique double, la série obtenue si l'étude est réalisée à la fois sur deux caractères différents x et y.
Elle est donc formée de deux séries simples qui peuvent être groupées ou non ; ou l'une peut être groupée et l'autre non groupée.
On la note (x_i, y_j, n_{ij}). On la représente dans un tableau à double entrée appelé tableau de contingence.

Si $x = \{x_1, x_2, \ldots, x_p\}$ et $y = \{y_1, y_2, \ldots, y_q\}$, alors :

— n_{ij} est l'effectif du couple (x_i, y_j). L'effectif total sera $\boxed{N = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}}$

— $\boxed{n_{i\bullet} = n_{i1} + n_{i2} + \cdots + n_{iq} \quad \text{et} \quad n_{\bullet j} = n_{1j} + n_{2j} + \cdots + n_{pj}}$ sont respectivement les effectifs partiels sur la ligne i et sur la colonne j.

— $\boxed{f_{ij} = \frac{n_{ij}}{N}}$ est la fréquence du couple (x_i, y_j). $\boxed{f_{i\bullet} = \frac{n_{i\bullet}}{N} \quad \text{et} \quad f_{\bullet j} = \frac{n_{\bullet j}}{N}}$ sont les fréquences partielles sur la ligne i et la colonne j.

— Les fréquences conditionnelles : $\boxed{f_{x_i/y_j} = \frac{n_{ij}}{n_{\bullet j}} \quad \text{et} \quad f_{y_j/x_i} = \frac{n_{ij}}{n_{i\bullet}}}$

— On appelle **nuage de points** l'ensemble des points $M(x_i; y_j)$, que l'on notera M_{ij} dans un repère.
On appelle **point moyen** le point G, barycentre des points $(M_{ij}; n_{ij})$.

— On appelle **covariance** d'une série statistique de deux variables x et y, le réel noté :

$$\boxed{\mathbf{Cov}(x,y) = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_i y_j - \bar{x} \cdot \bar{y}}$$

Exemple : Le tableau ci-dessous représente les notes x en Maths et les notes y en PC de 10 élèves d'une classe de TS_2.

x_i \ y_j	**8**	**11**	**12**
9	2	0	0
10	0	3	1
11	0	1	2
12	0	0	1

1. Donner la valeur de n_{23}. Interpréter cette valeur.
2. Déterminer les séries marginales de x et y, puis donner $\bar{x}$ et $\bar{y}$.
3. Déterminer la série conditionnelle $z = x \,/\, _{y=12}$. Calculer sa moyenne puis l'interpréter.
4. Déterminer $f_{3\bullet}$ et $f_{\bullet 2}$.
5. Calculer $\text{cov}(x, y)$.

Solution

1. n_{23} est la valeur qui se situe sur la deuxième ligne et la troisième colonne.
   Donc $n_{23} = 1$. Ça veut dire qu'il y a 1 seul élève qui a 10 en Maths et 12 en PC.
2. La série marginale de x : (On extrait la série de x de la série double)

Notes de Maths x_i	9	10	11	12
Effectifs n_i	2	4	3	1

$$\bar{x} = \frac{1}{10}(2 \times 9 + 4 \times 10 + 3 \times 11 + 1 \times 12) = 10{,}3$$

La série marginale de y : (On extrait la série de y de la série double)

Notes de PC y_j	8	11	12
Effectifs n_j	2	4	4

$$\bar{y} = \frac{1}{10}(2 \times 8 + 4 \times 11 + 4 \times 12) = 10{,}8$$

3. La série conditionnelle $z = x \,/\, _{y=12}$:

$z_i = x_i/_{y=12}$	10	11	12
n_i	1	2	1

$$\bar{z} = \frac{1}{4}(1 \times 10 + 2 \times 11 + 1 \times 12) = 11.$$

C'est la moyenne en Maths des élèves qui ont 12 en PC.

4. Déterminer $f_{3.}$ et $f_{.2}$.

$$f_{3.} = \frac{n_{3.}}{10} \quad \text{et } n_{3.} = 0 + 1 + 2 = 3 \quad \Rightarrow \quad f_{3.} = \frac{3}{10}$$

$$f_{.2} = \frac{n_{.2}}{10} \quad \text{et } n_{.2} = 0 + 3 + 1 + 0 = 4 \quad \Rightarrow \quad f_{.2} = \frac{4}{10}$$

5. La covariance de x et y.
   $\text{Cov}(x, y) = \frac{(2\times8\times9+3\times11\times10+1\times12\times10+1\times11\times11+2\times12\times11+1\times12\times12)}{10} - 10{,}3 \times 10{,}8 = 1{,}06.$

2. Cas particulier : Série injective

Une série est dite injective si : $\boxed{n_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}}$. Elle est notée $\boxed{(x_i; y_i)}$ et se présente sous forme d'un tableau de deux lignes de même longueur.
Dans ce cas, l'effectif total N est le nombre de couples $(x_i; y_i)$.

— $\boxed{\bar{x} = \frac{1}{N}\sum_{i=1}^{p} x_i}$ et $\boxed{\bar{y} = \frac{1}{N}\sum_{i=1}^{p} y_i}$ si la série n'est pas groupée en classes.

— $\boxed{\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_i c_i}$ et $\boxed{\bar{y} = \frac{1}{N}\sum_{i=1}^{p} n_i c_i}$ si la série est groupée en classes.

— $\boxed{V(x) = \frac{1}{N}\sum_{i=1}^{p} x_i^2 - \bar{x}^2}$ et $\boxed{V(y) = \frac{1}{N}\sum_{i=1}^{p} y_i^2 - \bar{y}^2}$

Plus simplement on a : $\boxed{V(x) = \overline{x^2} - \bar{x}^2}$ et $\boxed{V(y) = \overline{y^2} - \bar{y}^2}$

— $\boxed{\sigma(x) = \sqrt{V(x)}}$ et $\boxed{\sigma(y) = \sqrt{V(y)}}$

— $\boxed{\mathrm{Cov}(x, y) = \frac{1}{N}\sum_{i=1}^{p} x_i y_i - \bar{x}\bar{y}}$ ou plus simplement : $\boxed{\mathrm{Cov}(x, y) = \overline{xy} - \bar{x} \cdot \bar{y}}$

Le *nuage de points* est l'ensemble des points $M(x_i; y_i)$ dans un repère.
Le **point moyen** est le point $\boxed{G(\bar{x}; \bar{y})}$, il sera toujours au centre du nuage (Isobarycentre).

Exemple : Série non groupée en classes

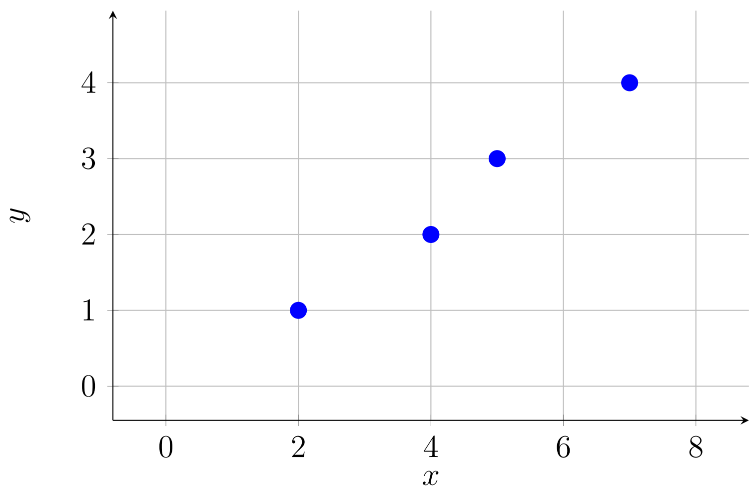
: On donne le tableau statistique ci-dessous :

x_i	2	4	5	7
y_i	1	2	3	4

1. Représenter le nuage de points de cette série statistique.
2. Calculer $\bar{x}, \bar{y}, V(x), V(y)$ et $\mathrm{Cov}(x, y)$.
3. Représenter le point moyen G dans le nuage.

Solution

1. Le nuage de points



— Si le nuage de points a la forme d'une droite, alors les variables x et y sont liées par une relation linéaire qu'on verra plus tard dans la suite du cours.

2.

					Total
x_i	2	4	5	7	18
y_i	1	2	3	4	**10**
x_i^2	4	16	25	49	**94**
y_i^2	1	4	9	16	**30**
$x_i y_i$	2	8	15	28	**53**

$$\bar{x} = \frac{18}{4} = 4{,}5 \quad ; \quad \bar{y} = \frac{10}{4} = 2{,}5 \quad ; \quad V(x) = \frac{94}{4} - (4{,}5)^2 = 3{,}25 \quad ; \quad V(y) = \frac{30}{4} - (2{,}5)^2 = 1{,}25$$

$$\mathrm{Cov}(x,y) = \frac{53}{4} - 4{,}5 \times 2{,}5 = 2$$

3. Le point moyen est donc $G(4{,}5\,;\,2{,}5)$. On peut le placer dans le nuage.

Exemple : Série groupée en classes

Durées de connexion (en minutes) de 40 élèves sur une plateforme :

Classe	$[0\,;\,10[$	$[10\,;\,20[$	$[20\,;\,30[$	$[30\,;\,40[$	$[40\,;\,50[$
Effectif n_i	5	8	12	10	5
Centre c_i	5	15	25	35	45

Effectif total : $N = 5 + 8 + 12 + 10 + 5 = 40$

$$\begin{aligned}\bar{x} &= \frac{1}{40}(5 \times 5 + 8 \times 15 + 12 \times 25 + 10 \times 35 + 5 \times 45)\\ &= \frac{5 \cdot 5 + 8 \cdot 15 + 12 \cdot 25 + 10 \cdot 35 + 5 \cdot 45}{40}\\ &= 25\end{aligned}$$

$$\begin{aligned}V(x) &= \frac{1}{40}(5 \cdot 5^2 + 8 \cdot 15^2 + 12 \cdot 25^2 + 10 \cdot 35^2 + 5 \cdot 45^2) - \bar{x}^2\\ &= \frac{5 \cdot 25 + 8 \cdot 225 + 12 \cdot 625 + 10 \cdot 1225 + 5 \cdot 2025}{40} - 625\\ &= 125\end{aligned}$$

$$\sigma(x) = \sqrt{V(x)} = \sqrt{125} \approx 11{,}18$$

Fréquence partielle de la classe $[20; 30[$: $\quad f_3 = \frac{12}{40} = 0{,}3$

IV. AJUSTEMENT LINÉAIRE

1. Coefficient de corrélation

On appelle coefficient de corrélation linéaire entre les variables x et y (le lien entre les deux variables) d'une série statistique double, le réel

$\boxed{r = \frac{\mathrm{Cov}(x,y)}{\sqrt{V(x) \times V(y)}}}$ ou encore $\boxed{r = \frac{\mathrm{Cov}(x,y)}{\sigma(x)\,\sigma(y)}}$

- On a toujours $\boxed{-1 \leq r \leq 1}$
- Si $|r| \geq 0{,}87$ ou $r^2 \geq 0{,}75$ alors la *corrélation entre x et y est forte.*
- Si $|r| < 0{,}87$ ou $r^2 < 0{,}75$ alors la *corrélation entre x et y est faible.*
- Si $r = -1$ ou $r = 1$ alors la *corrélation entre x et y est parfaite.*
- Si $r = 0$, alors la *corrélation entre x et y est nulle.*
  Dans ce cas, il n'y a aucune relation entre x et y.
  On dira que les variables x et y sont *indépendantes*.

Remarque : Quand la corrélation entre deux variables est forte, alors on peut faire une **estimation** d'une des valeurs connaissant l'autre à l'aide des droites de régression.

2. Droites de régression : Par la méthode des moindres carrés

On peut déterminer les droites de régression linéaires de la manière ci-dessous, appelée la méthode des moindres carrés :

$\boxed{(D_{y/x}) : y - \bar{y} = a(x - \bar{x}) \quad \text{avec} \quad a = \frac{\text{Cov}(x,y)}{V(x)}}$ est la droite de régression de y en x.

$\boxed{(D_{x/y}) : x - \bar{x} = a'(y - \bar{y}) \quad \text{avec} \quad a' = \frac{\text{Cov}(x,y)}{V(y)}}$ est la droite de régression de x en y.

- Après transformation, elles s'écrivent sous la forme :
  $(D_{y/x}) : y = ax + b \quad \text{et} \quad (D_{x/y}) : x = a'y + b'$
- On « peut » retrouver $D_{y/x}$ à partir de $D_{x/y}$ et réciproquement.
- Les droites de régression linéaires passent toujours par le point moyen.
- $\boxed{\text{On a toujours } aa' = r^2}$ (À démontrer).

Exercice d'application

D'après des études scientifiques, la croissance d'un arbre ne s'arrête jamais.
On considère un arbre dont la hauteur x en *mètres* et son âge y en *années* sont consignés dans le tableau ci-dessous :
Les résultats seront donnés à 1 chiffre après la virgule.

x_i	3	5	7,5	8
y_i	2	4	6	7

1. Déterminer le coefficient de corrélation linéaire entre x et y.
2. Justifier qu'on peut estimer la hauteur de cet arbre si on connaît son âge.
3. Quelle serait sa hauteur à l'âge de 10 ans ?
4. Si l'arbre mesure 11 mètres, estimer son âge en années.

Correction de l'exercice

1. **Détermination du coefficient de corrélation linéaire**

$$r = \frac{\text{Cov}(x,y)}{\sigma(x)\sigma(y)} \approx 0{,}995$$

Il y a donc une **corrélation linéaire très forte et positive** entre la hauteur et l'âge de l'arbre.

2. **Justification de l'utilisation d'une estimation linéaire**
   Comme $r \approx 1$, la corrélation est très forte. On peut donc utiliser la droite de régression pour estimer la hauteur de l'arbre à partir de son âge, ou inversement.
3. **Estimation de la hauteur à 10 ans**
   Droite de régression de y en fonction de x :
$$y = ax + b \quad \text{avec } a \approx 0{,}95 \quad \text{et } b \approx -0{,}83$$
$$y(10) = 0{,}95 \times 10 - 0{,}83 = \boxed{8{,}7 \text{ mètres}}$$
4. **Estimation de l'âge si l'arbre mesure 11 mètres**
   Droite de régression de x en fonction de y :
$$x = a'y + b' \quad \text{avec } a' \approx 1{,}042 \quad \text{et } b' \approx 0{,}924$$
$$y(11) = \frac{11 - 0{,}924}{1{,}042} \approx \boxed{12{,}4 \text{ années}}$$