# I - Définition, exemple

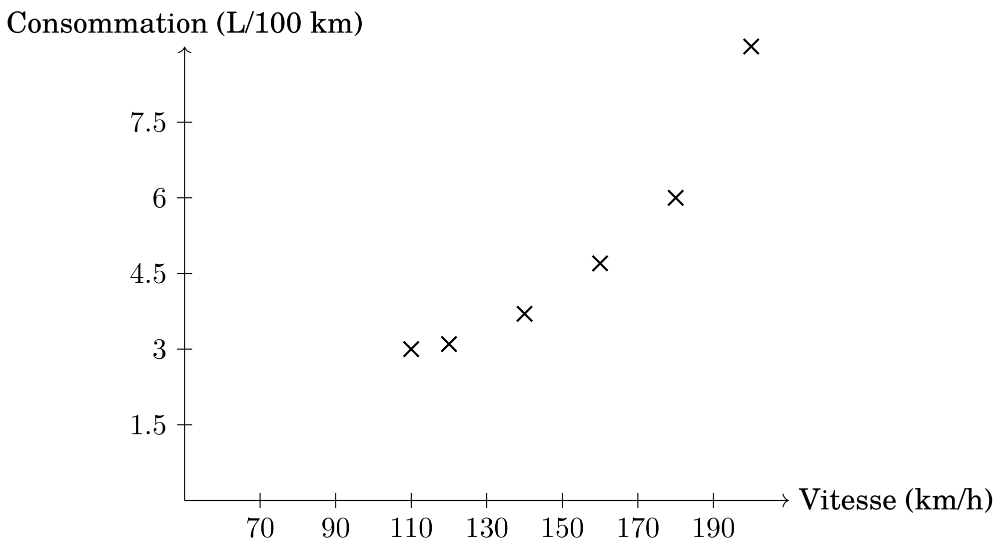
**Définition 1 :**
On appelle série statistique à deux variables (ou série statistique doubles) une série statistique à deux caractères étudiés simultanément.

**Exemple 1 :**
On a relevé, pour un modèle de voiture, la consommation en carburant (en L/100 km) pour différentes vitesses (en km/h) sur le cinquième rapport :

| Vitesse $x_i$ (en km/h) | 60 | 70 | 90 | 110 | 130 | 150 |
|---|---|---|---|---|---|---|
| Consommation $y_i$ (en L/100 km) | 3 | 3.1 | 3.7 | 4.7 | 6 | 9 |

**Définition 2 :**
Dans un repère orthogonal, l'ensemble des points $M_i$ de coordonnées $(x_i, y_i)$ s'appelle le nuage de points.

**Exemple 2 :**
Les points de l'exemple précédent forment le nuage de points suivant représenté dans le repère orthogonal ci-dessous.




# II - Point moyen

**Définition 3 :**
Le point moyen d'un nuage de points $G$ de coordonnées $(\overline{x}, \overline{y})$ où :

- $\overline{x}$ représente la moyenne des $x_i$ :
$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- $\overline{y}$ représente la moyenne des $y_i$ :
$$\overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

**Exemple 3 :**
Le point moyen de l'exemple 1 est :

$$\overline{x} = \frac{60 + 70 + 90 + 110 + 130 + 150}{6} = 101.66$$

$$\overline{y} = \frac{3 + 3.1 + 3.7 + 4.7 + 6 + 9}{6} = 4.75$$

## III - Ajustement affine par la méthode des moindres carrés

**Définition 4 :**
On appelle covariance de $x$ et de $y$ le nombre :

$$\text{cov}(x, y) = \frac{1}{N}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \left(\frac{1}{N}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\,\overline{y}$$

**Définition 5 :**
La variance du caractère $x$ est :

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

**Définition 6 :**
La variance du caractère $y$ est :

$$V(y) = \frac{1}{N}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \left(\frac{1}{N}\sum_{i=1}^{n} y_i^2\right) - \overline{y}^2 = \text{cov}(y, y)$$

**Définition 7 :**
L'écart type de $x$ est :

$$\sigma(x) = \sqrt{V(x)}$$

L'écart type de $y$ est :

$$\sigma(y) = \sqrt{V(y)}$$

**Exemple 4 :**
Calculons dans l'exemple 1 : $\text{cov}(x, y)$, $\text{cov}(x, x)$, $\text{cov}(y, y)$, $\sigma(x)$, $\sigma(y)$.
On a :

$$\overline{x} = 101.66, \quad \overline{y} = 4.91$$

| $x_i$ | $y_i$ | $x_i y_i$ | $x_i^2$ | $y_i^2$ |
|---|---|---|---|---|
| 60 | 3 | 180 | 3600 | 9 |
| 70 | 3.1 | 217 | 4900 | 9.61 |
| 90 | 3.7 | 333 | 8100 | 13.69 |
| 110 | 4.7 | 517 | 12100 | 22.09 |
| 130 | 6 | 780 | 16900 | 36 |
| 150 | 9 | 1350 | 22500 | 81 |
| $\sum$ | | 3377 | 68100 | 171.39 |

$$\text{cov}(x,y) = \frac{1}{N}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y} = \frac{3377}{6} - 101.66 \times 4.91 = 63.68$$

$$V(x) = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 = \frac{68100}{6} - (101.66)^2 = 1015.2444$$

$$V(y) = \frac{1}{N}\sum_{i=1}^{n} y_i^2 - \overline{y}^2 = \frac{171.39}{6} - (4.91)^2 = 4.4569$$

$$\sigma(x) = \sqrt{V(x)} = \sqrt{1015.2444} = 31.86$$

$$\sigma(y) = \sqrt{V(y)} = \sqrt{4.4569} = 2.11$$

## IV - Théorème : Ajustement affine par la méthode des moindres carrés

**Théorème 1 :**
Lors d'un ajustement affine par la méthode des moindres carrés :

1. La droite de régression $D$ de $Y$ en $X$ a pour équation :

$$D(Y/X) : Y = aX + b$$

   où :

$$a = \frac{\text{cov}(x,y)}{V(x)}, \quad b = \overline{Y} - a\overline{X}$$

   Cette droite passe par le point moyen du nuage $G(\overline{X}, \overline{Y})$.

2. La droite de régression $D$ de $X$ en $Y$ a pour équation :

$$D(X/Y) : X = a'Y + b'$$

   où :

$$a' = \frac{\text{cov}(x,y)}{V(y)}, \quad b' = \overline{X} - a'\overline{Y}$$

   Cette droite passe également par le point moyen du nuage $G(\overline{X}, \overline{Y})$.

**Exemple 5 :**
Calculons dans l'exemple 1 les droites de régression :

- Droite de régression $D(Y/X)$ :

$$a = \frac{\text{cov}(x,y)}{V(x)} = \frac{63.68}{1015.2444} = 0.0627$$

$$b = \overline{Y} - a\overline{X} = 4.91 - 0.0627 \times 101.66 = -1.46$$

   Donc :

$$D(Y/X) : Y = 0.0627X - 1.46$$

- Droite de régression $D(X/Y)$ :

$$a' = \frac{\text{cov}(x,y)}{V(y)} = \frac{63.68}{4.4569} = 14.287$$

$$b' = \overline{X} - a'\overline{Y} = 101.66 - 14.287 \times 4.91 = 31.51$$

Donc :

$$D(X/Y) : X = 14.287Y + 31.51$$

## V - Coefficient de corrélation linéaire

**Définition 8 :**
Le coefficient de corrélation linéaire d'une série statistique à deux variables $x$ et $y$ est le nombre $r$ défini par :

$$r = \frac{\text{cov}(x,y)}{\sqrt{V(x)}\sqrt{V(y)}} = \frac{\text{cov}(x,y)}{\sigma(x)\sigma(y)}$$

**Remarque 1 :**

- $-1 \leq r \leq 1$.
- Si $r = 1$ ou $r = -1$, il y a une corrélation positive ou négative parfaite entre $X$ et $Y$, et les points $(x_i, y_i)$ sont tous sur la droite de régression.
- Si $r = 0$, il n'y a pas de corrélation entre $X$ et $Y$, et les points $(x_i, y_i)$ sont dispersés au hasard.
- Si $0 < r < 1$, il y a une corrélation positive faible, moyenne ou forte entre $X$ et $Y$.
- Si $-1 < r < 0$, il y a une corrélation négative faible, moyenne ou forte entre $X$ et $Y$.

**Exemple 6 :**
Calculons dans l'exemple 1 le coefficient de corrélation linéaire :

$$\text{cov}(x,y) = 63.68, \quad \sigma(x) = 31.86, \quad \sigma(y) = 2.11$$

Donc :

$$r = \frac{\text{cov}(x,y)}{\sigma(x)\sigma(y)} = \frac{63.68}{31.86 \times 2.11} = 0.947$$

Il y a une corrélation positive forte entre $X$ et $Y$.

# VI - Exercices d'application

**Exercice 1 :**

Dans la série statistique suivante, $X$ représente le nombre de jours d'exposition au soleil d'une feuille et $Y$ le nombre de stomates aérifères au millimètre carré :

| $X$ | 2 | 4 | 8 | 10 | 24 | 40 | 52 |
|---|---|---|---|---|---|---|---|
| $Y$ | 6 | 11 | 15 | 20 | 39 | 62 | 85 |

1. Tracer le nuage des points.
2. Calculer le coefficient de corrélation linéaire entre $X$ et $Y$. Conclusion ?
3. Déterminer l'équation de la droite de régression de $Y$ en fonction de $X$.
4. Si on expose au soleil une feuille 15 jours, quel est le nombre de stomates aérifères peut-on prévoir ?

**Exercice 2 :**

Dans la série statistique suivante, $X$ représente le nombre d'heures d'étude par semaine et $Y$ la note obtenue à un examen (sur 20) :

| $X$ | 5 | 7 | 10 | 12 | 15 | 18 | 20 |
|---|---|---|---|---|---|---|---|
| $Y$ | 8 | 10 | 12 | 14 | 16 | 18 | 20 |

1. Tracer le nuage des points.
2. Calculer le coefficient de corrélation linéaire entre $X$ et $Y$. Conclusion ?
3. Déterminer l'équation de la droite de régression de $Y$ en fonction de $X$.
4. Si un étudiant étudie 14 heures par semaine, quelle note peut-on prévoir ?