STATISTIQUES À DEUX VARIABLES

I INTRODUCTION

1 STATISTIQUES À UNE VARIABLE

En Statistiques, on étudie des caractères sur une population.
Dans la partie « statistiques à une variable », on étudie **un seul** caractère sur la population.
Un caractère peut être quantitatif (mesurable) ou qualitatif (non mesurable).

EXEMPLES :

- On étudie le sport préféré des élèves d'une classe de 35 élèves.
  Population : Les élèves **Caractère étudié :** Le sport.
  Ici, le caractère ne se mesure pas, on dit qu'il est qualitatif.
- On étudie la hauteur des arbustes dans une pépinière.
  Population : Les arbustes **Caractère étudié :** La taille.
  Ici, le caractère se mesure, on dit qu'il est quantitatif.

ÉTUDE D'UNE SÉRIE STATISTIQUE À UNE VARIABLE :

Dans cette partie des statistiques, on peut calculer différents paramètres du caractère :

- **Des paramètres de position** : moyenne, médiane.
- **Des paramètres de dispersion :** écart inter-quartile, écart-type, étendue.
- **Bilan :** On représente souvent un diagramme en boîte pour récapituler ces données.
- **Représentation graphique** : On peut représenter la série en diagramme circulaire, histogramme, ...

2 STATISTIQUES À DEUX VARIABLES

Définition :
Dans la partie « statistiques à deux variables », on étudie deux caractères quantitatifs sur une même population.

EXEMPLES :

- On étudie le poids et la taille de nouveaux nés dans une maternité.
  Population : Les nouveaux nés **Caractères étudiés :** La taille et le poids
- On étudie le Chiffre d'affaire et le budget communication d'une entreprise.
  Population : L'entreprise **Caractères étudiés :** le Chiffre d'affaire et le budget communication
- On étudie le Chiffre d'affaire d'une entreprise sur plusieurs années.
  Population : L'entreprise **Caractères étudiés :** le Chiffre d'affaire et les années.

II GÉNÉRALITÉS SUR LES SÉRIES STATISTIQUES À DEUX VARIABLES

1 DÉFINITION

Définition :
On considère deux variables statistiques x et y observées sur une même population de n individus.
On note $x_1, x_2, \ldots, x_n$ les valeurs relevées pour la variable x et $y_1, y_2, \ldots, y_n$ les valeurs relevées pour la variable y.
Les couples $(x_1; y_1), (x_2; y_2), \ldots, (x_n; y_n)$ forment une série statistique à deux variables.
Dans ce chapitre, on va s'intéresser au lien qui peut exister entre ces deux variables.

2 NUAGE DE POINTS :

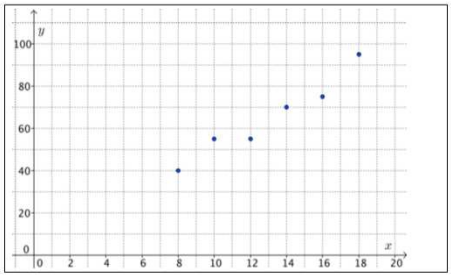
Définition : Dans un repère orthogonal, l'ensemble des points M_i de coordonnées $(x_i ; y_i)$, avec $1 \leq i \leq n$, est appelé le nuage de points associé à la série statistiques $(x_1 ; y_1), (x_2 ; y_2), \ldots, (x_n ; y_n)$ à deux variables.

EXEMPLE :

Le tableau suivant présente l'évolution du budget publicitaire et du chiffre d'affaire d'une société au cours des 6 dernières années :

Budget publicitaire en milliers d'euros x_i	8	10	12	14	16	18
Chiffre d'affaire en milliers d'euros y_i	40	55	55	70	75	95

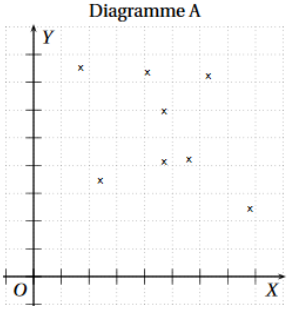
On peut placer les points du nuage correspondant à la série statistique ci-dessus.
Il y a six points $M_1(8;40), M_2(10;55), M_3(12;55), M_4(14;70), M_5(16;75)$ et $M_6(18;95)$.



3 INTRODUCTION À LA NOTION D'INTERPOLATION

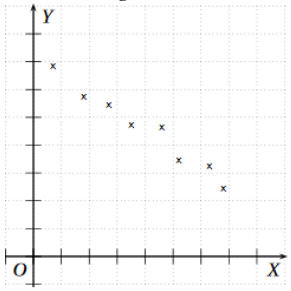
Un nuage peut avoir des formes différentes. L'objet des statistiques à deux variables est de les étudier et d'essayer de les modéliser quand cela est possible.

Diagramme A



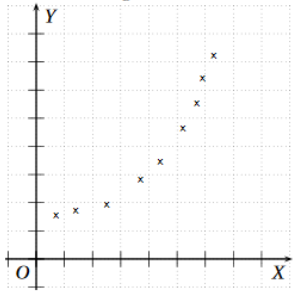
Dans ce cas, les points semblent dispersés, sans que l'on puisse imaginer de relation entre les deux variables.

Diagramme B



Dans ce cas, les points semblent relativement alignés.
On pourrait imaginer qu'une droite ne passe « pas très loin » de chaque point.
Dans ce cas, un ajustement semble possible.

Diagramme C



Dans ce cas, les points semblent relativement suivre une courbe.
On pourrait trouver par ordinateur/calculatrice une relation entre les deux variables.
Un ajustement semble possible.
Mais plus compliqué que celui du B.

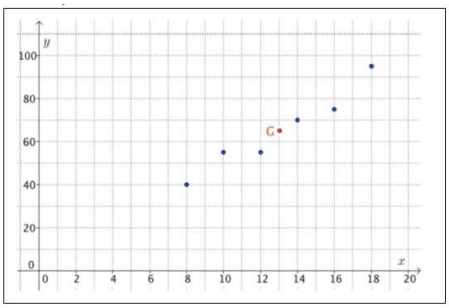
4 Point moyen

Définition :
Le point G de coordonnées $(\bar{x};\bar{y})$, où $\bar{x}$ et $\bar{y}$ sont les moyennes respectives des x_i et des y_i, est appelé le **point moyen** du nuage de points associé à la série statistique $(x_1;y_1),(x_2;y_2),\ldots,(x_n;y_n)$ à deux variables.

$$G\left(\frac{x_1+x_2+\ldots+x_n}{n};\frac{y_1+y_2+\ldots+y_n}{n}\right)$$

Exemple :

En reprenant l'exemple précédent de l'entreprise, on pourrait calculer les moyennes respectives du budget publicitaire et du Chiffre d'Affaire :
$\bar{x}=\dfrac{8+10+12+14+16+18}{6}=13$ et $\bar{y}=\dfrac{40+55+55+70+75+95}{6}=65$
et placer le point obtenu dans le nuage de point : $G(13;65)$



5 Définitions :

Ajustement affine : :

A partir d'un nuage de points, on cherche à établir un lien entre les deux caractères de la série statistique pour pouvoir :

- Préciser des valeurs inconnues dans la série statistique.
- Faire des pronostics sur les valeurs en dehors de la série statistique.

Pour cela, on essaie d'approcher le nuage de points à l'aide une droite.
Cette droite ne passant pas exactement par les points du nuage, on dit qu'on obtient un **ajustement affine**.

Interpolation :

Une interpolation est un calcul réalisé dans le domaine d'étude fourni par les valeurs de la série, pour déterminer des valeurs inconnues.

Extrapolation :

Une extrapolation est un calcul réalisé en dehors du domaine d'étude fourni par les valeurs de la série, pour déterminer des valeurs inconnues.

6 EXEMPLES :

- On donne une série exprimant la population d'une ville en fonction des années et on souhaite faire des prévisions pour les années à venir.
  Les prévisions sortent du domaine d'étude de la série, on parle dans ce cas d'extrapolation.
- On donne une série exprimant la température extérieure et la consommation électrique correspondante.
  Les températures étudiées s'échelonnent entre $-10°C$ et $10°C$ avec un pas de $4°C$.
  Sans faire de nouveaux relevés, on souhaite estimer la consommation électrique pour toutes les températures entières comprises entre $-10°C$ et $10°C$.
  Les calculs sont dans le domaine d'étude de la série, on parle dans ce cas d'interpolation.

DROITE D'AJUSTEMENT :

Lorsque les points d'un nuage sont sensiblement alignés, on peut construire une droite, appelé droite d'ajustement (ou droite de régression), passant « au plus près » de ces points.

III MÉTHODES D'AJUSTEMENTS :

1 MÉTHODE GRAPHIQUE AU JUGÉ :

On trace «au jugé » une droite **passant par le point moyen du nuage** qui « semble résumer » le nuage de points.
C'est une méthode simple mais peu rigoureuse.

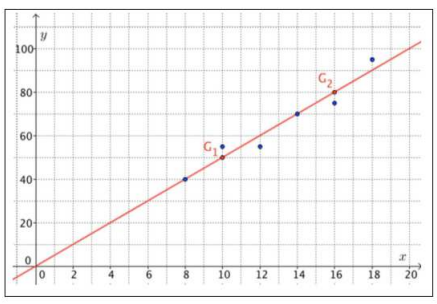
2 MÉTHODE DE MAYER :

On sépare le nuage en deux sous nuages et on calcule les coordonnées des points moyens des deux sous nuages obtenus.
La droite de MAYER est la droite passant par ces deux points. Elle passe aussi par le point moyen du nuage.

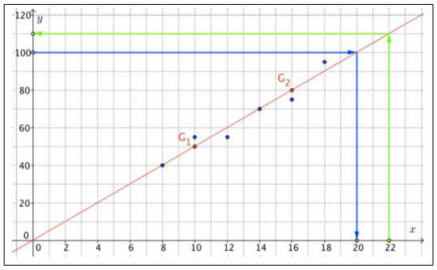
EXEMPLE :

On poursuit avec l'exemple de l'entreprise :

- On calcule les coordonnées du point moyen G_1 à partir des trois premiers points de la série statistique, puis les coordonnées du point moyen G_2 à partir des trois derniers points.
  On obtient : $\bar{x_1} = \dfrac{8+10+12}{3} = 10$ et $\bar{y_1} = \dfrac{40+55+55}{3} = 50$.
  Les coordonnées du point moyen sont G_1 sont $G_1(10;50)$.
  $\bar{x_2} = \dfrac{14+16+18}{3} = 16$ et $\bar{y_2} = \dfrac{70+75+95)}{3} = 80$.
  Les coordonnées du point moyen G_2 sont $G_2(16;80)$.
- La droite (G_1G_2) forme un ajustement affine du nuage de points.



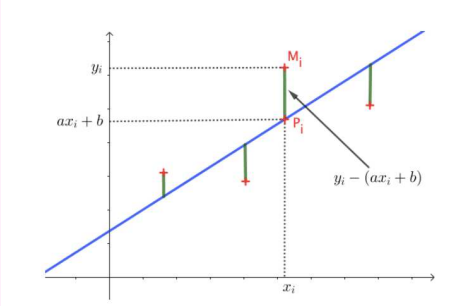
- Après avoir tracé cette droite, on peut, graphiquement, estimer :



- Le chiffre d'affaire à prévoir pour un budget publicitaire de 22000€.
- Le budget publicitaire à prévoir pour un chiffre d'affaire de 100000 €.

3 MÉTHODE DES MOINDRES CARRÉS

Cette méthode consiste à rechercher la position de la droite d'ajustement tel que la somme des carrés des longueurs donnant les distances respectives entre la droite et les points soit minimale.
Le principe consiste donc à déterminer les coefficients a et b d'une droite d'équation $y = ax + b$ de sorte qu'elle passe le "plus près possible" des points du nuage.
Pour chaque abscisse x_i, on calcule la distance M_iP_i entre le point du nuage et le point de la droite, soit : $M_iP_i = \left|y_i - (ax_i + b)\right|$



Dans la méthode des moindres carrés, on recherche a et b pour lesquels la somme des carrés des distances est minimale, soit :

$$M_1P_1^2 + \cdots + M_nP_n^2 = \big(y_1 - (ax_1 + b)\big)^2 + \cdots + \big(y_n - (ax_n + b)\big)^2$$

soit minimale.

4 PROPRIÉTÉ :

La droite d'ajustement de y en x a pour équation $y = ax + b$, avec :

$$a = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

$$b = \bar{y} - a\bar{x}$$

où

$$\text{cov}(x; y) = \frac{1}{n}\big((x_1 - \bar{x})\big(y_1 - \bar{y}\big) + \cdots + (x_n - \bar{x})\big(y_n - \bar{y}\big)\big)$$

est la covariance de (x, y) et

$$\text{var}(x) = \frac{1}{n}\big((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\big)$$

est la variance de x. - Admis -