

1 Nuage de points

Une série statistique à deux caractères numériques, dont les valeurs sont notées x_i et y_i , est appelée **série statistique à deux variables quantitatives**.

! **Remarque** : Les paramètres de position et de dispersion, étudiés en classe de 2^{de}, s'appliquent aux séries statistiques à deux variables.

Exemple

Le tableau ci-contre, correspondant à une série statistique à deux variables x et y , donne la masse mesurée y_i (en g) en fonction de l'épaisseur mesurée x_i (en mm) de plaques métalliques de même surface.

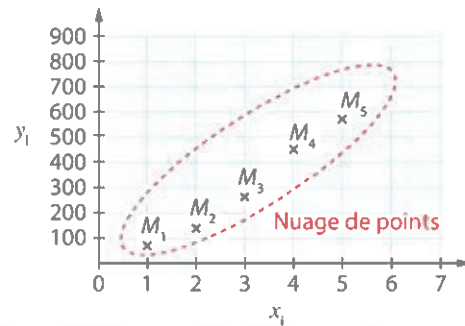
Épaisseur (en mm) : x_i	1	2	3	4	5
Masse (en g) : y_i	60	140	265	440	565

Dans un repère orthogonal, l'ensemble des points M_i de coordonnées $(x_i ; y_i)$ d'une série statistique constitue un **nuage de points**.

! **Remarque** : Le nuage de points peut être représenté à l'aide des fonctionnalités d'une calculatrice, d'un tableur-grapheur ou d'un logiciel de géométrie 2D.

Exemple

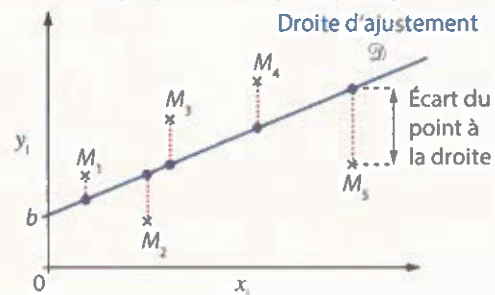
Pour représenter un nuage de points avec une calculatrice ou un tableur-grapheur, il faut saisir les valeurs x_i et y_i dans deux colonnes et utiliser les outils de représentation graphique propres à chaque outil. L'ensemble des points M_i de coordonnées $(x_i ; y_i)$ de cette série statistique constitue un nuage de points.



2 Ajustement affine de y en x

Pour une série statistique à deux variables, lorsque les points du nuage sont sensiblement alignés, un **ajustement affine** peut être réalisé. La **droite d'ajustement**, qui passe « le plus près possible » de tous les points du nuage, est caractérisée par une équation de la forme $y = ax + b$ (où a est le coefficient directeur et b l'ordonnée à l'origine de la droite).

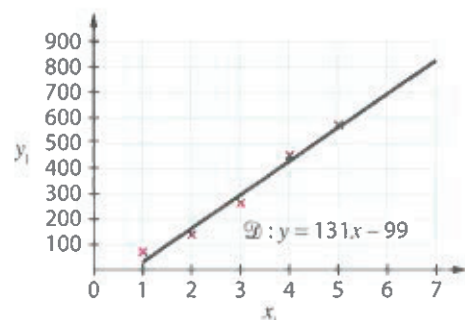
! **Remarque** : Les outils numériques (calculatrice, tableur-grapheur...) utilisent la méthode de calcul dite des **moindres carrés**, qui minimise la somme des écarts entre les points du nuage et les points de la droite d'ajustement \mathcal{D} .



Exemple

Dans le cas de la série du premier exemple, une fois le nuage de points représenté avec un tableur-grapheur, pour afficher l'équation $y = 131x - 99$ de la droite d'ajustement, il faut :

- sélectionner les points du nuage, faire un clic droit et choisir **Insérer une courbe de tendance** ;
- cocher les options **Linéaire** et **Afficher l'équation**,



Ajustement affine de y en x (suite)

- **Interpoler** consiste à calculer la valeur d'une des coordonnées d'un nouveau point appartenant à la droite d'ajustement et situé à l'intérieur du nuage de points.
- **Extrapoler** consiste à calculer la valeur d'une des coordonnées d'un nouveau point appartenant à la droite d'ajustement et situé à l'extérieur du nuage de points.

! **Remarque** : Dans les deux cas, on utilise l'équation de la droite d'ajustement pour déterminer la valeur manquante. Les coordonnées du nouveau point vérifient l'équation.

Exemples

Soit la droite d'ajustement \mathcal{D} de l'exemple précédent, d'équation $y = 131x - 99$.

- Calcul de l'abscisse x_A du point A d'ordonnée $y_A = 200$:

$$y_A = 131x_A - 99$$

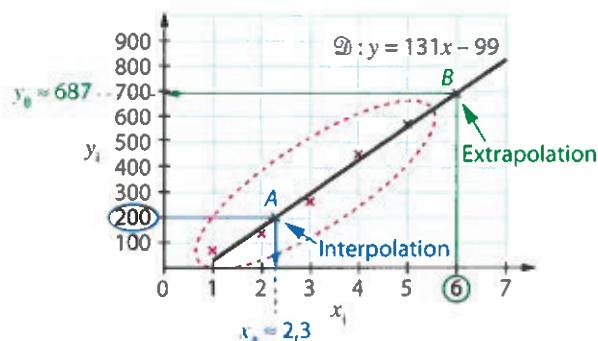
$$\Leftrightarrow 200 = 131x_A - 99$$

$$\Leftrightarrow x_A = \frac{200 + 99}{131} \approx 2,3$$

- Calcul de l'ordonnée y_B du point B d'abscisse $x_B = 6$:

$$y_B = 131x_B - 99$$

$$\Leftrightarrow y_B = 131 \times 6 - 99 = 687$$

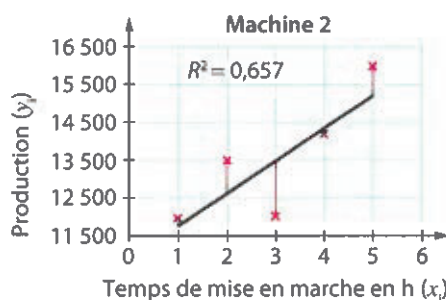
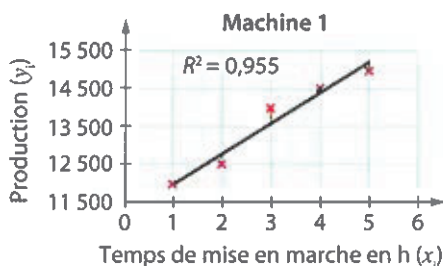


3 Coefficient de détermination

- Le **coefficient de détermination R^2** indique la qualité d'un ajustement affine : s'il est très proche de 1, les points du nuage sont presque alignés et il existe une forte **corrélation** entre les variables étudiées. Il est déterminé à l'aide des outils numériques.
- Deux variables peuvent être mathématiquement fortement corrélées ($R^2 \approx 1$) sans qu'il y ait nécessairement une relation de causalité entre l'une et l'autre dans la réalité (voir le problème 3 p. 19)

Exemples

Les graphiques ci-dessous représentent l'évolution du nombre de pièces produites par deux machines en fonction du temps (en h) de mise en marche.



- Machine 1 : $R^2 \approx 1$, les points du nuage sont sensiblement alignés : il y a une forte corrélation entre les variables x_i et y_i . La modélisation mathématique par une droite convient pour décrire la distribution des points : il est probablement intéressant de faire fonctionner cette machine au-delà de 5 h pour augmenter la production.
- Machine 2 : R^2 n'est pas proche de 1, les points du nuage sont dispersés : l'ajustement affine est peu pertinent, la corrélation entre les variables x_i et y_i est faible. La modélisation mathématique par une droite ne convient pas pour décrire la distribution des points : il n'est pas possible de prévoir la production au-delà de 5 h.