Statistique à deux variables

Introduction :

En statistique, on cherche à étudier l'effet d'un ou de plusieurs paramètres. Les années précédentes, il était question d'étudier en mathématiques une population avec des séries statistiques à une variable.
Cependant, dans de nombreux cas, les différents paramètres que l'on étudie pour une même population présentent des liens qu'il est important de pouvoir mettre en évidence, même s'il ne faut pas conclure trop vite à un lien de cause à effet. On parle alors de statistiques à plusieurs variables.

Ainsi, dans ce cours, nous allons nous intéresser aux séries statistiques à deux variables. Dans un premier temps, nous définirons ce qu'est une série statistique à deux variables, comment la représenter et quelles données caractéristiques on peut en déduire.
Ensuite, afin de mettre en avant les corrélations entre les deux variables, nous expliquerons comment faire un ajustement affine et comment en déduire la droite des moindres carrées, ce qui permettra d'interpoler ou d'extrapoler.
Enfin, nous montrerons comment se ramener à un ajustement linéaire avec un changement de variable.

1 Série statistique à deux variables

Pour étudier simultanément deux variables statistiques, il est possible de définir une série statistique double ou à deux variables.
Pour cela, on a pour une population donnée de n individus deux caractères quantitatifs :

- la variable x, pour laquelle les données relevées sont x_1, x_2, ..., x_n ;
- la variable y, pour laquelle les données relevées sont y_1, y_2, ..., y_n.

Chaque individu aura ainsi un couple de caractères associé $(x_i\ ;\ y_i)$ (avec $i \in \{1,\ 2,\ \ldots,\ n\}$). Notons que leurs unités seront souvent différentes (par ex., si on s'intéresse au lien entre la température extérieure et la consommation électrique, ou l'évolution d'une population quelconque en fonction du temps).

➜ L'ensemble de ces couples constitue une **série statistique à deux variables**.

a. Tableau de données et nuage de points

On représente généralement les couples sous forme de tableau avec une colonne (ou une ligne) pour le caractère x et une colonne (ou une ligne) pour le caractère y.

Exemple

Nous nous intéressons à l'éventuel lien entre la teneur en carbone, en pourcent, d'un objet et la charge de rupture, c'est-à-dire la charge, en kilogramme, qui provoquera la rupture de l'objet.

➜ 10 essais ont été faits en laboratoire, et nous obtenons les résultats suivants :

Teneur en carbone x_i (en %)	**Charge de rupture y_i** (en kg)
64	77
68	81
61	72
71	86
66	79
74	93
63	74
70	86
60	70
62	71

La plupart du temps, nous représentons ces données par un **nuage de points**, qui nous permet de mieux visualiser les données.

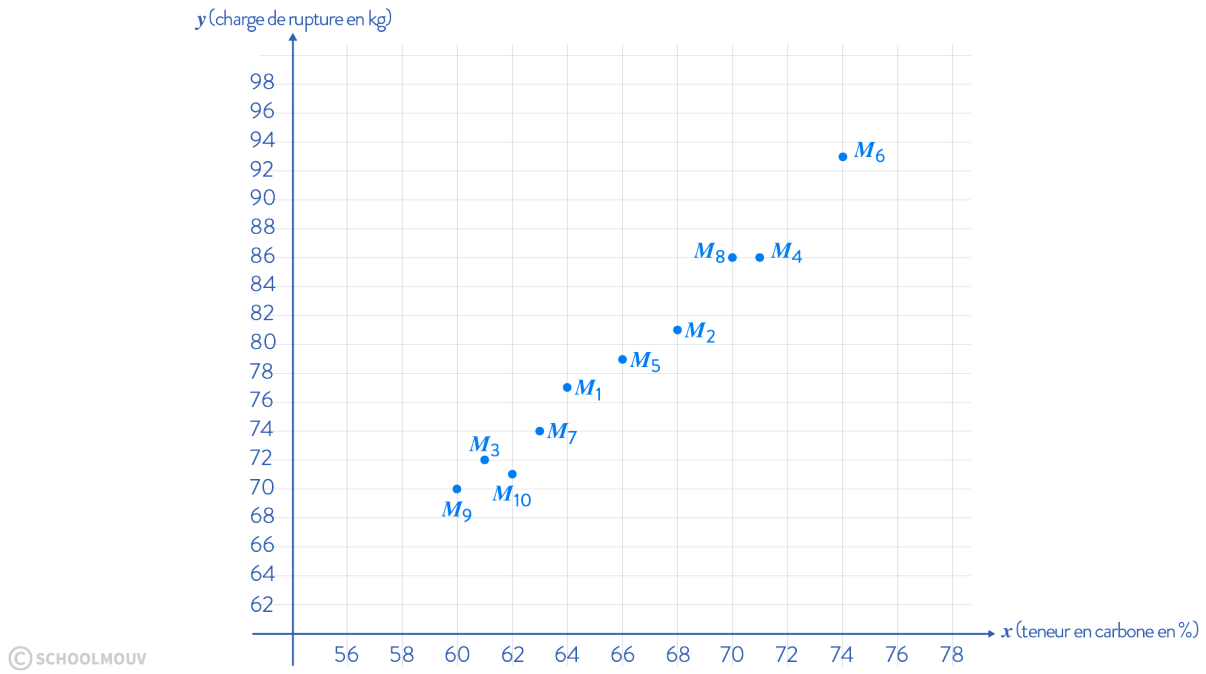
Définition

Nuage de points :

On représente une série statistique à deux variables x et y par un nuage de points dans un repère orthogonal $(O\ ;\ I,\ J)$, constitué de points M_i de coordonnées $(x_i\ ;\ y_i)$, x_i et y_i étant respectivement les valeurs des variables x et y pour l'individu i (i allant de 1 à n, avec n la taille de la population).

Exemple

Donnons le nuage de points correspondant aux données de notre exemple :



Teneur en carbone et charge de rupture

Le nuage de points que nous venons de représenter montre que les points ne semblent pas répartis au hasard.

➜ Les deux variables semblent avoir une corrélation : en effet, quand la teneur en carbone augmente, la charge de rupture paraît augmenter aussi.

Mais comment rendre cette corrélation plus évidente et, surtout, comment la quantifier ?

b. Point moyen

Tout d'abord, nous pouvons calculer la moyenne des deux variables, ce qui nous permet de définir le **point moyen**.

Définition

Point moyen :

Soit une série statistique à deux variables x et y, représentée par un nuage de points dans un repère $(O\ ;\ I,\ J)$.
On définit le point moyen de ce nuage comme le point G, de coordonnées $(\bar{x}\ ;\ \bar{y})$, où :

- $\bar{x}$ est la moyenne arithmétique des valeurs x_i associées à la variable x ;

- $\bar{y}$ est la moyenne arithmétique des valeurs y_i associées à la variable y.

Nous savons calculer ces moyennes grâce aux formules suivantes.

Rappel

Soit x_1, x_2 ..., x_n les n valeurs de la variable x, et y_1, y_2, ..., y_n les n valeurs de la variable y.

Nous avons alors :

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$
$$\bar{y} = \frac{y_1 + y_2 + y_3 + \ldots + y_n}{n}$$

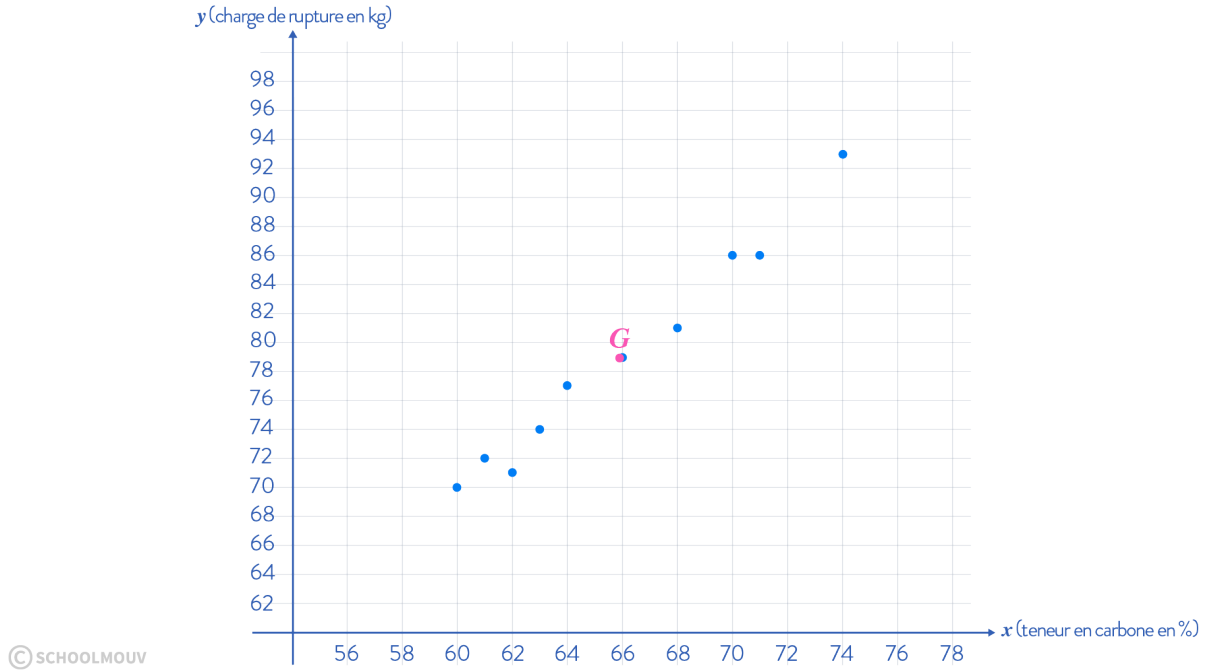
Exemple

Continuons de filer notre exemple, et calculons les moyennes de x et y :

$$\bar{x} = \frac{64 + 68 + 61 + 71 + 66 + 74 + 63 + 70 + 60 + 62}{10}$$
$$= \frac{659}{10}$$
$$= 65,9$$

$$\bar{y} = \frac{77 + 81 + 72 + 86 + 79 + 93 + 74 + 86 + 70 + 71}{10}$$
$$= \frac{789}{10}$$
$$= 78,9$$

➜ Le point moyen G a donc pour coordonnées $(65,9\ ;\ 78,9)$.



Point moyen

C. Covariance et coefficient de corrélation

Dans les classes précédentes, nous avons appris à calculer un indicateur qui permet de mesurer la dispersion des données d'une série statistique autour de sa moyenne.

➜ Il s'agit de la **variance**.

Rappel

Soit une série statistique à une variable x, d'effectif n : $(x_1,\ x_2,\ \ldots,\ x_n)$, de moyenne $\bar{x}$.
La variance de x, que nous notons ici $\text{var}(x)$, est donnée par la formule :

$$\begin{aligned}\text{var}(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2)\end{aligned}$$

Rappelons aussi que nous définissons l'écart-type de x, noté $\sigma(x)$, ainsi :

$$\sigma(x) = \sqrt{\text{var}(x)}$$

La **covariance**, elle, est une notion nouvelle, qui vient avec la notion de statistique à deux variables. Comme son nom l'indique, elle mesure la façon dont évoluent conjointement les deux variables considérées.

Définition

Covariance de $(x\ ;\ y)$:

Soit une série statistique à deux variables x et y, d'effectif n : $((x_1\ ;\ y_1),\ (x_2\ ;\ y_2),\ \ldots,\ (x_n\ ;\ y_n))$, respectivement de moyennes $\bar{x}$ et $\bar{y}$.
La covariance de $(x\ ;\ y)$, notée ici $\text{cov}(x\ ;\ y)$, est donnée par la formule :

$$\begin{aligned}\text{cov}(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\\ &= \frac{1}{n}((x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\ldots+(x_n-\bar{x})(y_n-\bar{y}))\end{aligned}$$

Nous le voyons, pour calculer la variance, nous effectuons une somme de produits entre deux grandeurs qui ne sont pas, la plupart du temps, exprimées dans la même unité.

➜ Nous allons donc définir un **coefficient de corrélation**, qui sera plus explicite.

Définition

Coefficient de corrélation :

Soit une série statistique à deux variables

- x, de variance $\text{var}(x)$ et d'écart-type $\sigma(x)$,
- et y, de variance $\text{var}(y)$ et d'écart-type $\sigma(y)$.

Soit $\text{cov}(x\ ;\ y)$ la covariance de x et y.
Le coefficient de corrélation r, aussi noté ρ_{xy}, est alors défini par :

$$\begin{aligned}r &= \frac{\text{cov}(x\ ;\ y)}{\sqrt{\text{var}(x)\text{var}(y)}}\\ &= \frac{\text{cov}(x\ ;\ y)}{\sigma(x)\sigma(y)}\end{aligned}$$

À retenir

Ce coefficient indique le lien, linéaire, qui existe entre les variables x et y :

- il appartient à l'intervalle $[-1\ ;\ 1]$;
- plus il est proche des bornes de l'intervalle -1 et 1, plus la corrélation linéaire entre x et y est forte ;

- en revanche, deux variables indépendantes ont un coefficient de corrélation proche de 0 ;
- s'il est positif, alors x et y varient « dans le même sens » (plus les valeurs de x grandissent, plus celles de y grandissent) ;
- s'il est négatif, alors x et y varient « en sens contraires » (plus les valeurs de x grandissent, plus celles de y diminuent).

Nous allons maintenant calculer tous ces indicateurs dans le cas de notre exemple et en tirer une première conclusion.

Exemple

1. Rappelons que nous avons trouvé : $\bar{x} = 65,9$ et $\bar{y} = 78,9$.

2. Calculons d'abord les variances et écarts-types de x et y :

$$\begin{aligned}
\text{var}(x) &= \frac{1}{10} \times ((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_9 - \bar{x})^2 + (x_{10} - \bar{x})^2) \\
&= \frac{1}{10} \times ((64 - 65,9)^2 + (68 - 65,9)^2 + \ldots + (60 - 65,9)^2 + (62 - 65,9)^2) \\
&= \frac{1}{10} \times ((-1,9)^2 + 2,1^2 + \ldots + (-5,9)^2 + (-3,9)^2) \\
&= \frac{1}{10} \times (3,61 + 4,41 + \ldots + 34,81 + 15,21) \\
&= \frac{198,9}{10} \\
&= 19,89 \\
\sigma(x) &= \sqrt{19,89} \\
&\approx 4,4598
\end{aligned}$$

$$\begin{aligned}
\text{var}(y) &= \frac{1}{10} \times ((y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \ldots + (y_9 - \bar{y})^2 + (y_{10} - \bar{y})^2) \\
&= \frac{1}{10} \times ((77 - 78,9)^2 + (81 - 78,9)^2 + \ldots + (70 - 78,9)^2 + (71 - 78,9)^2) \\
&= \frac{520,9}{10} \\
&= 52,09 \\
\sigma(x) &= \sqrt{19,89} \\
&\approx 7,2173
\end{aligned}$$

3. Intéressons-nous à la covariance de x et y :

$$\begin{aligned}\text{cov}(x\ ;\ y) &= \frac{1}{10}((x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\ldots+(x_{10}-\bar{x})(y_{10}-\bar{y}))\\ &= \frac{1}{10}(3,61+4,41+\ldots+30,81)\\ &= \frac{318,9}{10}\\ &= 31,89\end{aligned}$$

4 Nous pouvons calculer maintenant le coefficient de corrélation :

$$\begin{aligned}r &= \frac{\text{cov}(x\ ;\ y)}{\sqrt{\text{var}(x)\text{var}(y)}}\\ &= \frac{31,89}{\sqrt{19,89\times 52,09}}\\ &\approx 0,9907\end{aligned}$$

C Nous pouvons en déduire que x et y ont une corrélation linéaire forte ($r \approx 1$) ; en outre, quand les valeurs prises par x croissent, celles prises par y croissent également ($r > 0$).

2 Ajustement affine

Ce qui va suivre a pour but de donner des outils pour effectuer des prévisions pour des valeurs inconnues, que celles-ci soient dans le domaine d'étude ou en dehors.

➜ On parle alors d'**interpolation** et d'**extrapolation**.

Définition

Interpolation et extrapolation :

- Lorsque l'on s'intéresse à des valeurs inconnues mais qui font partie du domaine couvert par les données fournies par l'étude, alors on effectue une interpolation.
- Si l'on travaille hors de ce domaine, alors on effectue une extrapolation.

Exemple

Pour la teneur en carbone de nos objets et la charge de rupture associée :

- nous pouvons vouloir estimer la charge de rupture d'un objet dont la teneur de carbone est de $69,2\,\%$, valeur de x qui est bien comprise entre le minimum ($60\,\%$) et le maximum ($74\,\%$) de la série,

  ➜ nous effectuerons alors une interpolation ;

- nous pouvons aussi avoir besoin d'avoir une approximation de la charge de rupture pour un objet de teneur $50\,\%$;

- nous pouvons encore souhaiter connaître la teneur en carbone de l'objet pour que la charge de rupture soit de $100\ \text{kg}$;

  ➜ dans ces deux derniers cas, nous ferons une extrapolation.

Attention

Effectuer une extrapolation peut être dangereux : en effet, rien ne démontre que le modèle déduit des données fournies reste vrai en dehors de ce domaine.

➜ Il conviendra donc toujours de prendre des précautions.

Par exemple, si un commerçant s'intéresse au chiffre d'affaires qu'il fait en fonction de l'heure de la journée et qu'il ne relève les données qu'entre 17 heures et 19 heures, alors, ce créneau correspondant à la sortie des bureaux et donc à ses heures de grande fréquentation, il ne pourra extrapoler le chiffre fait lors des horaires « creux », à 14 h 30, par exemple.

a. Ajustement affine

Lorsqu'un lien linéaire semble apparaître entre deux variables, et afin de pouvoir faire des interpolations et des extrapolations, il est intéressant d'ajuster le nuage de points au moyen d'une droite et de caractériser ainsi la relation affine entre les deux variables.

➜ On parle d'**ajustement affine**.

Définition

Ajustement affine :

Le principe de l'ajustement affine est de tracer, lorsque les points d'un nuage semblent globalement alignés, une droite passant « au plus près » de ces points.

➜ Cette droite est alors appelée droite d'ajustement, ou droite de régression.

Remarquons que « au plus près » est une formulation assez vague. Il existe plusieurs techniques.

➜ Nous allons en présenter deux : une à partir de la notion de point moyen, la **méthode de Mayer**, et l'autre, très utilisée, dite **méthode des moindres carrés**.

b. Méthode de Mayer

Cette méthode, aussi appelée méthode des points moyens, consiste tout simplement à relier deux points moyens du nuage. Elle n'est guère fiable, car elle est notamment sensible aux valeurs extrêmes, mais elle a le mérite d'être simple et rapide.

À retenir

Méthodologie :

1. On divise le nuage en 2 groupes de points de même effectif (ou l'un avec un point supplémentaire, si l'effectif est impair).
2. On calcule le point moyen de ces 2 groupes.
3. On relie ces 2 points moyens pour obtenir la droite d'ajustement.
4. On peut aussi, si besoin, connaissant les coordonnées de 2 points, déterminer l'équation de la droite.

Appliquons-la rapidement à notre exemple.

Exemple

1. Nous considérons les données par ordre croissant de la teneur en carbone, et scindons donc les points en deux groupes de 5 couples :

   Groupe 1 : (60 ; 70), (61 ; 72), (62, 71), (63, 74), (64, 77)
   Groupe 2 : (66 ; 79), (68 ; 81), (70, 86), (71, 86), (74, 93)

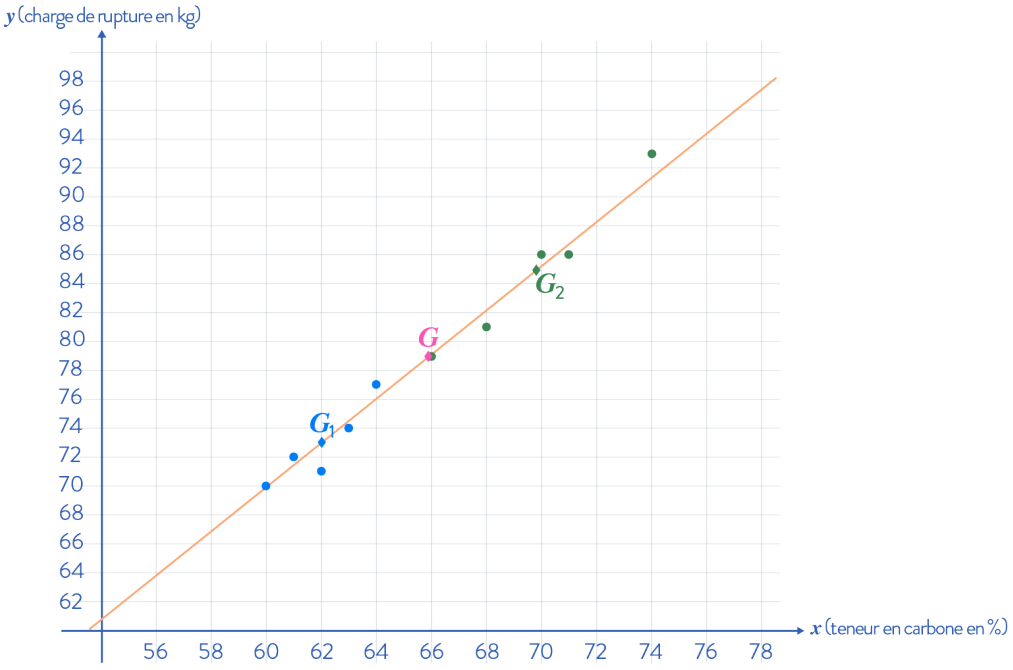
2. Nous calculons les coordonnées des points moyens G_1 et G_2, respectivement des groupes 1 et 2 :

$$\bar{x}_1 = \frac{60 + 61 + 62 + 63 + 64}{5}$$
$$= 62$$
$$\bar{y}_1 = \frac{70 + 72 + 71 + 74 + 77}{5}$$
$$= 72,8$$

$$\bar{x}_2 = \frac{66 + 68 + 70 + 71 + 74}{5}$$
$$= 69,8$$
$$\bar{y}_2 = \frac{79 + 81 + 86 + 86 + 93}{5}$$
$$= 85$$

➜ Nous obtenons donc G_1 de coordonnées $(62\ ;\ 72,8)$ et G_2 de coordonnées $(69,8\ ;\ 85)$.

3 Nous représentons ces points et les relions par une droite, qui sera donc notre droite d'ajustement.



Ajustement affine par la méthode de Mayer

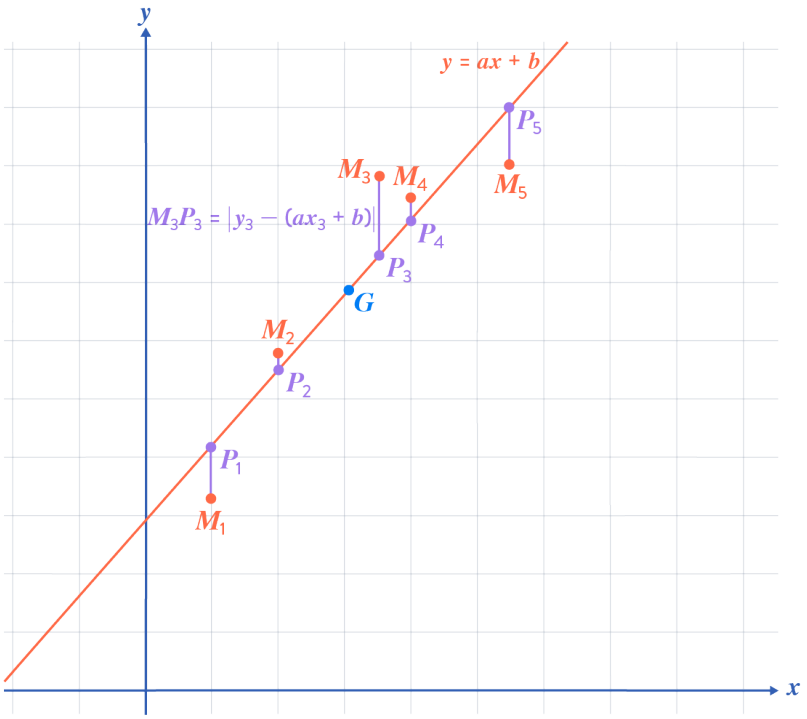
➜ Nous pouvons remarquer que le point moyen G appartient à cette droite.

C. Droite des moindres carrés

Nous allons maintenant aborder la méthode la plus utilisée pour effectuer un ajustement affine : la **méthode des moindres carrés**.

Pour bien la comprendre, considérons un nuage simple de points $M_i\,(x_i\,;\,y_i)$, représentons aussi une droite, qui passe par le point moyen G et dont l'équation est de la forme $y = ax + b$.

- Nous considérons en outre les points P_i de la droite, d'abscisse x_i et d'ordonnée $ax_i + b$.



Ajustement affine par la droite des moindres carrés

Ce qui nous intéresse, c'est la distance entre les points M_i et P_i associés.

- Dans le schéma ci-dessus, nous avons explicité cette distance entre les points M_3 et P_3. De la même façon, pour tout i compris entre 1 et 5, nous avons :

$$M_iP_i = |y_i - (ax_i + b)|$$

À retenir

Soit un nuage de n points, qui représente une série statistique à deux variables.
Déterminer la **droite des moindres carrés** consiste à trouver la droite qui minimise le carré des distances M_iP_i $(i \in \{1, \ldots, n\})$.

- Il s'agit donc de déterminer les réels a et b tels que la somme suivante soit minimale :

$$\sum_{i=1}^{n}(y_i-(ax_i+b))^2=(y_1-(ax_1+b))^2+\ldots+(y_n-(ax_n+b))^2$$

➜ Nous parlons aussi de **droite d'ajustement de y en x**.

Pour cela, nous allons admettre la propriété suivante.

Propriété

Soit une série statistique à deux variables :

- x, de moyenne $\bar{x}$, de variance $\text{var}(x)$ et d'écart-type $\sigma(x)$,
- et y, de moyenne $\bar{y}$, de variance $\text{var}(y)$ et d'écart-type $\sigma(y)$.

Soit $\text{cov}(x\ ;\ y)$ la covariance de x et y.
La droite des moindres carrés, ou droite d'ajustement de y en x, a pour équation $y=ax+b$ où :

$$\begin{aligned}a&=\frac{\text{cov}(x\ ;\ y)}{\text{var}(x)}\\&=\frac{\text{cov}(x\ ;\ y)}{\sigma^2(x)}\\b&=\bar{y}-a\bar{x}\end{aligned}$$

Résumons ce qui précède en donnant une méthodologie à suivre, lorsqu'un exercice demande de déterminer la droite d'ajustement par la méthode des moindres carrés.

À retenir

Méthodologie d'ajustement affine par la méthode des moindres carrés :

Soit une série statistique à deux variables x et y.

1. Si nécessaire, représenter le nuage de points $(x_i\ ;\ y_i)$ dans un repère orthogonal.
2. Calculer les moyennes $\bar{x}$ et $\bar{y}$ des deux variables.

➜ Placer le cas échéant le point moyen $G\,(\bar{x}\ ;\ \bar{y})$ dans la représentation.

3. Calculer les variances $\text{var}(x)$ et $\text{var}(y)$ des deux variables.
4. Calculer la covariance $\text{cov}(x\ ;\ y)$ des deux variables :

$$\text{cov}(x\ ;\ y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

5 Déduire l'équation de la droite d'ajustement de y en x :

$$y = \frac{\text{cov}(x\ ;\ y)}{\text{var}(x)} \cdot x + \bar{y} - a\bar{x}$$

6 Calculer le coefficient de corrélation :

$$r = \frac{\text{cov}(x\ ;\ y)}{\sqrt{\text{var}(x)\text{var}(y)}}$$

- En théorie, un ajustement affine est toujours possible, mais il est indispensable de mesurer sa pertinence ; le calcul du coefficient de corrélation est donc important pour pouvoir en juger.
- Plus r est proche en valeur absolue de 1, plus la corrélation linéaire est forte, et donc plus l'ajustement affine est pertinent.

7 Si l'ajustement s'avère suffisamment pertinent, alors on peut s'en servir pour effectuer :

- des *inter*polations (« *entre* » les données de la série),
- des *extra*polations (« *hors* » des données de la série).

Appliquons cette méthode, toujours à notre exemple.

Exemple

Nous avons déjà représenté le nuage de points (étape 1) et calculé les résultats pour les points 2, 3 et 4 :

Variable	Moyenne	Variance	Covariance
x	$\bar{x} = 65,9$	$\text{var}(x) = 19,89$	$\text{cov}(x\ ;\ y) = 31,89$
y	$\bar{y} = 78,9$	$\text{var}(y) = 52,09$	

5 Nous en déduisons les valeurs de a et b de l'équation réduite $y = ax + b$ de la droite des moindres carrés :

$$\begin{aligned} a &= \frac{\text{cov}(x\ ;\ y)}{\text{var}(x)} \\ &= \frac{31,89}{19,89} \\ &= \frac{1\,063}{663} \\ &\approx 1,603 \end{aligned}$$

$$\begin{aligned} b &= \bar{y} - a\bar{x} \\ &= 78,9 - \frac{1\,063}{663} \times 65,9 \\ &\approx -26,759 \end{aligned}$$

➜ Nous en déduisons l'équation de la droite d'ajustement, en arrondissant à 10^{-3} près :

$$y = 1,603\,x - 26,759$$

6 Nous avons aussi trouvé le coefficient de corrélation entre x et y :

$$r \approx 0,9907$$

➜ Nous l'avons déjà dit, mais reprécisons que l'ajustement affine est ici pertinent ; en outre, quand x grandit, y grandit aussi.

7 Nous allons maintenant nous servir de cette équation pour effectuer quelques prévisions.

Calculons d'abord la charge de rupture prévue par notre modèle pour un objet dont la teneur en carbone a été mesurée à $x' = 69,2\,\%$.

➜ Il s'agit d'une interpolation.

Il suffit de remplacer, dans l'équation, x par la valeur donnée (pour plus de rigueur, nous utilisons les expressions exactes, et non les arrondis) :

$$\begin{aligned} y' &= \frac{1\,063}{663} \times 69,2 + 78,9 - \frac{1\,063}{663} \times 65,9 \\ &\approx 84,191 \end{aligned}$$

➜ Le modèle prévoit une charge de rupture d'environ $84\ \text{kg}$.

Nous souhaitons maintenant que la charge de rupture de notre objet soit de $y'' = 100\ \text{kg}$. Quelle teneur en carbone doit-il avoir ?

➜ Nous voyons que $100\ \text{kg}$ est hors de notre domaine d'étude, il s'agit donc d'une extrapolation.

Nous remplaçons cette fois y par la valeur donnée :

$$\begin{aligned} 100 &= \frac{1\,063}{663} \times x'' + 78,9 - \frac{1\,063}{663} \times 65,9 \\ \Leftrightarrow x'' &= \frac{663}{1\,063} \times \left(21,1 + \frac{1\,063}{663} \times 65,9\right) \\ &= \frac{663}{1\,063} \times 21,1 + 65,9 \\ &\approx 79,060 \end{aligned}$$

➜ Avec une teneur de $80\ \%$, nous pouvons supposer que l'objet résistera à une charge de $100\ \text{kg}$.

Remarquons que, si nous avions posé la même question pour une charge de $200\ \text{kg}$, nous aurions trouvé une teneur en carbone d'environ $141\ \%$... Ce qui, en l'occurrence, serait un non-sens physique.

➜ Nous nous heurtons ici à une des limites de l'extrapolation.

Ci-dessus, nous avons effectué « manuellement » les calculs complets (à quelques points de suspension près), car il est important de bien comprendre le principe (et aussi parce que des exercices le demandent). À cet effet, nous avons travaillé avec un nombre restreint de données (effectif de 10).

En pratique, pour pouvoir effectuer des prévisions dignes de confiance et afin d'avoir un modèle mathématique d'ajustement le plus précis possible, nous devons disposer de beaucoup de données.
Bien sûr, dans de tels cas, calculer « manuellement » les moyennes, les variances, la covariance, etc., serait une tâche titanesque (et parfaite pour les erreurs de calcul...).

➜ Nous nous servons alors de notre calculatrice ou, mieux encore, d'un tableur pour déterminer directement les indicateurs et même l'équation de la droite des moindres carrés.

Astuce

Servons-nous de notre exemple pour montrer les fonctions concernant les statistiques à deux variables, et ce sur les tableurs les plus utilisés : Calc d'OpenOffice et Microsoft Excel.
Sur une feuille, nous avons au préalable entré les $(x_i\ ;\ y_i)$ sur deux colonnes (nous aurions aussi pu les mettre sur deux lignes) :

- les valeurs prises par x sont dans les cellules A1 à A10 ;
- les valeurs prises par y sont dans les cellules B1 à B10.

1. Pour calculer les principaux indicateurs :

Variable	x	y
Moyenne	MOYENNE(A1:A10)	MOYENNE(B1:B10)
Variance	VAR.P(A1:A10)	VAR.P(B1:B10)
Covariance	COVARIANCE(A1:A10 ; B1:B10)	
Coef. de corrélation r	COEFFICIENT.CORRELATION(A1:A10 ; B1:B10)	
Droite d'ajustement $y = ax + b$	a : PENTE(B1:B10 ; A1:A10)	
	b : ORDONNEE.ORIGINE(B1:B10 ; A1:A10)	

2. Pour représenter le nuage de points :
   - sélectionner les plages de données ;
   - insérer le graphique « Nuages de points » :
     - avec Calc : Insertion / Diagramme / XY (dispersion) / Points seuls,
     - avec Excel : Insérer / Graphique / XY (nuage de points).
3. Pour tracer la droite de régression, une fois le nuage de points réalisé :
   - avec Calc, le diagramme étant sélectionné (double-clic dessus, si nécessaire) : Insertion / Courbe de tendance / Linéaire, et cocher : Afficher l'équation ;
   - avec Excel : Clic droit sur un point du graphique / Ajouter une courbe de tendance / Linéaire, et cocher : Afficher l'équation sur le graphique.

3 Ajustement affine par changement de variable

Dans certains cas, les points du nuage ne peuvent être considérés comme alignés, car l'approximation serait de manière évidente beaucoup trop grande.

➜ Il ne faut pas pour autant conclure qu'il n'y a pas de corrélation entre les deux variables : il peut y avoir un lien, mais qui n'est pas linéaire.

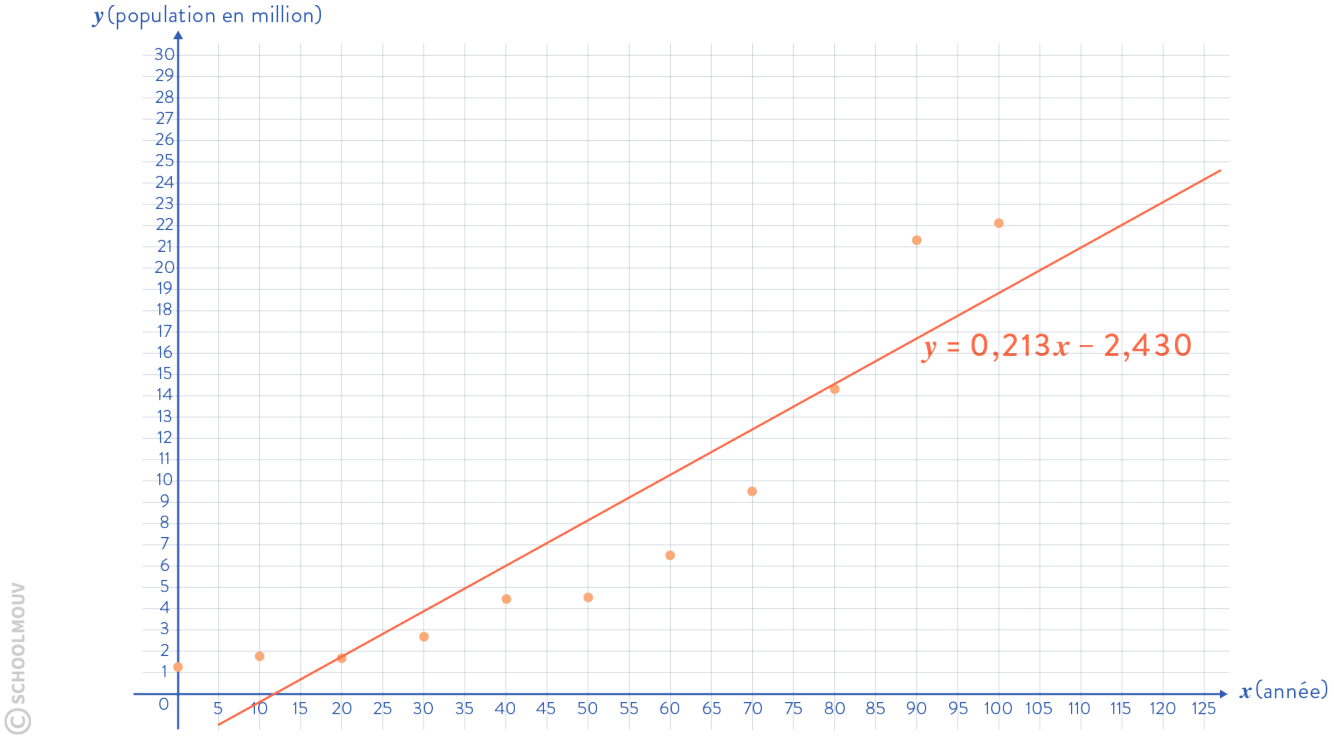
Parfois, il est toutefois possible d'étudier ce lien non linéaire au moyen d'un ajustement affine, et ce grâce à un **changement de variable**.
Dans cette dernière partie, nous allons donc montrer cette méthode à travers l'exemple de l'évolution d'une population, que vous avez déjà abordée, ou que vous aborderez bientôt, en enseignement scientifique (nous considérerons dans ce cours que le temps est une donnée continue, tandis que, dans le cours d'enseignement scientifique, nous travaillons par palier entier, et donc de manière discrète).

a. Un nuage de points non alignés

Nous disposons, pour une région, du recensement décennal de la population, sur tout le XX^{e} siècle, exprimé en million et arrondi à la dizaine de milliers (la précision est meilleure après les années 50) :

Année x_i	**Population** y_i (en million)
0	1, 210
10	1, 840
20	1, 810
30	2, 890
40	4, 430
50	4, 730
60	6, 542
70	9, 552
80	14, 264
90	21, 252
100	22, 035

2 Servons-nous d'un tableur pour représenter le nuage de points et tracer la droite d'ajustement de y en x.



Droite d'ajustement de y en x

3 Nous voyons que les points du nuage peuvent difficilement être considérés comme alignés. De plus, les distances entre les points et la droite semblent assez grandes.

➜ Donnons néanmoins le coefficient de corrélation, obtenu avec le tableur :

$$r \approx 0,923$$

Celui-ci n'est pas très éloigné de 1, mais nous ne pouvons pas dire non plus qu'il est « presque égal » à 1.

C Un ajustement affine par la méthode des moindres carrés ne semblent pas bien adapté.

Nous pouvons tout de même faire une première remarque.
Nous constatons une diminution de la population entre les années 10 et 20, et une stagnation entre les années 40 et 50.

➜ Des événements historiques peuvent expliquer ces points.

b. Changement de variable

Si les points ne sont pas alignés, nous reconnaissons quand même l'« allure » caractéristique de la courbe représentative de la fonction exponentielle.

➜ Nous allons donc considérer une nouvelle variable z définie, pour tout $i \in \{1, \ldots, 10\}$, par :

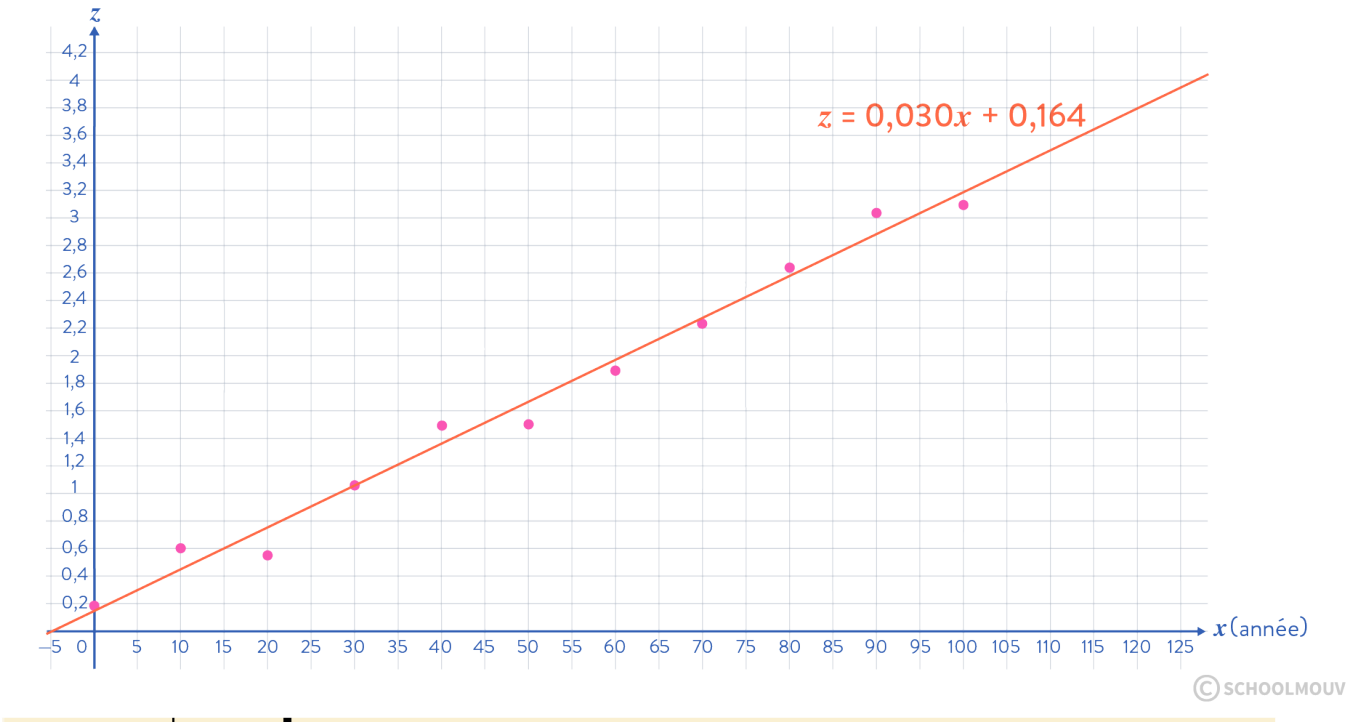
$$z_i = \ln(y_i)$$

Si l'évolution de la population suit effectivement une croissance exponentielle, nous pourrons alors mettre en évidence une corrélation linéaire entre x et z.
Notons que, dans cet exemple, nous nous intéressons aux effectifs d'une population, donc la variable y prendra des valeurs strictement positives, et nous pouvons donc travailler directement avec la fonction logarithme népérien.

1. Complétons notre tableau de données avec les valeurs de z, arrondies à 10^{-3} près :

Année x_i	**Population** y_i (en million)	$z_i = \ln(y_i)$
0	1, 210	0, 191
10	1, 840	0, 610
20	1, 810	0, 593
30	2, 890	1, 061
40	4, 430	1, 488
50	4, 730	1, 554
60	6, 542	1, 878
70	9, 552	2, 257
80	14, 264	2, 658
90	21, 252	3, 056
100	22, 035	3, 093

2. Représentons le nuage de points $(x_i \ ; \ z_i)$ et traçons la droite d'ajustement de z en x :



Droite d'ajustement de z en x

3 Nous nous rendons compte que le nuage a une allure bien plus « allongée » et que les points sont raisonnablement proches de la droite.

➜ Pour confirmer cette impression, donnons le coefficient de corrélation de $(x\ ;\ z)$:

$$r' \approx 0,993$$

Le coefficient est cette fois très proche de 1.

4 Nous décidons de modéliser l'évolution de la population par un ajustement réalisé grâce à ce changement de variable : $z = \ln(y)$.

L'équation de la droite d'ajustement de z en x est alors :

$$z = 0,030x + 0,164$$

Nous en déduisons :

$$\ln(y) = 0,030x + 0,164 \Leftrightarrow \mathrm{e}^{\ln(y)} = \mathrm{e}^{0,030x+0,164}$$

[par stricte croissance de exp]

$$\Leftrightarrow y = \mathrm{e}^{0,030x} \times \mathrm{e}^{0,164}$$

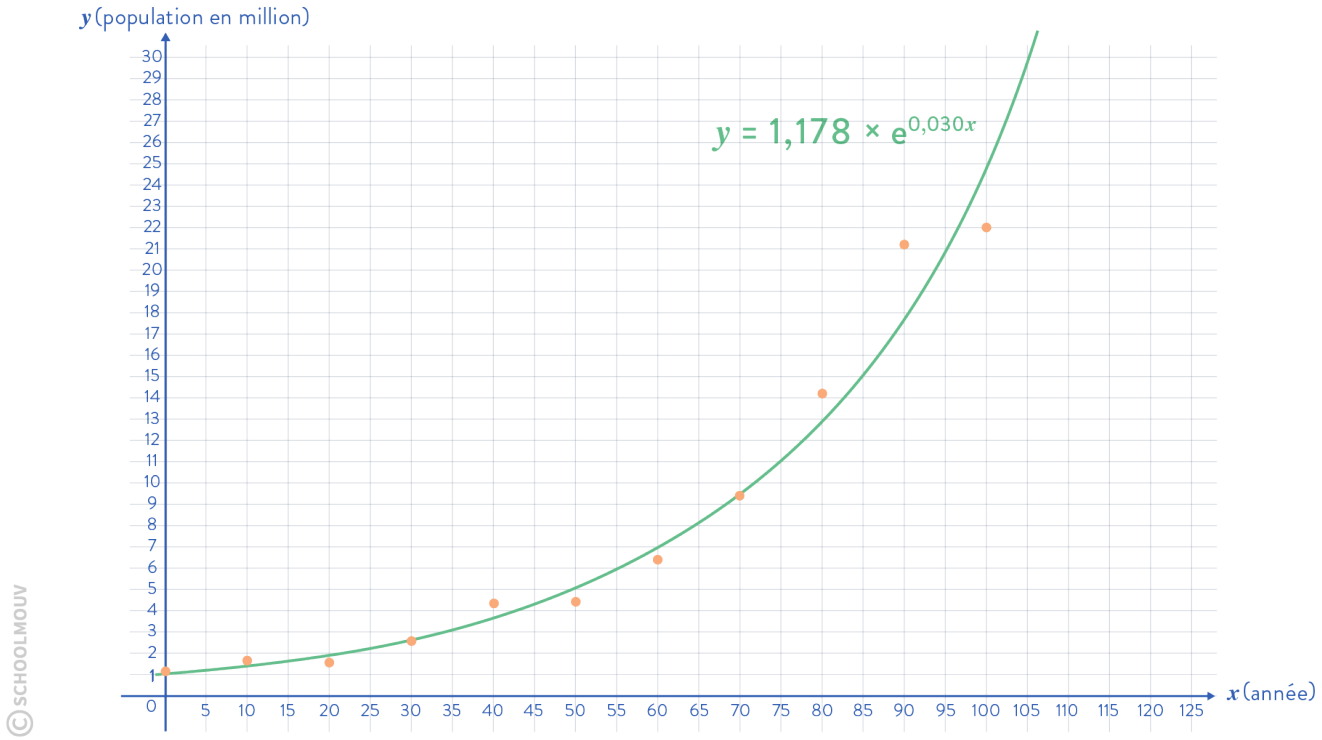
[car exp et ln sont réciproques et $\mathrm{e}^{a+b} = \mathrm{e}^{a} \times \mathrm{e}^{b}$]

En arrondissant à 10^{-3} près, nous obtenons finalement :

$$y = 1,178 \times e^{0,030x}$$

➜ La fonction $f : x \mapsto 1,178 \times e^{0,030x}$ permet d'ajuster le nuage de points $(x_i\ ;\ y_i)$.

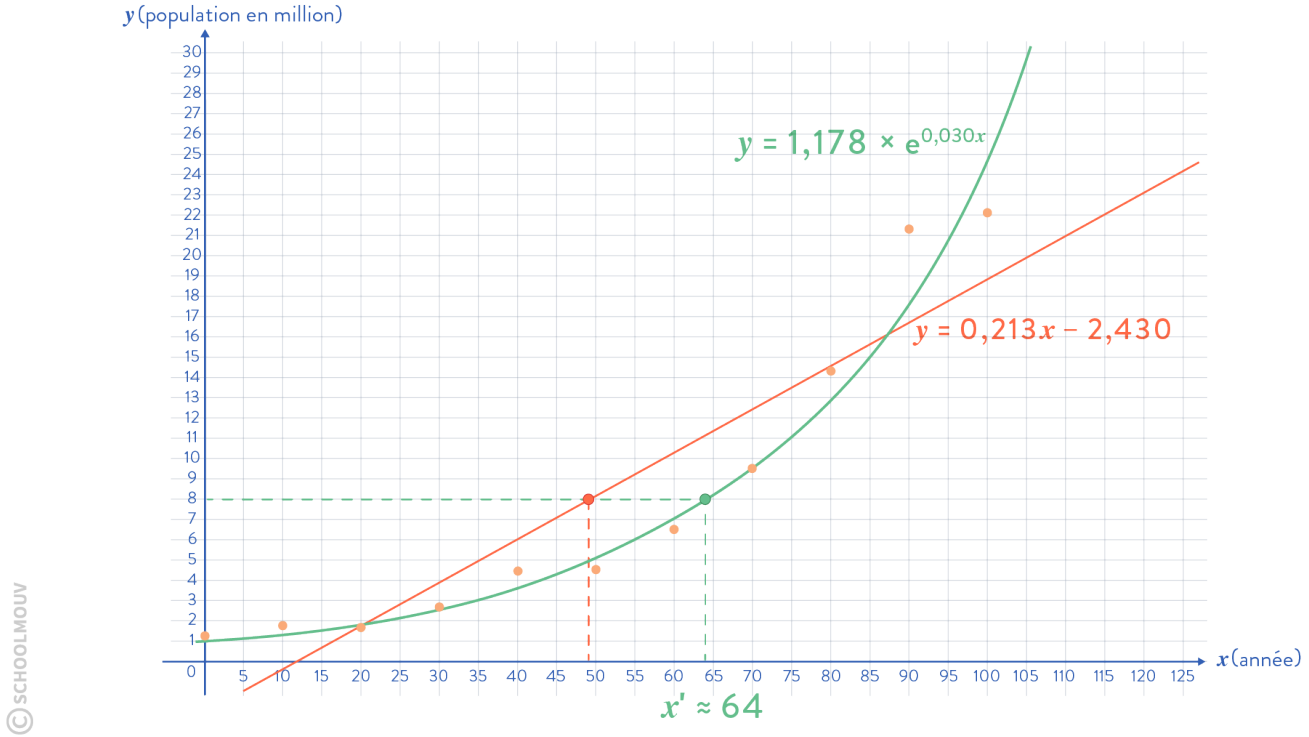
Traçons sa courbe représentative :



Courbe représentative de la fonction f

5 Servons-nous de ce modèle pour d'abord faire une interpolation.

Nous cherchons à savoir approximativement en quelle année x' la barre des 8 millions d'habitants a été franchie. Nous décidons de le faire graphiquement. Pour cela, nous déterminons graphiquement l'abscisse x' du point de la courbe représentative de f d'ordonnée $y' = 8$:



Courbe représentative de la fonction f et interpolation

Nous trouvons : $x' \approx 64$.

➜ La barre des 8 millions d'habitants a été franchie approximativement à l'année 64.

Remarquons que si nous avions utilisé la modélisation sans changement de variable, nous aurions trouvé environ 49 ans... ce qui aurait été en contradiction avec les données.

6 Effectuons enfin une extrapolation.

Nous nous demandons maintenant quelle sera la population y'' après 120 ans.
Nous nous servons bien sûr de la fonction dont nous disposons pour calculer $f(120)$:

$$\begin{aligned} y'' &= 1,178 \times e^{0,030 \times 120} \\ &\approx 43,113 \end{aligned}$$

➜ Si nous supposons que ce modèle reste valable hors du domaine des données, nous pouvons prévoir une population d'environ 43 millions de personnes.

Attention

Nous avons précisé que nous supposions le modèle valable hors du domaine de la série statistique, et ceci est toujours indispensable.

Dans notre exemple, nous constatons un infléchissement de la croissance à l'année 100.
Est-elle simplement conjoncturelle, comme les événements historiques à l'origine de la diminution et du ralentissement que nous avons déjà constatés ? Ou cet infléchissement a-t-il des raisons plus profondes et le modèle perd-il sa pertinence pour toute extrapolation ?
En effet, en enseignement scientifique, vous avez découvert, ou découvrirez, que la croissance d'une population peut être exponentielle durant des périodes brèves, mais qu'elle se heurte à la limite des ressources disponibles. Ainsi, une population aura tendance à tendre vers un maximum.

Ici, nous n'avons pas assez d'informations pour savoir ce qu'il en est précisément.

➜ Des données supplémentaires seront à relever au fil du temps, afin d'affiner le modèle.

C. Récapitulatif

Voici une méthodologie à suivre lorsqu'un changement de variable s'impose.

À retenir

Méthodologie :

- En fonction de l'allure du nuage de points, décider d'un changement de variable.
  - ➜ Généralement, les exercices vous guideront dans ce choix. Mais précisons qu'on fera appel surtout aux fonctions usuelles : logarithme, exponentielle, carré, racine carrée…
- Calculer les nouvelles valeurs déduites du changement de variable.
- Représenter le nouveau nuage de points et tracer la droite d'ajustement.
  - ➜ Calculer le coefficient de corrélation correspondant, afin de confirmer la pertinence du changement de variable.
- À partir de la définition de la nouvelle variable, en déduire la fonction d'ajustement des données initiales.
  - ➜ Représenter dans le nuage initial la courbe représentative de cette fonction, si l'on souhaite faire graphiquement des interpolations et des extrapolations.

➜ La définition de cette fonction permet aussi de faire, par le calcul, des interpolations et des extrapolations.

Notons aussi qu'une calculatrice ou un tableur sont aussi capables de déterminer très rapidement la fonction d'ajustement, selon le modèle que vous choisirez.

Astuce

Dans notre exemple de croissance exponentielle, une fois le nuage représenté dans un tableur :

- avec Calc, le diagramme étant sélectionné (double-clic dessus, si nécessaire) : Insertion / Courbe de tendance / Exponentielle, et cocher : Afficher l'équation ;
- avec Excel : Clic droit sur un point du graphique / Ajouter une courbe de tendance / Exponentielle, et cocher : Afficher l'équation sur le graphique.

Enfin, pour conclure ce cours, le long duquel nous avons parlé de lien ou de corrélation entre deux variables, il est indispensable d'ajouter un avertissement.

Attention

Il ne faut pas confondre corrélation entre deux variables et lien de cause à effet : mettre en évidence un lien ne suffit absolument pas à conclure que l'évolution d'une variable est la cause de l'évolution de l'autre. Une étude rigoureuse et la plus exhaustive possible est indispensable pour déterminer un lien de causalité.

Par exemple, il y aurait sans doute une corrélation entre le nombre de ventilateurs achetés et la quantité de glaces consommées, mais l'un n'est évidemment pas la cause de l'autre : la véritable cause est plutôt la température extérieure, qui influe sur les deux variables !
En revanche, ce sont des études approfondies qui montrent que les anomalies météorologiques, comme les canicules, sont sans doute dues au réchauffement climatique, lui-même causé par l'activité humaine...

Conclusion :

Dans ce cours, nous avons donc ajouté une dimension importante aux statistiques sur lesquelles nous avons travaillé jusqu'ici. En effet, mettre en évidence la corrélation entre deux variables est un aspect fondamental dans l'étude de données, même s'il n'est pas suffisant pour conclure à un lien de cause à effet.
Nous avons ici travaillé sur le lien entre deux variables, mais il peut y avoir bien sûr corrélation entre de multiples variables, compliquant la tâche. Il existe en statistique divers outils pour traiter de telles données, que certains découvriront durant leurs études supérieures.