Droite de régression

DroiteRégression01

En sciences, il faut parfois modéliser un phénomène en déterminant la relation entre deux variables. Les mathématiciens ont développé diverses méthodes de modélisation, comme la « droite de régression ».

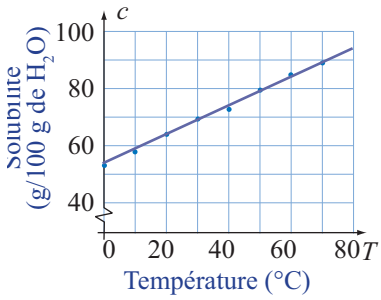
Droite de régression

La droite de régression est la droite qui représente le mieux la relation entre deux variables dont le nuage de points semble former une droite. C'est la droite pour laquelle la somme des carrés des distances aux points du nuage est minimale. Lorsque la variable est le temps, cette droite prend le nom de « droite de tendance ».

Le tableau ci-contre donne les mesures prises en laboratoire de la solubilité du bromure de potassium dans l'eau en fonction de la température de l'eau. La température T est donnée en degrés centigrades et la concentration c est donnée en grammes de soluté par cent grammes d'eau. Pour étudier plus à fond ce phénomène, il est intéressant de connaître l'équation de cette droite. Cette équation, de la forme $y = ax + b$, est également appelée « modèle affine ». Voyons la façon de déterminer les paramètres a et b de cette équation.

Solubilité du bromure de potassium

T_i (°C)	c_i (g/100 g de H_2O)
0	52,78
8	56,88
19	63,10
28	67,64
39	73,21
46	76,72
57	82,65
68	88,32



Soit un ensemble de couples $\{(x_1; y_1), (x_2; y_2), ..., (x_n; y_n)\}$ dont le nuage de points suggère un comportement affine. On souhaite décrire ces données par un modèle de la forme $y = ax + b$. Si tous les points étaient parfaitement alignés, on aurait le système suivant :

$$\begin{cases} y_1 = ax_1 + b \\ y_2 = ax_2 + b \\ \vdots \\ y_n = ax_n + b \end{cases}$$

En exprimant cette condition sous forme matricielle, on obtient :

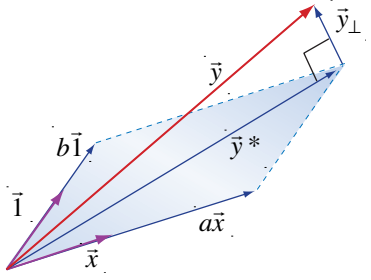
$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} ax_1 \\ ax_2 \\ \vdots \\ ax_n \end{pmatrix} + \begin{pmatrix} b \\ b \\ \vdots \\ b \end{pmatrix} \text{ ou } \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = a \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} + b \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Chacune des matrices colonnes de cette équation est un vecteur à n composantes et on peut écrire l'équation sous la forme :

$$\vec{y} = a\vec{x} + b\vec{1}$$

où $\vec{1}$ est le vecteur dont les n composantes sont égales à 1.

Cela signifie que si tous les points $(x_i; y_i)$ appartenaient à une même droite, le vecteur $\vec{y}$ serait combinaison linéaire des vecteurs $\vec{x}$ et $\vec{1}$ et ces trois vecteurs seraient coplanaires. Puisque les points $(x_i; y_i)$ ne sont pas tous sur une même droite, le vecteur $\vec{y}$ n'est pas dans le même plan que $\vec{x}$ et $\vec{1}$ et on a plutôt l'équation :



$$\vec{y}^* = a\vec{x} + b\vec{1}$$

où $\vec{y}^*$, $\vec{x}$ et $\vec{1}$ sont des vecteurs coplanaires. Parmi les vecteurs du plan déterminé par les vecteurs $\vec{x}$ et $\vec{1}$, le vecteur $\vec{y}^*$ est le plus proche du vecteur $\vec{y}$. Le vecteur $\vec{y}^*$ est le vecteur projection du vecteur $\vec{y}$ dans le plan auquel appartiennent les vecteurs $\vec{x}$ et $\vec{1}$. Il existe donc un vecteur $\vec{y}_\perp$ perpendiculaire à ce plan et formant un triangle rectangle avec les vecteurs $\vec{y}^*$ et $\vec{y}$. Ainsi, on a :

$$\vec{y} = \vec{y}^* + \vec{y}_\perp$$

Puisque $\vec{y}^* = a\vec{x} + b\vec{1}$, on peut écrire :

$$\vec{y} = a\vec{x} + b\vec{1} + \vec{y}_\perp$$

En effectuant le produit scalaire de chacun des membres de cette équation par le vecteur $\vec{1}$, on obtient :

$$\vec{1} \bullet \vec{y} = a\vec{1} \bullet \vec{x} + b\vec{1} \bullet \vec{1} + \vec{1} \bullet \vec{y}_\perp$$

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + bn + 0$$

$$\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + bn$$

REMARQUE

Le symbole de sommation permet d'écrire une somme de termes sous une forme condensée.

$$\sum_{i=1}^{n} y_i = y_1 + y_2 + \ldots + y_n$$

En isolant le paramètre b dans l'équation $\sum_{i=1}^{n} y_i = a\sum_{i=1}^{n} x_i + bn$, on obtient :

$$b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

Si on effectue le produit scalaire de chacun des membres de cette équation par le vecteur $\vec{x}$, on trouve :

$$\vec{x} \bullet \vec{y} = a\vec{x} \bullet \vec{x} + b\vec{x} \bullet \vec{1} + \vec{x} \bullet \vec{y}_\perp$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + 0$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i$$

En remplaçant b par $\frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$ dans l'équation qui précède, on a :

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i^2 + \left(\frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} x_i$$

On veut isoler le paramètre a dans cette équation. On multiplie d'abord les deux membres de l'équation par n, puis on applique la distributivité :

$$n\sum_{i=1}^{n} x_i y_i = na\sum_{i=1}^{n} x_i^2 + \left(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i\right)\sum_{i=1}^{n} x_i$$

$$n\sum_{i=1}^{n} x_i y_i = na\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - a\left(\sum_{i=1}^{n} x_i\right)^2$$

On regroupe les termes en a dans le membre de gauche :

$$na\sum_{i=1}^{n} x_i^2 - a\left(\sum_{i=1}^{n} x_i\right)^2 = n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

On met le paramètre a en évidence dans le membre de gauche et on l'isole par division.

$$a\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right) = n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Les paramètres de la droite qui est le meilleur modèle affine de l'ensemble des couples $\{(x_1; y_1), (x_2; y_2), ..., (x_n; y_n)\}$ sont donc :

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad \text{et} \quad b = \frac{\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i}{n}$$

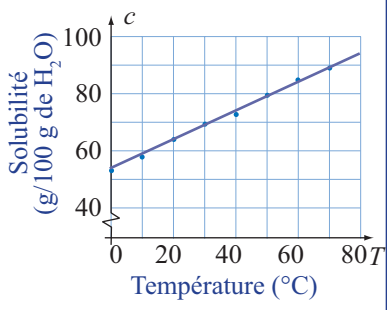
EXEMPLE 1

Le tableau ci-contre donne les mesures prises en laboratoire de la solubilité du bromure de potassium dans l'eau en fonction de la température de l'eau. La température T est donnée en degrés centigrades et la concentration c est donnée en grammes de soluté par cent grammes d'eau. On demande de construire un modèle décrivant la relation entre la température T et la solubilité c. Utiliser ce modèle pour estimer la solubilité à une température de 100°C.

DroiteRégression02

Solubilité du bromure de potassium

T_i (°C)	c_i (g/100 g de H_2O)
0	52,78
8	56,88
19	63,10
28	67,64
39	73,21
46	76,72
57	82,65
68	88,32



Solution

Pour déterminer les paramètres, il faut calculer les sommes des expressions donnant ces paramètres. C'est-à-dire la somme des températures, T_i, la somme des solubilités, c_i, la somme des carrés des valeurs de T_i et la somme des produits $T_i c_i$. Ces sommes sont effectuées dans le second tableau ci-contre.

T_i	c_i	T_i^2	$T_i c_i$
0	52,78	0	0,00
8	56,88	64	455,04
19	63,10	361	1 198,90
28	67,64	784	1 893,92
39	73,21	1 521	2 855,19
46	76,72	2 116	3 529,12
57	82,65	3 249	4 711,05
68	88,32	4 624	6 005,76
265	561,30	12 719	20 648,98

Dans cet exemple, on a pris 8 mesures, on a donc $n = 8$ et, en substituant les sommes dans les expressions permettant de calculer les paramètres, on obtient :

$$a = \frac{n\sum T_i c_i - \left(\sum T_i\right)\left(\sum c_i\right)}{n\sum T_i^2 - \left(\sum T_i\right)^2} = \frac{8\times 20\ 648{,}98 - 265\times 561{,}30}{8\times 12\ 719 - (265)^2}$$

$$= 0{,}521\ 690\ 614$$

$$b = \frac{\sum c_i - a\sum T_i}{n} = \frac{561{,}30 - 0{,}5216906...\times 265}{8} = 52{,}881\ 498...$$

En arrondissant a à cinq chiffres significatifs et b à deux décimales. on obtient le modèle affine :

$$c(T) = 0{,}521\ 69T + 52{,}88.$$

Pour estimer la solubilité à 100°C, on calcule l'image de 100 dans ce modèle, on obtient :

$$c(100) = 0{,}521\ 69 \times 100 + 52{,}88 = 105{,}049.$$

Puisque les mesure prises sont à quatre chiffres significatifs, on arrondit cette valeur à quatre chiffres significatifs et on estime la solubilité à 105,5 grammes par cent grammes d'eau.

PROCÉDURE

Calcul des paramètres d'une droite de régression

1. Représenter graphiquement les données afin de s'assurer que le modèle affine est approprié.
2. Pour simplifier le traitement et la gestion des données, construire un tableau en réservant une colonne à chacune des grandeurs n, x, y, xy et x^2. La dernière ligne du tableau contient les sommations utilisées dans les formules de a et de b.

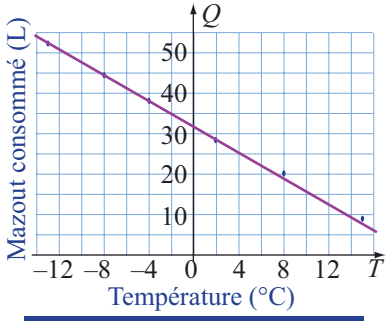
EXEMPLE 2

L'entrepreneur en construction pour lequel vous travaillez a décidé d'évaluer le coût de chauffage des maisons qu'il construit, car il veut inclure ce renseignement dans sa publicité. Il a fait relever, pour des périodes de 24 heures, la consommation moyenne de mazout en fonction de la température extérieure. Les relevés indiquent la température moyenne durant ces 24 heures. Le tableau ci-contre présente les données obtenues.

Trouver, par la méthode des moindres carrés, le modèle affine décrivant la relation entre la température et la quantité de mazout consommée.

Coût de chauffage

T_i	Q_i
–13	52,0
–8	44,0
–4	36,8
2	28,0
8	18,0
15	6,8



Solution

Identification des variables

La quantité de mazout consommé Q (L) dépend de la température extérieure T (°C). La représentation graphique des données est un nuage de points (présenté ci-contre) qui évoque une droite, même si les points ne sont pas parfaitement alignés.

Définition du lien entre les variables

Pour déterminer la valeur des paramètres de la droite, il faut calculer les produits des valeurs correspondantes et le carré des valeurs de la variable indépendante, puis faire la somme des données et de ces résultats. On peut présenter tous les calculs dans un même tableau, dont la dernière ligne est réservée aux sommes des valeurs inscrites dans les colonnes. En utilisant les formules des paramètres, on obtient :

$$a = \frac{n\sum T_iQ_i - \left(\sum T_i\right)\left(\sum Q_i\right)}{n\sum T_i^2 - \left(\sum T_i\right)^2} = \frac{6\times(-873{,}2) - 0\times 185{,}6}{6\times 542 - (0)^2} = -1{,}611\ldots$$

$$b = \frac{\sum Q_i - a\sum T_i}{n} = \frac{185{,}6 - (-1{,}611\ldots)\times 0}{6} = 30{,}93\ldots$$

Le modèle est donc $Q(T) = -1{,}611T + 30{,}93$.

Consommation de mazout

Valeurs observées			
T	Q	TQ	T^2
–13	52,0	–676,0	169
–8	44,0	–352,0	64
–4	36,8	–147,2	16
2	28,0	56,0	4
8	18,0	144,0	64
15	6,8	102,0	225
0	185,6	–873,2	542

Mesures de la précision du modèle

DroiteRégression03

Le modèle mathématique construit est-il fiable? Il existe des mesures qui permettent de répondre partiellement à cette question. Ce sont la somme des carrés des résidus, le coefficient de corrélation et le coefficient de détermination.

Calcul des résidus

On doit calculer pour chaque valeur de la variable indépendante, la différence entre la valeur observée et la valeur donnée par le modèle mathématique. De telles différences sont appelées « résidus » et la mesure de précision est la somme des carrés des résidus. On peut effectuer ce calcul de la somme des résidus à partir du tableau utilisé pour déterminer les paramètres du modèle affine. Ainsi, dans l'exemple 9.3.t, on obtient le tableau complémentaire suivant.

Consommation de mazout						
Valeurs observées				Valeurs du modèle	Résidus	Carrés des résidus
T_i	Q_i	T_iQ_i	T_i^2	$Q(T)$	R	R^2
–13	52,0	–676,0	169	51,877	0,123	0,015 069
–8	44,0	–352,0	64	43,822	0,178	0,031 722
–4	36,8	–147,2	16	37,378	–0,578	0,333 638
2	28,0	56,0	4	27,711	0,289	0,083 409
8	18,0	144,0	64	18,045	–0,045	0,002 005
15	6,8	102,0	225	6,767	0,033	0,001 070
0	185,6	–873,2	542			0,466 913

Coefficient de corrélation

Le **coefficient de corrélation** est une mesure de l'intensité du lien de linéarité entre deux variables. Il indique le degré de regroupement des points dans le voisinage de la droite. Il est défini par l'équation suivante :

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n\sum {x_i}^2 - \left(\sum x_i\right)^2}\ \sqrt{n\sum {y_i}^2 - \left(\sum y_i\right)^2}}$$

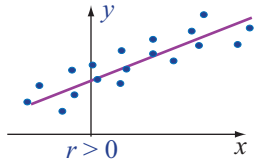
Ainsi, pour l'exemple 9.5.2, on a :

$$r = \frac{n\sum T_i Q_i - \left(\sum T_i\right)\left(\sum Q_i\right)}{\sqrt{n\sum {T_i}^2 - \left(\sum T_i\right)^2}\ \sqrt{n\sum {Q_i}^2 - \left(\sum Q_i\right)^2}}$$

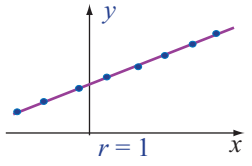
Le tableau de cet exemple donne quatre des sommes apparaissant dans la formule de r. Il manque seulement $\sum Q_i^2$. On peut donc facilement calculer le coefficient de corrélation :

$$r = \frac{6\times(-873{,}2) - 0\times 185{,}6}{\sqrt{6\times 542 - (0)^2}\ \sqrt{6\times 7\ 148{,}48 - (185{,}6)^2}} = -0{,}999\ 8$$

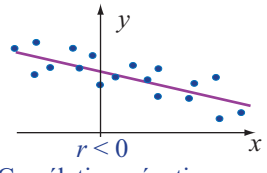
Le coefficient de corrélation linéaire r est un nombre compris entre –1 et 1 ($-1 \leq r \leq 1$). Lorsque $r = 0$ (corrélation nulle), le modèle affine n'est pas du tout approprié au phénomène. Lorsque r est proche de 1 ou de –1, le regroupement des points dans le voisinage de la droite est important. Si la valeur de r est positive, les variables varient dans un même sens, c'est-à-dire que la valeur de la variable dépendante augmente lorsque la valeur de la variable indépendante augmente. Si la valeur de r est négative, les valeurs des variables varient en sens inverse, c'est-à-dire que la valeur de la variable dépendante diminue lorsque la valeur de la variable indépendante augmente. L'exemple 9.3.7 illustre le dernier cas : la quantité de mazout



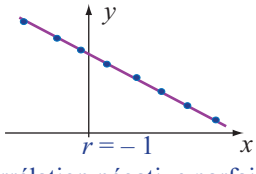
Corrélation positive



Corrélation positive parfaite



Corrélation négative



Corrélation négative parfaite

consommée diminue lorsque la température augmente. De plus, le coefficient r est –0,999 8, ce qui est très proche de –1. La corrélation est donc très forte.

Le coefficient de détermination est le carré du coefficient de corrélation. Il est une mesure de la pertinence d'utiliser un modèle affine en faisant abstraction du fait que la corrélation peut être positive ou négative. C'est une mesure de l'adéquation entre le modèle et les données observées.

DroiteRégression04

Droite de tendance

La droite de régression permet de construire un modèle simple, utilisé pour analyser des phénomènes ou décrire une tendance. On l'appelle alors « droite de tendance » lorsque la variable indépendante est le temps. On distingue deux cas dans l'analyse de tendance, selon que les valeurs estimées sont à l'intérieur ou à l'extérieur de l'ensemble des données observées : l'interpolation et l'extrapolation.

Lorsque les prévisions portent sur des valeurs à l'intérieur de l'intervalle des données, on procède à une **interpolation**. Généralement, les estimations par interpolation sont plutôt fiables.

Si les prévisions portent sur des valeurs à l'extérieur de l'ensemble des données, on procède à une **extrapolation**. Il est à noter que la fiabilité est plus grande lorsqu'on fait des prédictions pour des valeurs proches de l'ensemble des données observées. Une prédiction portant sur une valeur éloignée de cet intervalle donne une estimation qui, sans être à rejeter, doit être utilisée avec circonspection. Dans les deux cas, il ne faut pas s'attendre à ce que le modèle soit plus précis que les données qu'il décrit.

Laboratoire

Le laboratoire suggéré vous permet de programmer une feuille de calcul pour représenter graphiquement des données expérimentales, faire calculer les paramètres d'une droite de régression ainsi que le coefficient de corrélation.

La représentation graphique et le coefficient de corrélation vous permettront de juger de la fiabilité du modèle.

Feuille de route O8Moindres-carres-Affine

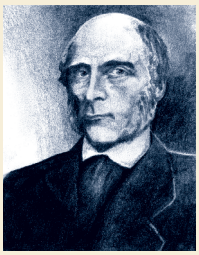
Vidéo : 08Moindres-carres

Note historique

FRANCIS GALTON

1822-1911

Sir Francis Galton était un homme de science britannique. Il a été anthropologue, explorateur, géographe, inventeur, météorologue, proto-généticien, psychométricien et statisticien. Il est considéré comme le fondateur de la psychologie différentielle ou comparée. C'est lui qui a mis en place de façon systématique la méthode d'identification des individus par leurs empreintes digitales. Il fut anobli en 1909 et a reçu la médaille Copley, décernée par la Royal Society.



Influencés par les travaux de Charles Darwin (1809-1882), les statisticiens anglais de la fin du XIX^{e} siècle ont utilisé les statistiques dans des contextes plus proches de la biologie que de la sociologie, comme le faisaient les statisticiens du continent européen. Francis Galton, cousin de Darwin, s'est penché sur des questions statistiques liées à la génétique, l'hérédité et le comportement humain. Alors qu'Adolphe Quételet (1796-1874) avait réalisé des travaux sur des données biométriques de l'homme, comme le poids, la taille et le périmètre thoracique, et avait montré que ces données se répartissaient selon une courbe normale, Galton a mené des recherches sur la variabilité des caractères, les différences entre les individus et les moyens de conserver et de favoriser les meilleurs d'entre eux. Sa contribution majeure est la notion de corrélation et la mesure de celle-ci, le coefficient de corrélation.

Lors d'études sur l'hérédité, réalisées en 1877, Galton se rendit compte que des parents de petite taille avaient des enfants plus petits que la moyenne, mais plus grands que leurs parents. De même, des parents plus grands que la moyenne avaient des enfants plus grands que la moyenne, mais plus petits que leurs parents. Ce phénomène indique qu'il y a corrélation entre la taille des parents et celle des enfants, mais qu'il y a également une régression par rapport à la moyenne, d'où l'appellation « droite de régression ». La régression vers la moyenne est inversement proportionnelle à la corrélation. Dans ses travaux sur l'eugénisme, Galton a étudié la dispersion des résultats et a élaboré les notions de médiane et de quartile. À l'époque, les travaux de Galton étaient perçus comme une contribution importante dans la lutte de la science contre l'obscurantisme religieux. Malheureusement, certains les ont utilisés comme justification pour les exactions commises dans l'Allemagne nazie.

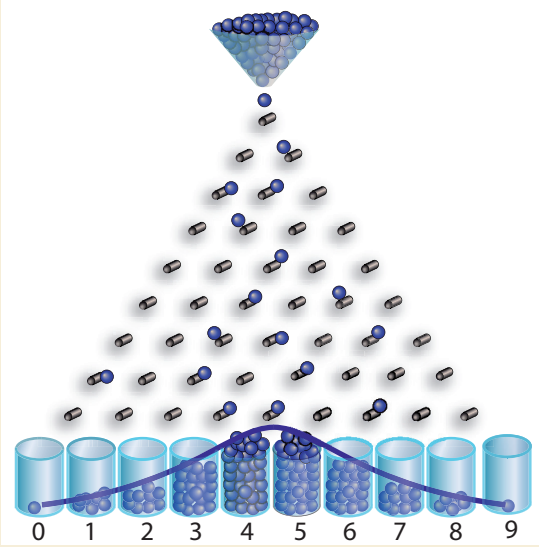
À partir de 1865, Galton s'est consacré à la statistique dans le but de quantifier les caractéristiques physiques, psychiques et comportementales de l'être humain, ainsi que leur évolution.

Darwin avait énoncé ses lois de l'évolution sans tenir compte du calcul des probabilités, mais ses théories ont assuré le triomphe d'une description probabiliste du monde. Galton a fait le lien entre la théorie de la sélection naturelle et la recherche mathématique, consacrant une large partie de son activité à la défense de la théorie de l'évolution et à montrer qu'elle permet de faire des prédictions susceptibles d'être vérifiées.

Planche de Galton

L'illustration ci-dessous représente une planche de Galton. Cette planche comporte neuf rangées de clous. On incline un peu la planche et on laisse tomber des billes sur le premier clou. Chaque bille suit une trajectoire qui la conduit dans l'un des récipients au bas de la planche.

Lorsqu'une bille frappe un clou, la probabilité qu'elle tombe à gauche de celui-ci est 1/2 et la probabilité qu'elle tombe à droite est également 1/2. Le calcul des probabilités permet de déterminer combien, en moyenne, il y aura de billes dans chacun des récipients. Le nombre de billes dans les récipients suit une loi binomiale de probabilités.



Exercices

1. Au cours d'une expérience sur la polarimétrie du sucrose, on a noté l'angle de rotation selon la concentration. Les données recueillies sont présentées dans le tableau ci-contre, où C est la concentration en g/100 ml et A est l'angle de rotation en degrés.

Polarimétrie du sucrose	
C_i (g/100 ml)	A_i (degrés)
4,0	5,64
6,5	9,10
8,0	11,36
12,0	16,80
15,5	22,01
16,5	22,94

   a) Construire un modèle mathématique de cette situation.
   b) Utiliser ce modèle pour calculer l'angle de rotation d'une concentration de 15,00 g/100 ml).
   c) Déterminer la concentration pour laquelle l'angle de rotation est de 10°, de 20°.

2. Le tableau suivant donne les mesures prises en laboratoire de la solubilité du bromure de potassium dans l'eau en fonction de la température de l'eau. La température T est donnée en degrés centigrades et la concentration c est donnée en grammes de soluté par cent grammes d'eau.

Solubilité du bromure de potassium			
T_i (°C)	c_i (g/100 g de H_20)	T_i (°C)	c_i (g/100 g de H_20)
0	52,78	39	73,21
8	56,88	46	76,72
19	63,10	57	82,65
28	67,64	68	88,32

   a) Représenter graphiquement les données pour vérifier s'il est plausible de décrire la relation entre les variables par l'équation d'une droite.
   b) Calculer les paramètres de la droite de régression.
   c) Calculer le coefficient de corrélation. Interpréter le résultat.

3. On réalise l'expérience suivante sur les échanges de chaleur. On plonge 25,0 g d'un alliage dans un bécher contenant 90,0 g d'eau à 25,82 °C. La température finale T_f (lorsque les températures ont atteint leur point d'équilibre) est fonction de la température T_a de l'alliage au moment où on le plonge dans l'eau. Les températures en Celsius, mesurées au cours de divers essais sont données dans le tableau suivant.

Échange de chaleur	
Température de l'alliage T_a (°C)	Température finale T_f (°C)
120	27,7
110	27,4
100	27,2
90	26,9
80	26,7
70	26,4

   a) Représenter graphiquement les données pour vérifier s'il est plausible de décrire la relation entre les variables par l'équation d'une droite.
   b) Calculer les paramètres de la droite de régression.
   c) Calculer le coefficient de corrélation. Interpréter le résultat.

4. Une association d'automobilistes a demandé à ses membres de noter la distance qu'ils ont parcourue et le coût d'utilisation de leur véhicule au cours de la dernière année, y compris les coûts de l'immatriculation, des assurances, de l'essence et de l'entretien. L'association a dressé le tableau suivant à l'aide des informations reçues pour la voiture la plus populaire auprès de ses membres.

Coût d'utilisation selon la distance parcourue			
Distance (km)	Coût ($)	Distance (km)	Coût ($)
5 000	3 950	20 000	6 600
10 000	4 860	25 000	7 520
15 000	5 740	30 000	8 460

   a) Construire un modèle mathématique décrivant la correspondance entre les deux variables.
   b) Donner une mesure de la précision du modèle en calculant les résidus.
   c) Prévoir, à l'aide du modèle, le coût d'utilisation de la voiture en question dans le cas où la distance parcourue en une année est de 45 000 km.

5. Dans une expérience sur la cinétique du lactose, on a mesuré l'absorbance pour différentes concentrations. Les données obtenues sont les suivantes :

Concentration et absorbance	
Concentration (mol/L)	Absorbance à 400 mn
0,003	0,75
0,005	0,61
0,008	0,45
0,009	0,30
0,015	0,16

a) Quelle est la variable indépendante et quelle est la variable dépendante de cette situation?
b) Calculer les paramètres de la droite de régression.
c) Calculer le coefficient de corrélation. Interpréter le résultat.

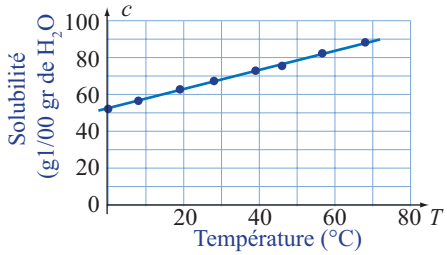
6. On a réalisé une expérience en mesurant le volume d'une mole de NH_3 à une température de 0°C en faisant varier la pression. Le but de l'expérience est d'étudier l'importance des déviations par rapport à la loi de Boyle-Mariotte. Les mesures obtenues sont compilées dans le tableau suivant :

Pression et volume	
Pression P (atm)	Volume V (L)
0,10	223,88
0,25	89,36
0,40	55,73
0,55	40,44
0,70	31,71
0,85	26,06
1,00	22,10

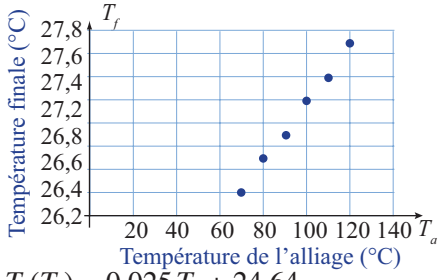
a) On veut déterminer s'il existe un lien affine entre P et PV. Établir le tableau des correspondances de ces deux variables.
b) Définir par la méthode des moindres carrés le meilleur modèle affine décrivant ces données.
c) La **constante idéale** du gaz pour la loi de Boyle-Mariotte est la valeur extrapolée obtenue en prolongeant la droite jusqu'à une pression nulle. Calculer cette constante.
d) Calculer le coefficient de corrélation. Quelle conclusion peut-on tirer de ce résultat?

Réponses

1. a) $A(C) = -1{,}400C + 0{,}062$
   b) 21,06°
   c) 7,1 g/100 mL et 14,2 g/100 mL

2. a)


Le modèle est $c(T) = 0{,}522T + 52{,}88$.
   c) Le coefficient de corrélation est 0,999 913 ..., ce qui signifie que le modèle affine est un excellent modèle pour décrire les données obtenues. puisque le coefficient est très près de 1.

3. a)


   b) $T_f(T_a) = 0{,}025\,T_a + 24{,}64$.
   c) Le coefficient de corrélation est de 0,998 488 474, la corrélation est positive et très forte.

4. a) $C(D) = 0{,}179\,4D + 3{,}04$
   b) Somme des carrés des résidus, 0,002 711. Le coefficient de corrélation est 0,999, ce qui indique une très forte corrélation.
   c) On peut douter de la fiabilité de ce résultat puisque la valeur 45 000 km est éloignée de l'intervalle des données, on obtient tout de même une estimation du coût.

5. a) La variable indépendante est la concentration C et la variable dépendante l'absorbance en millinewtons.
   b) Le modèle est $A = -49{,}52C + 0{,}85$.
   c) Le coefficient de corrélation est de –0,963 074 436, ce qui indique une corrélation négative et très forte.

6. b) Le modèle est $PV = -0{,}318P + 22{,}42$.
   c) La constante idéale est l'image de 0 par la relation. Compte tenu de la précision des données, on a 22,42 L·atm.
   d) Le coefficient de corrélation est $r = -0{,}999\,908\,548$. Le choix d'un modèle affine est très justifié.