Corrélation et causalité

Table des Matières

∽ Corrélation et causalité ∽

I. Introduction

☞Activité 1

Le tableau suivant donne l'implication des citoyens dans les sciences participatives liées à la biodiversité entre 2011 et 2017 :

année	2011	2012	2013	2014	2015	2016	2017
participants	21 143	30 363	35 209	30 940	35 403	45 251	53 732

biodiversité - les chiffres clés - 2018 - ministère du développement durable

1. Faire un graphique de la situation, l'origine du repère aura pour coordonnée (2010 ; 0).
   Échelle :
   - en abscisse 1cm pour une unité,
   - en ordonnée 1cm pour 10 000 participants.
2. Quel nombre de participants peut-on prévoir en 2025 ? Expliquer la démarche et votre modèle.

Pour tout le cours, on considère une série statistique à deux variables discrètes x et y prenant n valeurs :

x	x_1	x_2	...	x_i	...	x_n
y	y_1	y_2	...	y_i	...	y_n

II. Nuage de points et point moyen

Définition

Soient une série statistique à deux variables discrètes, x et y. On appelle nuage de point la représentation des points $M(x_i\,;\,y_i)$ dans un repère $\left(O\,;\,\vec{i}\,,\,\vec{j}\right)$.

Définition

Le point moyen G du nuage est le point de coordonnée $G\left(\overline{x}\,;\,\overline{y}\right)$.
$\overline{x}$ est la moyenne des valeurs de la variable x et $\overline{y}$ est la moyenne des valeurs de la variable y.

III. Approximation d'un nuage de points

Si le nuage de points est assez allongé, on cherche à approcher les points du nuage par une droite δ qui devra être la plus proche possible des points du nuage.

☞**Activité 2**

Soit la série statistique de variables discrètes x et y :

x	1	2	3	4	5	6	7
y	21	30	35	31	35	45	54

1. Faire le nuage de points dans un repère orthogonal $\left(O\,;\,\vec{i}\,,\,\vec{j}\right)$.
   Échelle :
   - en abscisse 1cm pour une unité,
   - en ordonnée 1cm pour 5 unités et l'axe des ordonnées comme à l'unité 20.
2. On approche le nuage de points par deux droites d_1 et d_2 dont on donne les équations :
   - $d_1 : y = 4,6x + 17,4$
   - $d_2 : y = 4,5x + 18,0$
3. Tracer les droites sur le graphique précédent.
4. (a) Compléter le tableau des erreurs absolues au carré en ordonnées pour chacune des droites :

x	1	2	3	4	5	6	7	
y	21	30	35	31	35	45	54	
y_{d_1}								
y_{d_2}								sommes
$\left(y_{d_1} - y\right)^2$								
$\left(y_{d_2} - y\right)^2$								

   (b) Quelle droite approxime le mieux le nuage ?

☞**Activité 3**

La méthode des moindres carrés

$\left(O\,;\,\vec{i}\,,\,\vec{j}\right)$ est un repère orthonormé du plan.
Soit un nuage de n points $M_i\left(x_i\,;\,y_i\right)_{1\leqslant i\leqslant n}$, G le point moyen de coordonnées $\left(\overline{x}\,;\,\overline{y}\right)$ et une droite δ d'équation réduite $y = mx + p$.
Le but de l'activité est de déterminer la droite qui minimise la l'erreur des carrés des écarts en ordonnées soit déterminer m et p tels que la somme $\sum\limits_{i=1}^{n}\left(y_i - \left(mx_i + p\right)\right)^2$ soit la plus petite possible.
Soient les fonctions f et g :

$$\begin{array}{rccl} f: & \mathbb{R} & \to & \mathbb{R} \\ & p & \mapsto & f(p) = \sum\limits_{i=1}^{n}\left(y_i - \left(mx_i + p\right)\right)^2 \end{array}$$

$$\begin{array}{rccl} g: & \mathbb{R} & \to & \mathbb{R} \\ & m & \mapsto & g(m) = \sum\limits_{i=1}^{n}\left(y_i - \left(mx_i + p\right)\right)^2 \end{array}$$

1. Propriété de la fonction f :
   (a) Donner la nature de la fonction f ?
   (b) Justifier que le minimum de f est atteint pour $f'(p) = 0$ soit $\sum\limits_{i=1}^{n}\left(y_i - \left(mx_i + p\right)\right) = 0$.
   (c) En déduire que les coordonnées de G vérifient $y = mx + p$.
   Ainsi G appartient à la droite δ et $p = \overline{y} - m\overline{x}$.

2. Propriété de la fonction g :
   (a) Montrer que

   $$g(m) = \sum_{i=1}^{n}\left(y_i - \left(mx_i + p\right)\right)^2 = \sum_{i=1}^{n}\left(\left(y_i - \overline{y}\right) - m\left(x_i - \overline{x}\right)\right)^2 = m^2\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 - 2m\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) + \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

   (b) Donner la nature de la fonction g.
   (c) Montrer que la minimum de g est atteint en $m = \dfrac{\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$

   Remarque : $\dfrac{\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)^2} = \dfrac{\dfrac{\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{n}}{\dfrac{\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n}} = \dfrac{\sigma_{xy}}{\sigma_x^2}$.

   σ_x^2 est la variance de la variable x.

Définition

Soit une série statistique à deux variables x et y.
On appelle covariance de x et y, notée σ_{xy} le nombre :

$$\sigma_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n}$$

Définition

Soit une série statistique à deux variables x et y, on note $\overline{x}$ et $\overline{y}$ la moyenne respective de la variable x et de la variable y.
$\left(O\,;\,\vec{i}\,,\,\vec{j}\right)$ est un repère orthogonal du plan. On appelle droite de régression linéaire par la méthode des moindres carrés y en x, la droite qui minimise l'erreur absolue en ordonnée au carré, cette droite a pour équation $y = mx + p$ avec :

- $m = \dfrac{\sigma_{xy}}{\sigma_x^2}$
- $p = \overline{y} - m\overline{x}$

Remarque

La droite δ déterminée par la méthode des moindres carrés passe par le point moyen G du nuage.

Exercice 1

On reprend l'activité 1 :
Le tableau suivant donne l'implication des citoyens dans les sciences participatives liées à la biodiversité entre 2011 et 2017 :

année	2011	2012	2013	2014	2015	2016	2017
participants	21 143	30 363	35 209	30 940	35 403	45 251	53 732

1. Déterminer la droite de régression linéaire par la méthode des moindres carrés.
   Vérification par la calculatrice.
2. Avec ce modèle, déterminer l'implication des citoyens dans les sciences participatives liées à la biodiversité en 2030.

Exercice 2

Soit la série statistique à deux variables x et y qui donne l'évolution du nombre d'habitants de la population française (en milliers) suivant le temps :

Année x	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Population y	62 765	63 070	63 375	63 697	63 697	64 300	64 468	64 639	64 737	64 821	64 124

Source : Insée

1. À l'aide de la calculatrice, déterminer la droite de régression linéaire par la méthode des moindres carrés.
2. Avec ce modèle, déterminer la population française en 2030.