06-05 Corrélation et ajustements

Définition

Le **coefficient de corrélation linéaire** de deux séries de valeurs (x_1 ; x_2 ; … ; x_n) et (y_1 ; y_2 ; … ; y_n) est le nombre r défini par $r = \dfrac{\mathrm{cov}(x\,;y)}{\sigma(x)\sigma(y)}$

Exemple

x_i	161	141	127	150	159	129	148	135
y_i	21,5	23,3	17,2	15,6	18,8	10,1	14,1	15,7

……

……

Remarques

- Le coefficient de corrélation linéaire est un nombre compris entre (-1) et 1.
  Lorsqu'il est proche de (-1) ou 1, on dit que la corrélation est forte entre x et y.
- Effectuer un ajustement, c'est faire une **interpolation**.
  Utiliser un ajustement pour un calcul, c'est faire une **extrapolation**.
- Il est parfois possible de réaliser des ajustements non affines de nuages dont les points rappellent des courbes connues. On procède pour cela à un changement de variable.

Exemple

En utilisant l' ……………………… effectuée dans le cours 06-04, on peut émettre l' ……………………… suivante:

« La température moyenne d'une ville ensoleillée un jour sur deux sera aux alentours de ………….. ».

06-05 Exercices

Exercice 1

1. Calculer le coefficient de corrélation linéaire de la série à deux variables ci-dessous.

x_i	2	3,2	4,3	5	8,5	9,7	12	14
y_i	13,526	7,673	4,773	3,497	3,323	3,109	3,004	2,755

2. On pose $z = \dfrac{1}{y}$

a] Compléter le tableau ci-dessous.

x_i	2	3,2	4,3	5	8,5	9,7	12	14
z_i								

b] Calculer le coefficient de corrélation linéaire de la série (x ; z).

c] Déterminer l'équation de la droite des moindres carrés de cette série.

3. **a]** Déduire des questions précédentes une interpolation de y en fonction de x.

b] Calculer une extrapolation au dixième de la valeur y correspondant à x = 7.

Exercice 2

Épidémie de grippe
Calculer, modéliser, communiquer

Notions revues : Logarithme népérien (chap. 4), Exponentielle (1re)

Pour étudier la propagation d'une épidémie de grippe, une enquête est faite auprès d'un échantillon de 1000 personnes. Le tableau ci-dessous donne les valeurs N_i de la variable N représentant le nombre d'individus ayant été contaminés à la date t_i du mois de janvier (la variable t est exprimée en jours).

t_i	1	2	5	10	15	20
N_i	89	171	304	418	488	499

1. Tracer le nuage de points de cette série statistique, dans un repère orthogonal. On prendra comme unité 0,5 cm pour 1 jour sur l'axe des abscisses et 1 cm pour 50 individus sur l'axe des ordonnées.

2. Le nuage de points permet-il d'envisager un ajustement affine ?

3. Déterminer une valeur approchée à 10^{-2} près du coefficient de corrélation linéaire r de cette série statistique à deux variables. Interpréter le résultat en termes de corrélation. Peut-on intuitivement envisager un lien de causalité entre la date et le nombre d'individus contaminés ?

4. On effectue le changement de variable suivant :

$$y = -5\ln\left(1 - \frac{N}{500}\right).$$

a. Recopier et compléter le tableau suivant. Arrondir les valeurs au millième.

t_i	1	2	5	10	15	20
N_i	89	171	304	418	488	499
y_i						

b. Déterminer, avec la calculatrice, l'équation de la droite de régression de y en t. Arrondir les coefficients à 10^{-4} près. Dans la suite de l'exercice, on considère que $y = 1{,}508t - 2{,}236$.

c. Montrer que la variable N s'exprime en fonction de y par la relation : $N = 500\left(1 - e^{-\frac{y}{5}}\right)$.

d. En déduire une expression de N en fonction de t.

5. En utilisant l'expression précédente comme modèle du nombre de personnes contaminées, déterminer la date à laquelle le nombre d'individus contaminés dépasse pour la première fois le quart de la population étudiée.