Chapitre 6

Statistiques à deux variables quantitatives

1 Nuage de points et ajustement affine

Série statistique à deux variables quantitatives

Sur une population, on étudie deux caractères quantitatifs x et y (par exemple l'ancienneté et la prime d'un employé). Pour chacun des n individus de cette population, on note x_i et y_i les valeurs prises par chacun de ces deux caractères et on présente les données à l'aide de la **série statistique à deux variables** ci dessous :

valeur x_i	x_1	x_2	...	x_n
Valeur y_i	y_1	y_2	...	y_n

Définitions

- Dans un repère, le **nuage de points** associé à cette série statistique est l'ensemble des points $M_i\,(x_i\ ;\ y_i)$ pour $i = 1, 2, ..., n$.
- Le **point moyen** de cette série statistique est le point $M\,(\overline{x}\ ;\ \overline{y})$ où $\overline{x} = \dfrac{x_1 + x_2 + ... + x_n}{n}$ et $\overline{y} = \dfrac{y_1 + y_2 + ... + y_n}{n}$.

Ajustement d'un nuage de points

On se demande s'il existe une dépendance entre les deux caractères x et y.

Figure 1

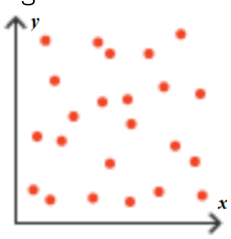


Figure 2

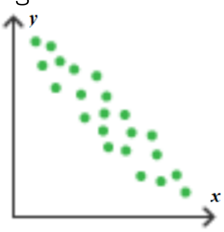
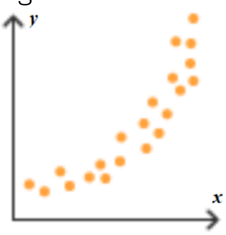


Figure 3



Selon la forme du nuage associé, on peut être incité à ne pas voir de dépendance (figure 1), ou bien à penser que les points se répartissent autour d'une droite (figure 2), ou bien autour d'un autre type de courbe (figure 3).

Définition

Pratiquer un **ajustement affine** d'un nuage de point consiste à tracer une droite qui passe le plus près possible des points du nuage.

On se demande :

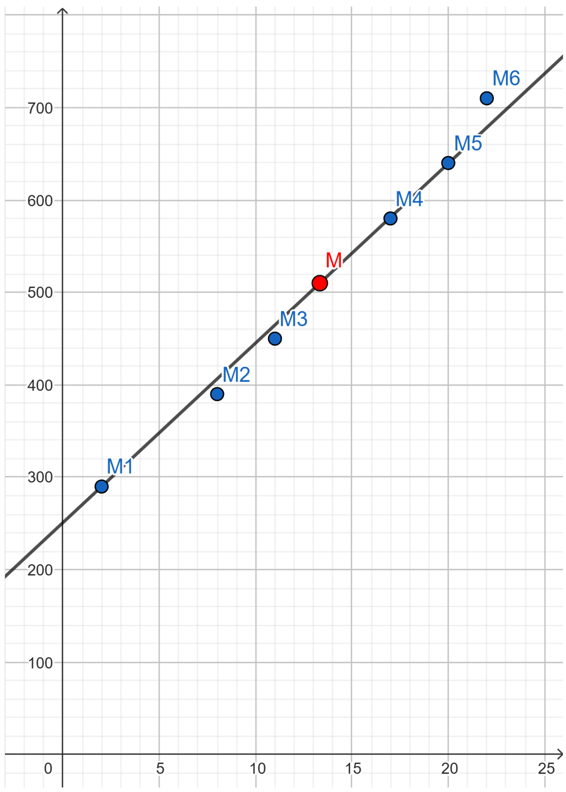
- Quelle droite tracer ?
- Y-en-a-t-il une meilleure que les autres ?
- Meilleure... selon quel critère ?

Exemple

On reprend l'exemple des employés du service informatique. On considère la série statistique à deux variables suivante :

x_i	2	8	11	17	20	22
y_i	290	390	450	580	640	710

On représente le nuage de points associé dans un repère orthonormé.



Le point moyen du nuage de points est $M\,(\overline{x}\ ;\ \overline{y})$ où :

$$\overline{x} = \frac{2+8+11+17+20+22}{6} \approx 13,33$$
$$\overline{y} = \frac{290+390+450+580+640+710}{6} = 510$$

Le nuage de point a une forme alongée, on peut l'ajuster par exemple par la droite (MM_1).

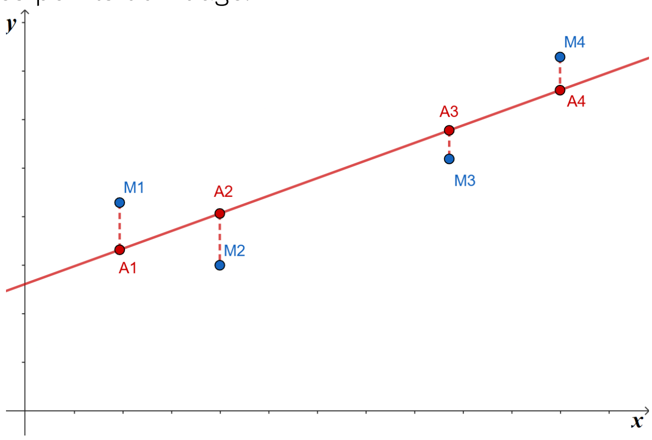
2 Droite des moindres carrés

Principe de la méthode des moindres carrés

Le principe de l'ajustement consiste à estimer les coefficients m et p de la droite d'équation $y = mx + p$ qui passe le plus près possible des points du nuage.

Pour cela, on cherche à minimiser la somme des carrés des distances entre les points du nuage et la droite.
Ainsi les nombres réels m et p sont choisis de sorte que la somme suivante soit minimale :

$$\sum_{i=1}^{n} M_iA_i^2 = \sum_{i=1}^{n} (y_i - (mx_i + p))^2$$



Définitions

Soient (x_i) et (y_i) deux séries statistiques composées de n valeurs.
On note $\overline{x}$ et $\overline{y}$ les moyennes de ces deux séries.

- La **variance** de la série (x_i) est : $V(x) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$;

  La **variance** de la série (y_i) est : $V(y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i - \overline{y})^2$

- L'**écart-type** de la série (x_i) est : $\sigma_x = \sqrt{V(x)}$

  L'**écart-type** de la série (y_i) est : $\sigma_y = \sqrt{V(y)}$

- La **covariance** de la série $(x_i; y_i)$ est : $\text{Cov}(x, y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$.

Propriété (admise)

Dans un repère orthonormé, la droite des moindres carrés associée au nuage de points $M_i\,(x_i\;;\;y_i)$:

- passe par le point moyen $M\,(\overline{x}\;;\;\overline{y})$;
- a pour équation $y = m(x - \overline{x}) + \overline{y}$ avec $m = \dfrac{\text{Cov}(x, y)}{V(x)}$.

Remarques

- On peut obtenir les coefficients m et p de la droite des moindres carrés à l'aide du menu « Régressions » de la calculatrice.
- La covariance $\text{Cov}(x, y)$ peut être positive ou négative. On la note aussi σ_{xy}.

Preuve

$$\begin{aligned}
S &= \sum_{i=1}^{n} A_iM_i^2 \\
&= \sum_{i=1}^{n} (y_i - (mx_i + p))^2 \\
&= \sum_{i=1}^{n} ((y_i - mx_i) - p)^2 \qquad \text{(*)} \\
&= \sum_{i=1}^{n} \left((y_i - mx_i)^2 - 2p(y_i - mx_i) + p^2\right) \\
&= \sum_{i=1}^{n}(y_i - mx_i)^2 - 2p\sum_{i=1}^{n}(y_i - mx_i) + \sum_{i=1}^{n} p^2 \\
&= \sum_{i=1}^{n}(y_i - mx_i)^2 - 2p\sum_{i=1}^{n}(y_i - mx_i) + np^2
\end{aligned}$$

- On suppose que m est fixé.
  S est donc un polynôme du second degré en p de la forme $S(p) = ap^2 + bp + c$ avec $a = n$.
  $a > 0$ donc S admet un minimum atteint lorsque sa dérivée s'annule.

$$\begin{aligned}
S'(p) = 0 \quad &\iff \quad 2np - 2\sum_{i=1}^{n}(y_i - mx_i) = 0 \\
&\iff \quad p = \frac{1}{n}\sum_{i=1}^{n}(y_i - mx_i) \\
&\iff \quad p = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{m}{n}\sum_{i=1}^{n} x_i \\
&\iff \quad p = \overline{y} - m\overline{x}
\end{aligned}$$

Donc
$$\begin{aligned}
S &= \sum_{i=1}^{n} (y_i - mx_i - \overline{y} + m\overline{x})^2 \qquad \text{d'après (*)} \\
&= \sum_{i=1}^{n} ((y_i - \overline{y}) - m\,(x_i - \overline{x}))^2 \\
&= \sum_{i=1}^{n} \left((y_i - \overline{y})^2 - 2m(y_i - \overline{y})(x_i - \overline{x}) + m^2(x_i - \overline{x})^2\right) \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2m\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) + m^2\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= nV(y) - 2mn\text{Cov}(x,y) + m^2nV(x)
\end{aligned}$$

- S est donc un polynôme du second degré en m de la forme $S(m) = am^2 + bm + c$ avec $a = nV(x) > 0$. Il admet donc un minimum atteint lorsque sa dérivée s'annule.

$$\begin{aligned}
S'(m) = 0 \quad &\iff \quad 2nV(x)m - 2n\text{Cov}(x,y) = 0 \\
&\iff \quad m = \frac{\text{Cov}(x,y)}{V(x)}
\end{aligned}$$

- On en déduit que la droite des moindres carrés a pour équation réduite :

$$y = mx + \overline{y} - m\overline{x} = m(x - \overline{x}) + \overline{y}$$

3 Coefficient de corrélation. Changement de variable

Coefficient de corrélation linéaire

La décision d'ajuster un nuage de points par une droite s'est prise jusqu'à présent à la seule vue du nuage de points (suivant que celui-ci est allongé ou non).
Les statisticiens ont mis au point un coefficient de corrélation linéaire qui permet de quantifier la dépendance entre deux caractères quantitatifs.

Définition

Le **coefficient de corrélation linéaire** entre les deux séries (x_i) et (y_i) est le nombre réel noté r et défini par :

$$r = \frac{\text{Cov}(x,y)}{\sigma_x \sigma_y}$$

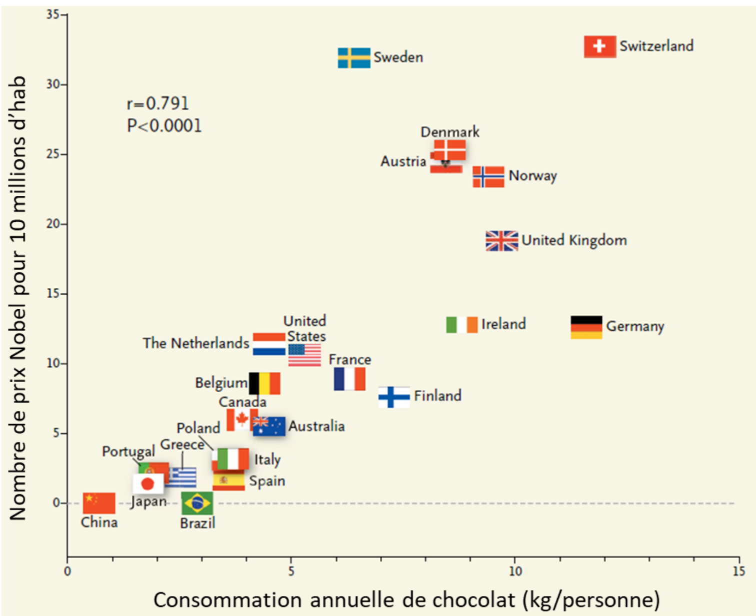
Propriété

Pour toute série statistique à deux variables quantitatives, on a : $-1 \leq r \leq 1$.

- plus $|r|$ est proche de 1, plus la dépendance entre les deux caractères est forte ;
- plus $|r|$ est proche de 0, plus la dépendance entre les deux caractères est faible ;
- si $|r| = 1$, les points du nuage de points sont alignés sur une droite ;
- si $r > 0$, la droite a une pente positive (la variable y augmente quand la variable x augmente) ;
- si $r < 0$, la droite a une pente négative.

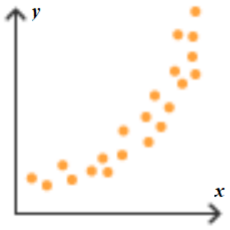
Remarque : Corrélation n'implique pas causalité

Une très forte corrélation peut exister entre deux variables sans qu'il y ait de relation de cause à effet entre elles.
Plusieurs exemples sur le site des décodeurs du Monde.

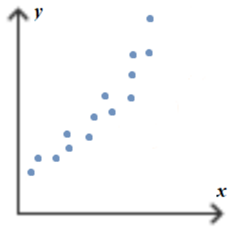


Ajustement se ramenant par changement de variable à un ajustement affine

Dans certains cas, les points du nuage de point semblent se répartir autour d'une courbe autre qu'une droite. Il est parfois possible de se ramener à un ajustement affine à l'aide d'un changement de variable.



Dans un tel cas, on peut penser qu'il existe une relation du type $y = ax^2 + b$ entre x et y.
En posant $u = x^2$, on se ramène à $y = au + b$ et on peut déterminer a et b avec la méthode des moindres carrés.



Dans ce cas, on peut penser qu'il existe une relation du type $y = ae^x + b$ entre x et y.
En posant $u = e^x$, on se ramène à $y = au + b$ et on peut déterminer a et b avec la méthode des moindres carrés.