Synthèse de cours (Terminale ES)
→ Séries statistiques à deux variables

Séries statistiques à deux variables

Définition

On appelle « série statistique à deux variables » la donnée de n couples (x_i, y_i) de valeurs réelles.

Nuage de points et point moyen

A chaque couple (x_i, y_i) on peut associer, dans un repère orthogonal, un point M_i de coordonnées (x_i, y_i). L'ensemble des points ainsi obtenus est appelé « nuage de points associé à la série statistique ».

On considère alors les moyenne statistiques : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Le point, noté G, de coordonnées $(\overline{x}, \overline{y})$ est appelé « point moyen » associé à la série statistique.

Ajustement affine par la méthode des moindres carrés

Objectif

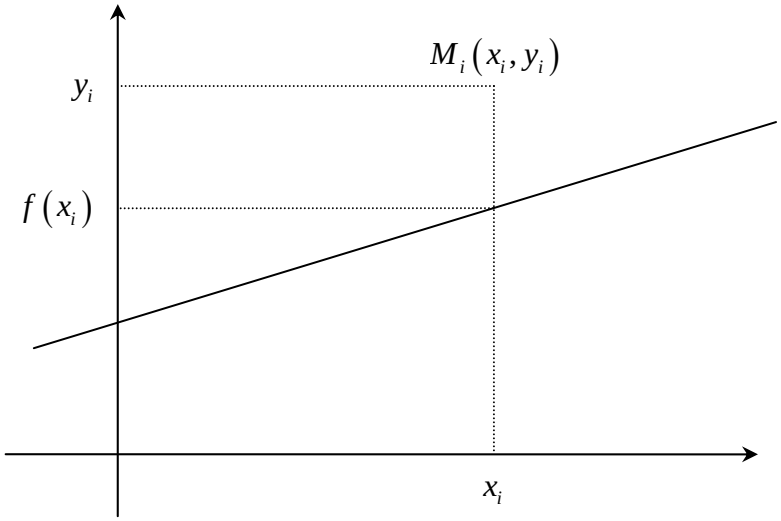
Effectuer un ajustement c'est chercher une fonction dont la représentation graphique décrive « au mieux » (le critère doit être précisé) le nuage de points associé à la série statistique considérée. Si la fonction cherchée est affine, on parle d'« ajustement affine ».

Si cette fonction affine est de la forme : $y = f(x) = ax + b$, on dit que l'on a effectué un « ajustement de y en x ». La variable « y » est alors appelée la « variable expliquée » et la variable « x », la « variable explicative ».

Disposant de f, on va pouvoir se donner une valeur de x quelconque et calculer $f(x)$:

- Si $x_{\min} < x < x_{\max}$, on dit que l'on effectue une « interpolation » ;
- Si $x < x_{\min}$ ou $x > x_{\max}$, on dit que l'on effectue une « extrapolation ».

Méthode des moindres carrés



→ Le principe général de cette méthode consiste à trouver une fonction *f* minimisant la quantité : $\sum_{i=1}^{n}\left(y_i - f\left(x_i\right)\right)^2$ (d'où son nom …)

Ajustement affine par la méthode des moindres carrés

On effectue un ajustement affine de *y* en *x*. On cherche *f* de la forme : $f\left(x\right) = ax + b$.

Dans ces conditions, on a :

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

Remarque :

$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$ est la covariance des deux séries x_i et y_i. On la note : $\mathrm{cov}\left(x, y\right)$;

$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$ est la variance de la série y_i. On la note : $\mathrm{var}\left(x\right)$

On peut donc écrire : $a = \frac{\mathrm{cov}\left(x, y\right)}{\mathrm{var}\left(x\right)}$.

Par ailleurs, la droite représentant la fonction *f* passe par le point moyen G. On a donc :

$$y - \overline{y} = a\left(x - \overline{x}\right)$$

Et on en déduit :

$$\boxed{b = -a\overline{x} + \overline{y}}$$

La droite obtenue est appelée « droite de régression de y en x ».

Rappels sur la variance et la covariance

On a :

$$\text{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}{x_i}^2 - \overline{x}^2$$

Pour retenir la seconde expression : « moyenne des carrés moins carré de la moyenne ».

On a :

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

Pour retenir la seconde expression : « moyenne des produits moins produit de la moyenne ».