Statistiques : Ajustements

I. Nuage de points

1) Série statistique à deux variables

On suppose que, suite à une étude faite, on s'intéresse à deux caractères quantitatifs (*ie* deux variables numériques) sur une population donnée.
À chaque individu de cette population, on associe donc un couple $(x_i ; y_i)$ de nombres réels où la variable x_i est la valeur de la première variable pour l'individu considéré et où la variable y_i est la valeur de la seconde variable.

> L'ensemble de ces couples forme une *série statistique à deux variables* ou encore *série statistique double*.

Les résultats peuvent être résumés dans un tableau :

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

Si la première variable est le temps, on parle alors de *série chronologique.*

2) Nuage de points

Nuage de points

> Dans un repère orthogonal bien choisi, l'ensemble des points M_i de coordonnées $(x_i ; y_i)$, avec $1 \leqslant i \leqslant n$, est appelé *le nuage de points* associé à cette série statistiques à deux variables.

Dans le cas d'une série chronologique, on relie souvent les points M_i par des segments de droite.

Point moyen

Notons $\bar{x}$ la moyenne des valeurs x_i et $\bar{y}$ la moyenne des valeurs y_i. On a :

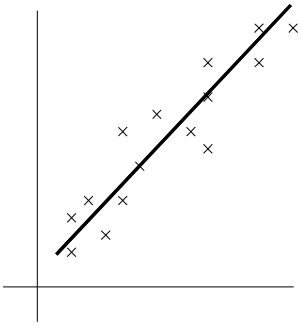
$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \text{ et } \bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$$

> Le point G de coordonnées $(\bar{x} ; \bar{y})$ est appelé *le point moyen* du nuage de points associé à cette série statistique à deux variables.

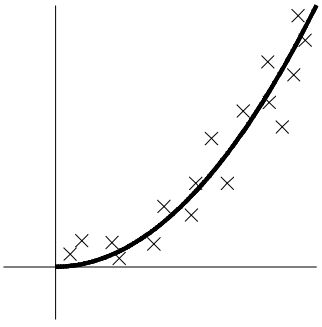
II. Ajustement affine par la méthode des moindres carrés

1) Le principe de l'ajustement

Suivant la forme du nuage de points, on peut essayer de trouver une fonction f qui modélise le lien entre les deux caractères (c'est-à-dire que la courbe d'équation $y = f(x)$ passe le plus près possible de chacun des points).



Ajustement par une droite (ou affine)



Ajustement par une parabole

Certaines formes de nuages de points ne permettent pas de tenter un ajustement…

2) Ajustement par la méthode des moindres carrés

- Considérons deux séries statistiques (x_i) et (y_i), telles que le nuage de points présente un « certain » alignement. On se propose de d'ajuster ce nuage par une droite ; on sait que des nombreuses droites sont possibles, laissées au choix de l'utilisateur.

- Mais une méthode, la *méthode des moindres carrés*, est principalement utilisée, parce qu'elle présente de nombreux intérêts théoriques et que les calculs qu'elle entraîne se font sans trop de difficulté.

L'idée est de déterminer les coefficients a et b d'une droite $\mathscr{D}$ d'équation $y = ax + b$ de sorte qu'elle passe le « plus près » possible des points du nuage.

Pour chaque abscisse x_i, on calcule la distance M_iP_i entre le point du nuage et le point de la droite, c'est-à-dire que :

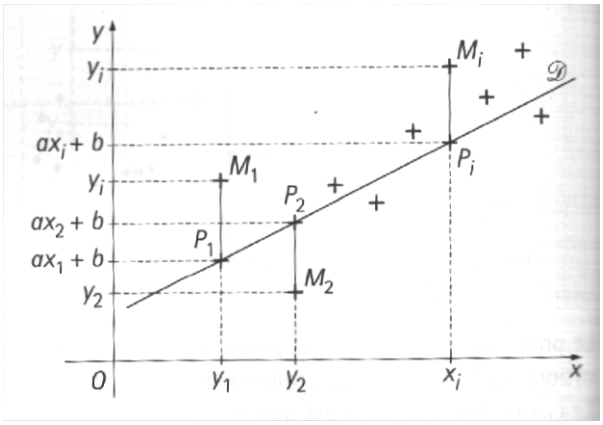
$$M_iP_i = \left| y_i - (ax_i + b_i) \right|.$$

Dans la méthode des moindres carrés, on recherche a et b pour lesquels la somme des carrés de ces distances

$$S = M_1P_1^2 + M_2P_2^2 + ... + M_nP_n^2$$

est *minimale.*

Il se trouve qu'il existe une *seule* droite répondant à cette condition.



- **Théorème** *(admis)*

La *droite de régression*[1] *de y en x* associée au nuage de points M_i de coordonnées (x_i ; y_i) avec $1 \leqslant i \leqslant n$ est la droite :

passant par le point moyen $G(\overline{x}\,;\,\overline{y})$ du nuage ;

de coefficient directeur a donné par la formule $a = \dfrac{\text{cov}(x,y)}{V(x)}$ avec

$V(x) = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ et $\text{cov}(x,y) = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$.

C'est la droite qui permet de minimiser la distance $S = M_1P_1^2 + M_2P_2^2 + ... + M_nP_n^2$.

Dans $V(x)$, on reconnaît la variance de la série statistique x.

Cov(x, y) est appelée *covariance* de x et de y. Cette grandeur peut être calculée avec la calculatrice (ainsi d'ailleurs que la variance).

Autre écriture de la variance :

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

Remarquon que l'équation de cette droite est finalement $y = a(x - \overline{x}) + \overline{y}$.

3) Utilités

Pour une valeur donnée x_0 de la variable x, l'ajustement permet de prévoir approximativement la valeur correspondante de y.

- si x_0 appartient à l'intervalle d'observation des valeurs de x, on dit que l'on fait une *interpolation*.
- si x_0 n'appartient pas à cet intervalle, on parle d'*extrapolation*, mais dans ce cas il faut faire l'hypothèse que le modèle reste plausible à l'extérieur de cet intervalle.

[1] EN 1886, Francis Galton examinait la taille des enfants en fonction de la taille moyenne des parents. Il nota que les enfants de parents de grande taille avaient tendance à être plus petits qu'eux, les enfants des parents de petite taille avaient tendance à être plus grands qu'eux. Il y avait donc régression du caractère "grande taille" vers la taille moyenne : on dit parfois que la droite d'ajustement de y en x est *la droite de régression de y en x*.