Séries statistiques à deux variables numériques. Nuage de point associé. Ajustement affine par la méthode des moindres carrés. Droite de régression. Applications. L'exposé pourra être illustré par un ou des exemples faisant appel à l'utilisation d'une calculatrice.

Chantal Menini

14 mai 2009

1 Introduction

Une définition très générale d'une étude statistique peut être : *Obtenir une information significative à partir de données observées.*
Nous nous interesserons ici à l'étude des valeurs prises simultannément par deux caractères quantitatifs d'une même population. Nous allons chercher à déterminer *une fonction d'ajustement*, c'est-à-dire une fonction qui aux valeurs d'un des caractères associe des valeurs voisines de celles prises par le second caractère dans un sens que nous allons préciser.

2 Vocabulaire. Séries statistiques à deux variables numériques.

On appelle *population* tout ensemble étudié par la statistique, nous le noterons Ω, un *individu* est un élément de Ω.
Un *caractère* est une propriété de la population qui peut être qualitatif (par exemple la couleur des yeux) ou quantitatif (par exemple la taille).
Nous nous placerons pour toute la suite de l'exposé dans le cas d'une population finie de cardinal n et de caractères quantitatifs.

Définition 2.1 *Une **série statistique à deux variables numériques** est une application qui à chaque individu associe la valeur prise par les deux caractères*

$$\begin{array}{rccc} (X,Y) : & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & (X(\omega), Y(\omega)) \end{array}$$

avec $\text{card}(X(\Omega)) \geq 2$ *et* $\text{card}(Y(\Omega)) \geq 2$.

Elle est usuellement donnée sous forme d'une suite $(x_i, y_i)_{1 \leq i \leq n}$ ou d'un tableau

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

.
Dans de nombreux exemples nous étudierons de cette façon des séries chronologiques, x_i sera alors la date à laquelle le caractère Y prendra la valeur y_i.
Rappels :

1. On note $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ la moyenne de $(x_i)_{1 \leq i \leq n}$.
2. On note $S_x = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$ l'écart type de $(x_i)_{1 \leq i \leq n}$.

Définition 2.2 *On appelle **covariance** de la série statistique* $(x_i, y_i)_{1 \leq i \leq n}$ *le réel* $C_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$.

Définition 2.3 *1. Dans un repère orthogonal l'ensemble des n points $M_i(x_i, y_i)$ constitue le* ***nuage de points*** *associé à la série statistique $(x_i, y_i)_{1\leq i\leq n}$.*

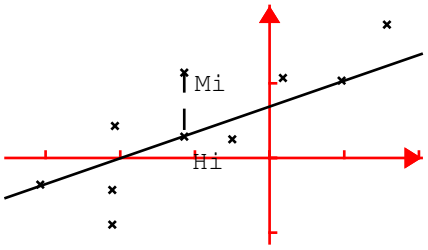
2. Le point $G(\bar{x}, \bar{y})$ est appelé ***point moyen*** *du nuage de points.*

Remarque 2.4 *G est l'isobarycentre du système de points $\{M_i, 1 \leq i \leq n\}$.*

3 Ajustement affine par la méthode des moindres carrés.

3.1 La méthode.

On cherche une fonction affine f telle que, si l'on note $\varepsilon_i = y_i - f(x_i)$ l'erreur commise lorsque l'on approche y_i par $f(x_i)$, alors, $\varepsilon(a,b) = \sum_{i=1}^{n} \varepsilon_i^2$ soit minimal. On dit qu'alors f *réalise un ajustement affine de Y en X par la méthode des moindres carrés.*



Minimiser $\varepsilon(a,b)$ s'interprète graphiquement comme la minimisation de $\sum_{i=1}^{n} M_i H_i^2$.

Théorème 3.1 *Etant donnée une série statistique $(x_i, y_i)_{1\leq i\leq n}$ il existe une unique fonction réalisant un ajustement affine de Y en X (resp. X en Y) par la méthode des moindres carrés. Elle est donné par $f(x) = ax + b$ avec*

$$a = \frac{C_{xy}}{S_x^2}, \quad b = \bar{y} - a\bar{x}$$

(resp. donnée par $g(y) = a'y + b'$ avec $a' = \frac{C_{xy}}{S_y^2}$ et $b' = \bar{x} - a\bar{y}$).

Définition 3.2 *La droite d'équation $y = ax + b$ (resp. $x = a'y + b'$) avec $a = \frac{C_{xy}}{S_x^2}$ et $b = \bar{y} - a\bar{x}$ (resp. $a' = \frac{C_{xy}}{S_y^2}$ et $b' = \bar{x} - a\bar{y}$) est appellée* ***droite de régression de*** *Y* ***en*** *X (resp. X en Y).*

Remarque 3.3 *1. Les droites de régression de Y en X et de X en Y passent par le point moyen du nuage de points.*

2. La droite de régression n'est pas modifiée par un changement d'origine du repère.

Preuve.
Une preuve astucieuse (mais pas tant que cela, voir à la fin dans les commentaires) mais utilisant des outils de niveau terminale.
On note $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i$ la moyenne des erreurs, alors

$$\begin{aligned}
\varepsilon(a,b) &= \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon} + \bar{\varepsilon})^2 \\
&= \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2 + n\bar{\varepsilon}^2 \\
&= \sum_{i=1}^{n} ((y_i - \bar{y}) - a(x_i - \bar{x}))^2 + n(\bar{y} - a\bar{x} - b)^2 \\
&\geq \sum_{i=1}^{n} ((y_i - \bar{y}) - a(x_i - \bar{x}))^2
\end{aligned}$$

pour tout couple (a,b), avec égalité si et seulement si $b = \bar{y} - a\bar{x}$.

Trouver (a,b) qui minimise $\varepsilon(a,b)$ équivaut à trouver a qui minimise $\varphi(a) = \sum_{i=1}^{n}((y_i - \bar{y}) - a(x_i - \bar{x}))^2$, b valant alors $\bar{y} - a\bar{x}$.

$$\varphi(a) = \left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] a^2 - 2\left[\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right] a + \sum_{i=1}^{n}(y_i - \bar{y})^2$$

est un polynôme de degré 2 et le coefficient de a^2 est positif (*pourquoi* $\sum_{i=1}^{n}(x_i - \bar{x})^2$ *est-il non nul ?*), il admet donc un minimum au point $a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ ce qui est le résultat annoncé.

Une preuve de niveau universitaire mais sans astuce (qui peut aussi servir à retrouver les coefficients de la droite de régression si on a un trou).

$\varepsilon(a,b)$ est une fonction de deux variables définie sur $\mathbb{R}^2$, si elle admet un minimum c'est nécessairement en un point où les dérivées partielles s'annulent soit

$$\begin{cases} \frac{\partial \varepsilon}{\partial a}(a,b) = -2\sum_{i=1}^{n} x_i(y_i - ax_i - b) = 0 \\ \frac{\partial \varepsilon}{\partial b}(a,b) = -2\sum_{i=1}^{n}(y_i - ax_i - b) = 0 \end{cases}$$

et l'on trouve un unique couple (a,b) solution de ce système.

Il reste alors à connaitre la nature de ce point, pour cela on calcule les dérivées partielles d'ordre 2.

$$r = \frac{\partial^2 \varepsilon}{\partial a^2}(a,b) = 2\sum_{i=1}^{n} x_i^2, \quad s = \frac{\partial^2 \varepsilon}{\partial a \partial b}(a,b) = 2\sum_{i=1}^{n} x_i, \quad t = \frac{\partial^2 \varepsilon}{\partial b^2}(a,b) = 2n$$

et $s^2 - rt = 4\left(\sum_{i=1}^{n} x_i\right)^2 - 4n\sum_{i=1}^{n} x_i^2$. Avec l'inégalité de Cauchy-Schwarz $s^2 - rt \leq 0$ et il y a égalité si et seulement si $x_1 = x_2 = \cdots x_n$ ce qui est exclu par hypothèse. $r > 0$ il s'agit donc d'un minimum local.

Reste à justifier que ce minimum local est en fait global, cela vient du fait que nous avons un polynôme de degré 2 (*écrivez la formule de Taylor pour vous en convaincre*). □

Remarque 3.4 *La valeur du minimum de $\varepsilon(a,b)$ est*

$$\sum_{i=1}^{n}\left((y_i - \bar{y}) - \frac{C_{xy}}{S_x^2}(x_i - \bar{x})\right)^2 = nS_y^2\left(1 - \left(\frac{C_{xy}}{S_x S_y}\right)^2\right).$$

3.2 Coefficient de corrélation.

Définition 3.5 *On appelle* ***coefficient de corrélation*** *la quantité $r_{xy} = \frac{C_{xy}}{S_x S_y}$.*

Remarque 3.6 *r_{xy}^2 est inchangé par un changement de variable affine.*

Proposition 3.7 *Le coefficient de corrélation r_{xy} appartient à $[-1,1]$ et il vaut 1 ou -1 si et seulement si les points du nuage sont alignés.*

Preuve. Cela découle de la remarque 3.4, r_{xy} appartient à $[-1,1]$ car $\varepsilon(a,b)$ est positif ou nul. r_{xy} vaut 1 ou -1 si et seulement si $\varepsilon(a,b)$ est nul soit si et seulement si les points du nuage sont alignés. □

Corrolaire 3.8 *$|C_{xy}| \leq S_x S_y$ et il y a égalité si et seulement si les points du nuage sont alignés.*

Proposition 3.9 *Les droites de régression de Y en X (D) et de X en Y (D') sont confondues si et seulement si le coefficient de corrélation r_{xy} vaut 1 ou -1.*

Preuve. D a pour équation $y = ax + b$, D' a pour équation $y = \frac{1}{a'}x - \frac{b'}{a'}$ lorsque a' est non nul. Elles passent toutes les deux par le point moyen, elles sont donc confondues si et seulement si $a = \frac{1}{a'}$ soit $r_{xy}^2 = 1$.

Si $a' = 0$ (et donc $a = 0$ aussi) elles sont perpendiculaires. □

4 Intéret et interprétation.

4.1 Corrélation et dépendance.

On dit qu'il y a une *forte corrélation* entre les caractères X et Y lorsque r_{xy} est proche de 1 en valeur absolue, et une *faible corrélation* lorsque r_{xy} est proche de 0.
Il faut faire attention au fait que corrélation ne veut pas dire dépendance, on fabrique très facilement une série statistique à deux variables fortement corrélées en ayant pris deux caractères dépendant linéairement du temps.
Par contre si l'on voit la série statistique comme une série de valeurs prises par deux variables aléatoires X et Y de façon équiprobable, l'indépendance de ces variables aléatoires implique que leur covariance $C(X,Y)$ (qui n'est autre que C_{xy} dans ce cas) est nulle. Rappelons que la réciproque est fausse, deux variables aléatoires peuvent avoir une covariance nulle sans qu'elles soient indépendantes.

4.2 Intéret.

Une fois que l'on a un ajustement "valable", soit, dans le cas de l'ajustement affine par la méthode des moindres carrés, si l'on a un nuage de points de forme allongée. On peut estimer la valeur prise par le caractère Y lorsque le caractère X prend une valeur x_0

- comprise entre les valeurs extrêmes de la suite $(x_i)_{1\leq i\leq n}$ on parlera alors *d'interpolation*,
- non comprise entre les valeurs extrêmes de la suite $(x_i)_{1\leq i\leq n}$ on parlera alors *d'extrapolation*.

Dans le deuxième cas il faudra au préalable s'assurer que le modèle reste valable.

5 Autres ajustements.

5.1 Droite de Mayer.

On sépare le nuage de points en deux nuages composés du même nombre de points (à 1 près), en général le premier est constitué des points d'abcisse inférieure à l'abcisse médiane et le deuxième des autres points. On considère G_1 le point moyen du premier nuage et G_2 le point moyen du deuxième nuage, la droite de Mayer est la droite (G_1G_2) (*elle passe par G le point moyen du nuage de points initial, pourquoi ?*).

5.2 Ajustement exponentiel.

On a l'impression que le nuage de points est "proche" de la courbe représentative d'une fonction exponentielle. Pour s'en convaincre soit, on placera les points du nuage sur une feuille munie d'un repère semi-logarithmique (ce qui évite des calculs supplémentaires), soit, on placera les points de coordonnées $(x_i, \ln y_i)$ dans un repère orthogonal. Dans les deux cas on doit obtenir un nuage ayant une forme allongée et que l'on va donc ajuster de façon affine par la méthode des moindres carrés. On obtient la relation $\ln y = ax + b$, soit $y = Ae^{ax}$.

5.3 Ajustement logarithmique.

Cette fois-ci on a l'impression que le nuage de points est "proche" de la courbe représentative d'une fonction logarithme, on fera donc un ajustement affine avec la suite $(\ln x_i, y_i)_{1\leq i\leq n}$ et on obtient la relation $y = a\ln x + b$.

5.4 Ajustement puissance.

On fait un ajustement affine avec la suite $(\ln x_i, \ln y_i)_{1\leq i\leq n}$ et on obtient la relation $\ln y = a\ln x + b$ soit $y = Ax^a$.

6 Commentaires.

- Cette leçon manque cruellement d'exemples, il faut en mettre au moins un que l'on suivra tout au long de l'exposé avec la calculatrice (calcul des moyennes, écart-type, covariance, affichage du nuage de points, tableur montrant les différentes erreurs en fonction du choix de a et b, droites de régressions, etc.). Bref il faut s'entrainer

en se construisant un exemple assez simple pour pouvoir le reproduire (ou un équivalent) le jour de l'oral sans document.

- La démonstration de l'existance d'un unique ajustement affine par la méthode des moindres carrés n'est pas au programme de Terminale ES, cependant c'est l'unique démonstration consistante de cet exposé. Cela semble donc difficile de faire l'impasse dessus. Faute de s'en souvenir il semble indispensable d'avoir une idée assez précise sur la façon de l'obtenir.
- Le choix a été fait ici de pas montrer dès le départ que $|C_{xy}| \leq S_x S_y$, ce choix est discutable et on peut très bien le faire avant de parler d'ajustement. En général pour le faire on utilise que le polynôme de degré deux en λ : $S^2_{\lambda x+y} = \lambda^2 S_x^2 - 2\lambda C_{xy} + S_y^2$ est positif ou nul donc de discriminant négatif ou nul et le discriminant vaut $4(C_{xy}^2 - S_x^2 S_y^2)$. Dans le cas d'égalité, le polynôme a une racine double, il existe donc λ tel que $S^2_{\lambda x+y} = 0$, on en déduit que les points sont alignés (au fait y a t-il une différence avec la démonstration de l'inégalité de Cauchy-Schwarz ?).
- On peut aussi introduire un exemple afin de discuter sur l'influence d'un point très éloigné du nuage dans la détermination de la droite de régression et le calcul du coefficient de corrélation.
- On peut interpréter géométriquement le coefficient de corrélation de la façon suivante. On considère dans $\mathbb{R}^n$ muni du produit scalaire usuel, les vecteurs $\tilde{X}$ et $\tilde{Y}$ de coordonnées respesctives $(x_1 - \bar{x}, \dots x_n - \bar{x})$ et $(y_1 - \bar{y}, \dots y_n - \bar{y})$ alors $C_{xy} = \langle \tilde{X}, \tilde{Y} \rangle$ le produit scalaire des vecteurs $\tilde{X}$ et $\tilde{Y}$; $S_x = \|\tilde{X}\|$ (resp. S_y) est la norme du vecteur $\tilde{X}$ (resp. $\tilde{Y}$). Ainsi r_{xy} est le cosinus d'une mesure de l'angle de vecteurs $(\tilde{X}, \tilde{Y})$.
- Si on a été capable de donner l'interprétation géométrique précédente, on doit être capable de poursuivre avec l'interprétation de l'ajustement affine par la méthode des moindres carrés. On note X, Y et I les vecteurs de $\mathbb{R}^n$ de coordonnées respesctives $(x_1, \dots x_n)$, $(y_1, \dots y_n)$ et $(1, \dots 1)$ alors minimiser ce que l'on a noté $\varepsilon(a, b)$ revient à minimiser $\|Y - aX - bI\|^2$, c'est-à-dire, trouver a et b tels que $aX + bI$ soit le projeté orthogonal de Y sur le sous-espace vectoriel de $\mathbb{R}^n$ engendré par I et X. Ceci sera réalisé si et seulement si $\langle Y - aX - bI, I \rangle = 0$ et $\langle Y - aX - bI, X \rangle = 0$, effectuez les calculs et vous verrez que c'est la façon la plus rapide pour déterminer a et b.
- "L'astuce" de la preuve pour la détermination des coefficients de la droite de régression est tout simplement d'utiliser (avec les notations précédentes) que $Y - aX - bI - \frac{1}{n}\langle Y - aX - bI, I \rangle I$ est orthogonal à I (cf. $\|I\|^2 = n$).

7 Lien avec les dossiers.

On peut re-exploiter cette leçon dans le dossier qui a pour thème : "Séries statistiques à deux variables" (et vice-versa ...), annales 10/07/2005 et 19/07/2007. En particulier les parties "Intéret et interprétation" et "Autres ajustements" peuvent être source d'inspiration pour trouver des exercices variés.

8 Bibliographie.

Des livres de terminales ES pour la trame et pour avoir un peu de recul (et trouver les démonstrations) par exemple "Statistique descriptive" de M. Janvier Ed. Dunod ou "Probabilités, analyse des données et Statistique" de G. Saporta Ed. Technip (assez complexe).