SERIES STATISTIQUES A DEUX VARIABLES
AJUSTEMENTS LINÉAIRES ET NON LINÉAIRES

Prérequis : Application réciproque. Moyenne. Nuage de points.

Soit Ω une population de N individus.

1. Rappels : séries statistiques à une variable

On appelle variable statistique ou numérique toute application $X : \Omega \to \mathbb{R}$. Pour tout $x \in \mathbb{R}$, si $X^{-1}(x)$ est non vide alors x est une modalité de X et le cardinal de $X^{-1}(x)$ s'appelle l' effectif de x. On appelle série statistique associée à X l'ensemble des couples (x_i, n_i), $i = 1, \ldots, p$ où n_i est l'effectif de la modalité x_i. On note

$$S_X = \{(x_i, n_i)\,;\, i = 1, \ldots, p\}.$$

Exemple 1 : On étudie le poids exprimé en kg de 877 jeunes adultes.

Intervalles	Effectif	Milieu de l'intervalle
$[50, 54[$	55	52
$[54, 58[$	170	56
$[58, 62[$	211	60
$[62, 66[$	233	64
$[66, 70[$	127	68
$[70, 74[$	57	72
$[74, 78[$	21	76
$[78, 82[$	3	80

Ici, Ω est l'ensemble de ces 877 individus. Le poids de tout individu ω appartient à un unique intervalle de la forme $[p-2, p+2[$ pour un certain p appartenant à $\{52, 56, \ldots, 80\}$. On définit alors la variable statistique X par $X(\omega) = p$. Ainsi

$$S_X = \{(52, 55), (56, 170), \ldots, (80, 3)\}.$$

2. Séries statistiques à deux variables numériques

Soit X, Y deux variables numériques définies sur Ω de modalités respectives $x_1, \ldots, x_p$ et $y_1, \ldots, y_q$. Etant donnés $i \in [1, p]$ et $j \in [1, q]$, on désigne par $n_{i,j}$ le nombre d'individus de Ω présentant la modalité x_i et y_j.

Définition 2.1. *On appelle série statistique à deux variables l'ensemble des triplets de la forme $(x_i, y_j, n_{i,j})$ pour lesquels $n_{i,j} > 0$.*

Exemple 1 (suite) : En considérant la taille (en cm) des 877 individus, on définit une variable statistique Y. La série statistique des variables X et Y est donnée par le tableau suivant.

$X\backslash Y$	163	166	169	172	175
52	13	25	17		
56	44	88	27	11	
60	41	84	57	27	2
64	18	54	113	33	15
68	5	33	31	33	25
72		3	9	22	23
76			4	7	10
80					3

Par exemple, 13 individus pèsent entre 50 et 54 kg et mesurent entre $161,5$ et $164,5$ cm.

3. Ajustement affine par la méthode des moindres carrés

L'ajustement affine a pour but de remplacer le nuage de points par une droite d'équation $y = \hat{a}x + \hat{b}$. L'interprétation graphique de la méthode des moindres carrés est très largement documentée (en particulier sur internet) et donc ne sera pas reprise ici. Soit X et Y deux variables statistiques.

3.1. **Cas où $n_{i,j} = 1$.** La série statistique des variables X et Y peut alors s'écrire sous la forme $(x_1, y_1), \cdots, (x_n, y_n)$. On note $\overline{x}$ et $\overline{y}$ les moyennes de X et Y.

Définition 3.1. *La variance de X est $V(X) = \frac{1}{n}\sum(x_i - \overline{x})^2$. La covariance de X, Y est $cov(X,Y) = \frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})$.*

Remarquons que $cov(X,X) = V(X)$.

Proposition 3.1. *Soit X et Y deux variables statistiques. Alors*
- (i) $V(X) = \frac{1}{n}\sum x_i^2 - \overline{x}^2$.
- (ii) $cov(X,Y) = \frac{1}{n}\sum x_i y_i - \overline{x}\,\overline{y}$.
- (iii) *Pour tout $\alpha \in \mathbb{R}$, $V(\alpha X + Y) = \alpha^2 V(X) + 2\alpha cov(X,Y) + V(Y)$.*

Démonstration. (i) et (ii) sont évidents. Pour (iii), utiliser (i) pour $V(\alpha X + Y)$, dévolopper les carrés de sommes. Puis utiliser (i) pour obtenir $V(X)$, $V(Y)$ et (ii).

Théorème 3.1. *[Méthode des moindres carrés] On suppose $V(X) \neq 0$. Pour tout $a, b \in \mathbb{R}$, on pose*

$$S(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2, \quad \hat{a} = \frac{cov(X,Y)}{V(X)}, \quad \hat{b} = \overline{y} - \hat{a}\overline{x}.$$

Alors

$$S(\hat{a}, \hat{b}) \leq S(a,b).$$

Démonstration. D'après (i) Proposition 3.1,

$$\frac{1}{n}\sum(y_i - ax_i - b)^2 = V(Y - aX - b) + \left(\frac{1}{n}\sum y_i - ax_i - b\right)^2.$$

Or $V(Y - aX - b) = V(Y - aX)$. Donc avec (iii) Proposition 3.1,

$$\frac{1}{n}S(a,b) = V(X)a^2 - 2cov(X,Y)a + V(Y) + (\overline{y} - a\overline{x} - b)^2.$$

La fonction $a \mapsto V(X)a^2 - 2cov(X,Y)a + V(Y)$ atteint son minimum en $\hat{a}$ car $V(X) \neq 0$. Par suite, pour tout $a, b \in \mathbb{R}$,

$$\frac{1}{n}S(a,b) \geq V(X)\hat{a}^2 - 2cov(X,Y)\hat{a} + V(Y) = \frac{1}{n}S(\hat{a}, \hat{b}).$$

Définition 3.2. *La droite d'équation $y = \hat{a}x + \hat{b}$ s'appelle la droite de regression de Y en X.*

Exemple : L'équation de la droite de régression de la série statistique $(1,1),(2,1),(3,2)$ est $y = \frac{1}{2}x + \frac{1}{3}$.

3.2. **Cas général $n_{i,j} > 0$, $\forall 1 \leq i \leq p$, $1 \leq j \leq q$.** Le nombre d'individus présentant la modalité x_i est

$$n_{i.} := \sum_{j=1}^{q} n_{i,j}.$$

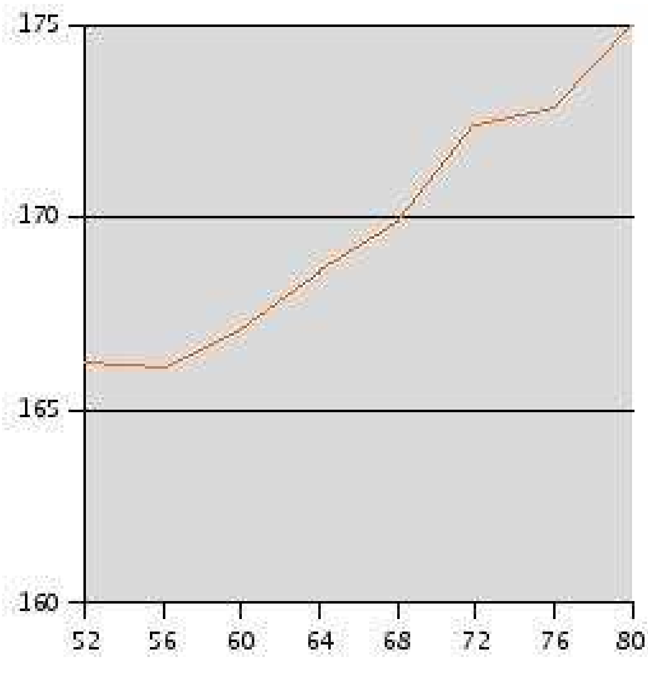
La population totale est $n = \sum_{i,j} n_{i,j}$. Avec les définitions et notations précédentes, on en déduit

$$\begin{aligned} \overline{x} &= \frac{1}{n} \sum_{i=1}^{p} n_{i.} x_i \\ V(X) &= \frac{1}{n} \sum_{i=1}^{p} n_{i.} x_i^2 - \overline{x}^2 \\ cov(X,Y) &= \frac{1}{n} \sum_{i,j} n_{i,j}(x_i - \overline{x})(y_i - \overline{y}). \end{aligned}$$

La méthode des moindres carrés s'énonce comme dans le théorème 3.1.

Exemple 1 (suite) : L'équation de la droite de régression de Y en X est, approximativement, $y = 0,33x + 147,83$. ($cov(X,Y) \simeq 10,72$.)

4. Ajustements non linéaires



4.1. **Ligne des moyennes conditionnelles.** Soit

$$S = \{(x_i, y_j, n_{i,j})\, ;\, 1 \leq i \leq p,\, 1 \leq j \leq q,\, n_{i,j} > 0\}$$

une série statistique à deux variables. Pour tout $i \in [1, p]$, la moyenne conditionnelle de Y sachant x_i est

$$\overline{y}_i = \frac{\displaystyle\sum_{j=1}^{q} n_{i,j} y_j}{\displaystyle\sum_{j=1}^{q} n_{i,j}}.$$

En outre, sans perte de généralité, on peut supposer $x_1 < \cdots < x_p$.

Définition 4.1. *La ligne brisée joignant les points de coordonnées* $(x_1, \overline{y}_1), \ldots, (x_p, \overline{y}_p)$ *s'appelle le ligne des moyennes conditionnelles de* Y *sachant* X*.*

Exemple 1 (suite) : Le graphique ci-dessus représente la ligne des moyennes conditionelles de la taille en fonction du poids.

4.2. **Ajustements exponentiel.** L'ajustement exponentiel a pour but de remplacer le nuage de points par une courbe d'équation $y = b.a^x$. Sous les hypothèses de la section 3, on pose $Z = \ln Y$. Par la méthode des moindres carrés, on obtient $Z = \hat{a}X + \hat{b}$. D'où, en posant $a = e^{\hat{a}}$ et $b = e^{\hat{b}}$, $Y = b.a^X$.

Arnaud Rougirel
IUFM Poitou Charentes
29 avril 2008
E-mail address: rougirel@math.univ-poitiers.fr