1 Statistique-Sociologie L1 – Couple de variables statistiques

Nuage statistique. Pour un couple de variables (X, Y) chaque individu de l'échantillon est représenté dans le plan par un point de coordonnées (X_i, Y_i) qui sont les valeurs des variables X, Y pour le i-ème individu. L'ensemble de ces points s'appelle *nuage statistique* de l'échantillon.

Barycentre ou Point Moyen. Dans le cas de deux variables, un indicateur de tendance centrale est le *barycentre* (appelé aussi *centre de gravité* ou *point moyen*) qui est l'analogue de la moyenne :

$$\textbf{Barycentre :}\quad G = (m(X), m(Y))\,.$$

1.1 La droite de régression de Mayer

La méthode de Mayer pour trouver une droite qui passe au plus près d'un nuage de points consiste à partager le nuage de points rangés dans l'ordre croissant de leurs abscisses en deux sous-groupes de même effectif. On calcule le barycentre (point moyen) de chaque sous-groupe. Si G_1 est le point moyen du premier sous-groupe et G_2 le point moyen du deuxième sous-groupe, la *droite de Mayer* est la droite passant par G_1 et G_2.
On utilise les coordonnées de G_1 et G_2 pour trouver l'équation de la droite de Mayer.

Exemple. Une entreprise veut faire des prévisions sur son chiffre d'affaires.

Année x_i	1	2	3	4	5	6	7	8
Chiffre d'affaires y_i	16	19	22	23	24	26	27	30

Ce tableau présente les chiffres d'affaires (en millions d'euros) depuis la création de l'entreprise.

Le calcul. On sépare les points du nuage en deux parties et on calcule leurs respectifs barycentres :

$$m(X_1) = \frac{1+2+3+4}{4} = 2,5, \quad m(Y_1) = \frac{16+19+22+23}{4} = 20 \quad \Longrightarrow \quad G_1 = (2,5\,;\,20).$$

$$m(X_2) = \frac{5+6+7+8}{4} = 6,5, \quad m(Y_2) = \frac{24+26+27+30}{4} = 26,75 \quad \Longrightarrow \quad G_2 = (6,5\,;\,26,75).$$

Le déplacement horizontal pour aller de G_1 à G_2 est égal à

$$\Delta_x = m(X_2) - m(X_1) = 6,5 - 2,5 = 4\,;$$

Le déplacement vertical pour aller de G_1 à G_2 est égal à

$$\Delta_y = m(Y_2) - m(Y_1) = 26,75 - 20 = 6,75\,.$$

La pente de la droite qui passe par G_1 et G_2 est égale à $a = \frac{\Delta_y}{\Delta_x} = \frac{6,75}{4} = 1,6875 \approx 1,69$

On trouve l'équation de la droite de Mayer avec la pente $a \approx 1,69$ et un des barycentres G_1 ou G_2, par exemple G_1 :

$$y = a(x - m(X_1)) + m(Y_1) \quad \text{c-à-d} \quad y = 1,69(x - 2,5) + 20 \quad \text{ce qui donne}$$

$$y = 1,69\,x + 15,775\,.$$

Voir le dessin sur la page 4.

Nous allons voir d'autres *droites de régression* qui approximent de manière beaucoup plus précise le nuage statistique. Mais avant de faire cela, il faut savoir déterminer si le nuage est "suffisamment aligné" pour qu'il soit modelé (décrit approximativement) par une telle droite de régression.

1.2 Y a-t-il toujours une corrélation linéaire ? Coefficient de Pearson

Dans l'exemple précédent, le nuage de points est "assez proche" d'être "aligné" et la droite de Mayer approxime (ou modèle) très bien le nuage de points.

On peut toujours calculer la droite de Mayer (ou les autres droites de régression) même si les points du nuage statistique sont loin d'être alignés, disons dispersés chaotiquement. Mais dans ce cas, les droites de régression ne pourront pas décrire (et modeler) la distribution du nuage statistique.

Pour déterminer si la corrélation entre deux caractères est *'suffisamment linéaire'*, nous utilisons le

Coefficient de corrélation linéaire de Pearson
Ce coefficient nous indique la présence ou l'absence d'une relation linéaire entre deux caractères quantitatifs. Pour calculer ce coefficient il faut tout d'abord calculer la covariance. La covariance est la moyenne du produit des écarts à la moyenne, mais qui se calcule avec la formule

$$\textbf{Covariance :}\quad cov(X,Y) = m(XY) - m(X)m(Y)\,.$$

Le *coefficient de corrélation linéaire* de deux caractères X et Y est égal à la covariance de X et Y divisée par le produit des écarts-types de X et Y :

$$\textbf{Coefficient de corrélation linéaire :}\quad r(X,Y) = \frac{cov(X,Y)}{s(X)\cdot s(Y)}.$$

Propriétés du coefficient $r(X,Y)$:
- $r(X,Y)$ varie entre -1 et $+1$;
- si r est proche de 0, il n'y a pas de relation linéaire entre X et Y ;
- si r est proche de -1, il existe une forte relation linéaire négative entre X et Y ;
- si r est proche de 1, il existe une forte relation linéaire positive entre X et Y.

La "valeur absolue" de r (proche 1) indique l'intensité de la relation c'est-à-dire la capacité à prédire les valeurs de Y en fonction de celles de X par une droite.
Le signe de r indique le sens de la relation, c-à-d le signe de la pente de la droite de régression.

Exemple. Le tableau de l'exemple précédent. Pour calculer la covariance on calcule d'abord

$$m(X) = \frac{1+2+\ldots+7+8}{8} = 4,5\,; \qquad m(Y) = \frac{16+19+\ldots+27+30}{8} = 23,375\,;$$

$$m(X^2) = \frac{1^2+2^2+\ldots+7^2+8^2}{8} = 25,5\,; \quad m(Y^2) = \frac{16^2+19^2+\ldots+27^2+30^2}{8} = 563,875\,; \quad \text{donc}$$

$$V(X) = 25,5-(4,5)^2 = 5,25 \quad \text{et} \quad V(Y) = 563,875-(23,375)^2 = 17,48\,;$$

$$s(X) = \sqrt{5,25} \approx 2,29 \quad \text{et} \quad s(Y) = \sqrt{17,48} \approx 4,18\,.$$

Il nous reste à calculer

$$m(XY) = \frac{1\times 16+2\times 19+3\times 22+\ldots+7\times 27+8\times 30}{8} = \frac{917}{8} = 114,625\,. \text{ Donc}$$

$$cov(X,Y) = 114,625 - 4,5\times 23,375 = 9,4375$$

Enfin, le coefficient de Pearson est égal à

$$r(X,Y) = \frac{cov(X,Y)}{s(X)\times s(Y)} = \frac{9,4375}{2,29\times 4,18} \approx 0,986\,.$$

Ce qui montre une correlation linéaire très forte (que l'on voit si on dessine le nuage statistique).

1.3 Comment obtenir les droites de régression ?

Si on a pu mettre en évidence l'existence d'une relation linéaire significative entre deux caractères quantitatifs X et Y, par le coefficient de Pearson, on peut modeler cette relation par l'équation d'une droite, écrite d'une des deux formes suivantes :

$$D_{Y|X} : \; Y = aX + b\,; \quad a = \frac{cov(X,Y)}{V(X)}\,; \quad b = m(Y) - a \cdot m(X)$$

$$D_{X|Y} : \; X = \widehat{a}Y + \widehat{b}\,; \quad \widehat{a} = \frac{cov(X,Y)}{V(Y)}\,; \quad \widehat{b} = m(X) - \widehat{a} \cdot m(Y)\,.$$

(1) $Y = a.X + b$: droite de régression de Y en fonction de X ;

(2) $X = \widehat{a}.Y + \widehat{b}$: droite de régression de X en fonction de Y ;

Barycentre. Il est utile de savoir que les deux droites, $D_{Y|X}$ et $D_{X|Y}$, passent par le barycentre $G = (m(X); m(Y))$ - c'est leur point d'intersection. Donc pour tracer ces droites, la première chose à faire est de tracer le barycentre.

Exemple. (1) Si on applique la formule de la droite $D_{Y|X}$ à l'exemple précédent, où nous avons déjà fait les calculs nécessaires, on obtient

$$a = \frac{9,4375}{5,25} \approx 1,798 \quad \text{et} \quad b = 23,375 - 1,798 \times 4,5 \approx 15,284\,.$$

L'équation de la droite de régression $D_{Y|X}$ (où Y s'exprime en fonction de X) est donc

$$y = 1,798\,x + 15,284\,.$$

Celle-ci donne une approximation plus fine que la droite de Mayer ($y = 1,69\,x + 15,775$).
Pour la tracer on peut prendre les points $G = (4,5\,;\, 23,375)$ et $P = (0\,;\, 15,284)$, obtenu de l'équation en faisant $x = 0$ (voir le dessin sur la page 4).

(2) Si on applique la formule de la droite $D_{X|Y}$ on obtient

$$\widehat{a} = \frac{9,4375}{17,48} \approx 0,54 \quad \text{et} \quad \widehat{b} = 4,5 - 0,54 \times 23,375 \approx -8,1\,.$$

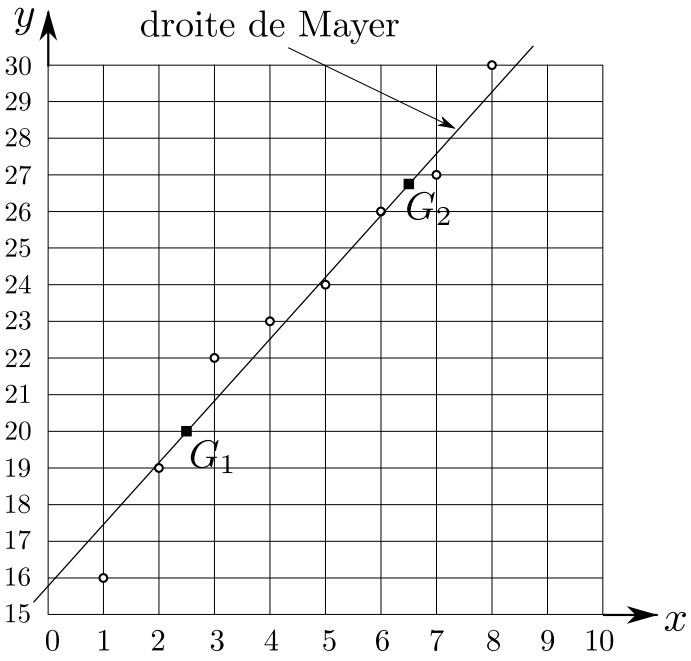
L'équation de la droite de régression $D_{X|Y}$ (où X s'exprime en fonction de Y) est donc

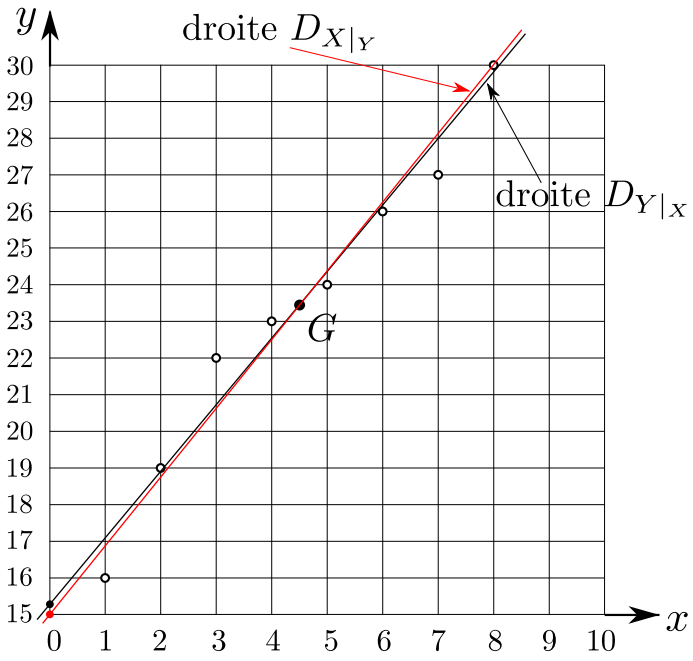
$$x = 0,54\,y - 8,1\,.$$

Pour la tracer on peut prendre les points $G = (4,5\,;\, 23,375)$ et $Q = (0\,;\, 15,04)$, obtenu de l'équation en faisant $x = 0$ et donc $y = \frac{8,1}{0,54} \approx 15,04$ (voir le dessin sur la page 4).
Les deux équations proposées ci-dessus correspondent à deux droites différentes, deux "résumés" différents du nuage statistique. La différence entre les deux droites vient du fait que les deux équations proposées correspondent à deux objectifs différents :

(1) **La droite de régression de Y en fonction de X** introduit l'hypothèse selon laquelle les valeurs de Y dépendent de celles de X. L'objectif est de minimiser la distance entre les valeurs Y_i^* (observées pour X_i) et les valeurs Y_i estimées par la relation $Y = aX + b$ (pour $X = X_i$).

(2) **La droite de régression de X en fonction de Y** introduit l'hypothèse inverse selon laquelle les valeurs de X dépendent de celles de Y. L'objectif, cette fois-ci, est de minimiser la distance entre les valeurs X_i^* (observées pour Y_i) et les valeurs X_i estimées par la relation $X = aY + b$ (pour $Y = Y_i$).


y
droite de Mayer
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
G_2
G_1
0 1 2 3 4 5 6 7 8 9 10
x


y
droite $D_{X|Y}$
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
droite $D_{Y|X}$
G
0 1 2 3 4 5 6 7 8 9 10
x