Analyse statistique de deux variables

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

Soient deux séries statistiques X et Y, qui prennent les valeurs $(x_1, y_1), \ldots, (x_n, y_n)$

La représentation graphique donne un *nuage* de points $M_i\,(1 \leqslant i \leqslant n)$.

⋆Les moyennes de X et Y sont respectivement $\overline{X} = \dfrac{1}{n}\,(x_1 + \cdots + x_n) = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ et $\overline{Y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$

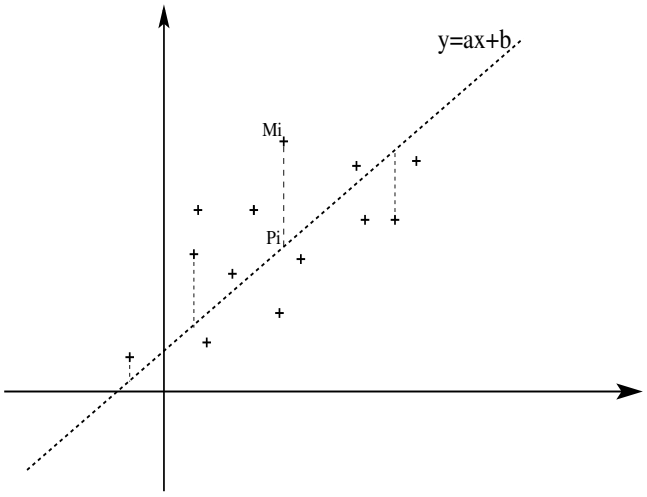
⋆Les variances sont $V(X) = \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right) - \overline{X}^2$ et $V(Y) = \dfrac{1}{n}\left(\sum_{i=1}^{n} y_i^2\right) - \overline{Y}^2$

⋆Enfin, $\mathrm{cov}(X,Y) = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})\,(y_i - \overline{Y}) = \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\,y_i\right) - \overline{X}\,\overline{Y}$

Ajustement linéaire

On cherche si Y dépend de façon affine de X, c'est à dire s'il existe une droite Δ, d'équation $y = a\,x + b$ qui approxime le nuage de points (M_i) de coordonnées (x_i, y_i) pour $i \in [\![1\,,\,n]\!]$.

Pour cela, une telle droite étant donnée, on définit les points P_i, projetés sur Δ des points M_i, parallèlement à l'axe des ordonnées. Donc si M_i a pour coordonnées (x_i, y_i), P_i a pour abscisse x_i (la même que M_i), et pour ordonnée $a\,x_i + b$ puisque P_i appartient à Δ. La droite cherchée –dite *droite de régression*– est celle qui minimise les distances M_iP_i, ou plus exactement $\sum_{i=1}^{n}(M_iP_i)^2$.



Cette somme peut être considérée comme une fonction f des deux variables a et b; en particulier, pour a donné, c'est un polynôme du second degré en b, qui admet un minimum au point $b_0 = \dfrac{\sum_{i=1}^{n}(y_i - a\,x_i)}{n} = \overline{Y} - a\,\overline{X}$. $M_iP_i = |y_i - a\,x_i - b|$,

donc $\sum_{i=1}^{n} M_iP_i^2 = \sum_{i=1}^{n}((y_i - a\,x_i) - b)^2$

$= \sum_{i=1}^{n}(y_i - a\,x_i)^2 - 2\,b\sum_{i=1}^{n}(y_i - a\,x_i) + b^2 n$

Cette valeur de b étant déterminée, l'expression de f devient $f(a, \overline{Y} - a\,\overline{X}) = \sum_{i=1}^{n}((y_i - \overline{Y}) - a\,(x_i - \overline{X}))^2 =$

$\sum_{i=1}^{n}(y_i - \overline{Y})^2 - 2\,a\sum_{i=1}^{n}(x_i - \overline{X})\,(y_i - \overline{Y}) + a^2\sum_{i=1}^{n}(x_i - \overline{X})^2$; c'est une fonction de la seule variable a.

On cherche à nouveau le minimum, atteint pour $a_0 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{X})\,(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2} = \dfrac{\mathrm{cov}(X,Y)}{V(X)}$, la droite de régression a donc pour équation $y - \overline{Y} = \dfrac{\mathrm{cov}(X,Y)}{V(X)}\,(x - \overline{X})$; on remarque qu'elle passe par le barycentre G du nuage de points de coordonnées $(\overline{X}, \overline{Y})$

Coefficient de corrélation linéaire

Pour $\lambda \in \mathbb{R}$, on étudie la variance de $\lambda\,X + Y$: $V(\lambda\,X + Y) = \lambda^2 V(X) + 2\,\lambda\,\mathrm{cov}(X,Y) + V(Y)$ cette valeur étant toujours positive, le discriminant $\Delta = 4(\mathrm{cov}^2(X,Y) - V(X)\,V(Y))$ est négatif. On en déduit que le coefficient $\rho(X,Y) = \dfrac{\mathrm{cov}(X,Y)}{\sqrt{V(X)\,V(Y)}} \in [-1\,,\,1]$.

D'autre part, $V(\lambda\,X + Y)$ est minimale pour $\lambda_0 = -\dfrac{\mathrm{cov}(X,Y)}{V(X)} = -\,a_0$. La droite de régression calculée minimise la variance de la variable aléatoire $\lambda\,X + Y$.

Lorsque le discriminant est nul (c'est à dire lorsque $\rho(X,Y) = 1$ ou -1, le minimum est nul. Dans ce cas, $\lambda_0\,X + Y$ est une variable aléatoire de variance nulle, elle est donc constante (égale à $K \in \mathbb{R}$). Donc $Y = -\lambda_0\,X + K$; Y est exactement fonction affine de X, les points M_i sont alignés et appartiennent tous à la droite de régression.

Conclusion

Plus le coefficient de corrélation linéaire $\rho(X,Y)$ est proche de 1 en valeur absolue, meilleure est l'approximation affine. Un coefficient trop proche de 0 remet en question la dépendance de Y par rapport à X. Le signe de a (donc de $\rho(X,Y)$ ou encore de $-\lambda_0$) représente le sens de variation *en moyenne* de Y en fonction de X.

Exercices d'application

Exemple 1 Deux techniques différentes «Biorad» (X) et «Isolab» (Y) ont été utilisées dans les mêmes conditions pour doser la quantité d'hémoglobine glycosylée dans 10 échantillons de sang provenant de 10 malades présumés. Les densités optiques obtenues sont consignées dans le tableau ci-dessous :

Biorad (x_i)	7,1	7,5	7,6	7,9	8,0	8,1	8,1	8,3	8,5	8,9
Isolab (y_i)	6,4	7,0	6,7	7,5	7,9	7,5	8,0	8,1	8,1	8,0

1. Déterminer l'équation de la droite de régression linéaire de Y en fonction de X.
2. Calculer le coefficient de corrélation linéaire, conclusion ?
3. Déterminer l'équation de la droite de régression linéaire de X en fonction de Y, remarque ?

Exemple 2 Mêmes questions concernant l'étude de deux caractères sur une population de dix individus :

x_i	208	185	206	170	188	199	184	203	196	161
y_i	113	83	99	66	88	98	102	96	100	56