Cours 07
Variance, covariance, corrélation

1 Définitions et propriétés élémentaires

Définition : Soient $x = (x_1, x_2, \ldots, x_n)$ et $y = (y_1, y_2, \ldots, y_n)$ deux échantillons de même taille n et de moyenne respective $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^n$.
On appelle *covariance* (estimée [1]) de x avec y et on note $\mathrm{cov}\,(x, y)$ le nombre

$$\mathrm{cov}\,(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

On appelle *variance* (estimée) de x et on note $\mathrm{Var}\,(x)$ le nombre

$$\mathrm{Var}\,(x) = \mathrm{cov}\,(x, x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

et $\sqrt{\mathrm{Var}\,(x)}$ s'appelle *l'écart-type* de l'échantillon x. On appelle *corrélation* de x avec y et on note $\mathrm{cor}\,(x, y)$ le nombre

$$\rho = \mathrm{cor}\,(x, y) = \frac{cov(x, y)}{\sqrt{\mathrm{Var}\,(x)}\sqrt{\mathrm{Var}\,(y)}}.$$

Proposition 1 *Soient $x,$, x', x'', y, y', et y'' des échantillons de même taille n. Notons $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ et $\bar{y} = \frac{1}{n}\sum_{i=1}^n$ la moyenne de x et de y. Soient λ, λ' et λ'' des nombres. Les relations suivantes sont satisfaites :*

1. $cov(\lambda' x' + \lambda'' x'', y) = \lambda' cov(x', y) + \lambda'' cov(x'', y)$ *(Bilinéarité : linéarité à gauche)*
2. $cov(x, \lambda' y' + \lambda'' y'') = \lambda' cov(x, y') + \lambda'' cov(x, y'')$ *(Bilinéarité : linéarité à droite)*
3. $cov(y, x) = cov(x, y)$ *(Symétrie)*
4. $Var(\lambda x) = \lambda^2 Var(x) \geq 0$ *(Homogénéité de degré 2)*
5. *Formules de Huygens :*

$$cov(x, y) = \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \frac{n}{n-1}\bar{x}\bar{y} \text{ ,et } Var(x) = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}^2.$$

6. *Inégalité de Cauchy-Schwarz :* $(cov(x, y))^2 \leq Var(x)\,Var(y)$, *et donc* $\rho^2 \leq 1$.

2 Relation avec la régression linéaire, interprétation géométrique

Lors du cours 06 nous avons déterminé l'équation $y = ax + b$ de la droite de régression linéaire de l'échantillon y sur l'échantillon de même taille x. Les valeurs de la pente a de cette droite et son ordonnée à l'origine b sont donnés par $a = \dfrac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$ et $b = \bar{y} - a\bar{x}$, où $\bar{x}$ et $\bar{y}$ désignent la moyenne de `x` et `y`. En divisant par $n-1$ le numérateur et le dénominateur de l'expression donnée de a et en appliquant la formule de Hygens, on voit que $\boxed{a = \frac{\mathrm{cov}\,(x,y)}{\mathrm{Var}\,(x)}.}$

Vision géométrique : $\mathrm{Var}\,(x)$ mesure l'*inertie* $\sum_{i=1}^n (x_i - \bar{x})^2$ des n points x_i par rapport au point $\bar{x}$, l'inertie $\sum_{i=1}^n (x_i - x')^2$ par rapport à tout autre point x' étant nécessairement supérieure. Cette inertie n'est nulle que si tous les x_i sont égaux, et ce qu'il est convenu d'appeler l'*échantillon centré* ${}^c x = x - \bar{x}$ caractérise la dispersion de x, alors que la moyenne $\bar{x}$ est le nombre unique résumant le mieux l'échantillon. L'écart-type $\sqrt{\mathrm{Var}\,(x)}$ mesure la dispersion ${}^c x$ de l'échantillon, de façon homogène par rapport au choix de l'unité utilisée pour former x, puisque pour tout $\lambda > 0$, $\sqrt{\mathrm{Var}\,(\lambda x)} = \lambda\sqrt{\mathrm{Var}\,(x)}$ (en passant du mètre au centimètre, l'écart-type est multiplié par le même nombre, 100, que les mesures relevées). On se débarrasse de la question du choix de l'unité en considérant l'échantillon *centré-réduit* ${}^{cr} x = {}^c x / \sqrt{\mathrm{Var}\,(x)}$: c'est une direction dans $\mathbf{R}^n$ indépendante des unités choisies, et $\rho = \mathrm{cor}\,(x, y) = \mathrm{cor}\,({}^{cr} x, {}^{cr} y)$ est le cosinus de l'angle que font les deux direction ${}^{cr} x$ et ${}^{cr} y$ dans le plan de $\mathbf{R}^n$ qu'elles déterminent. Si $\rho = 1$ c'est que ces deux variabilités ont des directions parfaitement égales (et $\rho = -1$ c'est que ces deux variabilités ont des directions parfaitement opposées). Dans les deux cas la connaissance de x équivaut à la connaissance de y : les résidus $\epsilon_i = y_i - ax_i - b$ pour la régression de y sur x sont tous nuls !

1. Nous adoptons ici les définitions pratiquées par les commandes `var`, `sd`, et `cov` de `R` fondée sur l'estimateur sans biais de ces grandeurs pour un échantillon dont l'espérance (estimée par `mean`) n'est pas connue à priori. Ceci explique que le lecteur pourra aussi trouver parfois un n à la place du dénominateur $n-1$ des définitions adoptées ici.

3 Matrice de variance-covariance

On appelle matrice de variance-covariance de x et y la matrice symétrique $\begin{pmatrix} \operatorname{Var}(x) & \operatorname{cov}(x,y) \\ \operatorname{cov}(y,x) & \operatorname{Var}(y) \end{pmatrix}$. On appelle matrice de corrélation de x et y la matrice symétrique $\begin{pmatrix} 1 & \operatorname{cor}(x,y) \\ \operatorname{cor}(y,x) & 1 \end{pmatrix}$.

Si on a m échantillons x^1, …, x^m de même taille on définit leur matrice de variance-covariance Σ et leur matrice de corrélation R par

$$\Sigma = \begin{pmatrix} \operatorname{Var}(x^1) & \cdots & \operatorname{cov}(x^1,x^k) & \cdots & \operatorname{cov}(x^1,x^m) \\ \vdots & \ddots & \vdots & & \vdots \\ \operatorname{cov}(x^k,x^1) & \cdots & \operatorname{Var}(x^k) & \cdots & \operatorname{cov}(x^k,x^m) \\ \vdots & & \vdots & \ddots & \vdots \\ \operatorname{cov}(x^m,x^1) & \cdots & \operatorname{cov}(x^m,x^k) & \cdots & \operatorname{Var}(x^m) \end{pmatrix}$$

$$\text{et } R = \begin{pmatrix} 1 & \cdots & \operatorname{cor}(x^1,x^k) & \cdots & \operatorname{cor}(x^1,x^m) \\ \vdots & \ddots & \vdots & & \vdots \\ \operatorname{cor}(x^k,x^1) & \cdots & 1 & \cdots & \operatorname{cor}(x^k,x^m) \\ \vdots & & \vdots & \ddots & \vdots \\ \operatorname{cor}(x^m,x^1) & \cdots & \operatorname{cor}(x^m,x^k) & \cdots & 1 \end{pmatrix}.$$