DU BII - MODULE 3 - CORRÉLATION, RÉGRESSION

G ACHAZ

Ce document décrit brièvement les principes de la covariation entre deux Variables Aléatoires, VA dans la suite. L'exemple illustré sur les figures est le jeu de données utilisé en cours.

Moyenne, variance

On s'intéresse ici à une population d'objets aléatoires caractérisés par 2 variables aléatoires, X et Y. On peut par exemple penser à des individus caractérisés par leur taille et leur poids.

Chacune de ces variables aléatoires possède une distribution *marginale*, souvent caractérisée par deux valeurs quantitatives : leurs espérances, $\mathbb{E}[X]$ et $\mathbb{E}[Y]$, et leurs variances, $\mathbb{V}\text{ar}[X]$ et $\mathbb{V}\text{ar}[Y]$. Ces distributions marginales ne "considèrent pas" les valeurs de l'autre variable aléatoire : elles sont intégrées sur tout l'espace de l'autre VA.

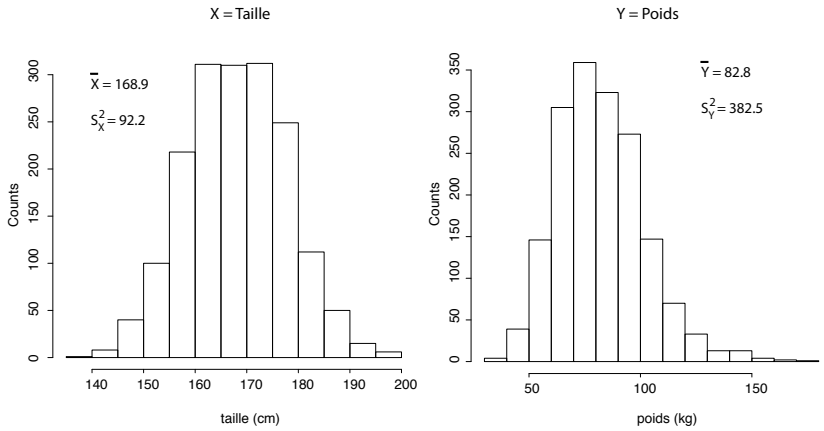

Figure 1. Représentation des deux distributions empiriques marginales de X (taille) et Y (poids) pour un échantillon de taille $n = 1732$. Sont également indiquées leurs moyennes et variances dans cet échantillon.

On rappelle que pour un échantillon, on peut calculer des estimateurs des moyennes, $\bar{X}$ et $\bar{Y}$, et des variances, s_X^2 et s_Y^2, comme :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

Covariance

Il existe, au delà de ces deux distributions marginales, une distribution jointe des deux VA que l'on peut représenter aisément dans un espace bidimensionnel.

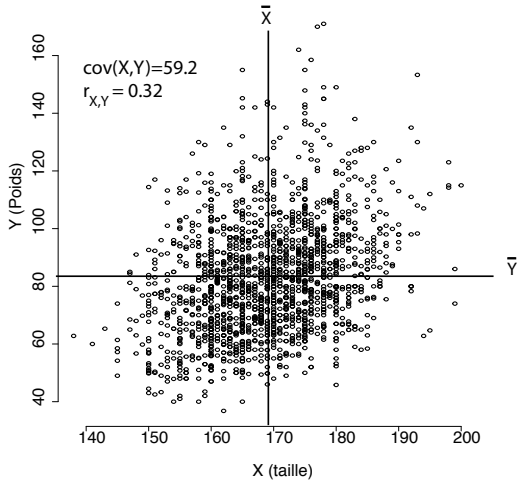

Figure 2. Représentation des deux VA dans un plan (0,x,y). Sont indiquées également les valeurs de covariance et le coefficient de corrélation (*vide infra*).

Une quantité qui caractérise la distribution jointe des deux VA est leur covariance ($\mathbb{C}\text{ov}[X, Y]$). Pour un échantillon de taille n, elle est estimée comme :

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

On note immédiatement que $\text{cov}(x, x) = s_X^2$, c'est à dire que la covariance est une généralisation de la variance pour deux VA. On notera aussi que la covariance d'un échantillon peut se lire graphiquement sur le plan en le découpant en 4 quarts, définis par les moyennes empiriques $\bar{X}$ et $\bar{Y}$ (cf. figure 2). Les points (x_i, y_i) dont les deux valeurs sont supérieures aux deux moyennes contribuent positivement à la covariance : $(x_i - \bar{X}) > 0, (y_i - \bar{Y}) > 0 \implies (x_i - \bar{X})(y_i - \bar{Y}) > 0$. De même pour les points dont les deux VA sont situées en dessous des moyennes $(x_i - \bar{X}) < 0, (y_i - \bar{Y}) < 0 \implies (x_i - \bar{X})(y_i - \bar{Y}) > 0$. Ainsi, en première approximation, on peut considérer que si les points sont principalement répartis dans ces deux quarts, la covariance est positive. S'ils sont situés plutôt dans les deux autres quarts, la covariance est négative. S'ils sont équi-répartis dans les 4 quarts, la covariance est proche de 0.

La covariance caractérise donc par son signe la relation entre les deux variables aléatoires. Covarient-elles négativement ou positivement ? On peut montrer que la covariance d'un échantillon est maximale lorsque les deux valeurs ont une relation linéaire, c-a-d qu'elles forment une droite. Dans ce cas, la valeur absolue de la covariance est $\sqrt{s_X^2 s_Y^2}$.

Corrélation

La corrélation entre deux VA aléatoires X et Y est caractérisée par le coefficient de corrélation, noté $r_{x,y}$ ou encore $\text{cor}(x, y)$ définit comme :

$$\begin{aligned} r_{x,y} &= \text{cov}(x,y)/\sqrt{s_X^2 s_Y^2} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{X})^2 \sum_{k=1}^{n}(y_k - \bar{Y})^2}} \end{aligned}$$

Ainsi, le coefficient de corrélation est une covariance normalisée par sa valeur absolue maximale. Les valeurs possibles de ce coefficient de corrélation sont donc $r_{x,y} \in [-1, 1]$. Il vaut 0 lorsque la covariance est nulle. Il vaut 1 lorsqu'elle est positive et maximale, signifiant que les couples x_i, y_i sont sur une droite parfaite de pente positive. Il vaut -1 lorsqu'elle est négative et maximale (une droite de pente négative).

Si les deux variables aléatoires sont indépendantes ($\mathbb{C}\text{ov}[X, Y] = 0$), il est cependant possible qu'un échantillon de ces variables aléatoires aient un covariance estimée non-nulle et donc un $r_{x,y} \neq 0$. Par exemple, dans le cas d'un échantillon de taille 2 $\{(x_1, y_1), (x_2, y_2)\}$, il est toujours possible de faire passer une droite parfaite de pente non nulle (sauf exception) entre ces deux points. Pourtant les deux VA pourraient être complètement indépendantes. Il existe donc une statistique permettant de tester l'indépendance entre les deux VA à partir de $r_{x,y}$ et de n.

Si les deux variables aléatoires sont indépendantes, leur coefficient de corrélation est d'espérance nulle (en moyenne, il vaut zéro) mais peut s'en écarter dans un échantillon de taille finie. On peut montrer dans ce cas que, la VA T_r (cf. formule ci-dessous), qui est une transformation de ce coefficient de corrélation, suit asymptotiquement une loi de Student à $n - 2$ degrés de libertés. Pour une valeur de $r_{x,y}$, noté simplement r ci-dessous, on calcule t_r comme :

$$t_r = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}$$

On regarde ensuite où se situe t_r dans la distribution de Student pour tester l'indépendance et/ou calculer une p-valeur. Dans l'exemple ci-dessus, on calcule $t_r = 13.81$; la valeur seuil à 5% est $t_{5\%,df=1730} = 1.96$ et la p-valeur associée $p \sim 1.58 \times 10^{-41}$, la corrélation est donc hautement significative.

Régression

Bien que la covariance et la corrélation permettent de caractériser quantitativement et statistiquement la relation entre deux VA, on pourrait souhaiter produire un modèle qui, à partir d'une valeur x_i prédirait sa valeur y associée, notée $\hat{y}_i$. Une des méthodes possibles pour déterminer les paramètres optimaux d'un modèle est la méthode des moindres carrés. Dans cette méthode, on calcule la somme des carrés des écarts entre les valeurs prédites $\hat{Y}$ et les valeur observées Y, calculée comme :

$$S = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

On cherche un modèle dont les prédictions sont au plus proches des données observées, c-a-d une modèle pour lequel cette somme est minimale. Un des modèles naturels possibles est celui d'une relation affine (linéaire à un centrage près) entre les deux variables $Y = aX + b$. Dans ce cas, S devient :

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

La somme est indicée par (a,b) pour indiquer que sa valeur change en fonction de la pente (a) et de l'ordonnée à l'origine (b) choisies. Dans la méthode des moindres carrés, on choisira ces deux valeurs telles que $S_{a,b}$ soit minimale.

Dans ce cas, on peut montrer (en calculant les valeurs de a et b pour lesquelles les dérivées partielles de cette somme s'annulent) que ces deux valeurs $(\hat{a}, \hat{b})$ sont données par :

$$\begin{aligned} \hat{a} &= \frac{\mathrm{cov(x,y)}}{s_X^2} \\ \hat{b} &= \bar{Y} - \hat{a}\bar{X} \end{aligned}$$

La pente est donc du même signe que la covariance et la droite passe par le point d'intersection des deux moyennes, comme illustré sur la figure 3 :

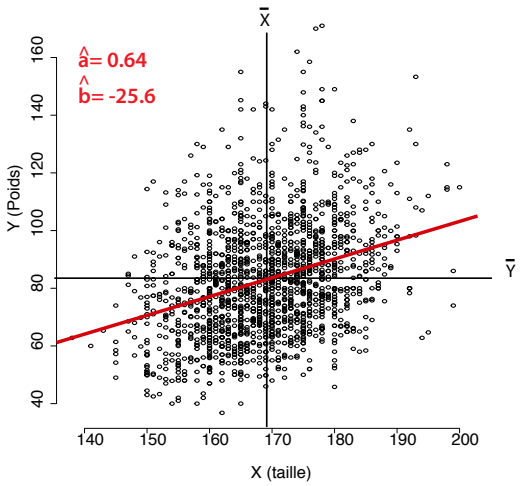

FIGURE 3. Régression estimée de Y sur X : le modèle prédit Y à partir de aX+b. Sont indiquées les estimations par moindres carrés de la pente, $\hat{a}$, et de l'ordonnée à l'origine, $\hat{b}$.