ÉTUDE SIMULTANÉE DE DEUX CARACTÈRES STATISTIQUES

Dans cette partie du cours, on considère sur un échantillon de N individus, deux variables statistiques quantitatives $X = (X_1, \ldots, X_N)$ et $Y = (Y_1, \ldots, Y_N)$. Lorsqu'on observe deux variables quantitatives sur les mêmes individus, on peut s'intéresser à une liaison éventuelle entre ces deux variables.

1. COVARIANCE ET CORRÉLATION

La covariance est une mesure numérique, qui permet d'évaluer la dépendance de deux variables, et, plus précisément, le sens de variation simultanée de ces deux variables.

Définition 1 (Covariance)**.** *On appelle covariance de X et Y la quantité*

$$Cov(X, Y) = \frac{1}{N} \sum_{n=1}^{N} (X_n - \bar{X})(Y_n - \bar{Y}) \, .$$

La proposition suivante donne un moyen pratique de calculer la covariance de deux variables statistiques.

Proposition 1.

$$Cov(X, Y) = \frac{1}{N} \sum_{n=1}^{N} X_n Y_n - \bar{X}\bar{Y}$$

Démonstration.

$$\begin{aligned}
\text{Cov}(X, Y) &= \frac{1}{N} \sum_{n=1}^{N} (X_n - \bar{X})(Y_n - \bar{Y}) \\
&= \frac{1}{N} \sum_{n=1}^{N} \left\{ X_n Y_n - \bar{X} Y_n - \bar{Y} X_n + \bar{X}\bar{Y} \right\} \\
&= \frac{1}{N} \sum_{n=1}^{N} X_n Y_n - \bar{X} \frac{1}{N} \sum_{n=1}^{N} Y_n - \bar{Y} \frac{1}{N} \sum_{n=1}^{N} X_n + \bar{X}\bar{Y} \\
&= \frac{1}{N} \sum_{n=1}^{N} X_n Y_n - \bar{X}\bar{Y}
\end{aligned}$$

∎

La covariance possède quelques propriétés élémentaire résumée dans la proposition suivante

Proposition 2 (propriétés)**.**

– *La covariance est symétrique :*

$$Cov(X, Y) = Cov(Y, X) \, .$$

– *La covariance d'une variable avec elle même correspond à sa variance :*

$$Cov(X, X) = \sigma_X^2 \ .$$

– *On a l'inégalité suivante :*

$$|Cov(X, Y)| \leq \sqrt{\sigma_X^2 \sigma_Y^2} \ .$$

La covariance étant une mesure du sens de variation simultanée de deux variables, on dira :
– $\text{Cov}(X, Y) > 0 \Leftrightarrow X$ et Y varient dans le même sens ;
– $\text{Cov}(X, Y) < 0 \Leftrightarrow X$ et Y varient en sens contraire.

De plus, on rappelle que, en probabilité, si X et Y dont deux variables aléatoires indépendantes, alors $\text{Cov}(X, Y) = 0$. Cependant, le fait que $\text{Cov}(X, Y) = 0$ **n'implique pas** que X et Y sont indépendants.

Grâce à la covariance, on peut définir le coefficient de corrélation linéaire, définit ci-après

Définition 2 (coefficient de corrélation linéaire)**.** *On appelle coefficient de corrélation linéaire la quantité notée r :*

$$r = \frac{Cov(X, Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} \ .$$

Quelque soit l'ordre de grandeur de X et Y, le coefficient de corrélation linéaire est un nombre sans unité, tel que $-1 \leq r \leq 1$. Il mesure la dépendance *linéaire* qui peut existe entre X et Y. En particulier, s'il existe $a \in \mathbb{R}$ et $b \in \mathbb{R}$ tels que

$$Y = aX + b \ ,$$

alors $r = 1$ si $a > 0$, ou $r = -1$ si $a < 0$.

De façon générale, on pourra affirmer
– Si $|r| = 1$, alors il existe une relation linéaire entre X et Y.
– Si $r = 0$, il y a indépendance *linéaire* entre X et Y (mais il peut exister une autre forme de dépendance).
– $0 < |r| < 1$ traduit une dépendance linéaire d'autant plus forte que $|r|$ est grand.

Quand le coefficient de corrélation est proche de 1 ou -1, les caractères sont dits "fortement corrélés". Il faut prendre garde à la confusion fréquente entre corrélation et causalité. Que deux phénomènes soient corrélés n'implique en aucune façon que l'un soit cause de l'autre. Très souvent, une forte corrélation indique que les deux caractères dépendent d'un troisième, qui n'a pas été mesuré. Ce troisième caractère est appelé "facteur de confusion". Qu'il existe une corrélation forte entre le rendement des impôts en Angleterre et la criminalité au Japon, indique que les deux sont liés à l'augmentation globale de la population. Le prix du blé et la population des rongeurs sont négativement corrélés car les deux dépendent du niveau de la récolte de blé. Il arrive qu'une forte corrélation traduise bien une vraie causalité, comme entre le nombre de cigarettes fumées par jour et l'apparition d'un cancer du poumon. Mais ce n'est pas la statistique qui démontre la causalité, elle permet seulement de la détecter. L'influence de la consommation de tabac sur l'apparition d'un cancer n'est scientifiquement démontrée que dans la mesure où on a

pu analyser les mécanismes physiologiques et biochimiques qui font que les goudrons et la nicotine induisent des erreurs dans la reproduction du code génétique des cellules.

2. RÉGRESSION LINÉAIRE

Afin de se faire une idée sur «comment» les variables X et Y peuvent être liées, il est naturel de les représenter sur un graphique en dessinant le nuage de points (X_n, Y_n). Lorsque ce nuage semble «rectiligne», on peut envisager d'exprimer la liaison entre X et Y sous forme de fonction affine $y = ax + b$. On cherche alors la *droite de régression* qui passe «au mieux» du nuage de point.

Cette droite sera trouvée par le modèle des moindres carrés. On cherche à expliquer la variable Y à l'aide de la variable X, de sorte qu'on ait

$$Y_n = aX_n + b + \varepsilon_n \; \forall n$$

où $a \in \mathbb{R}$ et $b \in \mathbb{R}$ et les ε_n sont des variables aléatoires indépendantes et identiquement distribuées selon une loi normale (i.e. $\varepsilon_n \sim \mathcal{N}(\mu, \sigma^2) \; \forall n$).

Une illustration de régression linaire est donnée sur la figure 1.

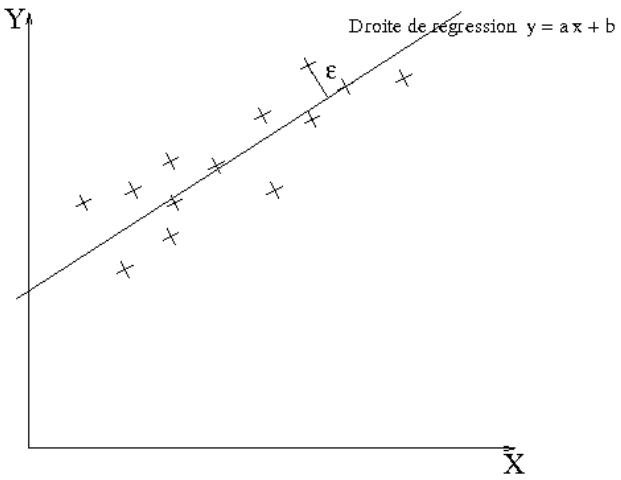

FIG. 1. Exemple d'une droite de régression.

La pente a de la droite de régression est donnée par

$$a = \frac{\mathrm{Cov}(X, Y)}{\sigma_X^2} \; ,$$

et l'ordonnée à l'origine b est donnée par

$$b = \bar{Y} - a\bar{X} \; .$$

De façon similaire, on aurait pu vouloir expliquer X à l'aide de la variable Y afin d'avoir

$$Y_n = a' X_n + b' + \varepsilon'_n \; \forall n$$

où $a' \in \mathbb{R}$ et $b' \in \mathbb{R}$ et les ε'_n sont des variables aléatoires indépendantes et identiquement distribuées selon une loi normale.

a' et b' sont alors données par :

$$a' = \frac{\text{Cov}(X,Y)}{\sigma_Y^2} \quad \text{et} \quad b' = \bar{X} - a\bar{Y} \ .$$

Il suffit en fait simplement d'inverser les rôles de X et de Y.

Proposition 3. *On pourra remarquer qu'on a*

$$r^2 = aa'$$

et

$$r = a\sqrt{\frac{\sigma_X^2}{\sigma_Y^2}} = a'\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}} \ .$$