LE COEFFICIENT DE CORRÉLATION

d'après le chapitre 10 du livre *Éléments de statistique*,
Jean-Jacques Droesbeke, Ellipses Marketing

1. Un peu d'histoire

Le coefficient de corrélation que nous allons présenter ci-dessous est l'un des paramètres majeurs de l'analyse statistique bivariée et multivariée. Son interprétation et les conditions de son utilisation méritent une attention particulère.

La formulation actuelle de ce coefficient est due à Karl Pearson qui, dans ses écrits historiques, a attribué la paternité de la notion de corrélation au physicien français Auguste Bravais sur base des travaux effectués en 1856 par ce dernier à propos de l'étude des erreurs dans les tirs d'artillerie. Pearson revint ultérieurement sur ce problème en rendant à Francis Galton ce qu'il avait attribué à Bravais.

Il n'en reste pas moins qu'actuellement, une certaine tradition s'est instaurée d'appeler le paramètre en cause "**coefficient de corrélation de Bravais-Pearson**".

2. Covariance d'une série statistique

Considérons une série statistique bivariée $\{(x_i; y_i), i = 1 \cdots, n\}$. Pour introduire le concept de coefficient de corrélation, nous allons considérer au préalable la **covariance**, définie par l'expression :

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \tag{2.1}$$

où $\overline{x}$ et $\overline{y}$ désignent respectivement la moyenne de la série marginale x et celle de y.

Remarque : Ce coefficient peut-être positif ou négatif selon la position des observations par rapport au **centre de gravité G**, de coordonnées $(\overline{x}; \overline{y})$.

3. Le coefficient de corrélation de Bravais-Pearson

Le coefficient de corrélation de Bravais-Pearson, habituellement désigné par r, est défini comme suit :

$$r = \frac{s_{xy}}{s_x s_y}, \tag{3.1}$$

où s_{xy} désigne la covariance de x et de y définie en (2.1) et s_x et s_y désignent les écarts-types des distributions marginales en x et en y définis respectivement par :

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{3.2}$$

et par :

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2. \tag{3.3}$$

4. Test d'hypothèse relatif à un coefficient de corrélation

Il est aisé d'introduire le concept de corrélation de la loi de probabilité bivariée d'un vecteur aléatoire $(X;Y)'$ par l'intermédiaire du **coefficient de corrélation de Bravais-Pearson**

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \tag{4.1}$$

où le numérateur de (4.1) est la **covariance**

$$\sigma_{xy} = \mathbb{E}[(X - \mu_x)(Y - \mu_Y)], \tag{4.2}$$

où μ_x et μ_y désignent les moyennes des distributions marginales et sont définies respectivement par :

$$\mu_x = \mathbb{E}[X] = \sum_x x p_{x.} = \sum_x x \left(\sum_y p_{xy}\right) \tag{4.3}$$

et par :

$$\mu_y = \mathbb{E}[Y] = \sum_y y p_{.y} = \sum_y y \left(\sum_x p_{xy}\right) \tag{4.4}$$

où le dénominateur de (4.1) est le produit des écarts-types marginaux σ_x et σ_y définis respectivement par :

$$\sigma_x^2 = \mathbb{E}[(X - \mu_x)^2] \tag{4.5}$$

et par :

$$\sigma_y^2 = \mathbb{E}[(Y - \mu_y)^2]. \tag{4.6}$$

Un problème d'inférence statistique peut se poser si l'on prélève dan la population définie par la loi bivariée évoquée ci-dessus, un échantillon aléatoire simple de taille n. Ce dernier se présente sous forme d'une série bivariée ou donne lieu à un tableau de contingence ; leur analyse descriptive permet de calculer **le coefficient de corrélation observé** r qui constitue ainsi **un estimateur de** ρ.

La construction d'un intervalle de confiance pour ρ ou la réalisation d'un test d'un hypothèse sont intéressantes dans le cas où la population est supposée normale. En particulier si l'on se trouve dans le cas, X et Y indépendantes si et seulement si leur coefficient de corrélation ρ est nul.

Tester l'hypothèse entre les 2 variables aléatoires X **et** Y revient dès lors à tester l'hypothèse (que nous considérons sous sa forme bilatérale, à ce propos il est conseillé de relire le rappel de cours qui a été fait sur le sujet et qui s'intitule "Les tests d'hypothèse")

$$H_0 : \rho = 0 \tag{4.7}$$

contre

$$H_1 : \rho \neq 0.$$

Si nous disposons d'un échantillon aléatoire simple d'effectif n prélevé dans la population, on peut montrer que, si H_0 est vraie, alors on a :

$$\frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}} \sim t_{n-2}, \tag{4.8}$$

où t_{n-2} représente une variable de Student à $n-2$ degrés de liberté.

Cette distribution a permis de calculer les quantiles $r_{\nu,\alpha}$ de r, que nous présentons dans la table numérotée 1 pour diverses valeurs de $\nu = n-2$ et du risque de première espèce α (l'ordre p du quantile est égale à $1 - \alpha/2$). On montre que la règle de décision du test de (4.7) est la suivante :

$$\begin{cases} RH_0 & \text{si} \quad |r| > r_{\nu,\alpha} \\ \overline{R}H_0 & \text{si} \quad |r| \leq r_{\nu,\alpha}. \end{cases}$$

Au vu de la table numérotée 1, nous constatons que si l'échantillon est petit, le coefficient r doit être relativement élevé, en valeur absolue, pour pouvoir rejeter H_0.