

MATRICE DE CENTRAGE (PAR RAPPORT A LA MOYENNE) (A3, F1)

(22 / 11 / 2019, © Monfort, Dicostat2005, 2005-2019)

La notion pratique de **matrice centrée pr à la moyenne** peut être utilisée dans plusieurs contextes (**corrélacion**, **analyse des données**, **modèle de régression**).

(i) Si $(\Omega, \mathcal{F}, P)$ est un **espace probabilisé** et $X : \Omega \mapsto \mathbf{R}^N$ un **échantillon** (aléatoire). Les N **écarts** des coordonnées X_n de X pr à leur **moyenne empirique** $\bar{X}_N = e_N' X / e_N' e_N = N^{-1} \sum_{n=1}^N X_n$ se décomposent selon :

$$(1) \quad X_n - \bar{X}_N = -N^{-1} \sum_{\alpha=1}^{n-1} X_\alpha + (1 - N^{-1}) X_n - N^{-1} \sum_{\beta=n+1}^N X_\beta ,$$

ce qui n'est autre que la n -ième coordonnée du vecteur :

$$(2) \quad Y = X - \bar{X}_N e_N = P X,$$

dans lequel $e_N = (1, \dots, 1)' \in \mathbf{R}^N$ est le **premier vecteur bissecteur** de $\mathbf{R}^N$ et P est une (N, N) -**matrice** d'élément général :

$$(3) \quad p_{\alpha\beta} = \delta_{\alpha\beta} - N^{-1}, \quad \forall (\alpha, \beta) \in \{1, \dots, N\}.$$

On dit que P est la **matrice de centrage par rapport à la moyenne** (empirique), ou la **matrice de passage aux écarts par rapport à la moyenne** (empirique). On l'appelle, plus simplement, **matrice de centrage**, ou **matrice de passage aux écarts**, ou encore **matrice des écarts**.

(ii) P vérifie les propriétés suivantes :

$$P = I_N - (e_N e_N' / e_N' e_N),$$

$$(4) \quad \begin{aligned} P' &= P \in S_N(\mathbf{R}) \quad (\text{matrice symétrique}), \\ P^2 &= P \in I_N(\mathbf{R}) \quad (\text{matrice idempotente}), \\ \text{rg } P &= N - 1 \quad (\text{matrice singulière}). \end{aligned}$$

(iii) Le **modèle de régression multiple** linéaire $y = X b + u$, avec $y (\in) \mathbf{R}^N$ et $X \in M_{NK}(\mathbf{R})$, fournit un exemple de mise en oeuvre de la notion précédente. On utilise souvent, en effet, les transformations suivantes :

$$(4) \quad y \mapsto P y = y - \bar{y}_N e_N ,$$

où $\bar{y}_N = N^{-1} \sum_{n=1}^N y_n$ (moyenne empirique des y_n), et :

$$(5) \quad X \mapsto P X = (x_1 - \bar{x}_1, \dots, x_K - \bar{x}_K),$$

où x_k désigne le k -ième vecteur colonne de X et où $x_k - \bar{x}_k = P x_k$, $\forall k \in N_K^*$ (écarts par rapport à la moyenne empirique des x_k).