MDI 220

1 Analyse statistique des données

On note Ω l'univers et $\mathcal{F}$ la tribu des événements. On considère une variable aléatoire X, appelée observation, définie sur $(\Omega, \mathcal{F})$ et à valeur dans l'espace des observations $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, ou $\mathcal{B}(\mathcal{X})$ est une tribu composée de parties de $\mathcal{X}$.

Def. Modèle statistique : famille de probabilités $\mathcal{P}$ sur $\mathcal{B}(\mathcal{X})$. Si Θ est un ensemble quelconque tel que $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ alors Θ est appelé **espace des paramètres** du modèle.

Rem. L'existence d'une paramétrisation est toujours acquise, quitte à prendre $\Theta = \mathcal{P}$.

Si Θ peut être choisi comme sous-ensemble d'un espace euclidien, le modèle est dit **paramétrique**. Si $\Theta \subset \Theta_1 \times \Theta_2$ où Θ_1 est inclus dans un espace euclidien, le modèle est dit **semi-paramétrique**.

Def. Une **statistique** est une variable aléatoire s'écrivant commme une fonction mesurable des observations, de type $\varphi(X)$ où $\varphi\colon (\mathcal{X}, \mathcal{B}(\mathcal{X})) \to (\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$ est mesurable.

Def (Identifiabilité)**.** Un modèle statistique $\mathcal{P}$ décrit par un paramètre $\theta \in \Theta$ est dit **identifiable** si $\theta \mapsto P_\theta$ est injective. Plus généralement, une fonction g de θ est dite identifiable si $(P_{\theta_1} = P_{\theta_2}) \implies (g(\theta_1) = g(\theta_2))$.

Rem. Avec $\Theta = \mathcal{P}$ on sait qu'il existe toujours au moins une paramétrisation identifiable.

Def. Un modèle statistique est dit **dominé** s'il existe une mesure positive μ sur $\mathcal{B}(\mathcal{X})$ telle que pour tout $\theta \in \Theta$, $P_\theta \in \mathcal{P}$ admette une densité de probabilité p_θ par rapport à μ.