IDENTIFICATION (B1, C4, G2)
(01 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

La notion d'**identification** est un concept central en **Statistique** : on dit qu'un **modèle statistique** est un **modèle identifiable** ssi la **propriété d'identification** est vérifiée par la **famille** des lois décrivant ce modèle. Par extension, cette notion peut aussi porter sur une **caractéristique légale** des lois en question, notamment la **relation fonctionnelle**, la **fonction de régression** ou la **fonction d'interdépendance** associées à ces lois.

Identifier une loi, au sein d'une **famille de lois** donnée, est une finalité importante de l'**inférence statistique** : cette finalité consiste à déterminer la **« vraie » loi** qui a engendré les **observations**, donc la **vraie valeur** du **paramètre**, dans le cas d'une **représentation statistique** paramétrique (ou la vraie valeur de la **caractéristique légale**).

(i) Soit $(\mathcal{X}, \mathcal{B}, P_\theta^X)_{\theta \in \Theta}$ un **modèle image**. On dit que θ' et θ'' sont des **paramètres équivalents** (du point de vue de l'**observation** X) ssi :

(1) $P_{\theta'}^X = P_{\theta''}^X$.

En notant ~ cette relation (d'équivalence), on dit que Θ est un **ensemble de paramètres identifiable** ssi :

(2) $\forall (\theta', \theta'') \in \Theta^2, \ \theta' \sim \theta'' \ \Rightarrow \ \theta' = \theta''$.

On dit aussi que la famille $(P_\theta^X)_{\theta \in \Theta}$ est une **famille de lois identifiable** ssi elle vérifie (2).

Autrement dit, l'application $\theta \mapsto P_\theta^X$ est **injective** (cf aussi **valeur identifiable d'un paramètre**), ie :

(3) $P_{\theta'}^X = P_{\theta''}^X \ \Rightarrow \ \theta' = \theta''$,

ou encore :

(4) $\theta'' \neq \theta' \ \Rightarrow \ P_{\theta''}^X \neq P_{\theta'}^X$.

(ii) Plus généralement, si $(\Gamma, \mathcal{B}_\Gamma)$ est un **espace probabilisable** auxiliaire, $\mathcal{B}_\Theta$ une **tribu** sur Θ et :

(5) $g : \Theta \mapsto \Gamma$

une application $(\mathcal{B}_\Theta, \mathcal{B}_\Gamma)$-mesurable, on dit que g est une **fonction identifiable** du paramètre θ (ou que $\tau = g(\theta)$ est identifiable) ssi :

(6) $\forall (\theta', \theta'') \in \Theta^2, \ \theta' \sim \theta'' \ \Rightarrow \ g(\theta') = g(\theta'')$.

Autrement dit, l'application $g(\theta) \mapsto P_\theta^X$ est **injective** :

(7) $P_{\theta'}^X = P_{\theta''}^X \Rightarrow g(\theta') = g(\theta'')$.

La définition (6) (resp (7)) peut se ramener à la définition (2) (resp (3)) en posant $(\Gamma, \mathcal{B}_\Gamma) = (\Theta, \mathcal{B}_\Theta)$ et $g = id_\Theta$. Elle est cependant souvent utilisée sous cette dernière forme (eg dans l'étude du **modèle d'interdépendance**) (cf **identifiabilité**).

Si $(\mathcal{Y}, \mathcal{C})$ est un espace probabilisable auxiliaire, on appelle parfois **fonction identifiante** ou **statistique identifiante**, ou encore **fonction d'identification** ou **statistique d'identification**, toute application :

(8) $S : \Theta \mapsto \mathcal{Y}$

qui est à la fois $(\mathcal{B}_\Theta , \mathcal{C})$-mesurable et tq :

(9) $\forall (\theta', \theta'') \in \Theta^2,\ \theta' \sim \theta'' \Rightarrow S(\theta') = S(\theta'')$.

Autrement dit, l'application $S(\theta) \mapsto P_\theta^X$ est **injective**. Par suite, si $(\Gamma, \mathcal{B}_\Gamma)$ est un espace probabilisable donné et $g : \Theta \mapsto \Gamma$ une application mesurable donnée, g est identifiable ssi il existe une application mesurable $h : \mathcal{Y} \mapsto \Gamma$ tq $g = h \circ S$.

(iii) Lorsque le paramètre θ contient des **caractéristiques légales**, la terminologie peut être précisée. Ainsi, on parle :

(a) d'**identification au premier ordre** s'il s'agit d'une caractéristique de **centralité** : eg **espérance** ou **paramètre de position**. Dans le cas d'une espérance, celà implique que les variables considérées soient intégrables (ie appartiennent à des espaces de type L^1) ;

(b) d'**identification au second ordre** s'il s'agit, à la fois, d'une caractéristique de centralité et d'une caractéristique de dispersion ou d'échelle (cf **paramètre d'échelle**). Dans le cas d'une espérance et d'une variance, celà implique que les variables considérées soient intégrables à l'ordre 2 (ie appartiennent à des espaces de type L^2) ;

(c) etc (**identification à l'ordre p** et espaces de type L^p).

Ainsi, le modèle de **régression** linéaire multiple standard (écrit dans l'**espace d'observation $\mathbf{R}^N$**) $y = X b + u$, avec $E u = 0$ et $V u = \sigma^2 . I_N$, n'est pas (en général) un modèle paramétrique, sauf lorsque eg $y \sim \mathcal{N}_N (X b, \sigma^2 . I_N)$. Il est « identifiable au second ordre » ssi l'application $(b, \sigma^2) \mapsto (E y, V y))$ est injective, donc ssi $(b, \sigma^2) \mapsto (X b, \sigma^2 . I_N)$ l'est, donc ssi rg X = K. Ce modèle n'est donc pas identifiable s'il est singulier (ie tq rg X < K) (cf **singularité**).

(iv) L'importance du concept d'identification s'apprécie à travers les remarques suivantes.

Soit θ^* la **vraie valeur d'un paramètre**, ie celle associée à la **loi** qui régit effectivement le **phénomène** aléatoire considéré, donc le « comportement » de l'observation X.

Si θ^* était connu (eg en cas de **simulation**), l'ensemble du phénomène serait connu, du moins « à travers » le modèle censé le représenter. La propriété selon laquelle la famille $(P_\theta^X)_{\theta \in \Theta}$ est identifiable assurerait donc, en théorie, l'identification (ie le « repérage ») de $P_{\theta^*}^X$, la « vraie » **lp** de X.

Or θ^* n'est pas, en général, connu. On ne peut (au mieux) que chercher à définir un **estimateur** (eg ponctuel) T = t (X), jugé adéquat (eg **estimateur sans biais**, **estimateur convergent**, etc), de θ^*. Par suite, en pratique, on ne peut identifier que la loi P_T^X : cette loi sera « proche », au sens d'une **topologie** définie sur $(P_\theta^X)_{\theta \in \Theta}$, de $P_{\theta^*}^X$ ssi T est « proche » (au sens du biais, ou de la convergence, etc, considérés) de la valeur θ^*.

(v) L'**identifiabilité** d'une lp assure donc l'unicité de sa détermination approchée lorsqu'on connaît seulement des estimateurs de son paramètre.

La notion se relie aussi à celle de **robustesse** et à celle de **spécification de modèle**.