MODÈLE STATISTIQUE (G2, J, K, L, M, N)
(03 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

C'est sur la notion de **modèle statistique** que sont fondées la plupart des **procédures statistiques**. En effet, cette notion permet de définir le concept central de **problème statistique**, qui est la base de la **théorie de la décision** statistique (cf **décision statistique**).

Le vocable de modèle statistique n'est pas nécessairement le plus approprié, dans la mesure où il suggère une formalisation « idéale » ou un schéma « de référence », alors qu'il ne s'agit que d'une **spécification** visant à représenter au mieux la **structure**, le fonctionnement et l'évolution (déroulement) d'un **phénomène** (cf **système**, **niveau, répartition, évolution**). Les vocables de **structure statistique** ou de **représentation statistique** semblent préférables.

(i) Soit Ω un **ensemble** abstrait, $\mathcal{T}$ une **tribu de parties** de Ω et $\mathcal{P}$ une **famille**, a priori quelconque, de **mesures de probabilité** définies sur $\mathcal{T}$.

On appelle **modèle statistique**, ou **représentation statistique**, ou **structure statistique**, ou parfois encore **espace statistique**, le triplet $(\Omega, \mathcal{T}, \mathcal{P})$ dans lequel :

(a) Ω représente l'**ensemble fondamental** des « éventualités » (ou « **événements simples** »). Les éléments $\omega \in \Omega$ peuvent être des **unités statistiques** (individus, etc), ou des **événements** possédant un intérêt particulier, que l'on cherche à analyser ;

(b) $\mathcal{T}$ s'interprète comme la famille de tous les événements (simples ou complexes) auxquels s'intéresse le **statisticien** ;

(c) $\mathcal{P}$ désigne une famille de probabilités P possibles (probabilités « candidates ») : chacune d'elles indique la façon selon laquelle les événements (observations, résultats) sont générés dans $\mathcal{T}$ (ie à partir de Ω).

Le modèle précédent, qui se note aussi $(\Omega, \mathcal{T}, \mathcal{P})_{P \in \mathcal{P}}$, est appelé **modèle de base**, ou **modèle fondamental**, ou **modèle sous-jacent**, ou encore **modèle initial**.

(ii) En pratique, il est souvent commode (ou le statisticien est contraint) de faire référence à un autre modèle, dont la forme s'impose d'emblée, ou qui correspond à ce qu'il est possible d'observer (cf **observabilité**, **observation**).

Soit alors $(\mathcal{X}, \mathcal{B})$ un **espace mesurable** (ici un **espace probabilisable**) donné, appelé **espace d'observation**, ou **espace d'échantillonnage**, et $X : \Omega \mapsto \mathcal{X}$ une **variable aléatoire** ou une **statistique** (eg un **échantillon** de résultats $X(\omega)$ observés sur les unités $\omega \in \Omega$). A tout événement élémentaire $\omega \in \Omega$ la variable X

associe donc un élément $x = X(\omega) \in \mathcal{X}$, couramment appelée **observation** générée par le modèle. Ceci suppose le modèle « adapté », ie correctement spécifié (cf **spécification**).

On suppose que $\mathcal{B}$ est une tribu de parties de $\mathcal{X}$ (eg la **tribu engendrée** par $\mathcal{T}$ et X). Cette tribu est « probabilisée » par une famille $\mathcal{P}^X$ de **lois de probabilité** P^X, définies sur $\mathcal{B}$, qui sont resp les images des probabilités P par X : en effet, par définition d'une **mesure image** (ie d'un **transport de mesure**), on a $P^X = X(P)$, $\forall P \in \mathcal{P}$.

Chacune de ces lois P^X est une loi possible (ou loi « candidate ») pour X : elle est ainsi susceptible de régir le comportement de la **variable observable** X.

On appelle alors **modèle image**, ou **représentation image**, ou **modèle d'observation**, ou **structure image**, ou **modèle apparent**, ou **modèle dérivé**, ou encore **modèle final**, le modèle $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$ ainsi défini.

(iii) Souvent, on « oublie » le modèle de départ et l'on raisonne directement sur le seul modèle image. Mais ceci n'est pas toujours le cas et l'on peut avoir à raisonner sur les deux modèles (eg en **théorie des sondages** ou en **théorie des plans d'expérience**) : ainsi, le modèle de base peut décrire un tirage d'unités statistiques, et le modèle d'observation décrire une **distribution** statistique possible des **variables** mesurées sur ces unités (cf **mesure**).

Les deux types de modèles (initial et final) se confondent lorsqu'on pose, d'emblée, $(\mathcal{X}, \mathcal{B}) = (\Omega, \mathcal{T})$ et $X = id_\Omega$ (identité de Ω), car le second s'identifie alors au premier (et vice versa).

Souvent aussi, le cadre général précédent n'est pas toujours entièrement explicité en pratique : ainsi, lorsqu'il traite un **modèle de régression**, le **statisticien** considère directement l'équation de régression et les hypothèses stochastiques associées, sans se référer nécessairement au schéma « théorique » (ou probabiliste) de base dans lequel ce modèle s'insère (cf **relation fonctionnelle**, **régression**).

(iv) Les notions précédentes sont présentées dans un cadre « non paramétré » (cf **modèle paramétrique**) : on parle alors de **modèle statistique non paramétré** (resp **non paramétrique**), ou de **structure statistique non paramétrée** (resp **non paramétrique**), etc.

D'autres « contextes » conduisent à des présentations adaptées. Ainsi :

(a) la notion de **modèle (statistique) paramétré**, ou de **représentation (statistique) paramétrée**, ou encore de **structure (statistique) paramétrée** est une notion équivalente à la précédente. En effet, on peut toujours paramétrer un modèle statistique (eg image) $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$ avec la famille $\mathcal{P}^X$ elle-même, et ceci définit la notion de **modèle indicé**, ou **modèle indexé** : la **loi** P^X joue le rôle d'**indice** pour la

famille $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$, qui peut ainsi s'écrire $(\mathcal{X}, \mathcal{B}, P^X)_{(PX \in \mathcal{P}X)}$ ou encore $(\mathcal{X}, \mathcal{B}, P^X : P^X \in \mathcal{P}^X)$;

(b) dans certaines **situations statistiques**, $\mathcal{P}^X$ peut être indexée par un ensemble Θ selon $(P_\theta^X)_{\theta \in \Theta}$ en sorte que la connaissance du **paramètre** θ entraîne l'entière connaissance de P_θ (cas d'une **famille de lois identifiable**) : on définit ainsi la notion de **modèle paramétrique**, ou de **représentation (statistique) paramétrique**, ou encore de **structure (statistique) paramétrique** ;

(c) enfin, dans certains cas, la famille $\mathcal{P}^X$ est seulement « partiellement » paramétrique, ce qui conduit à la notion de **modèle semi-paramétrique**.

Lorsque chacune des lois P^X ou P_θ^X du modèle admet une **densité de probabilité** pr à une **mesure positive** (**mesure σ-finie**) μ donnée (cf **famille de lois dominée**), définie sur $\mathcal{B}$, on définit un **modèle dominé**. Ce type de modèle est fréquemment utilisé :

$(c)_1$ parce que la notion, corrélative, de **fonction de vraisemblance** conduit à une méthode statistique très importante (**méthode du mv**, **méthode des tamis**) (cf **vraisemblance**) ;

$(c)_2$ en raison de la commodité de certains calculs.

(v) En pratique, une situation courante est la suivante : $\mathcal{X} = \mathbf{R}^N$, $X = (X_1, ..., X_N)$ est un **échantillon** (ou une **statistique**) à valeurs dans $\mathbf{R}^N$ et, lorsque $\mathcal{P}^X = (P_\theta^X)_{\theta \in \Theta}$, l'ensemble Θ des valeurs du paramètre θ du modèle est une partie (eg un **ouvert** ou un **compact**) de $\mathbf{R}^Q$.

A titre d'exemple, si $\mathcal{X} = \mathbf{R}^N$ et $\Theta = \mathbf{R}^N \times S_N^+(\mathbf{R})$, où $S_N^+(\mathbf{R})$ désigne l'ensemble des **matrices symétriques** réelles semi-définies positives (cf **matrice définie positive**), on définit le **modèle gaussien**, ou **modèle normal**, avec $X \sim P_\theta^X = \mathcal{N}_N(\mu, \Sigma)$ et $\theta = (\mu, \Sigma)$.

(vi) Dans le cas paramétré, si $(\mathcal{X}, \mathcal{B}, P_\theta^\xi)_{\theta \in \Theta}$ désigne le « vrai » modèle statistique gouvernant l'observation d'un **phénomène** (**observable**), ie le modèle sans **erreur de spécification**, on doit distinguer trois notions :

(a) la **valeur « courante » θ du paramètre**, qui sert généralement dans les calculs analytiques ;

(b) la « **vraie » valeur** (en général inconnue) **du paramètre** $\theta^* \in \Theta$ (cf **vraie valeur d'un paramètre**), qui définit la **loi** gouvernant réellement le phénomène étudié, ie la loi que la **Nature** a « décidé » d'attribuer à ce phénomène (cf **loi scientifique**) ;

(c) l'**estimation** $\tilde{\theta} \in \Theta$ **de la vraie valeur** $\theta^* \in \Theta$, qui est la valeur prise par une fonction des observations (ou **statistique**) et qui vérifie certaines conditions (eg une **équation estimante**). La valeur ainsi estimée, ou **estimation**, est diversement notée, eg :

$(c)_1$ d'une façon « neutre », selon le symbole attribué à l'**estimateur** considéré : eg S_N ou T_N , etc, pour une estimation quelconque ;

$(c)_2$ en fonction de la méthode qui définit l'estimateur : eg $\hat{\theta}$ ou $\tilde{\theta} \in \Theta$ pour une **méthode de moindres carrés**, etc.

On indique souvent la taille de l'échantillon qui le définit (eg S_N ou $\tilde{\theta}_N \in \Theta$ pour un N-échantillon), notamment dans l'étude des **propriétés asymptotiques** des **statistiques** considérées.

(vii) Ainsi, avec un **modèle de régression linéaire** standard $y = X b + u$, avec $E\ y / X = 0$ et $V\ y / X = \sigma^2 . I_N$, on devrait noter la « vraie » équation de régression du modèle, telle qu'elle est « observée », selon $y = X b^* + u$, car c'est la vraie valeur $b^* \in \mathbf{R}^K$ qui définit la **perturbation aléatoire** u en fonction des observations (X, y). Si $b^* \in \mathbf{R}^K$ est estimée par la **méthode des moindres carrés ordinaires**, l'estimation obtenue $\hat{b} = (X' X)^{-1} X' y$ définit un **estimateur des mco** $\hat{b} = t_N (X, y)$, calculé en minimisant pr à b la **forme quadratique** habituelle $q (b) = ||u||^2 = ||y - X b||^2$.

Les notations sont souvent « flottantes » : la valeur estimée est parfois notée β, ou encore la vraie valeur est notée β et son estimation b, etc.

(viii) Un modèle statistique concret est souvent de nature complexe (cf **complexité**) :

(a) ainsi, les **observations** X peuvent posséder des particularités ou une **structure** explicitement prise en compte dans le modèle : eg **observations manquantes** (cf **lacune**), observations aberrantes (cf **aberration**), **contrainte sur les variables** ou **contrainte sur les observations**, traitement dissymétrique de deux sous-ensembles de variables, les unes considérées comme endogènes, les autres comme exogènes (cf **modèle de régression**, **modèle d'interdépendance**, **variable exogène**, **variable endogène**, etc) ;

(b) par ailleurs, les **paramètres** du modèle peuvent être traités de façon plus ou moins simple : **contrainte sur les paramètres**, paramètres aléatoires ou **restrictions a priori** sur les paramètres en **théorie bayésienne**, paramètres évoluant en fonction de certaines variables (cf **modèle avec coefficients variables**), etc (cf aussi **classification des modèles**).

Cette variété de **formalisation** (cf **modélisation**, **spécification**) reflète la diversité des situations rencontrées en pratique. Mais la modélisation de ces situations se ramène généralement au (ou peut être considérée comme « plongée » dans le) schéma précédent, qui possède ainsi une grande universalité (*« versatilité »* des anglo-saxons).

Ainsi, un **espace probabilisé** $(\mathcal{X}, \mathcal{B}, P^X)$ peut être considéré comme comme un modèle statistique particulier : celui où la famille $\mathcal{P}^X$ des probabilités se réduit à une probabilité unique $\{P^X\}$. Cette loi est :

(a) parfois connue : eg **méthodes de MONTE CARLO**, **simulation**, étude d'une loi donnée (eg à partir de sa dérivée, ou de certaines de ses propriétés mathématiques) ;

(b) parfois inconnue : eg **calcul des probabilités** en général, lorsqu'il n'y a pas nécessité de spécification d'une loi particulière.

(ix) Certaines parties de l'analyse **Statistique** (eg **Statistique descriptive**, **analyse des données**) sont parfois qualifiées de « **statistique sans modèle** » : l'approche adoptée dans ces contextes est réputée ne pas nécessiter l'existence (ou la pré-existence) d'un modèle tq le modèle de base $(\Omega, \mathcal{T}, \mathcal{P})$ ou le modèle dérivé $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$.

Ces parties s'attachent, de façon générale, à analyser directement les données observées : soit description, soit caractérisations diverses (classes et **classification**, **facteurs** sous-jacents, etc). Autrement dit, ces types d'analyse se fondent directement sur $\mathcal{X}$, voire sur $(\Omega, \mathcal{X})$, et en appliquent divers calculs algébriques (notamment d'algèbre linéaire) ou analytiques (notamment avec des concepts de « **distance** », **lissage**, **interpolation**, etc).

Cependant :

(a) d'une part, l'approche « sans modèle » peut, de façon simple, être « plongée » dans l'approche « avec modèle », puisque les notions principales d'unités statistiques et de variables statistiques leur sont communes. En outre, souvent la première précède, logiquement, la seconde ;

(b) d'autre part, ce plongement peut préparer, non seulement la **modélisation** statistique, mais aussi la **décision statistique**. Même dans une approche de type « descriptive », des choix sont à opérer ainsi que divers calculs d'**optimisation**.