SPÉCIFICATION (G2, I2)
(23 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

De façon générale, la **modélisation** d'un **phénomène** observable, donc sa **représentation statistique**, constitue, en elle-même, une hypothèse a priori. Autrement dit, toute « théorie » n'est qu'une hypothèse, car elle peut être modifiée ou infirmée par une autre théorie.

En effet, la loi de gouvernant un phénomène étant inconnue, la nécessaire **spécification** (ie **modélisation** ou formalisation) d'un **modèle** constitue, en elle-même, une **hypothèse**, souvent cruciale (cf **relation fonctionnelle**, **régression**).

En particulier, lorsqu'une **procédure statistique** (eg **estimation**, **test d'hypothèses**), est mise en oeuvre, c'est généralement dans un cadre défini par le modèle considéré.

(i) Chaque **domaine de connaissance** étudié par un **homme de l'art** comporte trois éléments fondamentaux :

(a) des **phénomènes particuliers**, généralement considérés comme aléatoires (cf **système aléatoire**) ;

(b) un **dispositif d'observation** de ces phénomènes (**dispositif expérimental**, etc), dont le but est d'en « mesurer » les « manifestations » (cf **système d'observation**, **production statistique**). Ceci se traduit par l'élaboration de **données** (**observations**) relatives à un certain nombre de **variables**, qui sont des **variables d'intérêt** décrivant ces phénomènes ;

(c) un **modèle théorique** censé représenter (ou décrire) le fonctionnement de chaque phénomène. Ce modèle théorique est généralement suggéré par la **simple description** du phénomène observé, ou encore par **association d'idées** avec des phénomènes de même nature (phénomènes comparables, ou « transposables ») (cf **représentation statistique**).

(ii) Le phénomène aléatoire peut (au moins conceptuellement) être représenté par une **expérience aléatoire**, ie par un **espace probabilisable** $(\Omega, \mathcal{T})$. Il peut posséder des caractéristiques particulières :

(a) **définition totale ou partielle**. En effet, tous les **événements** aléatoires $A \in \mathcal{T}$ peuvent, pour des raisons diverses, n'être pas connus ou n'être pris en compte que partiellement. Ceci revient à admettre, le plus souvent, à tort ou à raison, que ces événements sont négligeables pour la compréhension du phénomène.

Une autre démarche consiste à n'étudier que les événements A entièrement définis. Les autres évènements, plus « vagues », peuvent difficilement être mobilisés.

On peut aussi adopter un point de vue bayésien : la **famille** $\mathscr{P}$ des **lp** qui gouvernent le phénomène aléatoire $(\Omega, \mathscr{T})$ est alors muni d'une **tribu de parties** sur laquelle est définie une mesure de **probabilité a priori**. Un cas typique est celui où $\mathscr{P} = (P_\theta)_{\theta \in \Theta}$ et où la tribu $\mathscr{B}_\Theta$ de parties de Θ considérée est munie de la mesure a priori Π, parfois aussi appelée **mesure de BAYES** ;

(b) **observabilité totale ou partielle** (cf **censure**, **observable**, **troncature**). En effet, certains événements $A \in \mathscr{T}$ peuvent, par nature, ne pas être observables. Dans le cas d'observation partielle, l'observateur subit un **biais de sélection** portant sur les événements à étudier. Moyennant certaines hypothèses, certaines procédures statistiques peuvent être adaptées à cette situation ;

(c) **contrôle total ou partiel** (cf **production statistique**). Certains modes d'observation des phénomènes s'effectuent sans possibilité de choisir les modalités d'observation : eg phénomènes non répétitifs, ou observés sans méthodologie statistique appropriée au préalable. Par contre, la **théorie des plans d'expérience** et la **théorie des sondages** constituent les deux méthodes de base dans lesquelles le **statisticien** cherche à contrôler au mieux les modalités d'observation afin d'obtenir des données utilisables : dans ce sens, on peut dire que ces théories définissent des **modes de construction (ou de génération) des données** (cf aussi **processus générateur de données**) ;

(d) **répétitivité totale, partielle ou impossible** (unicité) (cf **répétition**). L'importante **propriété de répétitivité** correspond à la possibilité de renouveler une expérience aléatoire ou un sondage « à l'identique » (cf **renouvellement**) ; un objectif visant à obtenir des données indépendantes et équidistribuées correspond à un exemple très courant de répétitivité expérimentale (cf **échantillon indépendant équidistribué**). A l'autre extrême, un phénomène non répétitif (eg à caractère historique) nécessite, pour son étude scientifique, de se référer :

$(d)_1$ soit au cadre général de la **théorie des processus**. Dans ce cas, les résultats de la **théorie ergodique** ou les propriétés de **coïntégration** jouent un rôle fondamental : ainsi, lorsqu'on ne peut observer qu'une seule **trajectoire** d'un processus, la propriété d'**ergodicité** permet néanmoins d'en inférer, une fois testée, des propriétés statistiques utilisables ;

$(d)_2$ soit au cadre bayésien (cf **théorie bayésienne**). Dans ce cas, on dispose des observations (ou résultats) d'une expérience unique, et l'on adopte une certaine **probabilité a priori** : par construction, on considère ainsi une seule expérience aléatoire dont la famille de lois dépend d'un **paramètre** aléatoire, paramètre dont la loi n'est autre que cette probabilité a priori.

(iii) Le modèle théorique peut, au sens le plus large, être représenté par la famille $\mathscr{P}$ des **mesures de probabilité** susceptibles de « gouverner » le phénomène aléatoire. En général, on ne considère pas toutes les mesures de probabilité P pouvant être définies sur $\mathscr{T}$. On se restreint eg :

(a) à des probabilités dont l'image par une **va** donnée est absolument continue pr à une mesure (positive) donnée : eg **mesure discrète**, **mesure de LEBESGUE** (cf **famille de lois dominée**, **fonction absolument continue**) ;

(b) ou à une **famille de lois** de lois $(P_\theta)_{\theta \in \Theta}$ (avec eg $\Theta \subset \mathbf{R}^Q$) ;

(c) ou seulement à des **caractéristiques** intéressantes de ces lois, définies à travers leur paramètre θ (cf **paramètre principal**).

(iv) En pratique, une caractéristique importante de ces lois est leur **relation fonctionnelle** (dans le cas général), ou leur fonction de **régression** (dans le cas numérique), qui permettent d'aborder les problèmes de **causalité**.

Ainsi, en supposant donnés deux **espaces d'observation** $(\mathcal{X}, \mathcal{B})$ et $(\mathcal{Y}, \mathcal{C})$ et un **couple aléatoire** $\zeta = (\xi, \eta) : \Omega \mapsto \mathcal{X} \times \mathcal{Y}$, dont la **loi conjointe** est P^ζ, la **fonction de régression** (supposée définie) $x \mapsto \varphi(x) = E(\eta / \xi = x)$ est une façon commode très utilisée pour représenter un « modèle » (au sens où l'entend souvent l'**homme de l'art**). Ce modèle relie, ici au sens de l'**espérance** (ie de façon « moyenne » ou « centrale »), la liste η des **variables endogènes** à la liste ξ des **variables exogènes**.

En supposant que les opérations aient un sens, on peut, plus généralement, choisir comme **caractéristique légale** la fonction d'**interdépendance**, qui s'associe naturellement à la notion de **modèle d'interdépendance**.

(v) D'un point de vue terminologique :

(a) une **spécification au sens large** est définie par l'écriture d'un modèle statistique particulier $(\Omega, \mathcal{T}, \mathcal{P})$;

(b) dans un sens plus étroit, une **spécification** est souvent définie par la seule famille $\mathcal{P}$ des **lp**, ou seulement par un ensemble **caractéristique** Γ de cette famille.

Si $(\mathcal{X}, \mathcal{B}, \mathcal{P}^X)$ est le **modèle image** du précédent par une va $X : \Omega \mapsto \mathcal{X}$, on applique la même terminologie aux éléments analogues.

(vi) Les observations supposées générées par les modèles précédents permettent de définir des **tests de spécification**. Ces tests ont pour objet de valider l'**hypothèse de base** suivante : le « vrai » modèle (ie la « bonne » spécification du modèle) $\mathcal{P}^X$ appartient à une sous-famille donnée ${\mathcal{P}_0}^X \subset \mathcal{P}^X$ (avec ${\mathcal{P}_0}^X \neq \varnothing$). L'**alternative** peut être une **hypothèse d'emboîtement** (${\mathcal{P}_a}^X \subset {\mathcal{P}_0}^X$, ou inversement), une **hypothèse de séparation** (ou **disjonction**) (${\mathcal{P}_a}^X \cap {\mathcal{P}_0}^X = \varnothing$), ou encore une **hypothèse d'orthogonalité** (si les espaces d'observation sont des **espaces vectoriels mesurables** orthogonaux entre eux).

Si la spécification consiste en une **régression** (spécifié dans l'**espace des variables**), celle-ci définit eg un **modèle de régression multiple** non linéaire (dans l'**espace des observations**) $y = F(b) + u$, assorti d'hypothèses usuelles tq $E(y / X) = F(b)$ et $V(y / X) = \sigma^2 I_N$. On teste alors le modèle $y = F_0(b) + u_0$ contre le modèle $y = F_a(b) + u_a$ (où F_0 et F_a sont des fonctions analytiquement données), ie la spécification F_0 contre la spécification F_a .

(vii) La **spécification** d'un modèle peut, en théorie, être élaborée par l'homme de l'art et « communiquée » ensuite au statisticien, du moins si cette spécification n'est pas trop « floue ». La qualité de l'**adéquation** aux données procurée par des modèles alternatifs permet de retenir certains d'entre eux comme plausibles, avec des degrés de vraisemblance divers.

En général, un **test de spécification** n'indique pas comment spécifier un modèle, mais seulement quelles spécifications semblent préférables à d'autres, compte tenu des données. On l'appelle aussi **test de sélection de modèles**, ou **test de choix entre modèles** : cf eg **test de D.R. COX**, **test de J.A. HAUSMANN**, test de M.H. PESARAN, test de G. SCHWARTZ, etc.

(viii) Dans un cadre plus large, on appelle **spécification** d'un modèle la donnée, non seulement du modèle $(\Omega, \mathcal{T}, \mathcal{P})$ lui-même, mais aussi d'un **espace de décision** $(D, \mathcal{B}_D)$ et d'une **fonction de perte** L. En ce sens, spéficier un modèle signifie spécifier (ou « poser ») un **problème de décision** statistique.

Deux spécifications distinctes $(\Omega', \mathcal{T}')$ et $(\Omega'', \mathcal{T}'')$ d'un même phénomène aléatoire, ou d'une théorie relative à ce phénomène, conduisent à deux modèles $(\Omega', \mathcal{T}', \mathcal{P}')$ et $(\Omega'', \mathcal{T}'', \mathcal{P}'')$ distincts. La comparaison des modèles nécessite souvent eg de « plonger » ces modèles dans un modèle global $(\Omega, \mathcal{T}, \mathcal{P})$, avec $\Omega = \Omega' \cup \Omega''$, $\mathcal{T} = \sigma(\mathcal{T}' \cup \mathcal{T}'')$ (tribu engendrée par l'ensemble des événements aléatoires), $\mathcal{P}$ étant la famille des probabilités définies sur $\mathcal{T}$ et dont les restrictions à $\mathcal{T}'$ et $\mathcal{T}''$ sont resp des éléments de $\mathcal{P}'$ et de $\mathcal{P}''$ (cf **plongement**).

Une telle approche permet de définir des tests de spécification (parmi ceux cités ci-dessus) et peut s'étendre au cadre des problèmes de décision statistique précédents (cf aussi **robustesse**). La comparaison porte, dans ce cas, entre le problème de décision $\{(\Omega', \mathcal{T}', \mathcal{P}'), (D', \mathcal{B}_{D'}), L'\}$ et le problème de décision $\{(\Omega'', \mathcal{T}'', \mathcal{P}''), (D'', \mathcal{B}_{D''}), L''\}$.

Il importe de noter que le **choix d'une spécification** implique des conséquences pour la mise en oeuvre de certaines **procédures statistiques** : ainsi, en matière de **prévision**, on peut étudier l'influence d'une **erreur de spécification** sur une méthode de prévision (**robustesse** du prédicteur choisi pr aux hypothèses retenues).