HYPOTHÈSE STATISTIQUE (I)
(29 / 07 / 2021, © Monfort, Dicostat2005, 2005-2021)

« Hypotheses non fingo » (je n'admets pas d'hypothèse non soutenue par l'observation ») (Isaac NEWTON)
« Science n'est qu'hypothèse »

(i) L'expression d'« **hypothèse statistique** », ou parfois d'« **hypothèse probabiliste** », implique une distinction entre :

(a) d'une part, l'**ensemble** des présupposés qui précèdent ou entourent la mise en oeuvre d'une **procédure statistique**. En effet, toute procédure suppose l'acceptation ex ante de conditions qui peuvent fonder sa validité, eg :

$(a)_1$ **indépendance** ;

$(a)_2$ **homogénéité** (cf eg **homoscédasticité**) ;

$(a)_3$ **ergodicité** ;

$(a)_4$ **linéarité** d'une **caractéristique conditionnelle** (cf eg **régression linéaire**) ;

$(a)_5$ **normalité** (ou **normalité asymptotique**) de la **famille** des **lois** d'un **modèle** (cf eg **famille de lois**), etc.

Ce type d'hypothèses est de même nature que celui qui intervient dans la démonstration d'une propriété mathématique (cf **théorème**).

En effet, le **calcul des probabilités** et la **Statistique** formant des branches des mathématiques, les hypothèses « purement statistiques » précédentes sont donc principalement de nature probabiliste. On peut cependant les distinguer des hypothèses « purement mathématiques » (eg **continuité**, **différentiabilité**, **linéarité**, etc) ;

(b) d'autre part, la nécessité d'une validation (confirmation ou infirmation) des hypothèses (au sens précédent) à partir de l'**observation** du **phénomène** sous examen : ainsi, admettre une hypothèse d'indépendance, alors que les données sont autocorrélées temporellement, revient à « refuser » une procédure candidate.

Ce deuxième sens de l'expression « **hypothèse statistique** » concerne donc les propriétés de la **loi de probabilité** supposée régir le phénomène.

(ii) On note souvent un **modèle** sous une forme paramétrée $(\Omega, \mathcal{T}, P_\theta)_{\theta \in \Theta}$ retenue pour décrire le phénomène à analyser. L'ensemble fondamental Ω, donné a priori, est muni d'une **tribu de parties** $\mathcal{T}$ sur laquelle on définit une **(mesure de) probabilité** P_θ susceptible de générer des **observations** (cf **loi scientifique**).

Cette probabilité constitue une famille $(P_\theta)_{\theta \in \Theta}$ lorsque θ parcourt l'ensemble de ses valeurs Θ.

Une **hypothèse statistique** est alors définie comme une proposition de la forme $\theta \in \Lambda$, où $\Lambda \subset \Theta$ est une partie (non vide) donnée de l'ensemble des **paramètres** Θ, et qui est en général distincte de Θ. Une hypothèse est généralement notée selon :

(1) $H : \theta \in \Lambda$.

Il est équivalent de supposer que « H est vraie » ou de supposer que $\theta \in \Lambda$. Souvent (eg dans la **théorie des tests** de **NEYMAN-PEARSON**), H est une **hypothèse privilégiée** et on la note plutôt H_0 (Λ étant alors notée Θ_0). On l'appelle alors **hypothèse de base**, ou **hypothèse fondamentale**, ou même, dans certains contextes (**hypothèse linéaire**, etc), **hypothèse nulle** (d'où la notation H_0).

(iii) On peut aussi considérer des modèles formalisés sous forme non paramétrée (Ω, $\mathscr{T}$, $\mathscr{P}$), dans laquelle $\mathscr{P}$ est une famille a priori quelconque de probabilités P définies sur $\mathscr{T}$. On appelle alors **hypothèse statistique** toute partie non vide $\mathscr{L} \subset \mathscr{P}$ en général distincte de $\mathscr{P}$, et l'on note :

(2) $H : P \in \mathscr{L}$.

Si $\mathscr{L} = \mathscr{P}_0$ est une partie privilégiée de $\mathscr{P}$, on appelle **hypothèse privilégiée**, ou **hypothèse fondamentale**, voire **hypothèse nulle**, l'hypothèse de base $H = H_0$ associée à $\mathscr{P}_0$.

(iv) Par suite, une hypothèse statistique peut concerner :

(a) le paramètre $\theta \in \Theta$ d'une famille $(P_\theta)_{\theta \in \Theta}$ paramétrée par un ensemble Θ ;

(b) une probabilité $P \in \mathscr{P}$ d'une famille quelconque $\mathscr{P}$.

Cependant, entre ces deux « extrêmes », toute **caractéristique légale** associée aux éléments d'une famille $(P_\theta)_{\theta \in \Theta}$ ou $\mathscr{P}$ peut aussi faire l'objet d'une hypothèse plus ou moins complexe : simple paramètre scalaire, paramètre vectoriel, paramètre fonctionnel (densité, fr, fc, **relation fonctionnelle**, **fonction de régression** ou **fonction d'interdépendance**, etc).

(v) On dit parfois que l'hypothèse $H : \theta \in \Lambda$ est :

(a) une **hypothèse paramétrique**, ou parfois une **hypothèse paramétrée**, si θ est de nature numérique, ou encore dans le cas où l'on peut écrire $\mathscr{P} = (P_\theta)_{\theta \in \Theta}$ (Θ est alors une partie d'un **espace vectoriel** réel de dimension finie) ;

(b) une **hypothèse non paramétrique** (ou une **hypothèse non paramétrée**) si θ n'est pas de nature numérique (ie Θ ne peut s'identifier à une partie d'un **espace vectoriel** réel de dimension finie), ou encore dans le cas où $\mathscr{P}$ ne peut pas être explicitement indexée (ou indexable) par une telle partie Θ (ie on ne peut l'écrire sous la forme $\mathscr{P} = (P_\theta)_{\theta \in \Theta}$).

(vi) En privilégiant une hypothèse $H_0 : \theta \in \Theta_0$ (resp $H_0 : P \in \mathscr{P}_0$), le **statisticien** peut, selon la **situation statistique** rencontrée, vouloir la « mettre en concurrence » :

(a) avec une hypothèse non « spécifiée », ie « vaguement » décrite, à savoir l'**hypothèse complémentaire** $H_1 : \theta \in \Theta_1$, avec $\Theta_1 = \Theta_0^c = \Theta \setminus \Theta_0$ (resp $H_1 : P \in \mathscr{P}_1$, avec $\mathscr{P}_1 = \mathscr{P}_0^c = \mathscr{P} \setminus \mathscr{P}_0$). Cette hypothèse est parfois aussi appelée « **hypothèse omnibus** » ou « **hypothèse résiduelle** » (même si elle possède un « consistance » importante) ;

(b) avec une hypothèse précise particulière $H_a : \theta \in \Theta_a$, avec $\Theta_a \subset \Theta_0^c$ et $\Theta_a \neq \Theta_a^c$ (resp $H_a : P \in \mathscr{P}_a$, avec $\mathscr{P}_a \subset \mathscr{P}_0^c$ et $\mathscr{P}_a \neq \mathscr{P}_0^c$). Une hypothèse tq H_a (souvent aussi notée H_1) est dite **hypothèse alternative**, ou **hypothèse non privilégiée**, ou même **hypothèse secondaire** (même si elle possède un intérêt en termes de candidature alternative).

(vii) On appelle **test significatif** un test dont la procédure de test (ie la **règle de décision** associée) conclut à l'acceptation de l'hypothèse de base H_0 , ie conduit à la décision $\theta \in \Theta_0$ (resp $P \in \mathscr{P}_0$). On tient alors l'hypothèse de base pour valide, ou vérifiée, ou « vraie » : cette hypothèse résume l' « **état de la question** » à un moment donné. Cette attitude du scientifique est temporaire, et se justifie, faute de mieux, en l'absence d'**information** supplémentaire : nouvelle observation, nouvelle hypothèse (ie théorie).

(viii) Dans le cadre d'un **modèle** de la forme $(\Omega, \mathscr{T}, \mathscr{P})$, lorsqu'on teste une hypothèse $H_0 : P \in \mathscr{P}_0$ contre une alternative $H_a : P \in \mathscr{P}_a$, on dit que :

(a) H_0 et H_a sont des **hypothèses disjointes** ssi $\mathscr{P}_0 \cap \mathscr{P}_a = \varnothing$;

(b) H_0 et H_a sont des **hypothèses emboîtées** ssi $\mathscr{P}_0 \subset \mathscr{P}_a$ (ou inversement), avec $\mathscr{P}_0 \neq \mathscr{P}_a$. Dans ce dernier cas, on dit que le modèle $(\Omega, \mathscr{T}, \mathscr{P}_0)$ est une **spécification particulière**, ou une **spécification plus précise**, du modèle considéré $(\Omega, \mathscr{T}, \mathscr{P})$.

(ix) Soit (Ω, $\mathscr{T}$, $P_\theta)_{\theta \in \Theta}$ un modèle dans lequel Θ est une variété linéaire d'un espace vectoriel réel donné (cf **variété affine**), et où Θ est muni d'un **produit scalaire**. Si Θ_0 et Θ_a sont deux sous-variétés orthogonales de Θ (cf **orthogonalité**), les hypothèses $H_0 : \theta \in \Theta_0$ et $H_a : \theta \in \Theta_a$ (avec $\Theta_0 \perp \Theta_a$) sont appelées **hypothèses orthogonales**. Leur intersection se réduit à l'élément neutre (nul).

(x) La **théorie des tests** a pour objet l'étude du choix entre hypothèses : eg acceptation ou rejet d'une hypothèse de base, ou encore indécision entre diverses hypothèses testées. On peut alors remarquer :

(a) que l'écriture d'un modèle tq (Ω, $\mathscr{T}$, $P_\theta)_{\theta \in \Theta}$ ou (Ω, $\mathscr{T}$, $\mathscr{P}$) constitue, en elle-même, une hypothèse (cf **spécification**). Le choix de la **population** Ω ou des **événements** éléments de $\mathscr{T}$ peut être plus ou moins « extensif ». De même, soit ($\mathscr{X}$, $\mathscr{B}$, $(P_\theta^X)_{\theta \in \Theta}$) (resp ($\mathscr{X}$, $\mathscr{B}$, $\mathscr{P}^X$)) le **modèle image** du premier (resp second) précédent par une va $X : \Omega \mapsto \mathscr{X}$, où ($\mathscr{X}$, $\mathscr{B}$) désigne un **espace d'observation** donné. Lorsque X est constitué des observations d'une « liste » ξ composée de K variables (ξ_1 ,..., ξ_K), l'omission, ou l'adjonction, de certaines variables modifie le modèle : la nature du modèle est donc changée ;

(b) que les modèles considérés sont parfois implicitement « plongés » dans des modèles plus généraux (quoique non nécessairement précisés) (cf **plongement**).

(xi) La **théorie de la sélection entre modèles** (cf **test de sélection de modèles**) a pour objet le **choix du modèle** : eg le choix entre un modèle (Ω', $\mathscr{T}'$, $P_\theta')_{\theta' \in \Theta'}$ et un modèle (Ω'', $\mathscr{T}''$, $P_{\theta''})_{\theta'' \in \Theta''}$. Cette théorie se ramène souvent à un problème de **test d'hypothèses**, au sens courant (cf aussi **robustesse**).

(xii) Enfin, ce qui précède se transpose directement à un **modèle image**, de la forme ($\mathscr{X}$, $\mathscr{B}$, $P_\theta^X)_{\theta \in \Theta}$ ou ($\mathscr{X}$, $\mathscr{B}$, $\mathscr{P}^X$) (cf (ix),(a)) : ce modèle est l'image d'un modèle de base précédent par une **va** ou par une **statistique** (eg **échantillon**) $X : \Omega \mapsto \mathscr{X}$, généralement **observable**.

(xiii) Une **« théorie »** est toujours une hypothèse particulière, ie une hypothèse vérifiant les deux conditions suivantes :

(a) elle est en **adéquation** suffisante avec l'observation ;

(b) elle réunit un consensus minimal parmi l'**ensemble** des **hommes de l'art**, ie l'ensemble des personnes concernées par le phénomène que cette théorie cherche à expliquer (« communauté scientifique »).

Dans ce cas, on parle aussi parfois de **« théorie dominante »**, ce qui peut donc aussi impliquer l'existence de **« théories concurrentes »**, ou **« théories alternatives »**.

Une théorie quelconque demeure donc toujours une hypothèse (cf *« doute scientifique »*), même si elle est très communément acceptée.