OBSERVATION MANQUANTE (G9)
(08 / 06 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'étude d'un **phénomène** nécessite la possibilité de l'observer sur les **unités statistiques** concernées par ce phénomène. Or, ce **contexte statistique**, théoriquement satisfaisant, n'est pas toujours vérifié. Divers types d'obstacles peuvent empêcher l'**observation complète** d'une « manifestation » de la **Nature**. On parle de situation d'**observation incomplète**.

Ces obstacles tiennent souvent aux limites du **système d'observation** (cf **système statistique**, **production statistique**) : ainsi (physique), dans l'exploration de l'« infiniment grand » (resp de l'« infiniment petit »), les téléscopes (resp microscopes) ne permettent qu'une observation limitée. Le progrès technique peut, dans une certaine mesure, et pour un certain temps, lever ce type de limites. Mais les développements de l'**activité scientifique** (cf **étude scientifique**) exigent, de plus en plus, des échelles d'analyse plus « fines » : ainsi (infiniment petit), une analyse macroscopique peut être suivie d'une analyse mésoscopique, puis d'une analyse microscopique, etc. Les **modélisations** phénoménales ainsi que leurs **représentations statistiques** peuvent être en avance pr à ces développement du système d'observation.

D'autres obstacles peuvent être fortuits : destruction (en tout ou partie) d'un **matériel expérimental**, absence individuelle lors d'une enquête par **sondage** (cf **non réponse**), **censure**, **troncature**, observation localisée dans certaines zones alors que le phénomène se déroule ailleurs, observation réalisée à certains instants alors que le phénomène se déroule à d'autres instants.

(i) Une **donnée** (cf **donnée**) est ici définie comme **observation** X_n d'une **variable** ξ effectuée sur une **unité statistique** n (cf aussi **variable statistique**, **variable aléatoire**). Elle peut donc manquer :

(a) soit lorsque l'unité n elle-même manque ou est absente ;

(b) soit lorsque ξ ne peut être « observée » sur cette unité (valeur X_n absente).

Les valeurs d'une variable ξ observée avec **erreur** peuvent, parfois, être aussi considérées comme des observations manquantes.

Une **observation manquante** peut ainsi être (cf aussi **lacune**) :

(a) soit une « donnée » perdue ;

(b) soit une donnée inutilisable ;

(c) soit une donnée non observable ou partiellement observable.

Ceci peut arriver eg :

(a) lorsqu'une **expérience** implique la destruction des **unités expérimentales**, eg : décès, panne, désagrégation physique ou biologique, etc ;

(b) lorsque les questions d'une enquête par **sondage** comportent des **non réponses** : silences, omissions.

Comme dans le cas d'une censure, ce **contexte statistique** implique une **situation d'observation incomplète**.

(ii) Soit $(\mathcal{X}, \mathcal{B}, (P_\theta^X)_{\theta \in \Theta})$ un modèle image tq (**modèle d'échantillonnage**) :

(a) $\mathcal{X} = \mathcal{X}_0^N$, $\mathcal{B} = \mathcal{B}_0^{\otimes N}$, avec eg $\mathcal{X}_0 = \mathbf{R}^L$ (L représentant eg le nombre des variables observées sur les **unités statistiques** $n \in N_N^*$) ;

(b) $P_\theta^X = (P_\theta^\xi)^{\otimes N}$, où $\xi : \Omega \mapsto \mathcal{X}_0$ est la **va parente** qui génère le N-**échantillon** $X = (X_1, ..., X_N) : \Omega \mapsto \mathcal{X}$ (**échantillon iid**).

On dit que l'on est en **situation d'observation manquante**, ou en **situation d'observation lacunaire**, lorsque X est tq, $\forall\, n \in N_N^*$, certaines « coordonnées » :

(1) $\{X_{n,\alpha(1)(n)}, ..., X_{n,\alpha(L(n))(n)}\}$

du **vecteur aléatoire** $X_n : \Omega \mapsto \mathbf{R}^L$ sont seules disponibles, avec $\{\alpha_1(n), ..., \alpha_{L(n)}(n)\} \subset N_L^*$ et $1 \leq L(n) \leq L$. Dans ce qui précède, les indices doubles sont notés (par commodité) $n,\alpha(L(n))(n)$ au lieu de $n,\alpha_{L(n)}(n)$.

Les **indices** $\alpha_l(n)$ $(l = 1,..., L(n))$ de ces coordonnées, comme leur nombre L (n), peuvent aussi dépendre de l'**observation** (unité) n. Chacune des L - L(n) autres coordonnées, indisponibles, est appelée **observation manquante**, ou **lacune**, relative à X_n (ou à l'unité n). Ces lacunes peuvent provenir du **hasard** (ie être aléatoires) ou non.

(iii) Diverses procédures statistiques portant sur θ (ou sur une fonction $g : \Theta \mapsto \mathbf{R}^Q$ de θ) sont fondées sur la **vraisemblance** :

(2) $L(x, \theta) = (dP_\theta^X / d\mu)(x)$,

où μ désigne une **mesure positive** dominant la famille $(P_\theta^X)_{\theta \in \Theta}$ (cf **famille de lois dominée**).

La vraisemblance L précédente dépend donc :

(a) de **variables observables** (ie présentes dans le « champ d'observation ») ;

(b) d'autres **variables inobservables** (ie absentes de ce champ).

Cette dépendance peut se produire de façon aléatoire ou non. On note alors eg : L (x , θ) = L (x_p , x_a , θ) (présence, absence) la vraisemblance.

(iv) Les procédures courantes sont généralement moins efficaces lorsque le **statisticien** doit « modéliser » en présence d'observations manquantes : **perte d'information**, moindre nombre de **degrés de liberté**, etc.

Dans ce cas, la **procédure statistique** mise en oeuvre doit être adaptée.

Dans certaines **situations statistiques**, les observations manquantes peuvent être assimilées à des **paramètres supplémentaires** (généralement des **paramètres importuns**). Il est parfois possible de les « estimer » (ou simplement de les évaluer) à partir de l'ensemble des données observées : eg par « prévision interne » (ie par « **interpolation** ») à partir du modèle estimé à l'aide des observations présentes x_p . Mais cette approche ne « crée » pas d'**information** nouvelle, voire même « réplique » une information existante, au risque de la sur-représenter.

On doit, en outre, distinguer :

(a) le cas où le nombre d'observations (manquantes) varie avec la taille N (le nombre de « paramètres » varie alors avec N) (études asymptotiques) ;

(b) le cas, plus favorable (mais exceptionnel), où il est constant.

(v) La présence d'**aberrations** dans un **ensemble** de **données** peut, dans certains cas, être traitée comme équivalente à la présence d'observations manquantes : en particulier, lorsque le **schéma probabiliste** générant les aberrations n'est pas élucidable, on peut considérer que ces valeurs aberrantes sont des **lacunes** d'**information**.

Pour leur repérage, un **test d'aberration** préalable peut être effectué sur ces données (eg échantillon), puis ces dernières sont traitées comme données manquantes.

(vi) En pratique, le symbolisme (1) signifie que certaines observations sont connues pour toutes les variables, d'autres pour certaines d'entre elles seulement.

De façon duale, certaines unités statistiques n répondent pour toutes les variables, d'autres pour certaines d'entre elles. Il est alors préférable d'utiliser l'information maximum, eg de calculer les **caractéristiques** empiriques utiles sur toutes les observations disponibles de chaque variable, au lieu de les calculer sur les seules observations communes à toutes les variables.

Ainsi, lorsque des données (eg « résultats », « relevés », etc) manquent à l'issue d'une **expérience planifiée** ou d'une enquête par **sondage** (destruction ou perte de l'information y afférente), les procédures habituelles doivent être modifiées. Dans ce sens, on peut :

(a) soit estimer les observations en question (**interpolation** ou **extrapolation**, **liens** ou **corrélations** entre observations, etc) ;

(b) soit, plus rarement, ne pas en tenir compte, si la procédure mise en oeuvre le permet : eg si le nombre d'observations présentes et observables est suffisamment important.

(vii) A titre d'exemple, lorsque $\mathcal{X}_0 = \mathbf{R}^L$ (ie lorsqu'il existe $L \geq 2$ variables réelles), et que l'on distingue une **variable endogène** et $K = L - 1$ **variables exogènes**, avec $K \leq N$, on peut se ramener à un problème de **régression avec observations manquantes** (régression avec lacunes).