VARIABLE OBSERVABLE (C1, C3, H1, J, K)
(09 / 10 / 2020, © Monfort, Dicostat2005, 2005-2020)

(i) Une **variable aléatoire observable** est une variable que le **statisticien** peut concrétiser (eg mesurer, qualifier, etc) (cf **observabilité**, **observation**). Cette situation est fondamentale pour permettre l'**inférence statistique**, et elle se présente dans tous les **domaines de connaissance**.

A l'inverse, une **variable inobservable**, parfois appelée **variable latente**, ou **variable cachée**, est une variable que l'on ne peut observer, et notamment mesurer, en particulier lorsqu'elle se trouve en dehors du **champ de l'expérimentation** : c'est le cas eg lorsqu'on se trouve en présence d'observations censurées, ou d'un échantillon censuré (cf **censure**).

La distinction entre ce qui est observable et ce qui ne l'est pas est donc une distinction de base en **Statistique** aussi bien que dans les diverses sciences.

(ii) Lorsque certaines des variables qui interviennent dans l'appréciation d'un **phénomène** (aléatoire) ne sont pas observables, le statisticien peut :

(a) soit adopter un ***comportement bayésien***, ie se donner la loi des variables en question, voire même la **loi conjointe** des variables inobservables et des variables observables. Les inobservables sont alors traitées comme des **paramètres** supplémentaires ;

(b) soit chercher à se ramener de l'inobservable à l'observable, le plus souvent en formulant des ***hypothèses de liaison*** (fonctionnelle, ou seulement statistique) entre les variables inobservables et d'autres variables qui sont observables.

(iii) Exemple de **liaison fonctionnelle**. On considère un **modèle statistique** $(\Omega, \mathcal{T}, \mathcal{P})$, un **espace probabilisable** $(\mathcal{X}, \mathcal{B})$ et un espace probabilisable $(\mathcal{Z}, \mathcal{D})$. La théorie qui conduit à la **spécification** de $(\mathcal{X}, \mathcal{B}, \mathcal{P}^{\xi})$, **modèle image** du précédent par une **variable aléatoire** inobservable $\zeta : \Omega \mapsto \mathcal{Z}$, doit être adaptée.

Ceci est notamment possible s'il existe une **application mesurable** $\varphi : \mathcal{Z} \mapsto \mathcal{X}$ (connue) tq la variable aléatoire $\xi = \varphi(\zeta)$ soit observable.

Le procédé qui consiste à remplacer ζ par le couple (ξ, φ) nécessite donc la connaissance de φ ; dans certains cas, il est possible d'estimer φ à partir des autres variables observables du modèle. L'inférence statistique se poursuit alors en opérant à l'aide du modèle $(\mathcal{X}, \mathcal{B}, \mathcal{P}^{\xi})$.

(iv) Exemple de **liaison statistique**. Pour estimer certains paramètres d'un modèle, une pratique courante consiste à utiliser des **corrélations** supposées exister entre des variables inobservables et des variables observables : une variable observable est alors supposée être une « **variable proche** », ou une « variable voisine », (en anglais : « ***proxy*** ») d'une variable inobservable à laquelle on l'associe.

Ceci est le cas lorsqu'un concept théorique est mesuré de façon approximative à l'aide d'une variable observable - alors que la vraie **mesure** du concept est inobservable - et que ces deux « mesures » sont supposées corrélées.

Cette approche est distincte d'autres approches, basées sur des corrélations supposées : eg la **méthode des variables instrumentales**, où les corrélations existent entre des groupes de variables toutes observables (cf **variable instrumentale**).

(v) La notion d'(in)observabilité peut aussi bien porter sur des variables aléatoires que sur des **paramètres**. La situation courante est celle dans laquelle les variables sont observables et les paramètres (inconnus) à déterminer (**estimation**).

Ainsi, dans un **modèle à équations simultanées** linéaire, écrit dans l'**espace des variables** (ξ, η) sous la forme $A \eta + B \xi = \varepsilon$, les matrices (paramètres) A et B sont inobservables (sauf dans le cas d'une **simulation**), de même que la **perturbation aléatoire** ε. Seul, le **couple aléatoire** (ξ, η) est, en principe, observable. C'est précisément le but de l'inférence statistique de déterminer ici des valeurs « plausibles » pour (A, B).