LACUNE (C8, G9)
(24 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

D'un point de vue conceptuel et terminologique, le terme de **lacune** peut se référer à l'une des deux **situations statistiques** suivantes :

(a) **variable omise**, ou **variable absente** : lorsqu'une **variable** est absente et ne permet pas de modéliser complètement un **phénomène** (ie de définir une **représentation statistique**), on parle d'**omission** ou de **variable omise**. Dans ce cas, c'est la variable qui est manquante lors de la **spécification** du modèle décrivant le phénomène ;

(b) **variable inobservable** : lorsqu'une **variable** n'est pas observée sur une **unité statistique** particulière (cf **observabilité**, **inobservabilité**), on parle plutôt de **lacune** ou d'**observation manquante**. Dans ce cas, c'est l'**observation** de la **variable aléatoire** sur l'unité considérée qui est manquante (ie non apparente ou indisponible). Cette situation peut conduire à l'impossbilité de calcul d'une **statistique** dépendant de cette observation.

C'est le second sens qui est le plus usuel.

(i) Une **lacune**, ou **observation manquante**, peut se rencontrer lors de la mise en oeuvre d'une source d'**information statistique**.

Elle peut résulter d'une absence d'information, partielle ou totale, relative à des **variables** (ou « descripteurs ») décrivant un **phénomène** donné. Autrement dit, le **statisticien** cherche à « mesurer » les « valeurs » de ces variables sur des unités statistiques, mais il ne peut pas observer certaines valeurs sur certaines unités (cf **observabilité**, **inobservable**).

Les raisons principales sont les suivantes :

(a) impossibilité pratique d'observation de certaines variables, ou seulement de certaines données : **système statistique** peu évolué (ne couvrant pas un champ complet souhaitable), instrument de **mesure** imprécis, etc ;

(b) **troncature** de la loi générant ces données ;

(c) **censure** des données, eg :

$(c)_1$ **unité de sondage** d'un **échantillon** obtenu par **sondage** : perte matérielle (« sondage destructif »), **non réponse** (complète, ou seulement partielle) des unités ;

$(c)_2$ **unité expérimentale** utilisée pendant une **expérimence aléatoire**. Sa **durée de vie** peut se trouver raccourcie pour des raisons expérimentales : expériences avec destruction (au sens physique ou biologique du terme).

(ii) De façon générale, soit $(\Omega, \mathscr{T}, P)$ un **espace probabilisé** et $\xi : \Omega \mapsto \mathscr{X}$ une va ξ, où $(\mathscr{X}, \mathscr{B})$ désigne l'**espace mesurable** des valeurs possible de ξ (ie un **espace**

d'observation). On note $\mathscr{L}(\xi)$ la **loi de probabilité** de ξ et $X = (X_1, ..., X_N)$ un N-**échantillon iid** issu de ξ. L'existence de lacunes signifie que l'on n'observe pas toutes les N « réalisations » X_n (n = 1 ,..., N) de ξ :

(a) certaines, en nombre N^o, sont effectivement observées et l'on connaît chaque indice n tq X_n est observé. On note alors X^o la suite de ces réalisations observables ;

(b) les $N^a = N - N^o$ autres réalisations ne sont pas observées : observations « absentes », ou « inobservations » (cf **inobservable**). On connaît chaque indice n tq X_n n'est pas observé. On note alors X^a la suite de ces réalisations inobservables.

On peut alors structurer X selon $X = (X^o, X^a)$, où X^o désigne la liste des observations effectives et X^a celle des lacunes.

Cette **situation statistique** peut interpréter comme suit :

(a) la loi $\mathscr{L}(\xi)$ de ξ permet, en principe, de générer une observation quelconque tq $\xi(\omega) = x$ (eg par **expérimentation** ou par **sondage**) ;

(b) mais il existe une variable dichotomique (**variable binaire**) α à valeurs dans {0, 1} qui est (implicitement) conjointement tirée selon une loi $\mathscr{L}(\alpha) = p \,.\, \mathbf{1}(\alpha = 1) + (1 - p) \,.\, \mathbf{1}(\alpha = 0)$ ou encore selon $\mathscr{L}(\alpha) = p \,.\, \delta_1 + (1 - p) \,.\, \delta_0$, où $p \in [0, 1]$ désigne la probabilité d'observation de ξ, 1 - p celle d'inobservation (lacune) et δ_x la **masse de DIRAC** placée en $x \in \mathscr{X}$.

Autrement dit, α « occulte » ξ au sens où, au lieu d'observer ξ selon la loi $\mathscr{L}(\xi)$ elle-même, on observe la valeur de ξ générée par $\mathscr{L}(\xi)$ seulement lorsque $\alpha = 1$, ie générée par la **loi conditionnelle** de ξ sachant α. Cette loi se déduit du **théorème des probabilités composées**, ie (sous forme « symbolique ») :

(1) $\quad \mathscr{L}(\xi / \alpha) = \mathscr{L}(\xi \cap \alpha) / \mathscr{L}(\alpha)$,

ie, de façon explicite, $\forall B \in \mathscr{B}$ et $\forall A \in \mathscr{B}$:

(2) $\quad \mathscr{L}([\xi \in B] / [\alpha \in A]) = \mathscr{L}([\xi \in B] \cap [\alpha \in A]) / \mathscr{L}([\alpha \in A])$,

$\mathscr{L}(\xi)$ est donc la « marge » de la **loi conjointe** $\mathscr{L}(\xi, \alpha)$ ou $\mathscr{L}(\xi \cap \alpha)$ (cf **loi marginale**, **censure**).

D'après (1) :

(3) $\quad \mathscr{L}(\xi \cap \alpha) = \mathscr{L}(\alpha) \,.\, \mathscr{L}(\xi / \alpha)$,

expression dans laquelle :

(a) $\mathscr{L}(\alpha)$ est connue ;

(b) les réalisations issues de $\mathscr{L}(\xi / \alpha)$ sont observables, et leur nombre N^o est connu.

Par suite, on peut estimer :

(a) p par la **proportion** $p^{\sim} = N^o / N$, ce qui conduit à l'estimateur $\mathscr{L}^{\sim}(\alpha) = p^{\sim} . \delta_1 + (1 - p^{\sim}) . \delta_0$ de $\mathscr{L}(\alpha)$;

(b) $\mathscr{L}(\xi / \alpha)$ par la **loi empirique** issue de X^o, et notée $P_N{}^o$.

(c) donc $\mathscr{L}(\xi \cap \alpha)$ par :

(4) $\mathscr{L}^{\sim}(\xi \cap \alpha) = \mathscr{L}^{\sim}(\alpha) . P_N{}^o$.

Un estimateur $\mathscr{L}^{\sim}(\alpha)$ de $\mathscr{L}(\xi)$ se déduit de (4) par **marginalisation** relativement à α.

(iii) La description précédente peut s'étendre à plusieurs types de variables (eg endogènes et exogènes) prises en compte dans un modèle.

Soit $(\Omega, \mathscr{T}, \mathscr{P})$ un **modèle statistique**, $\zeta = (\xi, \eta) : \Omega \mapsto (\mathscr{X}, \mathscr{Y})$ un **couple aléatoire**, où $(\mathscr{X}, \mathscr{B})$ désigne l'**espace mesurable** des valeurs possible de ξ (ie un **espace d'observation** ou un **espace d'états**) et $(\mathscr{Y}, \mathscr{C})$ celui des valeurs possibles de η. On note $\mathscr{L}(\xi, \eta)$ la **loi de probabilité** de ζ et $Z = (Z_1, ..., Z_N)$ un N-échantillon iid issu de ζ. On désigne alors par $X = (X_1, ..., X_N)$ l'échantillon partiel correspondant à ξ et par $Y = (Y_1, ..., Y_N)$ l'échantillon partiel correspondant à η.

Dans ce contexte, l'existence de lacunes signifie que l'on ne peut observer les N « réalisations » (X_n, Y_n) $(n = 1, ..., N)$ de ζ. Comme ces coordonnées sont associées (réalisations relatives à la même **unité statistique** n), on note :

(a) N^{oo} le nombre de coordonnées tq (ξ, η) est observé selon $(X_n{}^o, Y_n{}^o)$. On connaît les indices n correspondants. On note alors (X^o, Y^o) la suite des réalisations observables simultanément ;

(b) N^{oa} le nombre de coordonnées tq ξ seul est observé. On connaît les indices n correspondants. On note (X^o, Y^a) la suite des réalisations de ce type ;

(c) N^{ao} le nombre de coordonnées tq η seul est observé. On connaît les indices n correspondants. On note (X^a, Y^o) la suite des réalisations de ce type ;

(d) $N^{aa} = N - N^{oo} - N^{oa} - N^{ao}$ le nombre de coordonnées dont les observations sont absentes (« inobservations »). On connaît les indice n tq (ξ, η) n'est pas observé. On note (X^a, Y^a) la suite des réalisations inobservables simultanément.

On peut alors structurer Z selon $Z = \{(X^o, Y^o), (X^o, Y^a), (X^a, Y^o), (X^o, Y^o)\}$. L'interprétation est similaire à la précédente :

(a) la loi $\mathscr{L}(\xi, \eta)$ de (ξ, η) doit normalement donner lieu à des observations $(\xi(\omega), \eta(\omega)) = (x, y)$ de (ξ, η) ;

(b) le couple de variables dichotomiques (α, β) à valeurs dans $\{0, 1\}^2$ est conjointement tiré, α selon une loi $\mathscr{L}(\alpha / (\xi, \eta)) = p \,.\, \delta_1 + (1 - p) \,.\, \delta_0$, où $p \in [0, 1]$, et β selon une loi $\mathscr{L}(\beta / (\xi, \eta)) = r \,.\, \delta_1 + (1 - r) \,.\, \delta_0$, où $r \in [0, 1]$.

Autrement dit, (α, β) « occulte » (ξ, η), au sens où, au lieu d'observer (ξ, η) selon la loi $\mathscr{L}(\xi, \eta)$ elle-même, on observe seulement la valeur de (ξ, η) générée par la loi conditionnelle :

(5) $\mathscr{L}((\xi, \eta) / (\alpha, \beta)) = \mathscr{L}((\xi, \eta) \cap (\alpha, \beta)) / \mathscr{L}(\alpha, \beta)$,

La loi $\mathscr{L}(\xi, \eta)$ est la loi marginale de la loi jointe $\mathscr{L}((\xi, \eta) \cap (\alpha, \beta))$.

D'après (5) :

(6) $\mathscr{L}((\xi, \eta) \cap (\alpha, \beta)) = \mathscr{L}(\alpha, \beta) \,.\, \mathscr{L}((\xi, \eta) / (\alpha, \beta))$,

expression dans laquelle :

(a) $\mathscr{L}(\alpha, \beta)$ est connue, du moins si l'on admet l'hypothèse selon laquelle α et β sont indépendantes entre elles (l'occultation agit séparément sur ξ et sur η) :

(7) $\mathscr{L}(\alpha, \beta) = \{p \,.\, \delta_1 + (1 - p) \,.\, \delta_0\} \otimes \{r \,.\, \delta_1 + (1 - r) \,.\, \delta_0\}$.

(b) les réalisations issues de $\mathscr{L}((\xi, \eta) / (\alpha, \beta))$ sont observables, ainsi que les nombres N^{oo}, N^{ao}, N^{oa}, et N^{aa}.

(iv) En présence de lacunes, les **procédures statistiques** usuelles (estimation des paramètres, tests, prévisions, classifications, etc) doivent donc faire l'objet de **modification** ou d'adaptation.

Il en est de même lorsque les échantillons considérés ne sont pas iid.

(v) Ainsi, dans le cas d'un **problème d'estimation**, on étudie le **modèle statistique** $(\mathcal{X}, \mathcal{B}, (P_\theta^X)_{\theta \in \Theta})$, dans lequel X représente un **échantillon** (ou une **statistique** d'intérêt) et P_θ^X une **loi de probabilité** possible de X, dominée par une **mesure positive** μ (loi à **densité**). On note alors :

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

la **vraisemblance** associée au modèle.

Par suite, si X_p (resp x_p) désigne l'ensemble des valeurs « présentes » (ie observées) de X (resp dans la densité $\mathcal{L}$) et X_a (resp x_a) l'ensemble des valeurs « absentes » (ie non observées), plusieurs méthodes d'estimation de θ sont possibles.

(a) **méthode de « faux » maximum de vraisemblance (R.L. ANDERSON - M.S. BARTLETT)** dans laquelle le « **paramètre** » considéré, à estimer, est le couple (x_a , θ). Il s'agit alors de résoudre le **problème d'optimisation** suivant :

(9) $\sup_{(xa , \theta)} \mathcal{L}(x_p , x_a , \theta)$,

(où xa désigne x_a).

Il ne s'agit pas d'un « vrai » **maximum de vraisemblance** car, en général, le « paramètre partiel » x_a , parfois appelé **paramètre incident**, varie avec N (taille de l'échantillon X) ;

(b) **méthode de maximisation de la vraisemblance partiellement intégrée (M.H. de GROOT - K. GOEL)**. Si l'on note $\mathcal{X} = \mathcal{X}_p \times \mathcal{X}_a$ (ou, selon le cas, $\mathcal{X} = \mathcal{X}_p \cup \mathcal{X}_a$), l'ensemble des valeurs de X, représenté selon la présence et l'absence, ie si $x = (x_p , x_a)$, et si $\mu_{/\mathcal{X}a}$ est la **restriction** de μ à l'ensemble des valeurs absentes (on note $\mathcal{X}a$ pour désigner $\mathcal{X}_a$), la méthode consiste à maximiser pr à θ la « vraisemblance » (partiellement) intégrée :

(10) $\mathcal{L}(x_p , \theta) = \int L(x_p , x_a , \theta)\, d\mu_{/\mathcal{X}a}(x_a)$;

(c) **méthode bayésienne (S.J. PRESS - A.J. SCOTT)** consistant à définir une mesure de **probabilité a priori** $\Pi_{(xa , \theta)}$ relative au couple (x_a , θ), auquel on fait jouer le rôle de paramètre ;

(d) **méthode en trois temps** :

$(d)_1$ on considère d'abord un modèle « restreint » dans lequel l'échantillon est X_p et le paramètre θ. On estime alors provisoirement θ à partir de la vraisemblance associée à ce modèle (ie à partir de la seule donnée informative observable x_p). Cette étape n'est pas toujours réalisable ;

$(d)_2$ on revient au modèle « complet » et l'on estime x_a à l'aide d'une **caractéristique** de **centralité** (eg la **moyenne**) de la **loi** de ce modèle (caractéristique déjà estimée) ;

$(d)_3$ on applique la **méthode du maximum de vraisemblance** avec la vraisemblance complétée $\mathscr{L}(x_p , x_a , \theta)$, ce qui fournit l'estimateur recherché de θ.

(vi) Ainsi, l'analyse statistique des lacunes revient à considérer qu'il n'est possible d'observer que dans certaines parties de $\mathscr{X}$ (cf aussi **censure**). Autrement dit, si $s : \mathscr{X} \mapsto \mathscr{Y}$ est une **statistique** surjective (cf **application surjective**), on observe seulement les données incomplètes $Y = s(X) : \Omega \mapsto \mathscr{Y}$ au lieu d'observer complètement $X : \Omega \mapsto \mathscr{X}$.

Comme dans le cas d'une **aberration**, la notion de lacune peut donc aussi être présentée à partir de celle de **mélange de lois**. La situation est cependant plus favorable, car on connaît généralement les **indices** (ou « identifiants ») aussi bien que les observations des unités observées, mais aussi les indices (ou identifiants) des unités non observées.

Cette approche permet alors de définir directement la loi de ξ à l'aide des lois composant ce mélange :

(a) en effet, si ξ est la **variable aléatoire** considérée, si ξ_p est une variable **observable** (ie dont les observations sont toutes présentes) et si ξ_u est une variable **inobservable** (ie dont les observations sont toutes absentes) (*« unobservable »*), on peut définir la loi de ξ comme **combinaison linéaire convexe** de ces dernières, ie :

(11) $\quad P^{\xi} = \alpha_o \mathscr{L}(\xi_o) + \alpha_u \mathscr{L}(\xi_u),$

où $(\alpha_o , \alpha_u) \in S_2$ (**simplexe** de $\mathbf{R}^2$), P^{ξ} désigne la loi de ξ et $\mathscr{L}(\upsilon)$ la loi de υ (avec $\upsilon = \xi_o$ ou ξ_u) ;

(b) disposant d'observations X_n de ξ, avec $n = 1 ,..., N$, et en notant $X = (X_1 ,..., X_N)$ l'**échantillon**, on peut représenter la **loi empirique** P_N ou $\mathscr{L}_N(\xi)$ de ξ selon une équation de la forme :

(12) $\quad \mathscr{L}_N(\xi) = \alpha_o . \mathscr{L}_N(\xi_o) + \alpha_u . \mathscr{L}_N(\xi_u).$

Cette équation n'est cependant pas une **équation estimante** car elle ne permet pas, en elle-même, d'estimer chaque « **paramètre** » α.

Lorsque N_o observations sont présentes (observables) et N_u absentes (unobservables), avec $N_o + N_u = N$, et que X est organisé de façon conforme, selon (X^o, X^u), les données du problème sont (N_o , X^o, N_u).

En général, on peut estimer resp α_o et α_u à l'aide des **proportions** :

(13)
$$\alpha_o \tilde{} = N_o / N,$$
$$\alpha_u \tilde{} = N_u / N.$$

D'autre part, $\mathscr{L}(\xi_o)$ peut être estimée à l'aide de la fre $\mathscr{L}_{No}{}^{o}(\xi_o)$ définie à partir du sous-échantillon X^o de X (No désignant, par commodité, N_o).