CENSURE (F8, G9, H4, I5, J5)
(09 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

La **censure** est une propriété selon laquelle l'observation d'un **échantillon** est altérée d'une certaine façon.

Il s'agit d'une situation d'**observation incomplète** qui se présente lorsqu'on ne peut (ou ne veut) mesurer toutes les « composantes » ou « coordonnées » d'un **échantillon aléatoire**. Cette impossibilité peut tenir à des **situations statistiques** tq les suivantes :

(a) « *destruction* » d'**unités statistiques** en cours d'**observation**, eg :

(a)1 dans une **expérimentation** : panne physique d'un composant d'un **système**, décès d'une **unité** biologique, etc ;

(a)2 dans un **sondage** : absence ou **non réponse**, réponse par un tiers, etc ;

(b) utilisation d'un instrument de **mesure** dont la **précision** est insuffisante (cf **dispositif expérimental**) ;

(c) limitation des coûts, dont la conséquence est l'élimination de données « utiles » (cf **lacune**, **observation manquante**).

Elle intervient souvent dans les questions de **fiabilité** (d'un **système**) ou dans l'étude d'une **loi de survie** relative à des unités statistiques (cf **fonction de survie**).

(i) Soit $X : \Omega \mapsto \mathbf{R}^N$ un **échantillon aléatoire**.

On dit que X est un **échantillon censuré**, ou un **échantillon incomplet**, lorsqu'il n'est pas possible d'observer (ou lorsqu'on ne veut pas prendre en compte) certaines coordonnées $(X_{\alpha(1)}, ..., X_{\alpha(M)})$ (avec $1 \leq M \leq N$, où $(\alpha_1, ..., \alpha_M)$ est une sous-suite d'**indices** tq $1 \leq \alpha_1 \leq ... \leq \alpha_M \leq N$ (on note par commodité $\alpha(m)$ pour désigner α_m).

Si $X^{(\cdot)}$ est la **statistique d'ordre** associée à X, un exemple de censure (**censure bilatérale**) est celui où l'on n'observe que les $M = N - (L + K)$ coordonnées :

(1) $X^{(K+1)}, ..., X^{(N-L)}$

de $X^{(\cdot)}$, et non les coordonnées extrêmes.

Il y a **censure à gauche** ssi $L = 0$ et $K > 0$, et **censure à droite** ssi $K = 0$ et $L > 0$. Il existe alors un segment $[a, b] \subset \mathbf{R}$ tq $X_{\alpha(m)} \in [a, b]$, $\forall m \in \mathbf{N}_M^*$.

Si $b = - a = +\infty$, alors X est entièrement **observable** (absence de censure).

Si $b = a$ et si tous les $X_{\alpha(M)} \neq a$, alors X est entièrement inobservable (**censure complète**).

(ii) On appelle **censure de type I** une **censure en niveau**, dans laquelle l'intervalle (segment) [a, b] est donné, avec $X_{\alpha(m)} \in [a, b]$ ($\forall\ \alpha_m$). M n'est donc pas connu a priori. Plus généralement, on appelle censure de type I une censure tq, $B \subset \mathbf{R}^N$ étant donné, l'échantillon censuré X^s est une va observable seulement dans B.

Soit $p \in [0, 1]$. On appelle **censure de type II** une **censure en proportion** (ou en **fréquence**), dans laquelle une **proportion** p des coordonnées de X n'est pas observée. On n'observe ainsi seulement $X^s = (X_{\alpha(1)}, ..., X_{\alpha(M)})$, avec N - M = [p . N], où les indices α_m sont distincts entre eux. Ici, M n'est pas aléatoire. Plus généralement, on appelle censure de type II une censure tq, $B \subset \mathbf{R}^N$ étant donné, une proportion 1 - p des coordonnées de X est seule observable dans B.

Si les va X_n sont des **copies** iid (cf **suite équidistribuée**, **suite indépendante**) d'une **va** ξ dont la loi (sous forme paramétrique) est notée P_θ^X, et si $dP_\theta^X = f(. , \theta)\, d\mu$ en est la **densité**, la **vraisemblance** s'écrit :

(a) dans le cas de la censure de type I :

(2) $\quad L_1 (X, \theta) = c \left(\int_{-\infty}^{a} f(x, \theta)\, d\mu(x)\right)^K . \left(\Pi_{n=1}^{N-(L+M)} f(X_n, \theta)\right) . \left(\int_b^{+\infty} f(x, \theta)\, d\mu(x)\right)^L,$

où K + L = M et c est une **constante** de **normalisation** (qui dépend, en général, de (a, b)) ;

(b) dans le cas de la censure de type II :

(3) $\quad L_2 (X, \theta) = c \left(\int_{-\infty}^{X(K)} f(x, \theta)\, d\mu(x)\right)^K . \left(\Pi_{n=K+1}^{N-L} f(X_n, \theta)\right) . \left(\int_{X(N-L)}^{+\infty} f(x, \theta)\, d\mu(x)\right)^L.$

(iii) Plus généralement, on appelle **censure aléatoire (à droite)** la donnée d'une **suite** de **va** $(Y_n, \mathbf{1}(A_n))_{n=1,...,N}$ tq :

(4) $\quad Y_n \leq X_n, \ \forall\ n \in N_N^*,$

avec $A_n = [Y_n, X_n] = [\omega \in \Omega : Y_n(\omega) = X_n(\omega)] \in \mathcal{T}$, où $\mathcal{T}$ est une **tribu de parties** de Ω donnée. Autrement dit, X étant donné, on sait si $Y_n = X_n$ ou non.

Si les X_n sont des **« instants de censure »** (non négatifs) iid selon une **fr** F et si les Y_n sont des **temps de survie** (non négatifs) iid selon G, alors il y a **censure aléatoire (à droite)** ssi l'on observe $(Z_n, d_n)_{n=1,...,N}$, avec $Z_n = \min(X_n, Y_n)$, et $d_n = \mathbf{1}_{[Y(n) \leq X(n)]}$, $\forall\ n \in N_N^*$. Si les X_n et les Y_n sont indépendantes entre elles, $\forall$ n, alors les Z_n sont iid selon une fr H tq 1 - H = (1 - F) (1 - G), et les d_n sont iid selon une **loi de BERNOULLI** de paramètre $p = P(d_1 = 1) = \int_{\mathbf{R}+} (1 - F(x))\, dG(x)$ (cf **épreuve de BERNOULLI**).

On appelle **échantillon censuré au sens de E.L. KAPLAN - P. MEIER**, ou **échantillon incomplet au sens de E.L. KAPLAN - P. MEIER**, la donnée d'une suite $(Y_n)_{n=1,...,N}$ de **vars**, dites **variables observées**, défines par :

(5) $Y_n = \min(X_n, Z_n)$, $\forall n \in N_N^*$,

où les vars Z_n (qui peuvent dépendre des X_n) sont les **limites d'observation**, ou **seuils d'observation**, appelées **variables de censure**. Autrement dit, $\forall n \in N_N^*$, on sait :

(a) si $X_n \leq Z_n$, auquel cas on observe la va $Y_n = X_n$,

(b) ou si $X_n > Z_n$, auquel cas on observe la va $Y_n = Z_n$.

Si les **variables de censure** Z_n forment une **suite iid** distribuée selon la **fr** H (cf **suite équidistribuée**, **suite indépendante**), et si les variables X_n de l'échantillon X sont iid selon la fr F, alors la fr G des Y_n ($\forall n \in N_N^*$) est simplement tq :

(6) $1 - G(y) = (1 - F(y)) . (1 - H(y))$.

Si F (resp H) dépend d'un **paramètre** $\lambda \in \Lambda$ (resp $\mu \in M$), alors G dépend, en général, de $(\lambda, \mu) \in \Lambda \times M$, et l'estimation de (λ, μ) se fait généralement par la **méthode du maximum de vraisemblance**. Dans le cas contraire, on estime G par une **méthode non paramétrique** (cf **estimateur de KAPLAN-MEIER**).

(iv) Dans le cas d'une censure de type I, si les va X_n sont indépendantes et distribuées comme la **variable parente** ξ, et si l'on note $dP_\theta^\xi = f(. , \theta)\, d\mu$ la **dérivée de NIKODYM-RADON** (**densité**) de P_θ^ξ pr à une mesure dominante μ (cf **famille de lois dominée**), la vraisemblance s'écrit :

(7) $L_I(X, \theta) = c_{ab} . I_a . P_{LM} . I_b$,

avec :

c_{ab} = constante de normalisation,

(8)
$$I_a = \{\int \mathbf{1}_{]-\infty\, ,\, a]}\, f(x, \theta)\, d\mu(x)\}^K,$$
$$P_{LM} = \prod_{n=1}^{N-(L+M)} f(X_n, \theta),$$
$$I_b = \{\int \mathbf{1}_{]b\, ,\, +\infty]}\, f(x, \theta)\, d\mu(x)\}^L.$$

Dans le cas d'une censure de type II, et sous les mêmes hypothèses, la vraisemblance s'écrit :

(9) $L_{II}(X, \theta) = c_{ab} . I(X^{(K)}) . P_{KN} . I(X^{(N-L)})$,

avec :

c_{ab} = constante de normalisation,

(10)

$$I\,(X^{(K)}) = \{\int \mathbf{1}(]-\infty\,,X^{(K)}])\, f(x,\theta)\, d\mu\,(x)\}^K,$$

$$P_{KN} = \prod_{n=K+1}^{N-L} f(X_n\,,\theta),$$

$$I\,(X^{(N-L)}) = \{\int \mathbf{1}(]X^{(N-L)},+\infty])\, f(x,\theta)\, d\mu\,(x)\}^L,$$

où **1**(A) désigne la **fonction indicatrice** d'une partie A.

(v) Dans tous les cas de censure, la va X est théoriquement partout observable dans $\mathbf{R}^N$, mais l'on ne peut (ou ne veut) la mesurer à l'extérieur d'un certain ensemble de valeurs, noté B. Ainsi, on observe l'élément $\omega \in \Omega$, mais on ne peut (ou ne veut) mesurer que X (ω), pour tous les éléments $\omega \in$ U, où $U \subset \Omega$ et $U \neq \Omega$. Autrement dit, on n'observe X que pour des valeurs dans B = X (U) (image de U par X). Par suite :

(a) ou bien U est connu (situation souvent considérée), auquel cas on étudie la **loi** P^X de X à travers B ;

(b) ou bien U est inconnu (cas plus complexe), auquel cas on doit, en outre, estimer B (et, en particulier, sa **frontière** ∂ B).

(vi) Les différents problèmes que pose la censure en **Statistique** sont, dans le cas où X est un **échantillon iid** selon une **lp** P^ξ :

(a) l'**estimation** d'une **caractéristique** $\gamma = c\,(P^\xi) \in \Gamma$ de la loi qui a généré les observations ;

(b) l'étude de la **perte d'information** dûe à la censure ;

(c) divers **tests d'hypothèses** portant sur γ.

Les hypothèses faites sur P^ξ jouent un rôle important.