**ESTIMATEUR DE KAPLAN-MEIER (F, G, H4)**
(14 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

L'**estimateur de KAPLAN-MEIER** est un **estimateur** non paramétrique d'une **fonction de répartition** (ou d'une de ses **caractéristiques**), défini en cas d'**information** incomplète, ie lorsque l'échantillon observé subit une **censure**. Il est souvent utilisé dans l'étude des **durées de vie**.

(i) Soit $(\Omega, \mathcal{T}, P)$ un **espace probabilisé** et $(\xi, \zeta) : \Omega \mapsto \mathbf{R}_+ \times \bar{\mathbf{R}}_+$ un **couple aléatoire** constitué de **vars** positives $\xi$ et $\zeta$ indépendantes entre elles. La **variable aléatoire** $\xi$ s'interprète souvent comme une **durée de (sur)vie**.

On note F la **fr** de $\xi$, S = 1 - F sa **fonction de survie**, et H la fr de $\zeta$. On étudie la **va** $\xi$ à travers F (ou S).

Soit alors $((X_n , Z_n))_{n=1,...,N}$ un **échantillon iid** comme la **variable parente** $(\xi, \zeta)$ (cf **couple aléatoire**). On note $X = (X_n)_{n=1,...,N}$ et $Z = (Z_n)_{n=1,...,N}$ les échantillons resp associés à $\xi$ et à $\zeta$.

(a) en l'absence de censure (ie lorsque toutes les coordonnées $X_n$ de X sont observées), le nombre d'**unités statistiques** n « décédées » avant l'instant $x \in \mathbf{R}_+$ est défini par :

$$(1) \qquad N_x = \sum_{n=1}^{N} \mathbf{1}_{[X(n) < x]} .$$

en notant X(n) pour désigner $X_n$ .

Le **processus** $(N_x)_{x \in \mathbf{R}+}$ ainsi défini est un **processus ponctuel**. On peut considérer que X est censuré par un échantillon Z tq $Z_n = +\infty$ pour tout n = 1 ,..., N. Un estimateur classique de F (resp de S) est donc la **fonction de répartition empirique** $F_N$ (resp la **fonction de survie empirique**, définie selon : $S_N = 1 - F_N$) ;

(b) en présence de **censure aléatoire (au sens de E.L. KAPLAN - P. MEIER)**, la formule (1) et les estimateurs $F_N$ et $S_N$ doivent être modifiés. En effet, si la suite Z censure la suite X (cas d'une censure à droite), toute va $X_n$ peut être « tronquée » à droite par la va correspondante (ie de même indice) $Z_n$ . On observe donc seulement un « échantillon » $(Y_n , \mathbf{1}(A_n))_{n=1,...,N}$ (cf **fonction indicatrice**), dans lequel on note, $\forall$ n = 1 ,..., N :

$$(2) \qquad \begin{aligned} Y_n &= \min (X_n , Z_n) = X_n \wedge Z_n , \\ A_n &= [X_n \le Z_n] = \{\omega \in \Omega : X_n (\omega) \le Z_n (\omega)\}. \end{aligned}$$

Autrement dit, $\forall$ n $\in N_N^*$ :

$$(3) \qquad \mathbf{1}(A_n) = \begin{cases} 1 & \text{si } X_n \le Z_n \quad \text{(absence de censure)}, \\ 0 & \text{si } X_n > Z_n \quad \text{(censure)}. \end{cases}$$

Dans ce contexte, les indicatrices $\mathbf{1}(A_n)$ sont souvent notées $\delta_n$ .

Chaque variable $Z_n$ de la suite $(Z_n)_{n=1,\dots,N}$ est appelée **variable de censure**, ou **instant de censure**.

On connaît donc les valeurs $X_n$ des observations non censurées, ainsi que leurs indices n. Concrètement, la durée de vie $X_n$ de l'unité n n'est effectivement observée que si $X_n \leq Z_n$ . Par suite, le nombre d'unités n dont on peut observer le décès avant l'instant x est :

(4) $\quad N_x(Z) = \Sigma_{n=1}^{N} \mathbf{1}([X_n < x \wedge Z_n])$.

Le processus $(N_x(Z))_{x \in \mathbf{R}+}$ ainsi obtenu est encore un processus ponctuel. Par suite, lorsqu'il existe une censure du type précédent, la suite des $Y_n$ défini en (2) constitue un **échantillon iid**. Sa fr commune G vérifie :

(5) $\quad 1 - G(y) = (1 - F(y)) \,.\, (1 - H(y)), \qquad \forall\, y \in \mathbf{R}_+$ ,

et celle des va $\mathbf{1}(A_n)\, Y_n$ (observations non censurées) est :

(6) $\quad F^*(x) = P([Y_n < x] \cap [\mathbf{1}(A_n) = 1]) = \int_{]-\infty\,,\,x[} \{1 - H(t)\}\, dF(t)$.

(ii) En présence de censure (aléatoire à droite), on définit l'**estimateur (du produit limite) de E.L. KAPLAN - P. MEIER** de F comme l'estimateur $F_N^{\#}$ obtenu par maximisation de la **probabilité** des variables observées (**estimation non paramétrique** de F).

On montre qu'il est de la forme :

$$(7) \quad F_N^{\#}(x) = \begin{cases} 1 - \prod_{n \in Bn(x)} \{(N - R_n + 1)^{-1} (N - R_n)\}^{\mathbf{1}(An)} & \text{si } x < X^{(N)}, \\ 1 \text{ sinon,} \end{cases}$$

où $B_n(x) = \{n \in N_N^* : Y_n < x\}$ est noté Bn (x), et où $R_n$ désigne le **rang** du n-ième couple $(Y_N\ ,\ 1 - \mathbf{1}(A_n))$ dans la **suite** $(Y_n\ ,\ 1 - \mathbf{1}(A_n))_{n=1,\dots,N}$ supposée rangée selon l'**ordre lexicographique**.

L'estimateur du produit limite (7) s'écrit aussi sous la forme équivalente suivante :

$$(7)' \quad F_N^{\#}(x) = \begin{cases} 1 - \prod_{n \in Bn(x)} r_n\ , \quad \forall\, x \in \mathbf{R}_+\ , \\ 1 \text{ sinon,} \end{cases}$$

dans laquelle $r_n = \{R_n^{-1} (R_n - 1)\}^{\mathbf{1}(An)}$ et $R_n = \Sigma_{\alpha=1}^{N} \mathbf{1}[Y_\alpha \geq Y_n]$ (nombre d'instants $Y_\alpha$ tq $Y_\alpha \geq Y_n$).

(iii) En présence de censure aléatoire, la fonction de survie S = 1 - F est estimée naturellement à l'aide de $1 - F_N^{\#}$ (cf **statistique naturelle**).

(7)" $S_N^{\#}(x) = 1 - F_N^{\#}(x)$.

$F_N^{\#}$ (donc $S_N^{\#}$) possède ainsi des sauts aux seuls instants $X_n$ non censurés.

Sous certaines conditions, on montre que $F_N^{\#}$ (resp $S_N^{\#}$) est un estimateur fortement convergent de F (resp S) lorsque $N \to +\infty$ (cf **convergence forte**).

On peut encore estimer d'autres **caractéristiques** de F. Ainsi, la **fonction quantile** $F^{-1}$ peut s'estimer par :

(8) $q_N^{\sim}(p) = \inf \{x \in \mathbf{R}_+ : F_N^{\#}(x) \leq p\}, \qquad \forall\, p \in ]0, 1[$

(iv) Lorsque $\xi$ est censurée à droite par une variable de censure $\zeta$, on observe une suite aléatoire $(Y_n , \delta_n)_{n=1,...,N}$ , avec $Y_n = \min (X_n , Z_n)$ et $\delta_n = \mathbf{1}([X_n \leq Z_n])$. On suppose encore que les va $X_n$ forment une **suite iid** selon F et que les variables de censure $Z_n$ sont iid selon H, ces deux suites étant indépendantes entre elles.

L'**estimateur de E.L. KAPLAN - P. MEIER** de S s'écrit aussi :

(7)''' $$S_N^{\#}(x) = \begin{cases} \Pi_{n=1}^{N} \{(K(Y_n))^{-1} (K(Y_n) - 1)\}^{\mathbf{1}(An(x))}, & \forall\, x < Y^{(N)}, \\ 0, & \forall\, x \geq Y^{(N)}, \end{cases}$$

avec $K(x) = 1 + \Sigma_{\alpha=1}^{N} \mathbf{1}([Y_\alpha > x])$, $\forall\, x \geq 0$ , $Y^{(.)} = (Y^{(1)} ,..., Y^{(N)})$ (**statistique d'ordre** de Y, dans laquelle on suppose que $Y^{(n)} < Y^{(n+1)}$ (le cas d'observations multiples est négligeable si F et G sont continues), $A_n(x) = \{Y_n \leq x, \delta_n = 1\}$ et $\mathbf{1}(An(x))$ dénote $\mathbf{1}(A_n(x))$, où $\mathbf{1}(B)$ désigne la **fonction indicatrice** d'une partie B.

Si $N \to +\infty$, si F et H sont continues et si l'on note $T < T_G = \inf \{x \geq 0 : G(x) = 1\}$ (avec $G(T) < 1$), on montre que le processus : $S_N^*(x) = N \{F_N^{\#}(x) - S(x)\}$, avec $0 \leq x \leq T$, $N \in N_N^*$, converge en probabilité vers un **processus gaussien** $U = (U_x)_{x \in [0,T]}$ tq $E\, U_x = 0$ et $C(U_x , U_y) = g(x) S(x) S(y)$, $\forall\, (x, y) \in [0,T]^2_{\leq}$ , avec $g(.) = \int dF / (1 - F)^2 (1 - H)$, $\forall\, x < T_F$ , où $T_F = \inf \{x \geq 0 : F(x) = 1\}$ (cf **convergence en probabilité**).

(v) Pour calculer des intervalles de confiance (asymptotiques), divers estimateurs de la variance (asymptotique) de l'estimateur de KAPLAN - MEIER on été définis (l'un d'eux étant défini par la **formule de M. GREENWOOD**).

(vi) Les estimateurs précédents s'étendent à des va réelles non nécessairement positives (mêmes formules).

Si l'on note $S_N^{\#} = 1 - F_N^{\#}$ l'estimateur de la fonction de survie S et si l'on pose :

(10) $$C_x = \Sigma_{n=1}^{N} \mathbf{1}\,[Y_n \geq x], \qquad \forall\, x \in \mathbf{R}_+ ,$$
$$C(x) = E\, C_x ,$$

alors, sous des conditions assez générales (portant notamment sur Z), $S_N^{\#}(x)$ est (ponctuellement) asymptotiquement normale (cf **convergence en loi**, **normalité asymptotique**) :

(11) $\mathcal{L}\{S_N^{\#}(x) - S(x)\} \rightarrow_{N \rightarrow +\infty} \mathcal{N}_1(0, \sigma_\infty^2(x)), \quad \forall x \in \mathbf{R}_+$ ,

la variance asymptotique de $S_N^{\#}$ étant donnée par :

(12) $\sigma_\infty^2(x) = -\{S(x)\}^2 \int_{[0,x]} \{C(t) S(t)\}^{-1} dS(t)$.

Cette dernière est estimée par la **statistique** $s_N^2(x)$ obtenue en remplaçant dans (12) $S(x)$ par $S_N^{\#}(x)$, $-dS(Y_n) / S(Y_n)$ par le rapport $(1 - r_n) / r_n$ et $C(Y_n)$ par $R_n$ ($\forall n \in N_N^*$).

Les propriétés précédentes permettent de définir des **tests d'hypothèses** (asymptotiques) portant sur F, sur S ou sur les **quantiles** associés.

(v) Les notions précédentes se généralisent à plusieurs dimensions.

Dans le cas de deux dimensions, on note :

(13) $S(x_1, x_2) = P([\xi_1 > x_1] \cap [\xi_2 > x_2])$

la **fonction de survie bidimensionnelle** d'un **couple aléatoire** $(\xi_1, \xi_2)$ à valeurs dans $\mathbf{R}_+^2$, $X = (X_1, ..., X_N)$ un **échantillon iid** constitué de **copies** de $(\xi_1, \xi_2)$, avec $X_n = (X_{1n}, X_{2n})$, $\forall n \in N_N^*$, $(\zeta_1, \zeta_2)$ un couple censurant le précédent à droite, et $Z = (Z_1, ..., Z_N)$ un échantillon constitué de copies iid de $(\zeta_1, \zeta_2)$, avec $Z_n = (Z_{1n}, Z_{2n})$, $\forall n \in N_N^*$.

Par suite, on observe, $\forall n \in N_N^*$, les couples aléatoires :

(14) $Y_n = (Y_{1n}, Y_{2n})$, avec $Y_{kn} = \min(X_{kn}, Z_{kn})$, $(k = 1,2)$,

ainsi que les **variables de censure** :

(15) $\delta_n = (\delta_{1n}, \delta_{2n})$, avec $\delta_{kn} = \mathbf{1}([X_{kn} = Y_{kn}])$, $(k = 1,2)$.

Ainsi, lorsque $\delta_n = (0, 0)$, les deux coordonnées de $X_n$ sont censurées ; lorsque $\delta_n = (1, 1)$, elles ne le sont pas, et l'on observe alors $Y_n = X_n$ ; etc.

Si les $X_n$ sont indépendantes des $Z_n$ , on peut définir un **estimateur de KAPLAN-MEIER** de F (resp de S).