MÉTHODE DE EFRÖN (C12, F, G, H)
(21 / 04 / 2020, © Monfort, Dicostat2005, 2005-2020)

La **méthode de EFRÖN** est une méthode générale d'estimation par « **ré-échantillonnage »** effectué à partir de la **loi empirique**, ie par **simulation** basée sur celle-ci. Le **statisticien** estime ainsi une **caractéristique** relative à une **loi de probabilité** (théorique), puis assimile l'**estimation** obtenue à la « **vraie valeur** » de cette caractéristique.

(i) Soit $(\Omega, \mathscr{T}, \mathscr{P})$ un **modèle statistique** (sous forme non paramétrée), $(\mathscr{X}_0, \mathscr{B}_0)$ un **espace d'observation**, $\xi : \Omega \mapsto \mathscr{X}_0$ une **va** donnée, de **loi** P^ξ, et $X = (X_1, \ldots, X_N)$ un **échantillon iid** issu de ξ, ie un **échantillon indépendant** et de loi commune P^ξ. Soit $(\mathscr{X}, \mathscr{B}, \mathscr{P}^X) = (\mathscr{X}_0{}^N, \mathscr{B}_0{}^{\otimes N}, \mathscr{P}^X)$ le **modèle d'échantillonnage** (à distance finie), image du modèle de base initial par X, avec :

(1) $\quad P^X = (P^\xi)^{\otimes N} \in \mathscr{P}^X$.

On considère le **problème d'estimation** suivant. Soit $(\Gamma, \mathscr{B}_\Gamma)$ un espace de **caractéristiques** associé à $\mathscr{P}^X$ et T_N une **statistique** définie par l'**application mesurable** $t_N : \mathscr{X}_0{}^N \times \mathscr{P}^\xi \mapsto \Gamma$ selon :

(2) $\quad T_N = t_N (X, P^\xi), \qquad \forall P^\xi \in \mathscr{P}^\xi = \xi(\mathscr{P})$.

Pour estimer une caractéristique $\gamma = c(P^\xi)$ associée à la lp P^ξ, la **méthode de B. EFRÖN** (équivalent anglais : « *bootstrap method* ») consiste en la **procédure statistique** suivante :

(a) tirage avec remise d'un N-**échantillon aléatoire** $X^* = (X_1^*, \ldots, X_N^*)$ à l'aide de la **loi empirique** $P_N = N^{-1} \Sigma_{n=1}^N \delta(X_n)$ associée à X, où $\delta(a)$ désigne la **loi de DIRAC** placée au point a (cf **échantillon avec remise**, **tirage bernoullien**) ;

(b) calcul de l'**estimateur** de γ défini par la statistique :

(3) $\quad T_N^* = t_N (X^*, P_N)$.

(ii) A titre d'exemples :

(a) si $\mathscr{X} = \mathbf{R}$, et $P = \{\mathscr{N}_1(\gamma) : \gamma = (\mu, \sigma) \in \mathbf{R} \times \mathbf{R}_+^*\}$ (famille des **lois normales** scalaires), on peut estimer le couple γ (**moyenne** et **écart-type**) à l'aide de la statistique (cf **normalisation**, deuxième sens) $t_N (X, P^\xi) = (\overline{X}_N - \mu) / \sigma$, où $\overline{X}_N$ désigne la **moyenne empirique** associée à X ;

(b) soit $\Gamma = \mathscr{F}$ l'ensemble des **fr** F associées aux lois P^{ξ} et $p \in]0, 1[$. On peut calculer $t_N (X, P^{\xi}) = F_N^{-1} (p) - F^{-1} (p)$, où F_N désigne la **fonction de répartition empirique** de X (ie la fr associée à P_N) et F_N^{-1} (resp F^{-1}) l'inverse continue à droite de F_N (resp de F) (cf **fonction quantile**).

(iii) La méthode permet ainsi d'étudier une « statistique » T_N qui peut aussi dépendre de P^{ξ} (cf eg **fonctionnelle**, **fonction pivotale**). Elle s'applique aussi au cas où $\mathscr{P}$ est une famille paramétrée $(P_\theta)_{\theta \in \Theta}$, avec $\Theta \subset \mathbf{R}^Q$ (ie une famille « paramétrique ») : on remplace alors F_N par la **fonction de vraisemblance** estimée au point $\theta_N^{\sim}$ (cf **estimateur du maximum de vraisemblance**).

(iv) La lp de la statistique T_N est appelée **loi de B. EFRON** (en anglais : « *bootstrap distribution* »).

Si $\Gamma = \mathbf{R}$ et $s_N : \mathscr{X} \mapsto \Gamma$ est une fonction mesurable définissant une statistique $S_N = s_N (X)$ comme estimateur du paramètre $\gamma = c (P^{\xi})$, le choix de la statistique $t_N = s_N - \gamma$, ie de $t_N (X) = s_N (X) - c (P^{\xi})$, permet de calculer le **biais** de l'estimateur S_N , ie $B_{S(N)} = E\, s_N (X) - c (P^{\xi})$. Ceci est à rapprocher de la **méthode de QUENOUILLE** (« jackknife »).

Si $P^{\xi} = P_N$ (loi empirique), la statistique T_N^* suit la même loi que T_N .

(v) La **méthode de B. EFRON** peut aussi être présentée comme suit. Soit $(\mathscr{X}_0^N, \mathscr{B}_0^{\otimes N}, \mathscr{P}^X)$ le modèle statistique précédent, où $X = (X_1 ,..., X_N)$ est un **échantillon iid** selon P^{ξ}, et $(\mathscr{Y}, \mathscr{C})$ un **espace mesurable** auxiliaire. On définit une statistique S à l'aide d'une application mesurable $s : \mathscr{X} \mapsto \mathscr{Y}$ selon $S = s (X)$, avec $\mathscr{X} = \mathscr{X}_0^N$. On note alors P^S la loi de S, F la fr de $P^{\xi} \in \mathscr{P}^{\xi}$ et F_N la **fr empirique** associée à X.

Par suite, l'**estimateur de B. EFRON** (en anglais : « *bootstrap estimator* ») de $P^S = \mathscr{L}_N (S / F)$ (loi de S sachant F, loi qui dépend de X) est, par définition, $\mathscr{L}_N (S / F_N)$ (ie est obtenu en remplaçant F par F_N). Cet estimateur est « calculable » puisque, par hypothèse, X et F_N sont **observables** et que s est donnée.

(vi) L'interprétation de la méthode est la suivante. Si $X^* = (X_1^* ,..., X_N^*)$ est un échantillon iid selon la loi P_N (de fr associée F_N), alors $\mathscr{L}_N (S, F_N)$ n'est autre que la loi de $S^* = s (X^*)$.

En pratique, on peut approximer (et tabuler) cette loi par **simulation** (cf **méthodes de MONTE CARLO**) en tirant des échantillons tq X^*.

(vii) La méthode de EFRON permet aussi l'**évaluation d'une** **procédure statistique** lorsqu'il est possible de renouveler l'**échantillonnage** (cf **renouvellement**).

Ainsi, $(\mathscr{X}, \mathscr{B}, (P_\theta^X)_{\theta \in \Theta})$ étant un modèle statistique associé à un **échantillon aléatoire** $X = (X_1, ..., X_N)$ et $s : \mathscr{X} \mapsto \Theta$ une application mesurable définissant la statistique $S = s(X)$ (ie un **estimateur** de θ), on peut tirer (avec remise), dans l'ensemble $\{X_1, ..., X_N\}$, k échantillons aléatoires $X_i = (X_{i1}, ..., X_{iN})$ (de même taille N), iid selon la loi empirique P_N associée à X (où i = 1 ,...,, k).

Par suite, les statistiques $S_i = s(X_i)$ ($\forall\, i \in N_k^*$) permettent d'étudier (ou d'estimer) la **loi de l'estimateur** S lui-même. Dans ce contexte, $P^\kappa = k^{-1} \Sigma_{i=1}^k \, \delta\,(S_i)$ intervient souvent comme loi « empirique » associée aux S_i . En particulier, si $\Theta = \mathbf{R}$, on peut calculer :

(4) $\bar{S}_k = k^{-1} \Sigma_{i=1}^k S_i$ (**moyenne** des S_i),

(5) $V_k^2 = (k - 1)^{-1} \Sigma_{i=1}^k (S_i - \bar{S}_k)^2$ (**variance** des S_i).

Sous certaines hypothèses, un **test d'hypothèses** (asymptotique) peut se baser sur la **propriété asymptotique** :

(5) $\mathscr{L}((S - \bar{S}_k) / V_k) \to_{k \to +\infty} \mathscr{N}_1 (0, 1)$ (**loi normale réduite**).

(viii) De même, si X est un N-échantillon iid selon P_θ^X et si le **paramètre** θ est estimé à l'aide d'un estimateur $T = t(X)$, on peut étudier eg la **variabilité** de T pr à θ en procédant comme suit :

(a) estimation de la fr $F(. , \theta)$ associée à P_θ^X à l'aide de la fr empirique F_N ;

(b) tirage de k échantillons X_i de taille N, iid selon F_N , et calcul des estimateurs $T_i = t(X_i)$ qui en résultent, avec $i \in N_k^*$;

(c) calcul de la variabilité (eg **variance**) des T_i pr à T : cette variabilité estime alors celle de T pr à θ.