SONDAGE (M2)
(04 / 10 / 2019, © Monfort, Dicostat2005, 2005-2019)

(i) Soit $\Omega = \{\omega_1 ,..., \omega_M\}$ un ensemble fini (**population**), $(\mathcal{Y}, \mathcal{C})$ un **espace d'observation**, $\eta : \Omega \mapsto \mathcal{Y}$ une **variable** (ou un **caractère**) observable et $Y = (Y_1 ,..., Y_M)$ l'ensemble des valeurs prises par η dans la population Ω : ie $Y_m = \eta (\omega_m)$, $\forall$ m = 1 ,..., M.

Etant donné un **plan de sondage** Π sur Ω, on note $A = \{a_1 ,..., a_N\}$ un N-**échantillon aléatoire** tiré dans Ω selon Π et $y = (y_1 ,..., y_N)$ les valeurs prises par η sur l'échantillon A (ie $y_n = \eta (a_n)$, $\forall$ n = 1 ,..., N).

On cherche à estimer une fonction g (Y) du « **paramètre** » Y, où $g : \mathcal{Y}^M \mapsto \mathcal{Z}^L$ ($L \leq N$) est une fonction donnée à valeurs dans un **espace mesurable** $(\mathcal{Z}^L, \mathcal{D}^{\otimes L})$.

On appelle **sondage** dans (resp sur, resp de) Ω la donnée d'un couple (Π, t_N) constitué d'un plan de sondage Π et d'un **estimateur** T_N de g (Y) défini par une **application mesurable** $t_N : \mathcal{Y}^N \mapsto \mathcal{Z}^L$:

(1) $\quad T_N = t_N (y)$.

(ii) En pratique, on a souvent $\mathcal{Y} = \mathcal{Z} = \mathbf{R}$ et $\mathcal{C} = \mathcal{D} = \mathcal{B}_{\mathbf{R}}$. Par ailleurs, on note $\mathcal{L} (T_N)$ la **loi** de T_N , ie l'image de la loi de y par t_N , ou encore l'image de Π par $t_N \circ \eta_{/A}$, où $\eta_{/A}$ est la **restriction** de η à A.

(iii) On peut comparer deux sondages (Π' , t_N') et (Π'' , t_N'') sur Ω à l'aide d'un critère de coût (cf **fonction de coût**) combiné avec un critère d'**optimalité** statistique.

Ce dernier critère peut être un critère d'**efficacité relative** de t_N'' pr à t_N' (mesurée eg par l'**écart quadratique moyen**).

Ainsi, si $(A, P) \mapsto c (A, \Pi)$ représente une fonction de coût (nécessaire pour obtenir l'échantillon aléatoire (A, y)), on dit que le sondage (Π', t_N') est préférable au sondage (Π'', t_N''') ssi, à la fois :

(2)
$$c (A, \Pi') \leq c (A, \Pi'')$$
$$Q\, T_N' \leq Q\, T_N'', \qquad \forall\, Y \in \mathcal{Y}^M,$$

inégalités dans lesquelles $T_N' = t_N' (y)$, $T_N'' = t_N'' (y)$ et $Q\, T_N$ désigne l'**écart quadratique moyen** de T_N (ce qui suppose ici $\mathcal{Z} = \mathbf{R}$) :

(3) $\quad Q\, T_N = E_\Pi (T_N - g (Y)).(T_N - g (Y))' = \Sigma_{A \in \mathcal{A}} \Pi (A).(T_N - g (Y)).(T_N - g (Y))'$,

où z' est le transposé de z, E_Π est l'**espérance mathématique** calculée à l'aide de la probabilité Π et $\mathcal{A}$ désigne l'ensemble des échantillons (avec remise) de Ω (cf **théorie des sondages**) ;

(iv) On substitue souvent à la première inéquation (2) la suivante, qui traduit un critère de coût moyen minimum :

(4) $\quad m(\Pi') \leq m(\Pi'')$,

où :

(5) $\quad m(\Pi) = E_\Pi\, c(A, \Pi) = \Sigma_{A \in \mathcal{A}}\, \Pi(A).\, c(A, \Pi)$

dénote le coût moyen du sondage.

(v) Un sondage (P, t_N) est un **sondage sans biais** ssi :

(6) $\quad E_\Pi\, T_N$ ou $E_\Pi\, t_N(y) = \Sigma_{A \in \mathcal{A}}\, \Pi(A).\, t_N(y) = g(Y), \quad \forall\, Y \in \mathcal{Y}^M$.

(vi) Dans l'optique bayésienne (cf **école bayésienne**, **postulat de BAYES**, **principe bayésien**, **règle de BAYES**, **théorie bayésienne**), on suppose que $Y \sim Q^Y$ (**loi a priori**, élément d'une **famille** donnée $\mathcal{Q}^Y$), on remplace la deuxième inéquation (2) par l'inéquation :

(7) $\quad R_Q\, T_N' \leq R_Q\, T_N'', \quad \forall\, Q^Y \in \mathcal{Q}^Y$,

où $R_Q\, T_N = E_Q\, (Q\, T_N)$ représente le **risque de BAYES** associé à T_N .