SONDAGE SYSTÉMATIQUE (M3, M6)
(16 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

Un **sondage systématique** est un **sondage**, aléatoire ou non, effectué selon une « méthode systématique » : partant d'une première **unité statistique** (tirée au hasard ou non), on tire d'autres unités « équidistantes » de la première (dans l'espace ou dans le temps), ou équidistantes entre elles, jusqu'à atteindre une taille d'échantillon N donnée.

(i) En particulier, on appelle **sondage systématique** le sondage suivant. On étudie un ensemble fini $\Omega = \{\omega_1 ,..., \omega_M\}$ (eg **population**), dont les éléments (**unités statistiques**) sont dotés d'un « **caractère** » observable η. Autrement dit, il existe une application $\eta : \Omega \mapsto \mathcal{Y}$, où $(\mathcal{Y}, \mathcal{C})$ est un **espace d'observation** donné. On note $Y = (Y_1 ,..., Y_M)$ les valeurs observables sur la population, avec $Y_m = \eta (\omega_m)$ $(m = 1 ,..., M)$.

$N \in \mathbf{N}^*$ étant donné, soit alors $F = N / M$ un **taux de sondage** tel que $L = 1 / F = M / N$ soit un nombre entier (ie $L \in \mathbf{N}^*$).

On appelle **sondage systématique** dans Ω un sondage défini par le **plan de sondage** suivant :

(a) on prélève dans Ω un élément ω_l dont l'indice $l \in \{1 ,..., L\}$. Cet élément constitue le premier élément a_1 de l'échantillon A à extraire de Ω ;

(b) on prélève ensuite, systématiquement, les éléments $a_n \in A$ selon la suite :

(1) $a_2 = \omega_{L+l}$,..., $a_N = \omega_{(N-1)L+l}$.

La **partition** des indices relatifs aux unités de Ω est la suivante :

$\{1 ,..., L\}, \{L+1 ,..., 2 L\} ,..., \{(n-1) L + 1 ,..., n L\} ,..., \{(N-1) L + 1 ,..., M\}$.

Par suite, le sondage systématique est aléatoire ssi l est tiré au hasard uniforme dans $N_L^* = \{1 , ..., L\}$, ie ssi $p (\lambda) = P ([l = \lambda]) = L^{-1}$, $\forall \lambda \in N_L^*$.

Dans ce contexte, on appelle parfois **table de C.H.D. BUYS-BALLOT** le tableau associé à la matrice suivante (cf aussi **composante d'une série temporelle**, **série temporelle**) :

$$(2) \qquad Y = \begin{matrix} Y_1 & Y_{L+1} & ... & Y_{(n-1) L+1} & ... & Y_{(N-1) L+1} \\ ... & ... & ... & ... & ... & ... \\ Y_l & Y_{L+l} & ... & Y_{(n-1) L+l} & ... & Y_{(N-1) L+l} \\ ... & ... & ... & ... & ... & ... \\ Y_L & Y_{2L} & ... & Y_{n L} & ... & Y_{N L} \end{matrix}$$

Le plan de sondage systématique revient donc à tirer, selon une probabilité uniforme, une ligne du tableau Y précédent. On appelle parfois L la **période du tirage**, ou **période du plan**.

(ii) La relation suivante :

(3) $Y_{ln} = Y_{(n-1)\,L\,+\,l}$, $l = 1, ..., L$, $n = 1, ..., N$,

revient à remplacer l'indice double (l, n) par un indice simple (cf **indice**, **modèle multi-indicé**).

Par suite, l'**échantillon** $y = (y_1, ..., y_N)$ des valeurs observées sur l'échantillon A des unités tirées est défini selon $y_n = \eta\,(a_n)$ $(n = 1, ..., N)$. Cet échantillon est tq $y_n = \eta\,(\omega_{(n-1)\,L\,+\,l})$ $(\forall\ n)$.

On note alors :

(4)
$$Y_{l.} = N^{-1} \sum_{n=1}^{N} Y_{ln} \qquad \text{(**moyenne** de } \eta \text{ dans la ligne l)},$$
$$Y.. = M^{-1} \sum_{l=1}^{L} \sum_{n=1}^{N} Y_{ln} \qquad \text{(moyenne générale de } \eta \text{ dans le (tableau Y)}.$$

Le sondage est aléatoire si l est tiré au hasard (uniforme) dans N_L^*, ie si $l \sim \mathcal{U}(N_L^*)$ (**loi uniforme discrète**). Par suite, la moyenne $Y_{l.}$ est une **variable aléatoire** tq :

(5) $P\,(Y_{l.} = Y_{k.}) = L^{-1}$, $\forall\ k = 1, ..., L$,

où $Y_{k.}$ est une ligne donnée de Y.

De plus :

(6)
$$E\,Y_{l.} = Y..\ \text{(**estimateur sans biais**)}, \qquad \forall\ l = 1, ..., L$$
$$V\,Y_{l.} = (N^2\,L)^{-1} \cdot \sum_{l=1}^{L} \{\sum_{n=1}^{N} (Y_{ln} - Y..)\}^2.$$

(iii) On établit alors que :

(a) le sondage systématique est aussi efficace qu'un **sondage élémentaire** (bernoullien ou exhaustif) si la **covariance** (théorique) entre deux colonnes du tableau Y précédent est nulle (ie lorsque l'ordre selon lequel les unités sont rangées dans Y est purement aléatoire) ;

(b) par contre, le sondage systématique est moins efficace si cette même covariance est positive. Un cas limite est celui où le caractère η présente des variations périodiques entre les unités, la période étant égale à L : l'unité ω_l est identique à l'unité ω_{l+L} pour le caractère η considéré ;

(c) lorsque la covariance précédente est négative, le sondage systématique est plus efficace que le sondage élémentaire, les unités « voisines » (pour le

caractère η) ayant tendance à se ressembler. C'est dans cette situation que ce type de sondage est utilisé.