SONDAGE STRATIFIÉ (M3)
(12 / 02 / 2022, © Monfort, Dicostat2005, 2005-2022)

Un **sondage stratifié** est un **sondage** qui tient compte d'une **information** que l'on possède a priori sur les différentes **parties** de l'**ensemble** (**population**) étudié(e).

(i) Soit $\Omega = \{\omega_1, ..., \omega_M\}$ un ensemble fini et $\Pi_\Omega = \{\Omega_1, ..., \Omega_H\}$ ($H \leq M$) une **partition** de Ω en H classes, appelées **strates** de l'ensemble. On note $M_h = \text{Card}\ \Omega_h$ le nombre d'éléments de la strate Ω_h (avec $\Sigma_{h=1}^{H} M_h = M$) et ω_{hm} l'élément (**unité de sondage** ou individu) générique de cette strate, avec $m = 1, ..., M_h$ et $h = 1, ..., H$.

On extrait de Ω un **échantillon** $A = \{a_1, ..., a_N\}$ de taille N et l'on note $\Pi_A = \{A_1, ..., A_H\}$ la **partition trace** de Π_Ω sur A. On pose $N_h = \text{Card}\ A_h$ (nombre d'unités dans la strate d'échantillon h, avec $\Sigma_{h=1}^{H} N_h = N$) et $f_h = N_h / M_h$ les **taux de sondage par strate** h.

(ii) Si $(\mathcal{Y}, \mathcal{C})$ est un **espace d'observation**, on étudie un « **caractère** » η défini sur Ω, ie une application $\eta : \Omega \mapsto \mathcal{Y}$. On note $Y = (Y_1, ..., Y_M)$ la **suite** des valeurs observées sur l'ensemble Ω, avec $Y_m = \eta(\omega_m)$ ($m = 1, ..., M$) et $y = (y_1, ..., y_N)$ la suite des valeurs observées sur l'échantillon A, avec $y_n = \eta(a_n)$ ($n = 1, ..., N$). La suite finie Y joue ici un rôle de **paramètre**.

Par suite, (h, m) étant un bi-indice repérant l'unité m dans la strate h, on note aussi $Y_{hm} = \eta(\omega_{hm})$ et $y_{hn} = \eta(a_{hn})$ les valeurs observées correspondantes ($\forall$ h, $\forall$ m et $\forall$ n). On définit enfin, par strate Ω_h, des **caractéristiques** usuelles tq :

(1)
$$\bar{Y}_h = (M_h)^{-1} . \Sigma_{h=1}^{H} Y_{hm} \quad (\textbf{moyenne} \text{ théorique de strate}),$$
$$S_h^2 = (M_h - 1)^{-1} . \Sigma_{h=1}^{H} (Y_{hm} - \bar{Y}_h)^2 \ (\textbf{variance} \text{ théorique corrigée de strate}),$$

ainsi que les caractéristiques de strate correspondantes de l'échantillon A_h (ou y) :

(2)
$$\bar{y}_h = (N_h)^{-1} . \Sigma_{n=1}^{Nh} y_{hn} \quad (\text{moyenne empirique de strate})$$
$$s_h^2 = (N_h - 1)^{-1} . \Sigma_{n=1}^{Nh} (y_{hn} - \bar{y}_h)^2 \ (\text{variance empirique corrigée de strate}).$$

(iii) On appelle alors **tirage stratifié**, ou **sondage stratifié**, (selon Π_Ω), **sans remise et avec probabilités égales**, dans Ω un sondage dont le **plan de sondage** Π est tq :

(a) M_h et N_h sont donnés (fixés), $\forall$ h = 1 ,..., H ;

(b) le tirage des unités dans chaque strate est sans remise (cf **sondage exhaustif**) et s'effectue indépendamment des tirages dans les autres strates (cf **indépendance**).

L'**échantillon de NEYMAN** est un exemple d'échantillon stratifié avec taux de sondage variable d'une strate à l'autre.

(iv) Un **estimateur sans biais** de la moyenne théorique d'ensemble :

(3) $\bar{Y} = M^{-1} \Sigma_{m=1}^{M} Y_m = M^{-1} \Sigma_{h=1}^{H} \Sigma_{m=1}^{Mh} Y_{hm} = \Sigma_{h=1}^{H} (M_h / M) \bar{Y}_h$,

est l'**estimateur par stratification**, défini par :

(4) $T_N'' = \Sigma_{h=1}^{H} (M_h / M) \bar{y}_h$.

Sa variance, qui vaut :

(5) $V T_N'' = \Sigma_{h=1}^{H} (M_h / M)^2 (1 - f_h) . (S_h^2 / N_h)$,

est elle-même estimée sans biais par :

(6) $v T_N'' = \Sigma_{h=1}^{H} (M_h / M)^2 (1 - f_h) . (s_h^2 / N_h)$.

(v) Si $f_h = f$ ($\forall$ h = 1 ,..., H) (même **taux de sondage** dans les strates Ω_h), on a :

(7) $M_h / M = N_h / N$ ($\forall$ h = 1 ,..., H).

Dans ce cas, on dit parfois que l'échantillon A défini par Π_A est un **échantillon représentatif** de la population Ω. Par suite, (4) et (5) s'écrivent resp :

(4)' $T_N'' = N^{-1} . \Sigma_{h=1}^{H} \Sigma_{n=1}^{Nh} y_{hn}$

et :

(5)' $V T_N'' = N^{-1} (1 - f) . \Sigma_{h=1}^{H} (M_h / M) . S_h^2$.

(vi) On peut comparer les propriétés du sondage stratifié avec celles d'un sondage exhaustif non stratifié, de même taux de sondage. Ainsi, l'estimateur suivant de $\bar{Y}$:

(8) $T_N = N^{-1} . \Sigma_{h=1}^{H} \Sigma_{n=1}^{Nh} y_{hn}$

possède pour variance :

(9) $V T_N = N^{-1} (1 - f) . S^2$, avec $S^2 = (M - 1)^{-1} . \Sigma_{h=1}^{H} \Sigma_{m=1}^{Mh} (Y_{hm} - \bar{Y})^2$.

En remplaçant $M_h - 1$ et $M - 1$ resp par M_h et M, on obtient :

(10) $V T_N - V T_N'' = N^{-1} (1 - f) . \Sigma_{h=1}^{H} (M_h / M) (\bar{Y}_h - \bar{Y})^2 \geq 0$.

Ainsi, un échantillon stratifié représentatif est toujours plus efficace (au sens de la variance) qu'un échantillon non stratifié de même taille. Le gain de **dispersion** de l'estimateur est d'autant plus important que les strates sont plus hétérogènes entre elles (au sens des écarts pr à la moyenne d'ensemble) (cf **hétérogénéité**).

(vii) En pratique, si le sondage dans la strate Ω_h s'effectue avec un **coût unitaire** c_h (identique pour toutes les unités de Ω_h) (cf aussi **fonction de coût**), et si le budget total disponible pour le sondage s'élève à C, la répartition optimale de ce budget entre les strates est la solution du programme mathématique suivant (cf **programmation mathématique**) :

(11) $\min V T_N''$ sous $\Sigma_{h=1}^{H} N_h c_h = C$,

où $V T_N''$ est défini en (5) (ou (5)').

(vii) La définition des strates s'effectue souvent à l'aide de **caractères** $\xi_k : \Omega \mapsto \mathcal{X}_k$ observables sur les éléments $\omega \in \Omega$, où les $(\mathcal{X}_k , \mathcal{B}_k)$ (k = 1 ,..., K) sont des **espaces d'observation**. Les variables ξ_k sont appelées **variables de stratification**. Si ce sont des **variables quantitatives** (ie numériques, avec $\mathcal{X}_k = \mathbf{R}$, $\forall$ k) la partition Π_Ω induit sur $\mathbf{R}^K = \Pi_{h=1}^{H} \mathcal{X}_k$ une partition Π ($\mathbf{R}^K$). En particulier, si k = 1 et que l'on note ξ la variable de stratification, le choix de Π_Ω induit une partition :

(12) $\Pi_{\mathbf{R}}$ ou Π ($\mathbf{R}$) = $\{]-\infty, x_1[, [x_1 , x_2[,..., [x_{H-2} , x_{H-1}[, [x_{H-1} , +\infty[\}$

avec :

$\Omega_1 = \{\omega \in \Omega : \xi (\omega) < x_1\}$, ...,

$\Omega_h = \{\omega \in \Omega : \xi (\omega) \in [x_{h-1} , x_h[, h = 2 ,..., H-1\}$, ...,

$\Omega_H = \{\omega \in \Omega : \xi (\omega) \geq x_{H-1}\}$.

La connaissance de la **corrélation** entre η et ξ (ou, plus généralement, entre η et ξ_1 ,..., ξ_K) permet de déterminer les **limites de strates** (au moins de façon approximative). On peut eg étudier la **robustesse** des estimateurs retenus pr à des erreurs de « découpage » en strates de la population.

(viii) D'autres **caractéristiques** que la moyenne peuvent être étudiées à partir des définitions précédentes : notamment, le **total** $T = e_M' Y$ dans la population, ou encore la **proportion** p d'unités possédant une propriété donnée.

(ix) Lorsque l'on ne connaît pas de strates a priori, la notion de sondage stratifié doit être adaptée : ceci conduit notamment à la notion de **sondage post-stratifié**. Par distinction, le **sondage stratifié** décrit ici est un **sondage stratifié a priori** : il est souvent appelé **sondage pré-stratifié**, **sondage par pré-stratification**, ou **sondage par stratification a priori**.

(x) Enfin, on peut définir des sondages stratifiés plus complexes : sondage stratifié avec remise des unités, sondage stratifié avec probabilités inégales, sondages avec stratifications emboîtées, etc.