SONDAGE AVEC PROBABILITÉS INÉGALES (M3)
(04 / 10 / 2019, © Monfort, Dicostat2005, 2005-2019)

Un **sondage avec probabilité inégales** est un **sondage** dont l'objet est de sélectionner les **unités de sondage** dans une population en fonction de leur « importance » (« magnitude ») définie par un **caractère statistique**, celui faisant l'objet d'étude.

(i) Soit $\Omega = \{\omega_1 ,..., \omega_M\}$ une **population** finie, $(\mathcal{Y}, \mathcal{C})$ un **espace d'observation**, $\eta : \Omega \mapsto \mathcal{Y}$ un **caractère** observable sur chaque **unité de sondage** $\omega \in \Omega$, et $Y = (Y_1 ,..., Y_M)$ l'ensemble des valeurs $Y_m = \eta (\omega_m)$ $(m = 1 ,..., M)$ prises par η dans la population Ω.

Soit Π un **plan de sondage** de taille N sur Ω, ie une **mesure de probabilité** définie sur $\mathcal{P}(\Omega^N)$ et tq :

(1) $\quad \Pi = \Pi_{n=1}^{N} \Pi_n ,$

où Π_n est une mesure de probabilité sur le n-ième espace facteur Ω du produit Ω^N. On suppose que les probabilités Π_n sont identiques (ie $\Pi_n = \Pi_0$, $n = 1 ,..., N$) et tq (**probabilités inégales**) :

(2) $\quad \Pi_0 = \Sigma_{m=1}^{M} \Pi_{0m} . \delta (\omega_m).$

Π_{0m} est le **poids** de l'unité a_n considérée comme élément de Ω (ie il existe m tq $\Pi_{0n} = \Pi_{0m}$) et $\delta (\omega_m)$ désigne la **masse de DIRAC** placée au « point » ω_m .

On note $A = \{a_1 ,..., a_N\}$ un N-**échantillon aléatoire** tiré avec remise 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). Par suite, $\forall$ n = 1 ,..., N, la va y_n / Π_{0n} vérifie :

(3)
$$E (y_n / \Pi_{0n}) = \Sigma_{m=1}^{M} \Pi_{0m} . (Y_m / \Pi_{0m}) = \Sigma_{m=1}^{M} Y_m = T \quad \text{(espérance)},$$
$$V (y_n / \Pi_{0n}) = \Sigma_{m=1}^{M} \Pi_{0m} . \{(Y_m / \Pi_{0m}) - T\}^2 \quad \text{(variance)},$$

où T est le **total** du caractère η dans toute la population Ω.

Le sondage ainsi défini est appelé **sondage avec probabilités inégales**, ou **sondage non équiprobable**, de taille N avec remise.

(ii) Un **estimateur sans biais** de T est l'estimateur pondéré suivant :

(4) $\quad T_N = (1 / N) \Sigma_{m=1}^{M} (y_n / \Pi_{0n}),$

où les poids Π_{0n} sont supposés connus. Cet estimateur vérifie :

(5) $V\,T_N = (1/N) \sum_{m=1}^{M} \Pi_{0m} . \{(Y_m / \Pi_{0m}) - T\}^2 = N^{-1} V (y_n / \Pi_{0n})$.

Cette **variance** est estimée sans **biais** par :

(6) $v\,T_N = N^{-1} (N - 1)^{-1} \sum_{n=1}^{N} \{(y_n / \Pi_{0n}) - t\}^2$,

où $t = e_N' y = \sum_{n=1}^{N} y_n$ est le total de η dans l'échantillon A.

(ii) Par comparaison avec un **sondage bernoullien** de même taille N, dont l'estimateur de T est $T_N^{\sim} = M . \bar{y}_N$, on montre que (cf **efficacité relative**) :

(7) $V\,T_N^{\sim} - V\,T_N = N^{-1} . \sum_{m=1}^{M} (M - \Pi_{0m}^{-1}) . Y_m^2 \geq 0$.

Ainsi, $T_N^{\sim}$ est plus efficace que T_N si le caractère η est « corrélé » avec la distribution de poids Π_0 , ie si Y est positivement corrélé avec le vecteur $\Pi_0^v = (\Pi_{01}, ..., \Pi_{0M})$.

(iii) Il existe des formules analogues aux précédentes pour un **tirage exhaustif** (ie sans remise).