MODÈLE DE SONDAGE (M2)
(05 / 10 / 2019, © Monfort, Dicostat2005, 2005-2019)

Un **modèle de sondage** est un **modèle statistique** décrivant un **plan de sondage**, ie un mode d'**observation**, d'un **ensemble** (**population**) Ω composé d'**unités statistiques**, assorties de « descripteurs » divers η, considérés comme des **variables aléatoires**. Ce sondage est généralement destiné à étudier un **phénomène** naturel donné. Dans la situation la plus courante, Card $\Omega < +\infty$ (ensemble fini), et l'on peut noter $\Omega = \{\omega_1 ,..., \omega_M\}$, mais M peut être « grand » ($M >> 0$).

(i) L'**observation** est en général « restreinte », ie le **statisticien** n'observe pas l'ensemble Ω en entier (**recensement**), mais seulement une **partie** $A \subset \Omega$ judicieusement définie de cet ensemble, appelée **échantillon** (d'unités statistiques). La raison principale est d'ordre budgétaire (cf **fonction de coût**). Dans le cas usuel d'ensembles au plus dénombrables, A est plutôt définie comme élément du produit Ω^N : ainsi, lorsque Ω est fini (ie Card $\Omega = M$), la taille N de A vérifie $N \leq M$, et l'on peut alors noter $A = \{a_1 ,..., a_N\}$.

Le procédé mis en oeuvre est donc celui du **sondage**. La plupart des sondages (aléatoires) en grandeur réelle sont définis (ou « construits ») à partir de deux types de sondage classiques :

(a) le **sondage bernoullien** (cf **sondage avec remise**, **tirage bernoullien**) ;

(b) le **sondage exhaustif** (cf **sondage sans remise**, **tirage exhaustif**).

(ii) Les **descripteurs** η peuvent être des variables numériques (cf **variable quantitative**) ou non (cf **variable qualitative**).

On associe naturellement :

(a) à la population (d'unités) Ω tiré l'**ensemble des valeurs observées**, ou « **ensemble des observations** » $\eta(\Omega)$. Lorsque Ω est fini, on peut noter $\eta(\Omega) = \{\eta(\omega_1) ,..., \eta(\omega_M)\}$;

(b) à l'échantillon (d'unités) A tiré l'**échantillon des valeurs observées**, ou « **échantillon d'observations** » $\eta(A)$. Lorsque A est fini, on peut noter $\eta(A) = \{\eta(a_1) ,..., \eta(a_N)\}$.

(iii) L'**optimalité** d'un sondage résulte alors d'un arbitrage entre :

(a) des considérations économiques : l'accès à l'**information** « portée » par chaque **unité de sondage** doit être le moins coûteux possible. L'existence d'une **« barrière budgétaire »** interdit généralement une investigation exhaustive de la population. Ceci dépend de la **taille** de la population Ω (eg **grande base de données**) et de la taille de l'échantillon A, mais aussi des **« coûts d'accès »** de

toutes sortes : (in)existence de l'information, accès matériels ou impossibilités physiques, barrières psychologiques, etc ;

(b) des considérations statistiques : l'**estimateur** d'une grandeur d'intérêt (cf **variable d'intérêt**) doit posséder la meilleure **précision** possible.

Par suite, la conception d'un plan suit généralement deux démarches de base :

(a) soit fixation du budget (maximum) et recherche du sondage le plus efficace : taille d'échantillon, plan de sondage. Cette **situation** est courante eg en sociologie ;

(b) soit fixation d'une précision minimale exigible, et adaptation du budget de façon conforme. Cette situation peut se trouver notamment eg en biologie.

(iii) Les observations issues d'un sondage sont d'une double variabilité :

(a) la **variabilité « intrinsèque »** du phénomène considéré. Ainsi, les informations « exhaustives » (ie pouvant être obtenues de la population Ω), notées symboliquement η (Ω), correspondent à des unités de sondage plus ou moins diverses : les **attributs** (ou **caractères**) possèdent donc une dispersion entre elles. Par suite, la « loi » P_M régissant l'ensemble de ces données (ie la distribution η (Ω) = $\{\eta(\omega_1), ..., \eta(\omega_M)\}$) possède des paramètres qui permettent de « révéler » cette dispersion ;

(b) la **variabilité externe**, dûe au plan de sondage lui-même, puisque les unités de A sont extraites de Ω selon la probabilité Π du plan de sondage.