PLAN DE SONDAGE (M2)
(16 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

Un **plan de sondage** définit la **probabilité** de tirage d'un **échantillon aléatoire** dans un ensemble donné (« **population** »), ie la façon dont l'échantillon est « extrait » de cet ensemble.

(i) Soit Ω un **ensemble**, $\mathcal{Y}$ un ensemble d'observation et $\eta : \Omega \mapsto \mathcal{Y}$ une **variable** (eg un **caractère**) définie sur les éléments (« unités ») ω de Ω.

(a) on appelle **échantillon de taille N** $\in$ **N* extrait** de Ω un N-uple A = (a_1 ,..., a_N) constitué d'éléments $a_n \in$ A. Autrement dit, A $\in \Omega^N$. Chaque **unité statistique** a_n est alors appelée **unité de sondage** ;

(b) on appelle **échantillon de taille N** $\in$ **N* observé** sur $\mathcal{Y}$ le N-uple y = (y_1 ,..., y_N) constitué des éléments (ou « **observations** ») $y_n \in \mathcal{Y}$, avec $y_n = \eta(a_n)$, $\forall$ n $\in$ **N***, ie y = ($\eta(a_1)$,..., $\eta(a_N)$). Autrement dit, y $\in \mathcal{Y}^N$.

On doit ainsi distinguer entre :

(a) l'échantillon A constitué d'**unités de sondage** (ou éléments) a_n ;

(b) l'échantillon y constitué des **observations** (ou « mesures ») y_n effectuées sur ces éléments.

Lorsque Ω est fini (cas usuel), on appelle **plan de sondage aléatoire**, ou **tirage aléatoire**, de taille N sur Ω toute **mesure de probabilité** Π définie sur $\mathcal{P}(\Omega^N)$ (**ensemble** des **parties** de Ω^N). Une telle probabilité Π définit le tirage d'un N-uple A dans Ω. Par suite, y = ($\eta(a_1)$,..., $\eta(a_N)$), parfois notée η (A), est une **variable aléatoire** dont la **loi de probabilité** n'est autre que l'image de P par $\eta : \Omega^N \mapsto \mathcal{Y}^N$. Cette loi est une **loi discrète**.

(ii) La définition prend en compte, en particulier, deux types de sondage (ou « tirages ») classiques :

(a) le **sondage avec remise**, ou **sondage avec répétition**, dans lequel une unité ω peut être tirée plus d'une fois. La « diagonale » $\Delta(\Omega^N) = \{A \in \Omega^N : a_n = a_0, \forall n\}$ est alors la seule partie de Ω^N « chargée » par Π (cf **sondage bernoullien**, **tirage bernoullien**) ;

(b) le **sondage sans remise**, ou **sondage sans répétition**, dans lequel une unité ne peut être tirée qu'une fois au plus. La diagonale $\Delta(\Omega^N)$ n'est alors pas chargée par Π (probabilité nulle) (cf **tirage exhaustif**).

Dans ces deux sondages élémentaires, deux coordonnées distinctes y_α et y_β de y peuvent cependant être égales :

(a) soit parce que $\eta\ (a_\beta) = \eta\ (a_\alpha)$ pour deux éléments distincts ($\beta \neq \alpha$) de A (que le tirage soit avec ou sans remise) (eg lorsque l'ensemble des valeurs $\mathcal{Y}$ est un ensemble discret),

(b) soit parce que $a_\beta = a_\alpha$ (tirage avec remise seulement).

(iii) Lorsque Π est un plan de taille N donnée, on dit aussi que Π est un **plan de taille fixe**. Si Card Ω = M $< +\infty$, le rapport F = N / M est appelé **taux de sondage** (il est aussi noté T, f ou t).

Si Π est un plan de sondage sans remise et si A = Ω (ie si N = M = Card Ω), ce plan est appelé **recensement** : celui-ci consiste à extraire, une à une, sans les remettre, toutes les unités $\omega \in \Omega$.

(iv) Dans les deux exemples de base précédents, les plans de sondage associés sont resp les suivants :

(a) le **plan de sondage bernoullien** (ou **plan de sondage avec remise**) :

(1) $\quad \Pi\ (A) = \Pi\ (a_1, ..., a_N) = \prod_{n=1}^{N} \Pi_0\ (a_n) = (1 / M)^N = M^{-N}$.

En effet, dans ce cas, les tirages sont indépendants et l'on définit sur chaque espace facteur Ω de Ω^N une probabilité uniforme Π_0 (avec $\Pi_0\ (a_n) = M^{-1}$, $\forall\ n \in N_N^*$) (cf **loi uniforme discrète**). La formule (1) définit donc la probabilité de tout N-échantillon A, avec $\Pi = \Pi_0^{\otimes N}$. Un tel échantillon est un **échantillon iid** ;

(b) le **plan de sondage exhaustif** (ou **plan de sondage sans remise**) :

(2) $\quad \Pi\ (A) =$

$$\Pi\ (a_1, ..., a_N) = \Pi_1\ (a_1) \cdot \prod_{n=2}^{N} \Pi_n\ (a_n),$$

$$M^{-1} \prod_{n=2}^{N} \{M - (N - 1)\}^{-1} = (A_M^N)^{-1},$$

où A_M^N désigne le nombre d'**arrangements** sans répétition des N unités parmi les M. Ici, les tirages ne sont pas indépendants et l'on définit sur chaque espace facteurs Ω de Ω^N une probabilité Π_n , qui est une loi uniforme conditionnellement aux tirages antérieurs, ie : $\Pi_n\ (a_n) = \Pi_n\ (a_n / a_1, ..., a_{n-1}) = \{M - (N - 1)\}^{-1}$, $\forall\ n = 2, ..., N$, avec $\Pi_1\ (a_1) = 1 / M$.

Autrement dit :

$$\Pi\ (\omega) = \begin{cases} \{M\ (M - 1)\ ...\ (M - N + 1)\}^{-1} & \text{si les } \omega_m \text{ sont} \neq 2 \text{ à } 2, \\ 0 & \text{sinon.} \end{cases}$$

(v) Plus généralement, un **plan de sondage**, ou **tirage**, (aléatoire) dans Ω est une **mesure de probabilité** Π définie sur l'ensemble :

(3) $\quad \mathcal{E} = \cup_{N \in \mathbf{N}^*} \mathcal{P}(\Omega^N)$.

Dans un tel plan, non seulement A est un **échantillon aléatoire**, mais sa taille N aussi : la mesure Π « charge » (a priori) tous les échantillons extraits de Ω.

Il existe deux variantes alternatives de la définition d'un plan de sondage :

(a) un **plan de sondage** sur Ω (avec Card Ω = M $< +\infty$) est une mesure de probabilité sur $\mathcal{P}(\Omega)$. Pour se ramener à la définition précédente, il faut supposer que certains éléments ω de Ω sont multiples (sondages avec remise). Cette hypothèse n'est pas retenue ici, où tous les éléments d'un ensemble (fini) Ω sont considérés comme distincts deux à deux. En général, la notation mathématique A $\subset \Omega$ suppose, a priori, que l'échantillon A est tiré sans remise dans Ω, ie que tous les (indices des) éléments $a_n \in$ A sont différents ;

(b) un **plan de sondage** sur Ω est défini en considérant qu'un échantillon A de taille N est une application $\alpha : N_N^* \mapsto \Omega$ associant à tout n $\in N_N^*$ un élément $a_n = \alpha(n) \in \Omega$. C'est donc une **suite** (finie) sur Ω. L'ensemble (fini) N_N^* = {1 ,..., N} est appelé **ensemble des indices** de A. Par suite :

$(b)_1$ si α est une **application injective**, on dit que A est un **échantillon sans remise**, ou **échantillon exhaustif**, puisque $\beta \neq \alpha \Rightarrow a_\beta \neq a_\alpha$;

$(b)_2$ si α est une **application surjective**, on dit que A est un **échantillon avec remise**, ou **échantillon bernoullien**, puisque, $\forall\ \omega \in$ A, il existe (au moins) un indice n $\in N_N^*$ tq $\alpha(n) = \omega$.

On note alors $\mathcal{T}_N$ l'ensemble des échantillons sans remise (de taille N), $\mathcal{S}_N$ l'ensemble des échantillons avec remise (de taille N) et $\mathcal{A}_N = \mathcal{T}_N \cup \mathcal{S}_N$ l'ensemble des échantillons avec ou sans remise (de taille N). On pose :

(4) $\quad \mathcal{T} = \cup_{N=1}^{M} \mathcal{T}_N$, $\quad \mathcal{S} = \cup_{N=1}^{M} \mathcal{S}_N$, $\quad \mathcal{A} = \cup_{N=1}^{M} \mathcal{A}_N$.

Un **plan de sondage** est alors une (mesure de) probabilité définie sur $\mathcal{A} \cup \{\varnothing\}$.

(vi) Un **plan de sondage de taille fixe N** est une mesure de probabilité définie sur $\mathcal{A}_N \cup \{\varnothing\}$. Dans un plan de sondage de taille fixe N sans remise (resp avec remise), la mesure Π en question ne charge, en fait, que $\mathcal{T}_N$ (resp $\mathcal{S}_N$). De même, dans un plan de sondage sans remise (resp avec remise), Π ne charge que $\mathcal{T}$ (resp $\mathcal{S}$). Dans un plan sans remise (resp sans remise de taille N), Π est donc tq :

(5) $\Pi(A) = 0, \quad \forall A \subset \mathcal{T} \qquad (\text{resp } \forall A \subset \mathcal{T}_N)$.

Cette deuxième notion de plan de sondage donne, lorsque Ω est fini (card Ω = M) :

(6)
$$\text{Card } \mathcal{T}_N = A_M^N, \text{ si } N \leq M, \qquad \text{card } \mathcal{T}_N = 0 \text{ sinon ;}$$
$$\text{Card } \mathcal{A}_N = M^N, \ \forall (M, N)$$

(cf (1) et (2).

(vii) En pratique, un plan de sondage réel est plus ou moins complexe (cf **classification des sondages**).

L'étude des propriétés statistiques des diverses notions considérées en **théorie des sondages** (**estimateurs**, **tests**, **régions de confiance**, etc) est fondée sur celle des plans Π utilisés. Il en est de même des notions d'**efficacité** ou d'**optimalité** des plans de sondage (comparaison entre plans).

Des considérations de **coût** sont à prendre en compte : coût d'obtention des informations (**observation** des variables) sur les unités de A (cf aussi **fonction de coût**).

(viii) Lorsqu'aucune précision n'est donnée, on entend généralement par le terme de « sondage » un **sondage dans une population finie** Ω.

Cependant, il est possible d'étendre la notion de sondage au cas de populations Ω non nécessairement finies (cf **population continue**, **population infinie**). Un échantillon indexé par un ensemble quelconque T est une application $\alpha : T \mapsto \mathcal{A}$ associant à tout élément $t \in T$ un élément $a_t = \alpha(t) \in \Omega$. Celle-ci peut, comme précédemment, être injective, surjective, ou quelconque.

Un tel contexte intervient parfois en **théorie des processus** : étude de la « population » Ω constituée des **trajectoires** $t \mapsto X_t(\omega)$, où ω parcourt Ω.