THÉORIE DES SONDAGES (M)
(09 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

Pour analyser les **phénomènes** les plus divers, il existe trois sources fondamentales d'information (données, observations, etc) (cf **production statistique**) :

(a) les **données « historiques »**, ie **non conçues statistiquement** : écrits divers (chroniques, chartes), fichiers divers de gestion administrative ou autre, etc. Cette source d'information ne répond pas (nécessairement) à des fins scientifiques : sa finalité est la mémorisation de divers faits ou actes (histoire, histoires, etc). Elle n'est donc guère « contrôlée » par le **statisticien**. L'échelle de création de ce type de données peut être importante (actes souverains, histoire générale, etc) ou beaucoup plus limitée : écrits privés, chroniques locales, biographies (cf **monographie**). Les données météorologiques sont, en grande partie, de cette nature ;

(b) les **plans d'expérience** (cf **théorie des plans d'expérience**). Cette source consiste en l'organisation optimale d'expériences visant à démontrer l'existence d'effets (moyens, en variabilités, etc) entre les facteurs mis en oeuvre au cours d'une expérience et les « informations » résultant de celle-ci. Son échelle est généralement « localisée » (expérience en laboratoire, expérience agronomique, etc). Les données biologiques ou nutritionnelles sont généralement de ce type ;

(c) les **sondages** (cf **théorie des sondages**). Cette source consiste en des investigations partielles portant sur un ensemble, généralement appelé **population**. Elle produit divers relevés contenant les observations effectuées (cf **questionnaire**). Son échelle peut être plus ou moins étendue (populations humaines, mais aussi flore locale, faune aquatique, etc).

La **théorie des sondages** a ainsi pour objet la connaissance d'un **ensemble** donné (**population**) Ω à l'aide d'une **partie** seulement de cet ensemble : il s'agit d'un sous-ensemble A de Ω qui est constitué d'éléments (individus) ω de Ω (resp prélevés dans Ω).

En pratique, l'ensemble Ω peut être de nature très variée : populations humaine, ou animale (faune), ou végétale (flore), fabrication industrielle, univers physique, **système** économique, etc. Un élément ω (**unité statistique** alors appelée **unité de sondage**) est resp une personne physique, un animal, une plante, une pièce mécanique (ou autre objet physique), une planète, une entreprise, etc.

Cette théorie constitue un exemple-type de méthode d'**inférence statistique**, puisqu'elle vise à estimer des grandeurs d'ensemble à partir de grandeur analogues obtenues sur des champs restreints.

Chaque sondage porte, en général, sur une **variable** (ou **caractère**) η relatifs aux éléments $\omega \in \Omega$. Chaque élément ω fait ainsi l'objet de diverses **mesures** (ou **observations**) η (ω) de η.

Si η est à valeurs dans un ensemble $\mathcal{Y}$, ie si $\eta : \Omega \mapsto \mathcal{Y}$, on note $Y = \eta(\Omega)$ l'ensemble des valeurs de η observables sur les éléments de Ω et $y = \eta(A)$ l'ensemble des valeurs observées sur les éléments a de A. On note aussi $Y = (\eta(\omega))_{\omega \in \Omega}$ et $y = (\eta(a))_{a \in A}$. Il importe de remarquer que l'application η, bien que définie (on sait quelles variables sont à observer), n'est pas - en général - connue ni observable : le **statisticien** n'observe que certaines valeurs de η, celles notées y, puisqu'il n'observe que l'ensemble A et non Ω tout entier. A est appelé **échantillon** (**d'éléments** de Ω, ou d'**unités de sondage**) et, par extension, $y = (\eta(a))_{a \in A}$ est aussi appelé **échantillon** (d'éléments de $\mathcal{Y}$, ou d'observations proprement dites).

Par ailleurs, la variable η peut être « simple » (ou « scalaire ») aussi bien que « multiple » (ou « vectorielle ») (liste de G variables). Elle peut aussi être une **variable quantitative** (ie numérique), une **variable qualitative**, ou même une variable « mixte » (ie une « liste » comportant des variables quantitatives et des variables qualitatives, ou encore une variable à valeurs dans une partition $\mathcal{Y}$ de $\mathbf{R}^G$).

On peut toujours, au moins formellement, supposer que η se distribue sur $\mathcal{Y}$ (ie sur une **tribu** $\mathcal{C}$ de parties de $\mathcal{Y}$), selon une **distribution** inconnue P^η, encore notée $\mathcal{L}(\eta)$: P^η est alors considérée comme l'image par η d'une **mesure de probabilité** P définie sur une tribu $\mathcal{T}$ de parties de Ω. Ceci implique alors que η est assimilée à une **variable aléatoire** (ie un caractère $(\mathcal{T},\mathcal{C})$-mesurable). P^η est appelée **distribution** du caractère η dans la population Ω : c'est sur elle que porte l'essentiel du sondage.

(ii) On distingue usuellement entre deux types de sondages (cf **classification des sondages**) :

(a) le **sondage aléatoire** relève des méthodes de la **Statistique** « classique » : ce sondage suppose que A est extrait de Ω d'une manière probabiliste (ie **aléatoire**). Le tirage (probabiliste) de A est obtenu en définissant sur Ω un **plan de sondage** Π, plan aussi appelé **échantillonnage** ou **tirage aléatoire** ;

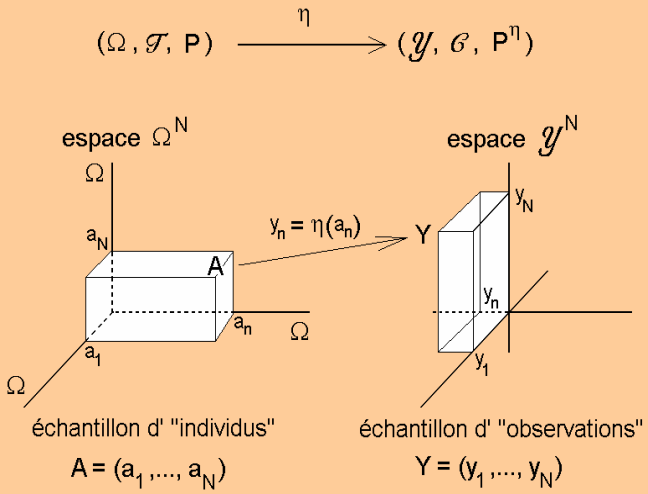
(b) le **sondage « par choix raisonné »** contient des unités tirées selon une information disponible a priori, et il peut relever des méthodes bayésiennes (modèles de **superpopulation**). Si A est déterminé par des considérations a priori ne faisant pas appel à un mécanisme de hasard, on parle de sondage par choix raisonné : une **enquête d'opinion** (sociologique, politique, ou économique pour l'étude du marché d'un produit, etc) fait souvent usage de tels sondages (cf **sondage par choix raisonné**, **sondage par quotas**, **sondage par unités-types**).

Dans ces deux types de sondage, l'objectif consiste à formuler un « jugement » sur Ω à partir de mesures y effectuées sur les éléments a de l'échantillon A. Une telle mesure s'interprète, concrètement, en fonction de la **nature de l'étude** qui a suscité le sondage : caractères physico-chimiques, trajectoires, caractères biologiques, caractères socio-économiques (migrations, fonctions économiques), etc.

(iii) L'ensemble Ω peut être a priori quelconque. En pratique, il est souvent fini et on le note $\Omega = \{\omega_1 ,..., \omega_M\}$ (avec card Ω = M). De même, A est souvent fini et on le note $A = \{a_1 ,..., a_N\}$ ou $A = (a_1 ,..., a_N)$ (avec card A = N). On pose alors $Y = \{Y_1 ,..., Y_M\}$ ou $(Y_1 ,..., Y_M)$, avec $Y_m = \eta (\omega_m)$ (où m = 1 ,..., M) et $y = \{y_1 ,..., y_N\}$, et $y = (y_1 ,..., y_N)$, avec $y_n = \eta (a_n)$ (où m = 1 ,..., N).

Dans certaines situations, et notamment dans l'étude des **propriétés asymptotiques**, on suppose souvent que Ω (resp A) est infini (en général, infini dénombrable).

Description d'un ensemble sondé



Lorsque Ω est fini, la distribution de η peut se représenter sous la forme d'une **loi empirique** (fictive) (cf **loi discrète**) :

(1) $P^{\eta} = M^{-1} \Sigma_{m=1}^{M} \delta_{\eta(\omega(m))}$ ou $M^{-1} \Sigma_{m=1}^{M} \delta_{Y(m)}$,

où δ_U désigne la **fonction indicatrice** d'une **partie** U, et $\omega(m)$ (resp Y(m)) désigne, par commodité, ω_m (resp Y_m).

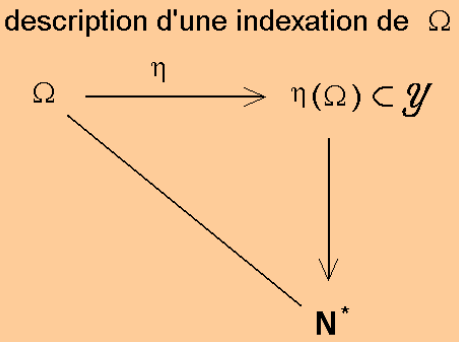
La loi emprique de l'échantillon y s'écrit, parallèlement :

(1) $P_N = N^{-1} \Sigma_{n=1}^{N} \delta_{\eta(a(n))}$ ou $M^{-1} \Sigma_{n=1}^{N} \delta_{y(n)}$.

l'aléatoire étant, dans ce deuxième formalisme, représenté par le tirage de A (ie des a_n) défini par le plan de sondage Π. Par analogie, la distribution P^{η} est aussi notée P_M .

Dans le cas fini, on identifie souvent Ω à $N_M^* = \{1 ,..., M\}$ et A à $N_N^* = \{1 ,..., N\}$. Les autres notations précédemment définies se modifient alors aisément. L'**application bijective** $\nu : \Omega \mapsto \mathbf{N}^*$ consiste à « numéroter » sans ambiguité (ie sans omission ni répétition) les éléments de Ω (ou **unités statistiques**), et l'image $\nu (\Omega)$ est souvent appelée **base de sondage** : c'est à l'aide d'une suite de **nombres au hasard** que les éléments de $\nu (\Omega)$ (donc ceux de Ω) sont généralement tirés.

Description d'un indexation de Θ



(iv) L'objet de la théorie est d'effectuer une inférence sur le **paramètre** inconnu qu'est la distribution P_M (ou seulement sur le paramètre Y si Ω est fini). Pour cela, un échantillon A est tiré « au hasard » selon le plan de sondage (ie la mesure de probabilité) Π. L'inférence porte sur une **caractéristique** $\gamma \in \Gamma$ de cette distribution, où Γ est un ensemble de caractéristiques, et $c(P^\eta) = \gamma$ représente les caractéristiques de P^η à étudier, eg :

(a) la **moyenne (arithmétique) théorique**, ou **moyenne** sur Ω (si $\mathcal{Y} = \mathbf{R}^G$, avec $G \geq 1$) :

(3) $\bar{Y}$ ou $\mu = E\,\eta = \int \eta\, dP = M^{-1} \Sigma_{m=1}^{M}\, Y_m = e_M' Y / M$;

(b) le **total théorique** (si Ω est fini et si $\mathcal{Y} = \mathbf{R}^G$, avec $G \geq 1$) :

(4) T ou $\tau = M \mu$;

(c) une **proportion théorique** :

(5) $p = P([\eta \in C]) = M^{-1} . \# \{Y_m \in \mathcal{Y} : Y_m \in C\}$,

où $C \in \mathcal{C}$ est donné et où # U désigne le cardinal (nombre d'éléments) d'une partie U.

Lorsque Ω est fini, le modèle général de la théorie est donc un **modèle statistique** paramétrique puisque P^η dépend du paramètre $Y = (Y_1, ..., Y_M)$ d'après (1).

(v) Si Π est un plan de sondage sur Ω, le caractère aléatoire de $y = \eta(A)$ vient donc de celui de A. La loi de y n'est donc autre que la loi image de Π par y, loi qui dépend donc aussi de Y (que l'on considère comme paramètre).

Par suite, avec N observations y on ne peut en général faire porter l'inférence que sur une fonction de moindre dimensionnalité que Y, eg g : $\mathcal{Y}^M \mapsto \mathcal{Z}^L$ de Y (avec $L \leq$

N), ie sur g ($\mathcal{Y}$) seulement. Par exemple, si $\mathcal{Y}$ = **R**, on étudie le total théorique g (Y) = e_M' Y = T, ou encore un **quantile** de P_M , etc.

Ainsi, l'intérêt majeur d'un sondage consiste à n'observer (ou mesurer) η que sur un ensemble restreint A, convenablement choisi (ie dont le plan de sondage générateur est « optimal »), pour en déduire des propriétés (ou caractéristiques) relativement à Ω (ou à η (Ω)) tout entier.

(vi) Selon les objectifs d'un sondage, il existe de nombreux plans de sondage adaptés (cf **classification des sondages**). Les deux plans de base à caractère aléatoire sont (a) le **« sondage avec remise »** (ou **sondage bernoullien**) et (b) le **« sondage sans remise »** (ou **sondage exhaustif**).

Il faut aussi noter que ce qui est une unités statistique pour un plan de sondage donné peut constituer une population pour un autre plan (et vice versa), notamment lorsque les plans sont « emboîtés » (cf eg **sondage emboîté**, **sondage en grappes**, **sondage stratifié**, etc) ; de même, un sondage aléatoire peut être défini à l' « intérieur » d'un échantillon A déjà extrait de Ω (échantillonnage double ou multiple, ou rééchantillonnage) (cf **conditionnement**).

(vii) La démarche générale d'un sondage peut donc se décomposer selon le schéma type suivant :

(a) **spécification d'un problème** donné et « traduction » statistique de celui-ci. Le problème dépend du **domaine de connaissance** considéré : il indique, notamment, la « liste » η des **variables à étudier**, voire même, parfois, une liste complémentaire ξ (variables « exogènes » destinées à préciser η) ;

(b) définition de l'ensemble Ω, appelé **base de sondage**, ou de son indexation ν (Ω), aussi appelée **base de sondage**, lesquels doivent permettre de « repérer », sans omission ni répétition, tous les éléments de Ω ;

(c) mise au point d'un **plan de sondage** Π qui décrit la façon d'extraire A de Ω. Cette mise au point se fait généralement à l'aide de nombres au hasard ;

(d) **étude statistique** proprement dite du problème considéré initialement à l'aide des observations disponibles et compte tenu de la définition de Π.

(viii) Certains sondages utilisent une **information a priori**, incorporée dans le cadre de la **théorie bayésienne**. Ainsi en est-il des sondages non aléatoires (sondage par « choix raisonné » ou sondage « bayésien »), des sondages à partir desquels sont définis l'**estimateur par le quotient** ou l'**estimateur par régression**, etc. Dans ce cadre, le paramètre Y est supposé être lui-même une **va** dont la **loi a priori** Q appartient à une famille donnée $\mathcal{Q}$ de lois de probabilité définies sur $\mathcal{G}^{\otimes M}$.

Ainsi, lorsque $\mathcal{Y}$ = **R**, on peut supposer que Y est « expliqué » selon un **modèle de régression multiple** :

(6) $E\,Y = X\,b$, avec $V\,Y = \Sigma$,

où $X \in M_{MK}(\mathbf{R})$ est une **matrice des observations** relatives à K variables exogènes ξ_1 ,..., ξ_k , X et Σ étant supposées connues.

La **méthode de prédiction bayésienne** (R.M. ROYALL) consiste à estimer eg le total $T = e_M' Y$ au vu d'un échantillon sans répétition A sur lequel on observe un vecteur noté y_A et une matrice notée X_A correspondant respectivement aux variables η et ξ_1 ,..., ξ_k .

On suppose que (y_A , X_A) vérifie (6) et l'on décompose alors (6) selon :

(7)
$$E\,y_A = X_A\,b, \quad \text{avec } V\,y_A = \Sigma_A,$$
$$E\,y_B = X_B\,b, \quad \text{avec } V\,y_B = \Sigma_B,$$

avec $B = \Omega \setminus A$.

La méthode consiste alors :

(a) à estimer b par la **méthode des moindres carrés généralisés** appliquée à la première équation de (7), d'où un estimateur $b_{A,g}^{\#}$ de b ;

(b) à décomposer T selon $e_N' y_A + e_{M-N}' y_B$;

(c) enfin, à estimer le second terme de T en estimant y_B par $y_{B,g}^{\#} = X_B\, b_{A,g}^{\#}$.

On montre que l'estimateur de T obtenu :

(8) $T_N = e_N' y_A + e_{M-N}' X_B\, b_{A,g}^{\#}$

possède des propriétés optimales.

(ix) En théorie des sondages, il importe de distinguer **trois « distributions »** :

(a) celle du **caractère** η dans la population Ω, distribution notée ici P^{η} ou P_M ;

(b) celle de l'**échantillon** aléatoire A, tiré dans Ω selon le plan de sondage Π. Ce plan tient donc lieu de **loi** pour A.

(c) celle de toute **statistique** d'intérêt se déduisant de A (resp de y) : total, proportion, etc. La loi de S est donc l'image de la loi de A (resp de la loi de y).

L'**aléatoire** provient donc du tirage des unités observées, à la différence de la théorie de la **régression**, dans laquelle l'aléatoire réside dans le **phénomène** étudié lui-même (aléatoire « intrinsèque »). Il arrive d'ailleurs, souvent, que les deux approches se combinent : modèles de régression estimés à partir de données issues d'un sondage (eg panels).

Lorsque les procédures d'inférence statistique sont appliquées aux données issues d'un sondage (estimation, tests, etc), les lois considérées sont alors les lois images de Π par diverses applications. Par suite, les caractéristiques usuelles (espérance, variance ou dispersion) sont calculées à partir de Π. De même, les notions d'optimalité, d'efficacité ou d'exhaustivité liées à un sondage sont relatives à Π (et non pas à P^{η} ou P_M).

(x) La théorie des sondages entretient des liens avec :

(a) la **théorie des plans d'expérience**. Comme celle-ci, elle conduit à des comparaisons entre plans, à choisir les meilleurs (plans de sondage optimaux) en tenant compte, le plus souvent, d'un budget économique, donc de considérations de coût (cf **fonction de coût**) ;

(b) la **théorie des processus**, notamment dans ses aspects asymtotiques (cf aussi **mesure produit**, **modèle produit**, **propriété asymptotique**).