La théorie des sondages

Michel LEJEUNE*

On présente ici les principes et enjeux simples de la théorie des sondages pouvant susciter des désirs d'approfondissement.

Introduction

La théorie des sondages ne fait pas partie, en France, des connaissances usuelles des statisticiens même si la pratique des sondages, quant à elle, est très répandue. Elle s'est développée à partir des années 1930 dans le monde anglo-saxon ainsi qu'en Inde. Bien que reposant sur les mêmes principes que la statistique mathématique classique elle en diffère sensiblement dans son esprit en raison d'objectifs spécifiques.

Cette théorie se consacre essentiellement au problème de la sélection de l'échantillon (voir ci-après la notion de plan de sondage qui fait le pendant des plans d'expériences en théorie classique) et à la recherche d'estimateurs. Du fait qu'elle porte sur des populations finies, d'existence bien concrète, elle ne peut ignorer les contraintes du monde réel, ce qui n'est peut-être pas sans lien avec le faible intérêt qu'elle suscite chez nous.

Plan de sondage et probabilités d'inclusion

On considère une *population* comprenant N *individus* parfaitement identifiés par un numéro d'ordre. Pour ce qui suit il nous suffira de ne retenir que ces numéros d'ordre et nous définissons ainsi la population $U = \{1,\ldots,k,\ldots,N\}$. Notons que les vocables de population et individus sont purement conventionnels. Les parties de cette population sont appelées *échantillons*. Dans cette présentation nous n'envisagerons que la situation où l'échantillon à sélectionner est de *taille* (cardinal) fixée, notée n, et désignerons simplement par S l'ensemble des échantillons de taille n. Commençons par donner quelques définitions.

Définition 1. *On appelle plan de sondage une loi de probabilité définie sur S.*

Concrètement le plan de sondage définit, pour chaque échantillon, la probabilité qu'il soit sélectionné *via* le mécanisme aléatoire utilisé.

Définition 2. *Soit un plan de sondage p et S_k l'ensemble des échantillons contenant l'individu k. On appelle probabilité d'inclusion de l'individu $k \in U$:*

$$\pi_k = \sum_{s \in S_k} p(s)$$

* *Université Pierre Mendès France, LABSAD (BSHM), 38040 Grenoble Cedex 09.*
michel.lejeune@upmf-grenoble.fr

Il s'agit donc de la probabilité que cet individu appartienne à l'échantillon sélectionné. Le fait d'avoir un échantillon de taille n se traduit par $\sum_{k\in U}\pi_k = n$. En effet, soit I_k la variable indicatrice de la sélection de l'individu k, on a $\pi_k = E(I_k)$ et :

$$\sum_{k\in U}\pi_k = \sum_{k\in U}E(I_k) = E(\sum_{k\in U}I_k) = n.$$

On définit de même la probabilité qu'un couple d'individus $\{k,l\}$ soit dans cet échantillon, en considérant l'ensemble S_{kl} des échantillons contenant ce couple.

Définition 3. *On appelle probabilité d'inclusion d'ordre 2 des individus k et l :*

$$\pi_{kl} = \sum_{s\in S_{kl}} p(s)\,.$$

Pour illustrer ces notions considérons le *plan simple sans remise* (PSSR), correspondant au cas où p est uniforme sur S :

$$\forall s \in S, p(s) = \binom{N}{n}^{-1}.$$

Comme il y a $\binom{N-1}{n-1}$ échantillons comprenant un individu donné, on a :

$$\forall k \in U, \quad \pi_k = \binom{N-1}{n-1}\binom{N}{n}^{-1} = \frac{n}{N}.$$

Ce résultat, à savoir que la probabilité d'être sélectionné, pour un individu donné, est égale au *taux de sondage* $\frac{n}{N}$, est tout à fait intuitif. On montre aisément que l'on a $\pi_{kl} = \frac{n(n-1)}{N(N-1)}$.

Remarque 4. *Notre définition d'échantillon exclut le cas du plan simple avec remise (PSAR), lequel correspond à la théorie classique de l'échantillonnage aléatoire. Les notions précédentes et la théorie générale peuvent aussi être développées dans cette situation. En pratique on n'effectue pas de plans avec remise car, on le sent bien intuitivement, le risque d'observer plusieurs fois le même individu constitue une perte d'information. La théorie montre, dans diverses situations, qu'un estimateur sans remise est meilleur qu'un estimateur avec remise et, même, qu'il est préférable, dans le cas d'un échantillon avec remise, de ne prendre en compte chaque individu qu'une seule fois.*

Les estimateurs

On s'intéresse maintenant à une certaine « variable » réelle observable sur chaque individu. Notons y_k la valeur prise par l'individu k. On souhaite estimer une caractéristique de la variable, le plus souvent son total (ou sa moyenne). Pour le total t_y, Horvitz et Thompson ont proposé un estimateur pour un plan quelconque. Soit $Y_1, \ldots, Y_n$ les variables aléatoires correspondant à la sélection de n individus selon le plan de sondage, cet estimateur est :

$$\widehat{t}_y^{HT} = \sum_{i=1}^{n}\frac{1}{\pi_i}Y_i\,.$$

Les $\frac{1}{\pi_i}$ sont appelés *poids de sondage.* Cet estimateur est sans biais. En effet :

$$\widehat{t}_y^{HT} = \sum_{k\in U} \frac{1}{\pi_k} I_k y_k, \text{ d'où } E(\widehat{t}_y^{HT}) = \sum_{k\in U} \frac{1}{\pi_k} E(I_k) y_k = \sum_{k\in U} y_k = t_y.$$

Il est généralement retenu car on montre qu'il est le seul à avoir cette propriété parmi les estimateurs fonctions linéaires des Y_i. De plus on sait exprimer sa variance et un estimateur sans biais de sa variance, *via* les probabilités d'inclusion d'ordres 1 et 2.

On peut évidemment étudier cet estimateur par une voie directe. Pour les plans PSAR et PSSR il est clair que $\frac{N}{n}\sum_{i=1}^{n} Y_i$ est sans biais pour t_y (chaque Y_i ayant une loi uniforme sur les y_k, son espérance est $m_y = \frac{1}{N} t_y$, la moyenne de la population). Notons que l'estimateur de Horvitz-Thompson pour la moyenne de la population est la moyenne de l'échantillon $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Le calcul de la variance montre les difficultés inhérentes à la théorie des sondages. Pour le plan PSAR les Y_i sont indépendantes et le calcul est immédiat. Pour le PSSR interviennent les covariances et le calcul demande une certaine adresse (le lecteur pourra tenter une démonstration en utilisant $\sum_{i=1}^{n} Y_i = \sum_{k\in U} I_k y_k$, $Var(I_k) = \pi_k(1-\pi_k)$ et $Cov(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$). On trouve :

$$Var(\overline{Y}) = \left(\frac{N-n}{N-1}\right)\frac{v^2}{n}$$

où v^2 est la variance de la population :

$$v^2 = \frac{1}{N}\sum_{k=1}^{N}(y_k - m_y)^2.$$

Un apport essentiel de la théorie des sondages est d'établir la *précision* d'une estimation (vulgairement : la fourchette). Celle-ci est définie par la demi-longueur d'un intervalle de confiance à 95 %. En vertu du théorème central limite qui donne une bonne approximation pour les tailles usuelles d'échantillons, cet intervalle repose sur la loi de Gauss (moyennant une deuxième approximation due à l'estimation de la variance v^2). Dans le cas PASR on adopte ainsi pour précision sur l'estimation de la moyenne :

$$1,96\sqrt{\left(1-\frac{n}{N}\right)\frac{s^2}{n}}$$

où s^2 est la variance de l'échantillon qui estime sans biais $\frac{N}{N-1}v^2$.

Le principe de stratification

Dans cette brève présentation il n'est pas possible de présenter les principaux plans de sondage. Nous nous concentrons sur le plan stratifié en raison de ses liens avec les pratiques très répandues d'utilisation de quotas et de redressement d'échantillons. Ces procédures reposent sur la même idée : toute partition de la population en *strates* fortement homogénes (variances internes aux strates faibles vis-à-vis de la variance globale) doit permettre d'accroître la précision.

Dans le plan stratifié (PSTRAT) on effectue des sondages, par exemple de type PSSR, indépendamment dans chaque strate et, *de facto*, se pose le problème du choix des tailles des « sous-échantillons ». L'estimateur naturel de t_y qui est aussi celui d'Horvitz-Thompson est obtenu, dans le cas de sous-échantillons de type PSSR, en recomposant les estimateurs des moyennes par strate, soit $\sum_{h=1}^{H} N_h \overline{Y}_h$ pour les H strates. Par l'indépendance mutuelle sa variance est immédiate. Un problème intéressant, mais purement théorique, est celui dit de « l'allocation optimale »,

à savoir de trouver les tailles n_h minimisant cette variance sous la contrainte $\sum_{h=1}^{H} n_h = n$ (pour le lecteur intéressé : les taux de sondage n_h/N_h sont alors proportionnels aux écarts-type des strates).

Généralement on stratifie à taux de sondage constants dans les strates ce qui garantit un gain de précision par rapport à un sondage PSSR. En effet la « fourchette » est multipliée par un facteur $(1-\eta^2)^{1/2}$ où η^2 est un coefficient bien connu des statisticiens, compris entre 0 et 1, qui mesure en quelque sorte le lien existant entre le critère de stratification et la variable d'intérêt (rapport de la variance intra-strates à la variance totale pour cette variable). Cette stratification, dite aussi « à la proportionnelle », est qualifiée par les praticiens comme produisant un échantillon *représentatif* vis-à-vis du critère de stratification. Notons que dans ce cas particulier les probabilités d'inclusion sont identiques pour tous les individus.

Ceci nous amène à la « *méthode des quotas* » dont la pratique est systématique dans les instituts. Elle consiste à sélectionner un échantillon qui soit un modèle réduit de la population sur certains critères de partitionnement dont on peut penser qu'ils ont un lien avec la thématique de l'enquête. Comme les effectifs des strates dans la population doivent être connus pour chaque critère d'une part et qu'il y a de fortes contraintes de mise en œuvre d'autre part, on se limite généralement à un critère géographique (région), à la classe d'âge, au sexe et, parfois, à la catégorie socio-professionnelle. La différence avec un plan PSTRAT tient au fait que, pour des raisons de faisabilité, la proportionnalité des effectifs de l'échantillon à ceux de la population ne concerne que les effectifs des critères « à la marge » et non des critères croisés entre eux.

Il n'est pas inutile, ici, pour les citoyens que nous sommes tous, d'ouvrir une parenthèse pratique. Dans la presse on déclare communément que les résultats d'un sondage proviennent d'un échantillon obtenu par la méthode des quotas et/ou d'un échantillon représentatif de la population selon la région, l'âge, etc. Ainsi il y a, chez les sondeurs, un véritable culte des quotas, la plupart d'entre eux étant convaincus qu'il suffit d'effectuer ce modèle réduit pour garantir de bonnes estimations. Or l'essentiel n'est pas là et, en fait, la mention consacrée dans la presse n'est d'aucun intérêt quant à juger de la qualité réelle de l'échantillon. Ce qui nous importe avant tout est de savoir dans quelle mesure un plan aléatoire, incluant ou non des contraintes de quotas, a-t-il été respecté.

L'approche théorique permet de déterminer le gain apporté par l'utilisation de quotas par rapport à un PSSR. Il est de même nature que celui présenté en stratification à ceci près que le coefficient η^2 repose ici sur la variance expliquée par un modèle à effets additifs des critères de quotas (pour les initiés : modèle d'analyse de variance sans interactions). Le constat empirique est bien décevant pour les critères usuels, le facteur de réduction de la fourchette $(1-\eta^2)^{1/2}$ ne passant que très rarement sous la valeur 0,90. Ce constat indique au passage que ces critères de quotas sont des déterminants extrêmement ténus du comportement ou des opinions des individus (il n'en reste pas moins que les quotas sont utiles, voire même nécessaires, pour limiter les biais importants découlant des obstacles de terrain).

La pratique des *redressements* relève des mêmes principes que les quotas mais elle intervient en aval du recueil de l'échantillon. Elle consiste, dans le but d'améliorer les estimations, à « caler » l'échantillon observé sur la population pour divers critères disponibles. Comme pour les quotas le calage se fait à la marge, critère par critère. Il s'effectue en attribuant des poids aux individus en perturbant le moins possible leurs poids de sondage. Mathématiquement il s'agit de déterminer les poids w_i tels que les contraintes de marges soient respectées pour les différents critères retenus et tels que :

$$\sum_{i=1}^{n} d\left(w_i, \frac{1}{\pi_i}\right)$$

soit minimal (où d est la distance choisie).

La résolution d'un tel problème nécessite une procédure itérative. En fait l'algorithme utilisé en pratique a été introduit tout à fait empiriquement : la recherche de poids correcteurs étant immédiate pour un seul critère, on calera successivement chacun des critères de façon cyclique jusqu'à quasi convergence vers les bonnes marges. Ce n'est que postérieurement que les travaux initiés par J.-C. Deville à l'INSEE ont donné un cadre théorique aux procédures de calage, notamment en introduisant la notion de distance mentionnée ci-dessus (l'algorithme usuel relève d'une distance implicite peu commune mais proche de la distance quadratique).

Les estimateurs par redressement n'ayant pas d'expression explicite leurs propriétés sont difficiles à établir. Ainsi, pour déterminer biais et variance, on doit se contenter d'approximations pour n grand. Brièvement disons que le

biais, pour des distances raisonnables, reste négligeable et que leur précision est très proche de celle découlant des estimateurs avec quotas (rapport égal à $1 + O(1/n^{1/2})$). Ainsi, comme pour ces derniers, le gain par rapport à un PSSR, pour une variable d'intérêt donnée, sera d'autant plus élevé que les critères retenus auront un lien fort avec elle. Ceci permet d'orienter le choix de ces critères sur la base des liens observés dans l'échantillon.

La modélisation des non-réponses

Il existe diverses sources d'erreur dans les sondages : erreur de mesure, biais de sélection, biais de couverture de la population et, souvent la plus importante, la *non-réponse*, à savoir qu'un individu dûment sélectionné par le plan de sondage n'a pu être observé pour une raison ou une autre. Les non-réponses, systématiquement présentes et même dans de fortes proportions, sont une cause potentielle de biais. Pour y remédier différents schémas du mécanisme de non-réponse ont été proposés pour lesquels s'appliquent des modèles appropriés. Par exemple, si le fait de répondre ou non à une enquête est indépendant de la variable d'intérêt conditionnellement à un ensemble de variables auxiliaires (par exemple âge, sexe, profession, statut matrimonial, etc.) le recours à un redressement sur ces variables est efficace. Il existe aussi des non-réponses partielles (seules certaines variables n'ont pu être observées ici ou là) débouchant sur des modèles *d'imputation*.

Le traitement des non-réponses totales ou partielles donne lieu actuellement à de nombreux développements théoriques.

Pour en savoir plus

Ardilly (P.), *Les techniques de sondage*, Technip, (1994).

Deville (J.-C.) et Sarndal (C.-E.), Calibration estimators in survey sampling, *Journal of the American Statistical Association*, vol. 87, p. 376-382, (1992).

Tillé (Y.), *Théorie des sondages*, Dunod, (2001).