POUR OU CONTRE LES PLANS DE SONDAGE A PROBABILITÉS INÉGALES ?

Léo Gerville-Réache

POUR OU CONTRE LES PLANS DE SONDAGE A PROBABILITES INEGALES ?

Léo Gerville-Réache[1]

[1]*Université de Bordeaux, CNRS, UMR 5251, France, leo.gerville@u-bordeaux2.fr*

Résumé. L'existence d'un estimateur « universel » pour la moyenne (ou la somme) d'une caractéristique d'une population finie, via un plan de sondage sans remise de taille fixe, est un problème fondamental de la théorie des sondages. Pour un plan à probabilités d'inclusion inégales, l'estimateur de Horvitz-Thompson est sans biais mais pas linéairement invariant alors que l'estimateur de Hàjek est linéairement invariant mais pas sans biais. Cependant, pour un plan à probabilités égales, ces deux estimateurs se simplifient, coïncident et sont alors sans biais et linéairement invariants. Pourquoi les plans de sondages à probabilités inégales connaissent-ils un développement aussi important ? Quelles sont leurs limites théoriques et pratiques ?

Mots-clés. Plan de sondage, probabilités d'inclusion inégales, estimateur de Horvitz-Thompson, estimateur de Hàjek, estimateur de Patel-Dharmadhikari.

Abstract. The existence of an "universal" estimator for mean (or sum) of a characteristic of a finite population, via sampling plan without replacement and fixed size, is a fundamental problem in sampling theory. For a plan with unequal probabilities of inclusion, the Horvitz-Thompson estimator is unbiased but not linearly invariant while the Hájek estimator is linearly invariant but not unbiased. However, with a equal probability plan, those two estimators are simpler, coincide and are then linearly invariant and unbiased. Why sampling plans with unequal probabilities know so important development? What are their theoretical and practical limits?

Keywords. Sampling plan, unequal probabilities of inclusion, Horvitz-Thompson estimator, Hajek estimator, Patel-Dharmadhikari estimator.

1 Introduction

La théorie des sondages et ses applications sont complexes à bien des égards. Tillé (2001) précise : « *Les statisticiens vont très vite se heurter à une difficulté de taille : dans les sondages en population finie, le modèle proposé postule l'indentifiabilité des unités. Cette composante du modèle rend non-pertinente l'application de la technique de réduction par exhaustivité et de la méthode du maximum de vraisemblance.* ».

Cet état de fait doit nous conduire à la plus extrême vigilance. Pour mémoire, Basu (1971) illustrait, à travers son intérêt pour les éléphants, son aversion pour l'estimateur de Horvitz-Thompson (1952). Nous pensons plus généralement que, c'est la question de l'utilisation des plans de sondage à probabilités inégales qui pose fondamentalement question. Au delà de la complexité à produire un algorithme général d'échantillonnage sous contrainte de probabilités d'inclusion fixées et inégales ou encore de la difficulté à produire une estimation de la variance de l'estimateur, Tillé (2006) fait le triste rappel que « *si l'on réussit à sélectionner un échantillon sans remise en respectant les probabilités inégales, on ne dispose pas nécessairement d'un bon estimateur. En effet, le π-estimateur d'une moyenne n'est pas invariant par translation. Le ratio de Hàjek est par contre invariant par translation mais il n'est pas sans biais.* ».

Ardilly (2006) évoque, à propos d'un estimateur de variance possiblement négatif, « *une situation statistiquement juste mais inadmissible (et donc philosophiquement complexe)* ».

Pourtant, le sondage à probabilités d'inclusion égales sans remise (stratifié ou pas) ne pose aucun problème : algorithme de sélection simplissime, estimateur sans problème. Plus important encore, la phase d'estimation ne nécessite pas l'indentifiabilité des unités. Un des messages des plans de sondages à probabilités égales est peut-être le suivant : sous quelles conditions un sondage à probabilités inégales vaut-il le coût ?

Il est clair qu'une « bonne » information auxiliaire justifie théoriquement la recherche d'un plan de sondage « optimisé ». Cependant, il convient d'apprécier nos capacités théoriques, techniques, algorithmiques, numériques et épistémologiques à utiliser une telle information.

2 Pourquoi des probabilités inégales ?

Le point de départ de la théorie des plans de sondages à probabilités inégales est simple, clair et légitime : Utiliser l'information auxiliaire disponible afin de construire un plan de sondage permettant d'estimer « au mieux » un paramètre de la population.

Quelques résultats théoriques ont alors permis le développement de cette approche. Le plus central est sans doute l'estimateur d'Horvitz-Thompson (1952) pour une somme (ou une moyenne) : estimateur sans biais, hyper-admissible dont le calcul ne dépend que des probabilités d'inclusion de premier ordre. En 1953, un estimateur « raisonnable » de la variance est alors proposé par Sen-Yates-Grundy.

Désormais, l'objectif est de définir les probabilités d'inclusion en fonction des informations auxiliaires. Lorsqu'une information auxiliaire sera connue pour chaque unité de l'échantillon, on se focalisera sur la « corrélation » entre cette information et la variable d'intérêt pour définir les probabilités d'inclusion de chaque unité. Si certaines informations auxiliaires sont globalisées (des moyennes et/ou des sommes), on pourra chercher un plan « équilibré ».

Dans tous les cas, les algorithmes de sélection (très nombreux) sont développés dans le souci de satisfaire des propriétés naturelles (Tillé 2001) :

- Les probabilités d'inclusion (d'ordre un) sont exactement celles prévues.
- La taille de l'échantillon est fixe.
- La méthode est applicable à tout ensemble de probabilités d'inclusion (d'ordre un) non nulles.
- Une unité ne peut être sélectionnée qu'une seule fois dans l'échantillon.

Ces considérations d'ordre général ne définissant pas un plan d'échantillonnage unique, des propriétés sur les probabilités de second ordre sont nécessaires. Parmi elles, on peut citer :

- Des probabilités de second ordre strictement positives.
- Des probabilités de second ordre vérifiant les conditions de Sen-Yates-Grundy.

Enfin, on cherchera :

- Un algorithme rapide, sans calculer les probabilités des échantillons.
- Un algorithme séquentiel qui ne nécessite qu'une seule lecture de la base de sondage.

Parmi les algorithmes développés, on peut noter, par exemple, ceux qui visent à maximiser l'entropie du plan de sondage.

Un autre argument de poids en faveur de la théorie des plans de sondages à probabilités inégales est

sa réciprocité. En effet, il existe une remarquable analogie avec les méthodes de redressement d'échantillon. Une fois le sondage réalisé, le hasard ne faisant jamais « parfaitement les choses », la théorie des sondages s'intéresse à la possibilité de minimiser la variance de l'estimateur conditionnellement aux données réellement observées.

L'idée principale est alors de modifier les poids de sondage de telle sorte que les poids « finaux » permettent une estimation plus précise du paramètre d'intérêt. Avec une approche qui consiste, une fois de plus, à utiliser les informations auxiliaires, on cherche des poids « finaux » qui vérifient certaines contraintes judicieusement choisies.

Des algorithmes de redressement sont alors développés, le plus connu en France est sans doute l'algorithme CALMAR de l'INSEE. Il faut noter que cet algorithme ne cherche que de nouveaux poids pour les individus de l'échantillon. Les temps de calcul d'un tel problème sont sans commune mesure avec ceux d'un algorithme de sélection d'échantillon (qui s'adresse à la totalité de la population).

Globalement tout semble se passer comme si on avait réalisé un plan de sondage avec des probabilités d'inclusion « légèrement » différentes de celles « utilisées » initialement.

Pour résumer, la théorie des plans de sondage à probabilités inégales regroupe des méthodes de sélection, des méthodes d'estimation et, par analogie, des méthodes de redressement qui utilisent des informations auxiliaires cherchant à « optimiser » l'estimation du paramètre d'intérêt.

3 Des limites aux probabilités inégales ?

Une théorie est rarement sans défaut, sans alternative, sans limite. La théorie des sondages a connu beaucoup de controverses, surtout à ses débuts. Si l'on s'intéresse à la pratique des sondages, les discussions sont encore bien plus nombreuses.

Pour autant, la théorie des sondages est une théorie statistique qui trouve son intérêt dans son applicabilité. Il convient donc de faire le point sur certains problèmes posés par la théorie des probabilités inégales au sein de la théorie des sondages. Pour cela, nous allons la mettre en perspective avec la théorie des sondages à probabilités égales.

Pour : les plans de sondage à probabilités d'inclusion égales

Le plan de sondage à probabilités égales est le plus naturel des plans de sondage. Il suit principalement le schéma ancestral de Bernoulli du tirage de boules dans une urne. En théorie des sondages, c'est un plan qui possède de remarquables propriétés.

- Méthode de sélection :
    - Les probabilités d'inclusion étant égales pour chaque unité, un simple algorithme de tri sur une variable pseudo aléatoire permet de sélectionner les *n* premiers de la liste.
    - Ce plan de sondage est à entropie maximale parmi l'ensemble des plans aléatoires de taille fixe.
    - Possibilité d'utiliser une information auxiliaire : un plan à strates proportionnelles à l'effectif suivi d'un tirage à probabilités égales dans chaque strate est un plan de sondage à probabilités d'inclusion égales.
- Méthode d'estimation
    - Les estimateurs d'une somme, d'une moyenne ou d'un quantile sont simples et explicites. Ces estimateurs sont sans biais et linéairement invariants.

- Si plusieurs variables d'intérêt sont mesurées, aucune variance d'estimateur n'est dégradée par les probabilités d'inclusion.
- La variance des estimateurs (somme ou moyenne) est explicite, sans biais et son calcul est exact.
- Aucun estimateur ne nécessite l'identification des unités dans l'échantillon et, de ce fait, la théorie statistique « classique » est applicable (exhaustivité, MLE…).

- Redressement
  - Techniquement, rien empêche une post stratification ou une pondération de type CALMAR.
  - Théoriquement, on s'interrogera sur la nécessité du redressement. Les distorsions de l'échantillon sont-elles dues à l'erreur d'échantillonnage ou à un biais (de couverture, de non réponse, …) ? Dans le premier cas, le redressement complique les choses et ne garantit pas la pertinence si on a plusieurs variables d'intérêt. Dans le second cas, on admet un biais qui nécessite une correction, c'est une toute autre affaire…

Contre : les plans de sondage à probabilités d'inclusion inégales

L'une des critiques les plus célèbres sur les plans de sondages à probabilités inégales est celle de Basu (1971), au travers de l'estimateur de Horvitz-Thompson. Son exemple sur le poids des éléphants d'un cirque illustre son aversion pour les probabilités inégales comme pour les propriétés mathématiquement justes mais concrètement inacceptables. Plus précisément, on peut lister les écueils suivants :

- Méthode de sélection
  - La méthode de sélection, basée sur les probabilités d'inclusion de premier ordre des unités, ne définit pas un unique plan de sondage. Un plan de sondage est défini par les probabilités des échantillons. Le principe du maximum d'entropie est alors la solution théoriquement raisonnable. Sa mise en œuvre est malheureusement concrètement plus que limitée.
  - Les autres algorithmes existant tentent de satisfaire aux propriétés nécessaires de sélection décrites dans la partie 2 mais sans les réaliser simultanément. Chaque algorithme peut s'avérer être ponctuellement « performant » et les recommandations fleurissent pour tel ou tel en fonction des multiples spécificités du sondage.
  - La détermination des probabilités d'inclusion est basée sur une hypothèse de « dépendance » entre la ou les variables auxiliaires et la ou les variables d'intérêts. Quelle intensité de dépendance garantit la pertinence ou la justesse de ces probabilités d'inclusion ?
  - Les calculs aboutissent parfois à des probabilités d'inclusion supérieures à un. Cela se traduit par la sélection non probabiliste de certaines unités. La relation entre les probabilités d'inclusion et la variable auxiliaire se trouve alors déformée.

- Méthode d'estimation
  - Aucun estimateur n'est satisfaisant, même pour une simple moyenne. L'un est sans biais mais non linéairement invariant et peut produire des estimations aberrantes (pour une proportion par exemple), l'autre est linéairement invariant mais pas sans biais et ses poids sont aléatoires…
  - Quid de l'estimation d'un quantile ou du calcul de la p-value d'un test d'hypothèse ?

- Les variances empiriques ne sont que des approximations d'approximation, quand elles ne sont pas négatives, ou encore, strictement positives alors même que dans la population, la variable d'intérêt est constante.

- Redressement (sans doute le point le plus inquiétant)
  - En théorie, on distingue bien la repondération pour erreur d'échantillonnage et la repondération pour erreur autre que d'échantillonnage. En pratique il est impossible de faire sérieusement la différence. Comment tester la part des choses avec un échantillon à probabilités d'inclusion inégales ?
  - La repondération est donc un fourre-tout dont on suppose, grâce aux propriétés théoriques obtenues pour le redressement d'un estimateur sans biais, qu'il traite l'essentiel des maux liés à la réalité effective du sondage.
  - On lit, ici ou là, que le redressement doit être « minime » ou « raisonnable » (les algorithmes développés vont souvent dans ce sens). Cela veut dire quoi ? Comment apprécier les choses ?
  - Quant à la « qualité » et/ou la « pertinence » statistique des variables auxiliaires, servant pour l'échantillonnage comme pour le redressement, ce sont des hypothèses qui nécessiteraient : analyse, quantification et intégration dans les calculs.

Au final, dans le cas le plus favorable d'une moyenne ou d'une somme, que peut-on dire de la qualité d'une estimation, initialement fragile, de variance approximative, ayant subi un redressement le plus souvent suspect ?

4 Discussion

En introduction, nous posions la question suivante : sous quelles conditions un sondage à probabilités inégales vaut-il le coût ? Les arguments évoqués contre les probabilités inégales sont, pour certains, liés à la théorie même des sondages, pour d'autres, ils sont liés à la pratique.

D'un point de vue théorique, l'existence d'un estimateur sans biais linéairement invariant (pour une moyenne ou une somme) nous semble essentielle. Il est bien connu qu'un tel estimateur existe pour les plans de sondages à probabilités égales. Ce qui est visiblement moins connu, c'est qu'il existe également pour certains plans de sondages à probabilités inégales. Les travaux de Patel-Dharmadhikari (1977) donnent les conditions nécessaires et suffisantes (CNS), sur les probabilités d'inclusion de premier et second ordre, de l'existence d'un tel estimateur. Les auteurs donnent également une définition implicite des poids de l'estimateur dilaté associé.

Comme souvent en théorie des sondages, les possibilités concrètes d'utilisation de cet estimateur sont extrêmement limitées. Mais l'enseignement fondamental n'est pas là. Il résulte dans le fait que si l'on ne demande pas « l'impossible » aux probabilités d'inclusion, en limitant leur distorsion alors certaines choses redeviennent théoriquement possibles.

D'un point de vue pratique, on peut se demander à qui sont destinés les nombreux développements autour des plans à probabilités inégales. Il semble que l'optimisation de la variance d'un estimateur via le plan de sondage (connaissant les innombrables sources d'augmentation non contrôlées, de la variance et du biais, qu'engendre la pratique) reste statistiquement anecdotique. En revanche cela donne une illusion de justification du redressement par repondération. Or, ne nous voilons pas la face, la raison du redressement est essentiellement une tentative de correction des erreurs autres que d'échantillonnage. Dans ce cadre, la théorie des sondages propose peu de critères de qualité de la correction, d'encadrement du biais. Tillé (2001) écrit pour la correction pour non-réponse que l'on ne peut généralement pas faire l'hypothèse de l'indépendance entre la non-réponse et les variables d'intérêt, mais ne développe dans son dernier chapitre que le cas ou l'on suppose l'indépendance.

En l'état, la théorie des sondages à probabilités inégales est orpheline. Comme Tillé (2001) le précise, « *si l'on réussit à sélectionner un échantillon sans remise en respectant les probabilités inégales, on ne dispose pas nécessairement d'un bon estimateur*». Le développement des travaux de Patel-Dharmadhikari (1977) et Patel and Patel (1993) sur les probabilités d'inclusion nous semble nécessaire.

D'autre part, le développement de méthodes « d'encadrement des biais » et leur intégration dans les calculs d'intervalle de confiance (ou de crédibilité) rendraient certains résultats de sondage plus réalistes (voir Gerville-Réache (2012) sur le cas (très) particulier des sondages d'intentions de vote).

En attendant, la qualité d'un sondage dépendant de la qualité de l'échantillon et de la qualité de l'estimation, il semble que les probabilités inégales posent actuellement, plus de problèmes, théoriques et pratiques, qu'elles n'apportent de solutions.

Bibliographie

[1] Ardilly P. (2006), *Les techniques de sondage*, Edition TECHNIP.

[2] Basu, D. (1971) An essay on the logical foundations of survey sampling, part one. In: Godambe and Sprott (eds) *Foundations of Statistical Inference*. Holt, Reinhart andWilson, 203–242.

[3] Chernoff H (1960), A compromise between bias and variance in the use of non representative samples, *Contributions to Probability and Statistics / Essays in Honor of H. Hotelling*, 153-167.

[4] Cochran W.G. (1977), *Sampling* techniques, 3rd edition, Wiley & Sons, NY.

[5] Deville J.C. (1993) « Les techniques de sondage, de Pascal Ardilly », *Courrier des statistiques* n° 67, 59-60.

[6] Deville, J.-C., and Tillé, Y. (2004). Efficient balanced sampling : the cube method, *Biometrika*, 91, 893-912.

[7] Gerville-Réache L., (2012) Sondages d'intention de vote : l'estimation des « marges d'erreur », *7ème colloque francophone sur les sondages*, Rennes, France.

[8] Horvitz D.G., Thompson D.J. (1952), A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association,* 47, 663–685

[9] Kruskal W., Mosteller F. (1979) Representative Sampling, III: The Current Statistical Literature. *International Statistical Review* Vol. 47, No. 3, 245-265

[10] Neyman J. (1934) On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection, *Journal of the Royal Statistical Society*, Vol. 97, No. 4, 558-625

[11] Patel H.C., Dharmadhikari S.W. (1977), On linear invariant unbiased estimators in survey sampling, *Sankhyā: The Indian Journal of Statistics*, Volume 39, Series C, Pt. 1, 21-27

[12] Patel J.A., Patel H.C. (1993). On Balanced Sampling Designs, *Sankhyā: The Indian Journal of Statistics*, Series B, Vol. 55, No. 2, 283-287.

[13] Riandey B., Widmer I. (2010), L'enseignement des sondages à l'usage du plus grand nombre: quelques réflexions tirées de l'expérience, *Statistique et enseignement,* Volume 1, n°1, 47-63.

[14] Shende, P.S., Ajgonkar, S.G. Prabhu (2002). A note on linear unbiased invariant estimator for some classes of estimators. *J. Indian Soc. Agric. Stat.* 55, No. 2, Article No. 1, 153-157.

[15] Tillé Y. (2001), *Théorie des sondages*, Edition DUNOD.