Échantillonnage probabiliste: Principes pour le choix du plan de sondage et équilibrage

Matthieu Wilhelm [1] & Yves Tillé [1]

[1] *Université de Neuchâtel, Institut de statistique, Avenue de Bellevaux 51, 2000 Neuchâtel, Suisse, matthieu.wilhelm@unine.ch*

Résumé. L'objectif de ce document est double. Tout d'abord, trois principes théoriques sont formalisés: randomisation, surreprésentation et restriction. Nous élaborons ces principes et donnons une justification de leur utilisation dans le choix systématique du plan d'échantillonnage. Dans le cadre assisté par modèle, la connaissance de la population est formalisée par la modélisation de la population et le plan d'échantillonnage est choisi en conséquence. Nous montrons comment les principes de surreprésentation et de restriction découlent naturellement de la modélisation de la population. L'échantillonnage équilibré apparat alors comme une conséquence de la modélisation. Deuxièmement, un examen de l'échantillonnage équilibré probabiliste est présenté dans le cadre du modèle. Pour certains modèles de base, l'échantillonnage équilibré peut s'avérer tre un plan d'échantillonnage optimal. L'accent est mis sur les nouvelles méthodes d'échantillonnage spatial et les modèles connexes. Un exemple illustratif montre les avantages des différentes méthodes. Tout au long du document, divers exemples illustrent comment les trois principes peuvent tre appliqués afin d'améliorer l'inférence.

Mots-clés. Échantillonnage équilibré, basé sur le modèle, inférence, entropie, méthode du cube.

Abstract. The aim of this paper is twofold. First, three theoretical principles are formalized: randomization, overrepresentation and restriction. We develop these principles and give a rationale for their use in choosing the sampling design in a systematic way. In the model-assisted framework, knowledge of the population is formalized by modelling the population and the sampling design is chosen accordingly. We show how the principles of overrepresentation and of restriction naturally arise from the modelling of the population. The balanced sampling then appears as a consequence of the modelling. Second, a review of probability balanced sampling is presented through the model-assisted framework. For some basic models, balanced sampling can be shown to be an optimal sampling design. Emphasis is placed on new spatial sampling methods and their related models. An illustrative example shows the advantages of the different methods. Throughout the paper, various examples illustrate how the three principles can be applied in order to improve inference.

Keywords. Balanced sampling, model-based, inference, entropy, cube method.

1 Quelques principes d'échantillonnage

Les statisticiens savent que la conception d'une enqute est une question complexe qui exige de l'expérience, une connaissance approfondie de la nature des variables d'intért et de la base de sondage. La plupart des manuels de sondages présentent une liste de méthodes d'échantillonnage. Toutefois, le choix du plan d'échantillonnage devrait tre le résultat de l'application de plusieurs principes. Dans ce qui suit, nous essayons d'établir quelques lignes directrices théoriques. Trois principes peuvent guider le choix du plan d'échantillonnage: le principe de randomisation, le principe de surreprésentation et le principe de restriction.

1.1 Le principe de la randomisation

Dans l'inférence fondée sur le plan de sondage, l'extrapolation des estimateurs de l'échantillon aux paramètres de la population est basée sur le plan de sondage, c'est-à-dire sur la façon dont l'échantillon est sélectionné. Ainsi, le premier principe consiste non seulement à sélectionner un échantillon au hasard mais à le sélectionner de manière aussi aléatoire que possible. La mesure commune de l'aléa d'un plan d'échantillonnage est son entropie, qui s'exprime de la manière suivante:

$$I(p) = -\sum_{s \in \mathcal{S}} p(s) \log p(s),$$

où on considère que $0 \log 0 = 0$..

Intuitivement, l'entropie est une mesure de la quantité d'information mais aussi une mesure du caractère aléatoire. Pour les plans d'échantillonnage complexes, les probabilités d'inclusion de second ordre sont rarement disponibles. Toutefois, lorsqu'on envisage des plans d'échantillonnage à grande entropie, on peut estimer la variance en utilisant des formules qui ne dépendent pas des probabilités d'inclusion du deuxième ordre.

1.2 Le principe de surreprésentation

L'échantillonnage consiste à sélectionner un sous-ensemble de la population. Toutefois, il n'y a aucune raison particulière de choisir les unités ayant des probabilités d'inclusion égales. Dans les enqutes liées aux entreprises, les établissements sont généralement sélectionnés avec des probabilités d'inclusion très différentes qui sont en général proportionnelles au nombre d'employés. Pour tre efficace, le choix des unités vise à réduire l'incertitude. Il est donc plus souhaitable de surreprésenter les unités qui contribuent davantage à la variance de l'estimateur.

L'idée de " représentativité " est donc complètement trompeuse et repose sur la fausse intuition qu'un échantillon doit tre semblable à la population pour qu'il y ait une inférence parce que l'échantillon est une " copie à l'échelle " de la population. En fait, la seule

exigence pour que l'estimateur soit non biaisé consiste à utiliser un plan d'échantillonnage avec une probabilité d'inclusion du premier ordre non nulle pour toutes les unités de la population.

L'échantillonnage à probabilités inégales peut tre utilisé pour estimer le total Y plus efficacement. L'idée principale est de suréchantillonner les unités qui sont plus incertaines parce que l'échantillon doit recueillir autant d'informations que possible auprès de la population. En général, le principe de surreprésentation implique qu'un plan d'échantillonnage devrait avoir des probabilités d'inclusion inégales si de l'informations auxiliaire est disponible.

1.3 Le principe de restriction

Le principe de restriction consiste à sélectionner uniquement des échantillons présentant un ensemble donné de caractéristiques, par exemple en fixant la taille de l'échantillon ou les tailles de l'échantillon dans des sous-catégories de la population (stratification). Il y a de nombreuses raisons pour lesquelles des restrictions devraient tre imposées. Par exemple, les catégories vides de l'échantillon peuvent tre évitées, ce qui peut tre très gênant lorsqu'il s'agit d'estimer des paramètres dans de petits sous-ensembles de la population. Il est également souhaitable que les estimations de l'échantillon soient cohérentes avec certaines connaissances auxiliaires. Ainsi, seuls les échantillons qui satisfont à une telle propriété devraient être considérés. Par cohérent, nous entendons que l'estimation à partir de l'échantillon d'une variable auxiliaire doit correspondre à un total connu. Ces échantillons sont dits équilibrés. L'échantillonnage équilibré nous permet d'éviter les " mauvais " échantillons, qui sont ceux qui donnent des estimations pour les variables auxiliaires qui sont loin des totaux connus de la population.

Cette communication est baée sur l'article cité dans la bibliographie.

Bibliographie

Tillé, Y. et Wilhelm, M. (2017). Probability Sampling Designs: Principles for Choice of Design and Balancing, *Statistical Science*, 32, pp. 176-189.
Hájek, J. (1981). Sampling from a Finite Population, Marcel Dekker: New-York,