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

Matthieu Wilhelm

Joint work with Yves Tillé

Université de Lausanne

Colloque Francophone sur les sondages, Lyon, octobre 2018

Dans cet exposé

- Introduction et notations
- Le principe de randomisation ;
- Le principe de surreprésentation ;
- Le principe de restriction ;
- Quelques illustrations ;

Introduction

Source de cet exposé

Cet exposé est basé sur la référence suivante :
Principles for Choice of Design and Balanced Sampling, Y. Tillé and M. W., *Statistical Science*, 2017.

Plan de sondage

On considère une population de N unités, $U = \{1, \ldots, N\}$. Un plan d'échantillonnage (sans remise) $p(.)$ est une distribution de probabilité définie sur l'ensemble des parties de $\mathcal{U}$,

$$p(s) \geq 0, \ s \subset U, \text{ such that } \sum_{s \subset U} p(s) = 1.$$

On note S l'échantillon aléatoire, de sorte que $\Pr(S = s) = p(s)$, et n désigne la taille d'échantillon.

Probabilités d'inclusion

La probabilité d'inclusion d'un individu k est la probabilité qu'il a d'appartenir à l'échantillon, c'est-à-dire

$$\pi_k = \Pr(k \in S) = \sum_{s \ni k} p(s).$$

La probabilité d'inclusion d'individus k et ℓ est la probabilité qu'il ont d'appartenir à l'échantillon simultanément, c'est-à-dire

$$\pi_{k\ell} = \pi_{\ell k} = \Pr(k \text{ and } \ell \in S) = \sum_{s \ni k,\ell} p(s).$$

Selon cette définition, on a $\pi_{kk} = \pi_k$.

L'estimateur de Horvitz-Thompson

L'estimateur de (Narain-) Horvitz-Thompson (Narain (1951), Horvitz and Thompson (1952) est défini comme

$$\widehat{T} = \sum_{k \in S} \frac{z_k}{\pi_k},$$

et est non-biaisé pour le total sur la population

$$T = \sum_{k \in U} z_k,$$

sous réserve que $\pi_k > 0$, $k \in U$.

Variance de l'estimateur de Horvitz-Thompson

Soit

$$\Delta_{k\ell} = \begin{cases} \pi_{k\ell} - \pi_k \pi_\ell & \text{if } k \neq \ell, \\ \pi_k(1 - \pi_k) & \text{if } k = \ell, \end{cases}$$

Alors la variance de l'estimateur de Horvitz-Thompson estimator est donnée par

$$\mathrm{var}\left(\widehat{T}\right) = \sum_{k \in U} \sum_{\ell \in U} \frac{z_k z_\ell}{\pi_k \pi_\ell} \Delta_{k\ell},$$

et peut être estimée de manière non-biaisée par

$$\widehat{\mathrm{var}}\left(\widehat{T}\right) = \sum_{k \in S} \sum_{\ell \in S} \frac{z_k z_\ell}{\pi_k \pi_\ell} \frac{\Delta_{k\ell}}{\pi_{k\ell}},$$

sous réserve que $\pi_{k\ell} > 0$, $k \neq \ell \in U$.

Principes pour le choix du plan de sondage et équilibrage

Trois principes

Nous suggérons d'être particulièrement attentifs aux trois aspects suivants pour construire un plan de sondage :

- Le principe de randomisation ;
- Le principe de surreprésentation ;
- Le principe de restriction.

Randomisation

Cela consiste à maximiser l'entropie (Hàjek, 1981) du plan de sondage, en tenant compte des diverses contraintes imposées (stratification, taille fixe, etc...).

Maximiser l'entropie est bénéfique pour les raisons suivantes :

- Convergence (plus rapide) des estimateurs vers leur distribution asymptotique normale (Berger, 1998a, 1998b) ,
- Approximation simplifiée de la variance estimator pour les plans à grande entropie (Brewer and Donadio, 2003).

Surreprésentation

L'idée fort répandue selon laquelle un "bon" échantillon est un échantillon à petite échelle de la population est erronée.
Au contraire,

- Le choix des unités vise à réduire l'incertitude
  $\Rightarrow$ il faut donc surreprésenter les individus qui contribuent le plus à l'erreur quadratique moyenne de l'estimateur.

Utilisation de modèle pour le choix des probabilités d'inclusions dans le cadre du paradigme "assisté par le modèle" (Särndal et al., 1992) .

Restriction et...

Il s'agit de limiter le support $\mathcal{Q}$ du plan (l'ensemble des échantillons avec une probabilité positive d'être sélectionné) aux échantillons "raisonnables".

- Il s'agit par exemple d'éviter que la taille d'échantillon soit trop petite dans certain petits domaines d'intérêt.
- Il s'agit par exemple d'assurer que l'échantillon est cohérent avec l'information auxiliaire que l'on possède.

... échantillonnage équilibré

Par définition, un échantillonnage équilibré satisfait

$$\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} = \sum_{k \in U} \mathbf{x}_k,$$

pour tout $S \in \mathcal{Q}$ et où $\mathbf{x}_k$ est un vecteur de variables auxiliaires pour l'individu k.

L'échantillonnage équilibré traduit donc en quelque sorte ce que "échantillon cohérent avec l'information auxiliaire" signifie.

Un modèle simple

Supposons que la variable d'intérêt z satisfait le modèle suivant $\mathbf{x}_k$:

$$z_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \tag{1}$$

où

- $\mathbf{x}_k = (x_{k1}, \ldots, \ldots, x_{kp})^\top$ est un vecteur de p variables auxiliaires,
- $\boldsymbol{\beta}$ est le vecteur des coefficients de régression,
- $\varepsilon_k \overset{\text{ind.}}{\sim} (0, \sigma_k^2)$.

Variance aniticipée

La variance anticipée de cet estimateur est donnée par

$$\begin{aligned}
\mathrm{AVar}(\widehat{Z}) &= \mathrm{E}_p \mathrm{E}_M (\widehat{Z} - Z)^2 \\
&= \mathrm{E}_p \left[\left(\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} - \sum_{k \in U} \mathbf{x}_k \right)^\top \boldsymbol{\beta} \right]^2 + \sum_{k \in U} (1 - \pi_k) \frac{\sigma_k^2}{\pi_k}
\end{aligned}$$

Minimisation de la variance anticipée

- $$\mathrm{E}_p \left[\left(\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} - \sum_{k \in U} \mathbf{x}_k \right)^\top \boldsymbol{\beta} \right]^2$$

  est nulle si l'échantillon est parfaitement équilibré.

- $$\sum_{k \in U} (1 - \pi_k) \frac{\sigma_k^2}{\pi_k}$$

  est minimisée si $\pi_k \propto \sigma_k, \forall k \in U$.

A priori, il n'existe pas de solution générale au fait de maximiser l'entropie sous des contraintes d'équilibrage et de probabilités d'inclusion.

Et en incluant de l'autocorrélation

Un modèle un tout petit peu plus général est donné par

$$z_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k, \tag{2}$$

où

- $\mathbf{x}_k = (x_{k1}, \ldots, \ldots, x_{kp})^\top$ est un vecteur de p variables auxiliaires,
- $\boldsymbol{\beta}$ est le vecteur des coefficients de régression,
- $\varepsilon_k \overset{\text{ind.}}{\sim} (0, \sigma_k^2)$.

et où

$$[\Sigma]_{k\ell} = (1 - \delta_{k\ell})\theta_{k\ell}\sigma_k\sigma_\ell + \delta_{k\ell}\sigma_k^2,$$

$\delta_{k\ell}$ étant le symbole de Kronecker.

Variance anticipée avec de la corrélation

Sous le modèle (2), a variance anticipée s'écrit comme

$$
\begin{aligned}
&\mathrm{AVar}(\widehat{Z}) \\
&= \mathrm{E}_p \left[\left(\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} - \sum_{k \in U} \mathbf{x}_k \right)^\top \boldsymbol{\beta} \right]^2 + \sum_{k \in U} \sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) \frac{\sigma_k \sigma_\ell \theta_{k\ell}}{\pi_k \pi_\ell} \\
&= \mathrm{E}_p \left[\left(\sum_{k \in S} \frac{\mathbf{x}_k}{\pi_k} - \sum_{k \in U} \mathbf{x}_k \right)^\top \boldsymbol{\beta} \right]^2 \\
&+ \sum_{k \in U} \frac{\sigma_k^2}{\pi_k} + \sum_{k \in U} \sum_{\substack{\ell \in U \\ \ell \neq k}} \frac{\sigma_k \sigma_\ell}{\pi_k \pi_\ell} \pi_{k\ell} \theta_{k\ell} - C
\end{aligned}
$$

Minimisation de la variance aniticipée

Supposons que $\theta_{k\ell}$ est une fonction décroissante de la distance entre les individus k et ℓ et que nous voulions minimiser la quantité suivante par rapport aux probabilités d'inclusion $\{\pi_k\}$, sous une contrainte d'espérance de taille n.
En particulier, on chercher à minimiser

$$\mathrm{AVar}(\widehat{Z})$$
$$\propto \mathrm{E}_p\left[\left(\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k}-\sum_{k\in U}\mathbf{x}_k\right)^\top\boldsymbol{\beta}\right]^2+\sum_{k\in U}\frac{\sigma_k^2}{\pi_k}+\sum_{k\in U}\sum_{\substack{\ell\in U\\ \ell\neq k}}\frac{\sigma_k\sigma_\ell}{\pi_k\pi_\ell}\pi_{k\ell}\theta_{k\ell}.$$

Quelques pistes pour l'optimisation

On remarque que

- $\mathrm{E}_p\left[\left(\sum_{k\in S}\frac{\mathbf{x}_k}{\pi_k}-\sum_{k\in U}\mathbf{x}_k\right)^\top\boldsymbol{\beta}\right]^2$ est minimisé si p est équilibré sur les variables $\mathbf{x}$,
- $\sum_{k\in U}\frac{\sigma_k^2}{\pi_k}$ est minimisé si $\pi_k \propto \sigma_k$,
- $\sum_{\substack{\ell\in U\\ \ell\neq k}}\frac{\sigma_k\sigma_\ell}{\pi_k\pi_\ell}\pi_{k\ell}\theta_{k\ell} = c_0^2\sum_{\substack{\ell\in U\\ \ell\neq k}}\pi_{k\ell}\theta_{k\ell}$ est réduit si deux individus proches ont une probabilité faible d'être échantillonnés simultanément $\Rightarrow$ Répulsion.

(Grafström et Tillé, 2013)

Quelques remarques

- L'équilibrage réduit la variance de n'importe quel modèle linéaire.
- La répulsion est intéressante s'il existe de l'autocorrélation dans la population (liée à la distance).
- La répulsion reste intéressante même en l'absence d'information auxiliaire.
- La répulsion diminue l'entropie de manière générale. On pourrait envisager de maximiser l'entropie sous des contraintes de type $0 < \delta_0 \leq \pi_{k\ell} \leq \delta_1$ pour tous les couples k, ℓ tels que $\|k - \ell\| \leq d_0$, avec δ_1 suffisamment petit ?

Conclusion

- Nous avons donc rappelé certains aspects importants dans le choix d'un plan de sondage.
- Nous avons explicité certains principes et justifié leur application.
- Nous avons montré comment implémenter ces principes en utilisant une approche basée sur le plan.

Bibliographie I

Y. G. Berger.

Rate of convergence for asymptotic variance for the Horvitz-Thompson estimator.

Journal of Statistical Planning and Inference, 74 :149–168, 1998a.

Y. G. Berger.

Rate of convergence to normal distribution for the Horvitz-Thompson estimator.

Journal of Statistical Planning and Inference, 67 :209–226, 1998b.

K. R. W. Brewer and M. E. Donadio.

The high entropy variance of the Horvitz-Thompson estimator.

Survey Methodology, 29 :189–196, 2003.

W. G. Cochran.

Sampling Techniques.

Wiley, New York, 1977.

Bibliographie II

A. Grafström and Y. Tillé.

Doubly balanced spatial sampling with spreading and restitution of auxiliary totals.

Environmetrics, 14 :120–131, 2013.

D. G. Horvitz and D. J. Thompson.

A generalization of sampling without replacement from a finite universe.

Journal of the American Statistical Association, 47 :663–685, 1952.

R. D. Narain.

On sampling without replacement with varying probabilities.

Journal of the Indian Society of Agricultural Statistics, 3 :169–174, 1951.

C.-E. Särndal, B. Swensson, and J. H. Wretman.

Model Assisted Survey Sampling.

Springer, New York, 1992.

Bibliographie III

Y. Tillé.

Sampling Algorithms.

Springer, New York, 2006.

Y. Tillé and M. Wilhelm.

Probability sampling designs : Principles for choice of design and balancing.

Statistical Science, 32 :176–189, 2017.