La convergence d'estimateurs et d'estimateurs de variance pour l'échantillonnage à deux degrés

Audrey-Anne Vallée [1] & Guillaume Chauvet [2]

[1] *Institut de Statistique, Université de Neuchâtel, Neuchâtel, Suisse, audrey-anne.vallee@unine.ch*

[2] *Ensai (Irmar), Campus de Ker Lann, Bruz, France, guillaume.chauvet@ensai.fr*

Résumé. Les plans de sondage à deux degrés sont communément utilisés pour des enquêtes sur les ménages et sur la santé. Un estimateur de base pour des totaux dans les populations est l'estimateur Horvitz-Thompson, accompagné de ses estimateurs de variance usuels. La convergence et la distribution asymptotique des estimateurs sont complexes à étudier dans les plans à deux degrés. Dans cette présentation, les propriétés asymptotiques de l'estimateur du total Horvitz-Thompson et de ses estimateurs de variance sont étudiées. Sous des hypothèses raisonnables, la convergence des estimateurs est prouvée. Dans le cas du plan de sondage réjectif au premier degré, la convergence d'un estimateur de variance simplifié du type Hájek est vérifiée et il est montré que l'estimateur du total a une distribution asymptotiquement normale. Les résultats d'une étude par simulations évaluant les propriétés asymptotiques seront présentés.

Mots-clés. Échantillonnage réjectif, estimateur de variance simplifié, estimateurs d'Horvitz-Thompson, normalité asymptotique, plans à grande entropie.

Abstract. Two-stage sampling designs are commonly used for household and health surveys. A basic estimator for population totals is the Horvitz-Thompson estimator, along with its usual variance estimators. The consistency and the asymptotic distribution of the estimators are complicated to study in two-stage designs. In this presentation, the asymptotic properties of the Horvitz-Thompson total estimator and its variance estimators are studied. Under reasonable assumptions, the consistency of the estimators is proven. In the case of rejective sampling at the first stage, the consistency of a simplified Hájek-type variance estimator is verified and the total estimator is shown to be asymptotically normally distributed. The results of a simulation study evaluating the asymptotic properties will be presented.

Keywords. Asymptotic normality, high entropy designs, Horvitz-Thompson estimators, rejective sampling, simplified variance estimator.

1 Introduction

Dans les sondages de grande envergure, comme les enquêtes sur les ménages ou sur la santé, la population est souvent répartie sur un grand territoire et la base de sondage est

rarement fournie. Les plans de sondage à deux degrés sont commodes dans ce genre de situation. En effet, le premier degré ne nécessite qu'une base de sondage composée d'unités groupées qui est généralement plus facile à créer. Le premier degré du plan de sondage, consiste à sélectionner un groupe d'unités primaires d'échantillonnage (ou unités de la population groupées). Au deuxième degré, des unités secondaires d'échantillonnage (ou unités de la population) sont choisies dans l'échantillon d'unités primaires. L'estimateur du total d'Horvitz-Thompson peut être utilisé avec un estimateur de sa variance habituel. Par contre, sa variance est normalement plus grande que celle obtenue avec plan de sondage où les unités primaires ont été échantillonnées directement. L'échantillonnage à plusieurs degrés est détaillé dans, par exemple, Cochran (1977), Särndal et coll. (1992) et Fuller (2011).

Certaines propriétés des estimateurs sont nécessaires dans les sondages : la convergence de l'estimateur et de son estimateur de variance et la validité d'un théorème central limite pour l'estimateur. Ces résultats ont été étudiés pour les plans de sondage à un degré. Hájek (1964) a prouvé la normalité asymptotique de l'estimateur d'Horvitz-Thompson dans un plan réjectif. Berger (1998a) et Berger (1998b) ont étudié la normalité asymptotique et la convergence d'estimateurs de variance dans les plans à grande entropie. Breidt et Opsomer (2000) ont étudié ces propriétés pour les estimateurs par la régression polynomiale locale sous des hypothèses qui sont difficilement extrapolées aux plans à deux degrés. Récemment, Boistard et coll. (2017) et Bertail et coll. (2017) ont établi un théorème central limite fonctionnel pour des processus empiriques d'Horvitz-Thompson. Dans les plans de sondage à deux degrés, les propriétés asymptotiques des estimateurs ne sont pas aussi clairement établies. Krewski et Rao (1981) ont étudié le cas où les unités primaires sont sélectionnées avec remise et Ohlsson (1989) a établi un théorème central limite général pour les plans à deux degrés. Récemment, Chauvet (2015) a considéré des méthodes de couplage pour obtenir un théorème central limite pour l'estimateur d'Horvitz-Thompson et la convergence d'un estimateur de variance par bootstrap lorsque le plan au premier degré est un plan simple stratifié.

Dans ce travail, les propriétés asymptotiques d'estimateurs du total et de la variance sont étudiées dans le cadre général des plans de sondage à deux degrés. La convergence de l'estimateur d'Horvitz-Thompson est montrée pour les plans à deux degrés. La convergence de l'estimateur de la variance d'Horvitz-Thompson et d'un estimateur de variance simplifié est aussi montrée pour le même contexte général. Lorsqu'un plan réjectif est utilisé au premier degré, il est montré que l'estimateur de variance simplifié de type Hájek est convergent et que l'estimateur d'Horvitz-Thompson est asymptotiquement normalement distribué. Ces propriétés asymptotiques sont satisfaites en émettant plusieurs hypothèses qui sont raisonnables en pratique. Les propriétés des estimateurs de variance du type Hájek sont évaluées dans une étude par simulations.

2 Estimateurs et résultats principaux

Soit une population finie U de N Unités Secondaires d'Échantillonnage (USEs) qui est partitionnée en une population U_I de N_I Unités Primaires d'Échantillonnage (UPEs). Un échantillon à deux degrés est sélectionné dans U pour estimer le total de la population de la variable y,

$$Y = \sum_{k \in U} y_k = \sum_{u_i \in U_I} Y_i,$$

où $Y_i = \sum_{k \in u_i} y_i$ est le total de la variable y dans l'UPE u_i. Au premier degré du plan de sondage, un échantillon S_I de n_I UPEs est sélectionné dans U_I. La probabilité d'inclusion de l'UPE u_i est notée π_{Ii} et la probabilité jointe d'inclusion des UPEs u_i et u_j dans S_I est π_{Iij}. Au deuxième degré du plan de sondage, un échantillon S_i de n_i USEs est choisi dans chaque UPE $u_i \in S_I$. La probabilité d'inclusion de l'USE k conditionnellement à la sélection de son UPE u_i est notée $\pi_{k|i}$ et la probabilité conditionnelle jointe dans S_i des USEs k et ℓ est $\pi_{k\ell|i}$. Le plan de sondage est dit invariant et indépendant (Särndal et coll., 1992) : la sélection de l'échantillon au deuxième degré est indépendant de S_I et les USEs sont sélectionnées indépendamment d'une UPE à l'autre.

L'estimateur d'Horvitz-Thompson de Y est

$$\widehat{Y}_\pi = \sum_{u_i \in S_I} \frac{\widehat{Y}_i}{\pi_{Ii}} \quad \text{avec} \quad \widehat{Y}_i = \sum_{k \in S_i} \frac{y_k}{\pi_{k|i}}.$$

La variance de $\widehat{Y}_\pi$ est

$$\begin{aligned} \mathrm{V}(\widehat{Y}_\pi) &= \sum_{u_i, u_j \in U_I} \Delta_{ij} \frac{Y_i}{\pi_{Ii}} \frac{Y_j}{\pi_{Ij}} + \sum_{u_i \in U_I} \left(\frac{1 - \pi_{Ii}}{\pi_{Ii}} \right) V_i + \sum_{u_i \in U_I} V_i \\ &= \mathrm{V}_1(\widehat{Y}_\pi) + \mathrm{V}_2(\widehat{Y}_\pi) + \mathrm{V}_3(\widehat{Y}_\pi), \end{aligned} \tag{1}$$

où

$$V_i = \mathrm{V}(\widehat{Y}_i) = \sum_{k, \ell \in u_i} \Delta_{k\ell|i} \frac{y_k}{\pi_{k|i}} \frac{y_\ell}{\pi_{\ell|i}},$$

$\Delta_{ij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$ et $\Delta_{k\ell|i} = \pi_{k\ell|i} - \pi_{k|i}\pi_{\ell|i}$. L'estimateur de variance d'Horvitz-Thompson est

$$\begin{aligned} \widehat{\mathrm{V}}_{HT}(\widehat{Y}_\pi) &= \sum_{u_i, u_j \in S_I} \frac{\Delta_{ij}}{\pi_{Iij}} \frac{\widehat{Y}_i}{\pi_{Ii}} \frac{\widehat{Y}_j}{\pi_{Ij}} + \sum_{u_i \in S_I} \frac{\widehat{V}_{HT,i}}{\pi_{Ii}} \\ &= \widehat{\mathrm{V}}_{HT,A}(\widehat{Y}_\pi) + \widehat{\mathrm{V}}_{HT,B}(\widehat{Y}_\pi), \end{aligned} \tag{2}$$

où

$$\widehat{V}_{HT,i} = \sum_{k,\ell \in S_i} \frac{\Delta_{k\ell|i}}{\pi_{k\ell|i}} \frac{y_k}{\pi_{k|i}} \frac{y_\ell}{\pi_{\ell|i}}.$$

Dans le cadre asymptotique habituel en sondage, la population U appartient à une séquence de populations finies imbriquées $\{U_t\}$ de tailles croissantes N_t. Dans notre contexte, il est supposé que le nombre d'UPEs $N_I \to \infty$ et que le nombre d'UPEs échantillonnées $n_I \to \infty$ lorsque $t \to \infty$. Pour vérifier les propriétés asymptotiques des estimateurs présentés dans ce travail, plusieurs hypothèses sont formulées. Brièvement, des restrictions sur les bornes des probabilités d'inclusion d'ordre un, deux et quatre, sur la variabilité de N_i et n_i, sur l'ordre de grandeur de n_I et sur des bornes pour le moment d'ordre quatre et la moyenne de la variable y sont émises. La majorité des hypothèses sont satisfaites avec plusieurs plans de sondage ou contrôlées par les enquêteurs.

Dans ce travail, la convergence de l'estimateur du total d'Horvitz-Thompson $\widehat{Y}_\pi$ est démontrée. Il est prouvé que le premier terme de l'estimateur de la variance d'Horvitz-Thompson dans l'équation (2), $\widehat{\mathrm{V}}_{HT,A}(\widehat{Y}_\pi)$, est un estimateur convergent vers $V_1(\widehat{Y}_\pi) + V_2(\widehat{Y}_\pi)$ de l'équation (1). Il est démontré que le terme $\widehat{\mathrm{V}}_{HT,B}(\widehat{Y}_\pi)$ dans (2) est un estimateur convergent vers $V_3(\widehat{Y}_\pi)$ dans l'équation (1). Pour éviter le calcul de la variance $\widehat{V}_{HT,i}$ dans toutes les UPEs échantillonnées, il est montré que le terme $\widehat{\mathrm{V}}_{HT,A}(\widehat{Y}_\pi)$ est un estimateur convergent vers la variance $\mathrm{V}(\widehat{Y}_\pi)$ si

$$\frac{V_3(\widehat{Y}_\pi)}{V_1(\widehat{Y}_\pi) + V_2(\widehat{Y}_\pi)} \to 0.$$

Cette condition est notamment respectée si le taux de sondage au premier degré est négligeable.

Pour les plans de sondage réjectifs, Hájek (1964) a construit une approximation asymptotique des probabilités d'inclusion d'ordre deux. Il a proposé un estimateur de variance simple qui ne requière pas les probabilités d'inclusion d'ordre deux et qui s'exprime en une seule somme. Quand un plan réjectif est utilisé au premier degré du plan de sondage, nous utilisons un estimateur de variance du type Hájek pour remplacer $\widehat{\mathrm{V}}_{HT,A}(\widehat{Y}_\pi)$ et nous prouvons qu'il est convergent. Nous montrons aussi que l'estimateur du total d'Horvitz-Thompson est asymptotiquement normalement distribué. Dans une étude par simulations, nous évaluons les propriétés asymptotiques de cet estimateur de variance du type Hájek. Le biais relatif Monte Carlo pour $\mathrm{V}_1(\widehat{Y}_\pi) + \mathrm{V}_2(\widehat{Y}_\pi)$ et pour $\mathrm{V}(\widehat{Y}_\pi)$ et les intervalles de confiance Monte Carlo associés sont calculés.

Références

Berger, Y. G. (1998a). Rate of convergence for asymptotic variance of the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, **74**, 149 – 168.

Berger, Y. G. (1998b). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, **67**, 209 – 226.

Bertail, P., Chautru, E. et Clémençon, S. (2017). Empirical processes in survey sampling with (conditional) poisson designs. *Scandinavian Journal of Statistics*, **44**, 97–111.

Boistard, H., Lopuhaä, H. P. et Ruiz-Gazen, A. (2017). Functional central limit theorems for single-stage sampling designs. *The Annals of Statistics*, **45**, 1728–1758.

Breidt, F. J. et Opsomer, J. D. (2000). Local polynomial regresssion estimators in survey sampling. *The Annals of Statistics*, **28**, 1026–1053.

Chauvet, G. (2015). Coupling methods for multistage sampling. *The Annals of Statistics*, **43**, 2484–2506.

Cochran, W. G. (1977). *Sampling techniques*. New York : Wiley, third edn.

Fuller, W. A. (2011). *Sampling statistics*. New York : Wiley.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, **35**, 1491–1523.

Krewski, D. et Rao, J. N. K. (1981). Inference from stratified samples : properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, **9**, 1010–1019.

Ohlsson, E. (1989). Asymptotic normality for two-stage sampling from a finite population. *Probability Theory and Related Fields*, **81**, 341–352.

Särndal, C.-E., Swensson, B. et Wretman, J. (1992). *Model assisted survey sampling*. New York : Springer.