Stratégies d'échantillonnage et d'estimation pour une variable d'écoute

Guillaume Chauvet *

27 novembre 2018

1 Notation

On considère une population U_J de N_J jours, et une population U_I de N_I individus. On s'intéresse à la population produit $U_I \times U_J$, de taille $N_I \times N_J$: autrement dit, l'unité statistique qui nous intéresse est un couple (individu , jour), noté (i, j). Dans le cas considéré dans la simulation, nous avons $N_J = 85$ et $N_I = 44\ 000\ 000$.

On considère la variable d'écoute y_{ij}, égale à 1 si l'individu i écoute une radio donnée le jour j, pour laquelle on mesure l'audience moyenne définie par

$$\mu = \frac{1}{N_I \times N_J} \sum_{(i,j) \in U_I \times U_J} y_{ij}. \quad (1)$$

On considère dans la Section 2 une stratégie d'échantillonnage stratifié par jour. On considère dans la Section 3 une stratégie d'échantillonnage d'individus, vus comme des grappes de jours. On considère dans la Section 4 des estimations composites associant les deux échantillons, et le choix d'un paramètre optimal de mélange des deux estimations.

2 Echantillon "déclaratif"

Dans cette stratégie d'échantillonnage, la population $U_I \times U_J$ est stratifiée par jour, et pour chaque jour j on tire indépendamment un échantillon s_j^a

*ENSAI/IRMAR, Campus de Ker Lann, 35170 Bruz, France. E-mail: chauvet@ensai.fr

de n_I^a individus parmi les N_I individus de U_I. Le poids d'un couple (i,j) est donc

$$d_{ij}^a = \frac{N_I}{n_I^a} \times 1 = \frac{N_I}{n_I^a}. \quad (2)$$

Par exemple, avec $N_I = 44\ 000\ 000$ et $n_I^a = 250$, on obtient $d_{ij}^a = 176\ 000$.

L'estimateur de l'audience moyenne μ peut s'écrire

$$\hat{\mu}^a = \frac{1}{N_I \times N_J} \sum_{j \in U_J} \sum_{i \in s_j^a} \frac{N_I}{n_I^a} y_{ij} = \frac{1}{N_J \times n_I^a} \sum_{j \in U_J} \sum_{i \in s_j^a} y_{ij}. \quad (3)$$

La variance de cet estimateur s'obtient à partir des formules pour le sondage aléatoire simple stratifié, ce qui conduit à

$$V(\hat{\mu}^a) = \frac{1}{(N_J)^2} \sum_{j \in U_J} \left(\frac{1}{n_I^a} - \frac{1}{N_I} \right) S_{\bullet j}^2 \quad (4)$$

avec

$$S_{\bullet j}^2 = \frac{1}{N_I - 1} \sum_{i \in U_I} (y_{ij} - \mu_{\bullet j})^2 \text{ la dispersion pour le jour } j, \quad (5)$$

$$\mu_{\bullet j} = \frac{1}{N_I} \sum_{i \in U_I} y_{ij} \text{ l'audience moyenne pour le jour } j.$$

Comme la variable y_{ij} est une indicatrice, on peut réécrire la formule (4) sous la forme

$$\begin{aligned} V(\hat{\mu}^a) &= \frac{1}{(N_J)^2} \left(\frac{1}{n_I^a} - \frac{1}{N_I} \right) \sum_{j \in U_J} \frac{N_I}{N_I - 1} \mu_{\bullet j}(1 - \mu_{\bullet j}), \quad (6) \\ &\simeq \frac{1}{(N_J)^2} \left(\frac{1}{n_I^a} - \frac{1}{N_I} \right) \sum_{j \in U_J} \mu_{\bullet j}(1 - \mu_{\bullet j}) \text{ car } N_I \text{ est grand}, \\ &\simeq \frac{1}{n_I^a (N_J)^2} \sum_{j \in U_J} \mu_{\bullet j}(1 - \mu_{\bullet j}) \text{ car } n_I^a \text{ est petit devant } N_I. \end{aligned}$$

On notera plus simplement cette variance $V\left(\hat{\mu}_y^a\right) \equiv V^a$ dans la suite.

3 Echantillon "panel"

Dans cette stratégie d'échantillonnage, un échantillon s_I^b de n_I^b individus est sélectionné parmi les N_I individus de U_I. On décrit dans la Section 3.1 le cas où tous les jours d'écoute de ces individus sont obtenus (stratégie par grappes), puis dans la Section 3.2 le cas où un échantillon seulement de jours d'écoute est utilisé (stratégie à deux degrés).

3.1 Stratégie par grappes

Dans le premier cas, on retient tous les jours d'écoute pour un individu i de s_I^b. Le poids d'un couple (i,j) est donc

$$d_{ij}^b = \frac{N_I}{n_I^b} \times 1 = \frac{N_I}{n_I^b}. \tag{7}$$

Par exemple, avec $N_I = 44\ 000\ 000$ et $n_I^b = 6\ 700$, on obtient $d_{ij}^b \simeq 6\ 567$.

L'estimateur de l'audience moyenne μ peut s'écrire

$$\hat{\mu}^b = \frac{1}{N_I \times N_J} \sum_{i \in s_I^b} \sum_{j \in U_J} \frac{N_I}{n_I^b} y_{ij} = \frac{1}{n_I^b} \sum_{i \in s_I^b} \mu_{i\bullet}, \tag{8}$$

$$\text{avec } \mu_{i\bullet} = \frac{1}{N_J} \sum_{j \in U_J} y_{ij} \text{ l'audience moyenne pour l'individu } i.$$

La variance de cet estimateur s'obtient à partir des formules pour un sondage aléatoire simple de grappes, ce qui conduit à

$$V(\hat{\mu}^b) = \left(\frac{1}{n_I^b} - \frac{1}{N_I} \right) S_{\bullet\bullet}^2 \quad \text{avec} \quad S_{\bullet\bullet}^2 = \frac{1}{N_I - 1} \sum_{i \in U_I} (\mu_{i\bullet} - \mu)^2. \tag{9}$$

3.2 Stratégie à deux degrés

Dans ce second cas, on sélectionne pour chaque individu i de s_I^b un sous-échantillon s_i^b de n_J^b jours. On fait les hypothèses habituelles d'indépendance et d'invariance pour un tirage à deux degrés, i.e.

- — indépendance : les tirages des sous-échantillons s_i^b sont indépendants, conditionnellement à s_I^b,
- — invariance : les tirages des sous-échantillons s_i^b sont indépendants de s_I^b.

Dans ce cas, le poids d'un couple (i,j) est donc

$$d_{ij}^b = \frac{N_I}{n_I^b} \times \frac{N_J}{n_J^b} = \frac{N_I N_J}{n_I^b n_J^b}. \tag{10}$$

Par exemple, avec $N_I = 44\ 000\ 000$ et $n_I^b = 6\ 700$, et en tirant $n_J^b = 42$ jours parmi les $N_J = 85$ jours, on obtient $d_{ij}^b \simeq 13\ 291$.

L'estimateur de l'audience moyenne μ peut s'écrire

$$\hat{\mu}^b = \frac{1}{N_I \times N_J} \sum_{i \in s_I^b} \sum_{j \in s_i^b} \frac{N_I\ N_J}{n_I^b\ n_J^b} y_{ij} = \frac{1}{n_I^b\ n_J^b} \sum_{i \in s_I^b} \sum_{j \in s_i^b} y_{ij}. \tag{11}$$

La variance de cet estimateur s'obtient à partir des formules pour un tirage avec sondage aléatoire simple à chaque degré, ce qui conduit à

$$V(\hat{\mu}^b) = \left(\frac{1}{n_I^b} - \frac{1}{N_I}\right) S_{\bullet\bullet}^2 + \frac{1}{n_I^b}\frac{1}{N_I}\sum_{i\in U_I}\left(\frac{1}{n_J^b} - \frac{1}{N_J}\right) S_{i\bullet}^2, \tag{12}$$

$$\text{avec } S_{i\bullet}^2 = \frac{1}{N_J - 1}\sum_{j\in U_J}(y_{ij} - \mu_{i\bullet})^2 \text{ la dispersion pour l'individu } i.$$

On notera plus simplement cette variance $V(\hat{\mu}^b) \equiv V^b$ dans la suite. Notons que les formules (8) et (9) sont des cas particuliers de (11) et (12), obtenus quand $n_J^b = N_J$.

Comme la variable y_{ij} est une indicatrice, on peut réécrire la formule (12) sous la forme

$$\begin{aligned} V(\hat{\mu}^b) &= \left(\frac{1}{n_I^b} - \frac{1}{N_I}\right) S_{\bullet\bullet}^2 + \frac{1}{n_I^b}\frac{1}{N_I}\sum_{i\in U_I}\left(\frac{1}{n_J^b} - \frac{1}{N_J}\right)\frac{N_J}{N_J - 1}\mu_{i\bullet}(1-\mu_{i\bullet}), \\ &\simeq \left(\frac{1}{n_I^b} - \frac{1}{N_I}\right) S_{\bullet\bullet}^2 + \frac{1}{n_I^b}\frac{1}{N_I}\left(\frac{1}{n_J^b} - \frac{1}{N_J}\right)\sum_{i\in U_I}\mu_{i\bullet}(1-\mu_{i\bullet}), \qquad (13) \\ &\simeq \frac{S_{\bullet\bullet}^2}{n_I^b} + \frac{1}{n_I^b}\frac{1}{N_I}\left(\frac{1}{n_J^b} - \frac{1}{N_J}\right)\sum_{i\in U_I}\mu_{i\bullet}(1-\mu_{i\bullet}) \text{ car } n_I^b \text{ est petit devant } N_I. \end{aligned}$$

4 Estimateur composite

On obtient un estimateur composite utilisant les deux estimateurs en donnant un poids α à l'échantillon "déclaratif" et un poids $1-\alpha$ à l'échantillon "panel". On obtient l'estimateur

$$\hat{\mu}(\alpha) = \alpha\hat{\mu}^a + (1-\alpha)\hat{\mu}_y^b, \tag{14}$$

dont la variance vaut

$$V\left\{\hat{\mu}(\alpha)\right\} = \alpha^2 V^a + (1-\alpha)^2 V^b. \tag{15}$$

En optimisant selon α, on obtient la valeur

$$\alpha_{opt} = \frac{V^b}{V^a + V^b} \tag{16}$$

qui conduit à l'estimateur $\hat{\mu}(\alpha)$ ayant la plus petite variance (estimateur de Hartley, 1962).

Références

Hartley, H.O. (1962). *Multiple frame surveys*. In Proceedings of the Social Statistics Section, American Statistical Association, 19(6), p. 2.