Procédures rapides pour la sélection d'échantillons à probabilités inégales à partir d'un flux

Yves Tillé
Université de Neuchâtel

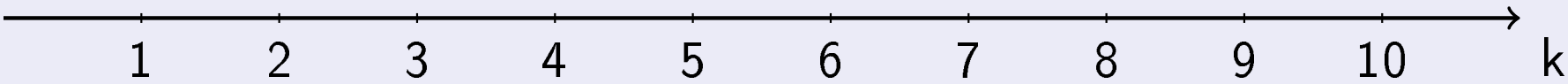
10éme Colloque francophone sur les sondages

Table of contents

1. Introduction et notations
2. Méthodes à taux fixe
3. Méthodes de taille fixe
4. Généralisation
5. Conclusions

Notation

Flux (stream)



- Le flux est tellement grand qu'on veut décider directement de la sélection d'une nouvelle unité quand elle apparaît.
- Les unités non-sélectionnées sont effacées.

Notations

- Suite de populations $U_1, U_2, \ldots, U_n, \ldots, U_i, \ldots, U_N$.
- Suite d'échantillons $S_1, S_2, \ldots, S_n, \ldots, S_i, \ldots, S_N$.
- $S_i \subset U_i$.
- Variable auxiliaire $x_k > 0, k \in U_N$.

Méthodes à taux fixe

- Probabilités d'inclusion $\pi_k = \min(\tau x_k, 1)$.
- τ est fixé.
- L'espérance de la taille de l'échantillon vaut $\sum_{k \in U_N} \min(\tau x_k, 1)$.

Méthodes à taux fixe : solutions

- Plan de Poisson.
  Distribution de n_S est Poisson-Binomiale
  (Hodges Jr. and Le Cam, 1960; Stein, 1990; Chen and Liu, 1997)
- Tirage systématique à probabilités inégales
  Taille d'échantillon presque fixe (Madow, 1949).
- Méthode du pivot ordonnée
  (Deville and Tillé, 1998; Grafström et al., 2012; Chauvet, 2012).
- Méthode de Fuller
  méthode du pivot avec une première unité fantôme $\pi_0 \sim Unif(0,1)$
  (Fuller, 1970; Tillé, 2018).
- Échantillonnage équilibré. Méthode du cube rapide (phase de vol uniquement). On n'applique la phase de vol que sur les $p+1$ premières unités avec des valeurs non-entières (Deville and Tillé, 2004; Chauvet and Tillé, 2006).

Méthodes de taille fixe

- Probabilités d'inclusion $\pi_k(U_i, n) = \min(\tau_i\, x_k, 1)$ tel que

$$\sum_{k \in U_i} \min(\tau_i\, x_k, 1) = n.$$

Méthodes de taille fixe

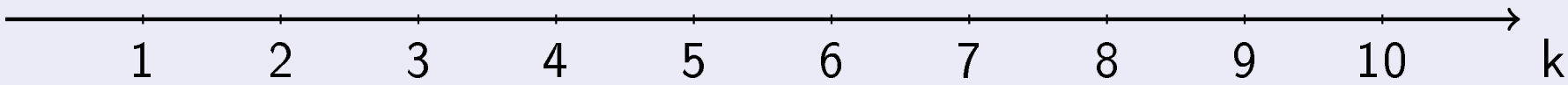
Exemple du calcul des $\pi_k(U_i, n)$ $N = 12, n = 5$

x	0.70	2.6	0.48	0.40	0.21	0.73	0.15	0.43	0.53	0.05	0.34	3.70
U_5	1	1	1	1	1							
U_6	1	1	0.88	0.73	0.39	1						
U_7	1	1	0.77	0.64	0.34	1	0.24					
U_8	0.91	1	0.62	0.51	0.28	0.94	0.19	0.56				
U_9	0.77	1	0.53	0.44	0.24	0.80	0.16	0.48	0.58			
U_{10}	0.76	1	0.52	0.43	0.23	0.79	0.16	0.47	0.57	0.05		
U_{11}	0.70	1	0.48	0.40	0.21	0.72	0.15	0.43	0.52	0.05	0.34	
U_{12}	0.53	1	0.36	0.30	0.16	0.54	0.11	0.32	0.39	0.04	0.25	1

Méthode de Chao

Méthode de Chao

- La méthode de Chao est une méthode à réservoir (Chao, 1982; Sugden et al., 1996).
- Au début le réservoir contient les $n = 5$ premières unités.



1	2	3	4	5

- Chaque fois qu'une nouvelle unité apparaît avec une probabilité, elle peut entrer dans le réservoir et une unité du réservoir est enlevée.

Méthode de Chao

Méthode de Chao

- À l'étape $i = n+1, \ldots, N$, l'unité i est incluse dans le réservoir avec une probabilité $\pi_i(U_i, n)$.
- Si l'unité i est sélectionnée, l'une des unités du réservoir est enlevée avec la probabilité :

$$a_{ki} = \frac{1}{\pi_k(U_i, n)} \left[1 - \frac{\pi_k(U_i, n)}{\pi_k(U_{i-1}, n)}\right], k = 1, \ldots, i-1.$$

- Il est en effet possible de prouver que

$$\sum_{k \in S_{i-1}} a_{ki} = 1.$$

- Cohen et al. (2009) ont montré que les unités non sélectionnées peuvent être définitivement oubliées.

Généralisation

Quasi-échantillons

- Phase de vol rapide (ffph pour fast flight phase).
- Sélection de quasi-échantillon :

$$\mathrm{ffph}(\pi_1, \ldots, \pi_k, \ldots, \pi_N) = \boldsymbol{\psi} = (\psi_1, \ldots, \psi_k, \ldots, \psi_N)^\top$$

de telle sorte que $0 \leq \psi_k \leq 1, k \in U_N$, $\mathrm{E}(\psi_k) = \pi_k$,
$\mathrm{card}\{0 < \psi_k < 1\} \leq p$ et

$$\sum_{k \in U_N} \frac{\psi_k \mathbf{z}_k}{\pi_k} = \sum_{k \in U_N} \mathbf{z}_k.$$

où $\mathbf{z}_k \in \mathbb{R}^p$.

Généralisation

Quasi-échantillons

- Sélection en deux phases avec des probabilités respectives $\pi_k^1 > \pi_k^2$ équilibré sur $\mathbf{z}_k$.

Proposition

Si le quasi-échantillon ψ_1 tiré avec les probabilités π_k^1 est équilibré sur $\mathbf{z}_k$ et qu'il existe un vecteur $\boldsymbol{\theta} \in \mathbb{R}^p$ tel que $\boldsymbol{\theta}^\top \mathbf{z}_k = v_k$, alors les probabilités d'inclusion π_k^2 et donc les probabilités de tirage ξ_k peuvent être calculées à partir de v_k sans connaître les unités telles que $\psi_k^1 = 0$, en résolvant dans τ_2 :

$$\sum_{k \in U_N} \min(1, \tau_2 v_k) \frac{\psi_k^1}{\pi_k^1} \pi_k^1 = m.$$

Généralisation

Plusieurs généralisations

- Méthode du réservoir équilibrée.
- Méthode de Chao par blocs (on ne considère plus une nouvelle unités mais un bloc de H nouvelles unités).
- Méthode du réservoir équilibrée par bloc.
- Méthode en deux passages.

Méthode en deux passages

Méthode en deux passages

- D'abord on sélectionne un grand échantillon équilibré sur $\mathbf{z}_k = (\pi_k^1, x_k)^\top$.
- $\pi_k^1 = \min\left(nx_k \frac{\sum_{k\in U_k} x_k}{\sum_{k\in U_{k+1}} x_k}, 1\right)$.
- Ensuite un sous-échantillon.
- $\pi_k^2 = \pi_k(U_N, n)$.

Méthode en deux passages

Méthode en deux passages

- Exemple avec des probabilités d'inclusion égales $x_k = 1$, $n = 5$.

k	1	2	3	4	5	6	7	8	9	10	11	12
x_k	1	1	1	1	1	1	1	1	1	1	1	1
π_k^1	1	1	1	1	1	$\frac{5}{6}$	$\frac{5}{7}$	$\frac{5}{8}$	$\frac{5}{9}$	$\frac{5}{10}$	$\frac{5}{11}$	$\frac{5}{12}$
π_k^2	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{6}{12}$	$\frac{7}{12}$	$\frac{8}{12}$	$\frac{9}{12}$	$\frac{10}{12}$	$\frac{11}{12}$	1

- La taille du premier échantillon vaut approximativement $n + n \ln \frac{N}{n}$

Conclusions

- Un résultat général permet de multiples implémentations.
- La méthode du réservoir est généralisable à des blocs.
- Les méthodes sont généralisables aux plans équilibrés.
- On peut concevoir des stratégies complexes en plusieurs phases.

Bibliography I

Chao, M.-T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69 :653–656.

Chauvet, G. (2012). On a characterization of ordered pivotal sampling. *Bernoulli*, 18(4) :1099–1471.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics*, 21 :9–31.

Chen, X.-H. and Liu, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, 7 :875–892.

Cohen, E., Duffield, N., Kaplan, H., Lund, C., and Thorup, M. (2009). Stream sampling for variance-optimal estimation of subset sums. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1255–1264. Society for Industrial and Applied Mathematics.

Deville, J.-C. and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85 :89–101.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling : The cube method. *Biometrika*, 91 :893–912.

Fuller, W. A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, B32 :209–226.

Grafström, A., Lundström, N. L. P., and Schelin, L. (2012). Spatially balanced sampling through the pivotal method. *Biometrics*, 68(2) :514–520.

Bibliography II

Hodges Jr., J. L. and Le Cam, L. (1960). The Poisson approximation to the Poisson binomial distribution. *Annals of Mathematical Statistics*, 31 :737–740.

Madow, W. G. (1949). On the theory of systematic sampling, II. *Annals of Mathematical Statistics*, 20 :333–354.

Stein, C. (1990). Application of Newton's identities to a generalized birthday problem and to the Poisson-Binomial distribution. Technical Report TC 354, Department of Statistics, Stanford University.

Sugden, R. A., Smith, T. M. F., and Brown, R. P. (1996). Chao's list sequential scheme for unequal probability sampling. *Journal of Applied Statistics*, 23 :413–421.

Tillé, Y. (2018). Fast implementation of Fuller's unequal probability sampling method. Technical report, University of Neuchâtel.