Bootstrap pour un tirage à plusieurs degrés avec tirage à probabilités inégales des unités primaires

Guillaume Chauvet

École Nationale de la Statistique et de l'Analyse de l'Information

Journées de Méthodologie Statistique
Cité internationale universitaire de Paris
31/03/2015

1. Le Bootstrap
2. Tirage à plusieurs degrés
3. Etude par simulations
4. Perspectives

Le Bootstrap

Principe du Bootstrap

Soit $(X_1, \ldots, X_n)$ un échantillon $\sim_{i.i.d.} \mathcal{L}(\mu, \sigma^2)$.
On estime $\theta = f(\mu)$ par $\hat{\theta} = f(\bar{X}_n)$ avec $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$.

On tire un rééchantillon $(X_1^*, \ldots, X_m^*)$ de façon i.i.d dans $\{X_1, \ldots, X_n\}$.
On répète le rééchantillonnage B fois, et on estime $V(\hat{\theta})$ par

$$v_{BOOT}(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{*b} - \frac{1}{B} \sum_{c=1}^{B} \hat{\theta}^{*c} \right)^2 .$$

Idée naïve pour un échantillon S tiré selon un plan de sondage $p(\cdot)$: on obtient une estimation de variance en tirant des rééchantillons avec remise dans les valeurs échantillonnées $\{y_1, \ldots, y_n\}$.

ENSAI École nationale de la statistique et de l'analyse de l'information

Principe du Bootstrap

Soit $(X_1, \ldots, X_n)$ un échantillon $\sim_{i.i.d.} \mathcal{L}(\mu, \sigma^2)$.
On estime $\theta = f(\mu)$ par $\hat{\theta} = f(\bar{X}_n)$ avec $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$.

On tire un rééchantillon $(X_1^*, \ldots, X_m^*)$ de façon i.i.d dans $\{X_1, \ldots, X_n\}$.
On répète le rééchantillonnage B fois, et on estime $V(\hat{\theta})$ par

$$v_{BOOT}(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{*b} - \frac{1}{B} \sum_{c=1}^{B} \hat{\theta}^{*c} \right)^2 .$$

Idée naïve pour un échantillon S tiré selon un plan de sondage $p(\cdot)$: on obtient une estimation de variance en tirant des rééchantillons avec remise dans les valeurs échantillonnées $\{y_1, \ldots, y_n\}$.

Cas d'une enquête

Dans le cas d'un sondage dans U de taille N, l'alea est porté par le vecteur $(I_1, \ldots, I_N)^\top$ des indicatrices de sélection :

- SRS sans remise : tirages non indépendants, identiquement distribués,
- Sondage stratifié : tirages indépendants, non identiquement distribués,
- Sondage à probabilités inégales : tirages souvent non indépendants, non identiquement distribués,

Il existe une vaste littérature pour adapter les méthodes de Bootstrap en Sondages (Shao and Tu, 1995 ; Davison and Sardy, 2007).
L'objectif est ici d'étudier une méthode simple de Bootstrap, via :

$$(I_1, \ldots, I_N) \quad \Rightarrow \quad (X_1, \ldots, X_{n_I}) \text{ approximativement i.i.d.}$$

ENSAI École nationale de la statistique et de l'analyse de l'information

Tirage à plusieurs degrés

Principe du tirage à plusieurs degrés

La population U est partitionnée en N_I grosses unités appelées **Unités Primaires** (UP). Les individus de U sont les **Unités Secondaires** (US) :

- Premier degré : un échantillon S_I d'UP est sélectionné.
- Second degré : un échantillon d'US est tiré dans chaque UP $\in S_I$.

Exemples d'enquêtes issues d'un plan à plusieurs degrés :

1. Enquête Panel Politique de la Ville (PPV) : sélection d'un échantillon de quartiers (UP), puis de logements (US), puis d'individus (UT) (Dieusaert et Henry, 2012).
2. Enquête épidémiologique : estimation de la contamination au plomb en tirant un échantillon d'hôpitaux (UP), puis d'enfants (US) dont les logements sont inspectés (Lucas, 2013).
3. Enquête PISA : sélection d'un échantillon de collèges (UP), puis d'un échantillon d'élèves de 15 ans (US).

ENSAI École nationale de la statistique et de l'analyse de l'information

Tirage multinomial des UP

L'échantillon S_I^{WR} est obtenu à partir de n_I tirages indépendants dans U_I.
Soit π_{Ii} le nombre moyen de sélections pour u_i.
L'estimateur de Hansen-Hurvitz du total Y vaut

$$\hat{Y}_{WR} = \sum_{j=1}^{n_I} \frac{\hat{Y}_{i(j)}}{\pi_{Ii(j)}} \equiv \frac{1}{n_I} \sum_{j=1}^{n_I} X_j. \quad (1)$$

On estime $\theta = f(Y)$ par $\hat{\theta}_{WR} = f(\hat{Y}_{WR})$.

L'échantillon $(X_1, \ldots, X_{n_I})$ est i.i.d. (Särndal et al., 1992).
On peut appliquer la méthode du bootstrap des UP (Rao et Wu, 1988) en rééchantillonnant de façon i.i.d dans $\{X_1, \ldots, X_{n_I}\}$.
On estime $V(\hat{\theta})$ par $v_{BOOT}(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{*b} - \frac{1}{B} \sum_{c=1}^{B} \hat{\theta}^{*c} \right)^2$. On capte la variance associée à l'ensemble des degrés de tirage.

Tirage sans remise des UP

L'échantillon S_I^R est obtenu par tirage sans remise dans U_I, avec π_{Ii} la probabilité d'inclusion de u_i dans S_I.
L'estimateur de Horvitz-Thompson du total Y vaut

$$\hat{Y}_R = \sum_{u_i \in S_I^R} \frac{\hat{Y}_i}{\pi_{Ii}} \quad \text{avec} \quad \hat{Y}_i \text{ un estimateur SB de } Y_i. \tag{2}$$

Idée : si la méthode de tirage dans U_I est proche d'un tirage multinomial, la méthode du Bootstrap des UP devrait être encore valide pour estimer la variance de $\hat{\theta}_R = f(\hat{Y}_R)$.

Avantage : méthode très simple. On a seulement besoin de bootstrapper les estimateurs $\hat{Y}_i$ par UP. Pas besoin d'estimateurs de variance dans les UP.

ENSAI École nationale de la statistique et de l'analyse de l'information

Principal résultat

Sous des hypothèses raisonnables

+ $n_I \to \infty$, $\frac{n_I}{\sqrt{N_I}} \to 0$ et $m \longrightarrow \infty$,
+ S_I^R est sélectionné par tirage réjectif (Hajek, 1964),
+ f homogène de degré α, différentiable sur $\mathbb{R}^q$ avec des dérivées partielles bornées et $f'(N^{-1}Y) \neq 0$,

alors il existe un couplage entre S_I^R et S_I^{WR} tel que :

$$\frac{V(\hat{\theta}_R)}{V(\hat{\theta}_{WR})} \longrightarrow 1, \tag{3}$$

$$\frac{V^*(\hat{\theta}_R^*)}{V^*(\hat{\theta}_{WR}^*)} \longrightarrow_{Pr} 1. \tag{4}$$

ENSAI École nationale de la statistique et de l'analyse de l'information

Etude par simulations

Population simulée

Nous générons une population contenant $N_I = 2,000$ UP, de façon à ce que le CV des tailles d'UP soit égal à 0.06. Dans chaque population, nous générons pour chaque US $k \in u_i$:

$$y_{1k} = \lambda + \sigma\ v_i + \{\rho^{-1}(1-\rho)\}^{0.5}\sigma\ (\alpha\ \epsilon_k + \eta_k), \quad (5)$$

$$y_{2k} = \lambda + \sigma\ v_i + \{\rho^{-1}(1-\rho)\}^{0.5}\sigma\ (\alpha\ \epsilon_k + \nu_k), \quad (6)$$

où $v_i, \epsilon_k, \eta_k \sim \mathcal{N}(0,1)$. Les paramètres λ, σ, ρ et α sont choisis de façon à obtenir

- un coefficient de corrélation intra-grappes approximativement égal à 0.1, 0.2 ou 0.3,
- un coefficient de corrélation entre y_1 et y_2 approximativement égal à 0.60.

ENSAI École nationale de la statistique et de l'analyse de l'information

Plan de sondage et paramètres d'intérêt

Dans chaque population, nous sélectionnons $B = 1,000$ échantillons selon un plan à deux degrés autopondéré, avec :

- un tirage réjectif au premier degré avec $0.01 \leq f_I \leq 0.25$,
- un tirage systématique de taille $n_0 = 5$ ou 20 au second degré.

On estime la variance des estimateurs par substitution des paramètres

$$
\begin{aligned}
R &= \frac{\mu_{y1}}{\mu_{y2}} \\
r &= \frac{\sum_{k\in U}(y_{1k} - \mu_{y1})(y_{2k} - \mu_{y2})}{\sqrt{\sum_{k\in U}(y_{1k} - \mu_{y1})^2}\sqrt{\sum_{k\in U}(y_{2k} - \mu_{y2})^2}},
\end{aligned}
$$

en utilisant le tirage avec remise des UP. La vraie variance est approximée en utilisant en utilisant un tirage indépendant de $C = 20,000$ échantillons.

ENSAI École nationale de la statistique et de l'analyse de l'information

Echantillon de taille $n_0 = 5$ au second degré

		$\rho = 0.1$		$\rho = 0.2$		$\rho = 0.3$	
		RB	L+U	RB	L+U	RB	L+U
Total	$f_I = 1\%$	0.02	93.1	0.02	92.9	0.01	93.8
	$f_I = 2.5\%$	-0.01	94.2	0.01	93.7	0.01	94.7
	$f_I = 5\%$	0.01	94.7	0.02	95.7	0.02	95.7
	$f_I = 10\%$	0.04	94.8	0.04	94.4	0.04	95.5
	$f_I = 25\%$	0.08	96.0	0.13	95.4	0.17	96.0
Ratio	$f_I = 1\%$	0.03	93.5	0.00	93.5	0.02	93.2
	$f_I = 2.5\%$	-0.01	93.2	0.01	94.5	0.02	94.2
	$f_I = 5\%$	0.02	95.6	0.00	93.9	0.00	95.8
	$f_I = 10\%$	0.02	94.8	0.01	95.2	0.00	95.3
	$f_I = 25\%$	0.02	95.7	0.04	94.6	0.02	94.4
Coef. of correlation	$f_I = 1\%$	-0.03	92.8	0.00	92.0	0.01	93.4
	$f_I = 2.5\%$	0.01	93.4	0.02	93.6	0.04	92.5
	$f_I = 5\%$	0.01	93.7	0.01	93.7	-0.03	94.1
	$f_I = 10\%$	0.01	95.9	0.01	94.9	0.03	94.7
	$f_I = 25\%$	0.03	95.7	0.06	95.5	0.04	95.3

Echantillon de taille $n_0 = 20$ au second degré

		$\rho = 0.1$		$\rho = 0.2$		$\rho = 0.3$	
		RB	L+U	RB	L+U	RB	L+U
Total	$f_I = 1\%$	-0.01	94.3	0.00	93.9	0.01	94.4
	$f_I = 2.5\%$	0.02	94.5	0.04	95.7	0.04	94.1
	$f_I = 5\%$	0.03	95.3	0.06	95.4	0.06	96.4
	$f_I = 10\%$	0.10	95.8	0.10	95.3	0.13	95.1
	$f_I = 25\%$	0.23	96.7	0.25	97.0	0.29	96.7
Ratio	$f_I = 1\%$	-0.01	93.0	-0.01	93.6	0.01	93.9
	$f_I = 2.5\%$	0.03	93.8	0.02	94.8	0.01	93.6
	$f_I = 5\%$	0.01	94.6	0.03	94.4	0.02	95.8
	$f_I = 10\%$	0.04	95.5	0.03	95.1	0.04	94.6
	$f_I = 25\%$	0.14	95.8	0.15	96.9	0.13	96.0
Coef. of correlation	$f_I = 1\%$	0.00	95.1	0.00	93.7	-0.02	93.3
	$f_I = 2.5\%$	0.02	94.3	0.02	94.7	0.01	94.2
	$f_I = 5\%$	0.01	95.1	0.04	93.2	0.03	94.8
	$f_I = 10\%$	0.06	94.1	0.07	95.5	0.05	95.8
	$f_I = 25\%$	0.14	97.3	0.15	96.3	0.19	96.6

Perspectives

Hypothèses principales :

+ $n_I \to \infty$, $\frac{n_I}{\sqrt{N_I}} \to 0$ et $m \longrightarrow \infty$,

Condition forte sur n_I : grande et fortement négligeable devant N_I. A affaiblir en $\frac{n_I}{N_I} \to 0$ comme dans le cas de probabilités égales (Chauvet, 2014).

Perspectives

Hypothèses principales :

+ S_I^R est sélectionné par tirage réjectif (Hajek, 1964),

Plans proches du plan réjectif ? Méthode de Hanurav-Vijayan (programmée par défaut dans SAS) ?

Perspectives

Hypothèses principales :

+ f homogène de degré α, différentiable sur $\mathbb{R}^q$ avec des dérivées partielles bornées et $f'(N^{-1}Y) \neq 0$,

Restreindre à un voisinage de μ_y, mais nécessaire de bien contrôler $\hat{Y}_R - Y$.

Références

- Chauvet, G. (2014). Coupling methods for multistage sampling. Soumis.
- Davison, A.C. and Sardy, S. (2007). Resampling variance estimation in Surveys with missing data. Journal of Official Statistics, 23, 371-386.
- Dieusaert, P. and Henry, M. (2014). L'enquête Panel Politique de la Ville. 8ème Colloque francophone sur les Sondages, Dijon.
- Hajek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics, 35, 1491-1523.
- Lucas, J-P. (2013). Contamination des logements par le plomb : prévalence des logements à risque et identification des déterminants de la contamination. PhD dissertation, Université de Nantes.
- Rao, J.N.K and Wu, C.F.J. (1988). Resampling inference with complex survey data. Journal of the american statistical association, 83, 231-241.
- Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). Model Assisted Survey Sampling. New-York, Springer-Verlag.
- Shao, J. and Tu, D. (1995). The Jackknife and the Bootstrap. New-York, Springer.