

Méthodes d'imputation aléatoires équilibrées

Guillaume Chauvet (Crest, Ensaï)

Travail joint avec J-C. Deville (Ensaï)
et D. Haziza (Univ. de Montréal)

Méthodes Avancées pour l'Analyse de Sondages Complexes
Université de Bourgogne, 09/11/2010

Résumé

Lors d'une enquête, les échantillons sont sélectionnés de façon équilibrée afin de réduire la variance des estimateurs. La méthode du Cube est un algorithme de tirage général permettant de sélectionner des échantillons équilibrés.

Une enquête est généralement entachée de non-réponse, qui diminue la taille de l'échantillon observé. L'imputation aléatoire permet de traiter la non-réponse partielle en préservant la distribution de la variable imputée, mais au prix d'une variabilité supplémentaire.

L'imputation aléatoire équilibrée permet de préserver la distribution de la variable imputée tout en limitant la variance d'imputation.

Plan de l'exposé

Méthodes d'échantillonnage

Introduction

La méthode du Cube

Traitement de la non-réponse partielle

Introduction

Imputation équilibrée

Etude par simulations

Méthodes d'échantillonnage

Notation

On considère une population finie d'individus

$$U = \{1, \dots, k, \dots, N\},$$

où chaque individu est supposé identifiable par son label k . On note y_k la valeur prise par une variable d'intérêt y sur l'individu k de U .

Un échantillon aléatoire S est sélectionné dans U au moyen d'un plan de sondage $p(\cdot)$. Les probabilités d'inclusion $\pi_k = \mathbb{P}(k \in S)$ sont supposées connues et non nulles.

Du point de vue de l'échantillonnage, la variable y est fixée et non aléatoire. L'alea provient de la sélection de S .

Paramètre

On s'intéresse à l'estimation du total

$$t_y = \sum_{k \in U} y_k,$$

que l'on peut estimer sans biais sous le plan de sondage par

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} d_k y_k.$$

On a donc :

$$E_p [\hat{t}_{y\pi}] = t_y.$$

Choix du plan de sondage

Le plan de sondage est choisi de façon à minimiser la variance des estimateurs, tout en respectant des contraintes de coût.

- ▶ stratification, tirage à probabilités inégales
⇒ réduction de la variance
- ▶ tirage multidegrés
⇒ réduction des coûts

La précision du plan repose sur des propriétés d'*équilibrage* : l'échantillon est sélectionné de façon à respecter une information connue.

Exemples :

- ▶ respect de structures âge-sexe (méthode des quotas),
- ▶ répartition par effectif salarié (stratification).

Exemples : propriété d'équilibrage

Le SAS sans remise permet d'estimer exactement la taille de U :

$$d_k = \frac{N}{n} \quad \Rightarrow \quad \sum_{k \in S} d_k = N$$
$$\hat{t}_{\mathbf{x}\pi} = t_{\mathbf{x}} \quad \text{avec} \quad \mathbf{x}_k = 1.$$

Le SAS stratifié permet d'estimer exactement les tailles de strates.
Pour une strate U_h , et $k \in U_h$:

$$d_k = \frac{N_h}{n_h} \quad \Rightarrow \quad \sum_{k \in S_h} d_k = N_h$$
$$\hat{t}_{\mathbf{x}\pi} = t_{\mathbf{x}} \quad \text{avec} \quad \mathbf{x}_k = [1 [k \in U_1], \dots, 1 [k \in U_H]]'.$$

Exemples : effet sur la variance

Pour le SAS sans remise :

$$V_p [\hat{t}_{y\pi}] = N^2 \frac{1-f}{n} S_e^2 \quad \text{avec} \quad e_k = y_k - \mathbf{x}'_k \alpha$$
$$\mathbf{x}_k = 1$$
$$\alpha = \mu_y$$

Pour le SAS stratifié à allocation proportionnelle :

$$V_p [\hat{t}_{y\pi}] \simeq N^2 \frac{1-f}{n} S_e^2 \quad \text{avec} \quad e_k = y_k - \mathbf{x}'_k \beta$$
$$\mathbf{x}_k = [1 [k \in U_1], \dots, 1 [k \in U_H]]'$$
$$\beta = [\mu_{y1}, \dots, \mu_{yH}]'$$

Echantillonnage équilibré

Un échantillon est dit *équilibré* sur un jeu de variables auxiliaires \mathbf{x} si

$$\hat{t}_{\mathbf{x}\pi} = t_{\mathbf{x}}.$$

Le total $t_{\mathbf{x}}$ est donc parfaitement estimé (variance nulle).

Par extension, un plan de sondage est dit équilibré sur les variables \mathbf{x} si seuls les échantillons équilibrés sur \mathbf{x} ont une probabilité non nulle d'être sélectionnés.

Question : existe t-il une méthode générale permettant de sélectionner des échantillons équilibrés, pour un vecteur \mathbf{x} de variables auxiliaires quelconque, et pour un jeu quelconque de probabilités d'inclusion ?

La méthode du Cube

Représentation du Cube

Deville et Tillé (2004) proposent un algorithme général pour la sélection d'échantillons équilibrés.

L'algorithme est basé sur une représentation géométrique du plan de sondage : un échantillon s est vu comme un sommet

$$(s_1, \dots, s_N) \in \{0, 1\}^N$$

du N -cube $C = [0, 1]^N$.

L'algorithme consiste à partir du vecteur π des probabilités d'inclusion pour aboutir sur un des sommets du Cube, à l'aide d'une marche aléatoire.

Etape de base : la phase de vol

Le résultat de l'échantillonnage est donné par le vecteur :

$$\pi(T) = \pi + \sum_{t=1}^T \delta(t),$$

$$\delta(t) = \begin{cases} +\lambda_1(t) u(t) & \text{avec la probabilité } \lambda_2(t) / [\lambda_1(t) + \lambda_2(t)] \\ -\lambda_2(t) u(t) & \text{avec la probabilité } \lambda_1(t) / [\lambda_1(t) + \lambda_2(t)] \end{cases}$$

Les paramètres utilisés sont :

- ▶ $\lambda_1(t), \lambda_2(t) > 0$
→ choisis afin qu'au moins une unité soit sélectionnée ou définitivement rejetée.
- ▶ $u(t) \in \text{Ker}((\mathbf{x}_k / \pi_k)_{k \in U})$
→ pour que les équations d'équilibrage soient respectées.
- ▶ le choix aléatoire assure que les probabilités d'inclusion sont respectées.

Etape de fin : la phase d'atterrissage

L'algorithme précédent s'arrête quand il n'est plus possible de trouver un vecteur $u(t)$ respectant les contraintes précédentes : c'est la fin de la *phase de vol*.

La *phase d'atterrissage* permet de terminer l'échantillonnage pour les unités restantes (au plus $dim(\mathbf{x})$). L'impact sur la variance peut généralement être négligé si le nombre de variables d'équilibrage est faible.

Le vecteur $\pi(T)$ obtenu à la dernière étape de l'algorithme donne le résultat de l'échantillonnage (vol + atterrissage).

Traitement de la non-réponse partielle

Le problème

En situation de *réponse totale*, les variables d'intérêt sont relevées sur tous les individus de l'échantillon.

En pratique, les valeurs prises par les variables d'intérêt ne sont observées que sur une partie des individus de l'échantillon. On se trouve alors en situation de *non-réponse*.

Le fichier de données d'enquête peut être schématiquement découpé en 3 parties :

- ▶ individus qui ont répondu à toutes les questions de l'enquête :
les répondants,
- ▶ individus qui ont répondu à une partie des questions :
les répondants partiels,
- ▶ individus qui n'ont répondu à aucune question :
les non-répondants.

Non-réponse partielle

Nous nous intéressons dans la suite au cas de la non-réponse partielle. Elle est généralement traitée par *imputation* : une valeur manquante y_k est remplacée par une valeur plausible y_k^* .

L'estimateur du total devient :

$$\hat{t}_{yI} = \sum_{k \in S_r} d_k y_k + \sum_{k \in S_{nr}} d_k y_k^*$$

Par rapport à la situation de réponse totale, deux mécanismes aléatoires supplémentaires interviennent :

- ▶ le *mécanisme de non-réponse*, qui conduit à l'échantillon de répondants effectivement observé pour la variable y ,
- ▶ le *mécanisme d'imputation* utilisé pour remplacer les valeurs manquantes de y .

Modèle d'imputation

Le mécanisme d'imputation est généralement motivé par un *modèle d'imputation* (par exemple, un modèle de régression) qui vise à prédire la variable y_k à l'aide d'une information auxiliaire \mathbf{z}_k disponible sur l'ensemble de l'échantillon.

$$m : y_k = \mathbf{z}'_k \beta + \epsilon_k,$$

$$E_m(\epsilon_k) = 0 \quad V_m(\epsilon_k) = \sigma^2 v_k \quad Cov_m(\epsilon_k, \epsilon_l) = 0$$
$$\equiv \sigma_k^2$$

Imputation déterministe

L'imputation par la régression déterministe est obtenue en prenant $y_k^* = \mathbf{z}'_k \hat{\beta}_r$, avec

$$\hat{\beta}_r = \left(\sum_{k \in S_r} \omega_k v_k^{-1} \mathbf{z}_k \mathbf{z}'_k \right)^{-1} \sum_{k \in S_r} \omega_k v_k^{-1} \mathbf{z}_k y_k$$

un estimateur du paramètre β inconnu, et ω_k désigne un poids d'imputation associé à l'unité k (Haziza, 2008).

Cas particuliers : imputation par la moyenne (par classe), imputation par le ratio.

Dans ce cas, l'estimateur imputé est égal à

$$\hat{t}_{yI} = \sum_{k \in S_r} d_k y_k + \sum_{k \in S_{nr}} d_k \left[\mathbf{z}'_k \hat{\beta}_r \right].$$

Imputation déterministe (2)

L'avantage de l'imputation déterministe est qu'elle n'occasionne pas de variabilité supplémentaire : les valeurs imputées

$$y_k^* = \mathbf{z}'_k \hat{\beta}_r$$

sont fixes, conditionnellement à l'échantillon S et à l'échantillon de répondants S_r .

Le principal inconvénient de l'imputation déterministe est qu'elle ne préserve pas la distribution de la variable imputée
 \Rightarrow inconsistante dans le cas d'un quantile.

Imputation aléatoire

L'imputation par la régression aléatoire est obtenue en prenant

$$y_k^* = \mathbf{z}'_k \hat{\beta}_r + \hat{\sigma}_k \epsilon_k^*,$$

i.e. en rajoutant à la prédiction de y_k un terme aléatoire.

Les résidus ϵ_k^* peuvent être générés selon une distribution paramétrique, ou tirés (généralement avec remise) dans les résidus standardisés calculés sur les répondants.

Cas particuliers : imputation par hot-deck (par classe), par le ratio aléatoire.

Dans ce cas, l'estimateur imputé est égal à

$$\hat{t}_{yI} = \sum_{k \in S_r} d_k y_k + \sum_{k \in S_{nr}} d_k \left[\mathbf{z}'_k \hat{\beta}_r \right] + \sum_{k \in S_{nr}} d_k \left[\hat{\sigma}_k \epsilon_k^* \right].$$

Imputation aléatoire (2)

L'inconvénient de l'imputation aléatoire est qu'elle occasionne une variabilité supplémentaire appelée *variance d'imputation* : les valeurs imputées

$$y_k^* = \mathbf{z}'_k \hat{\beta}_r + \hat{\sigma}_k \epsilon_k^*$$

changent si le mécanisme d'imputation est répété.

Le principal avantage de l'imputation déterministe est qu'elle préserve la distribution de la variable imputée

⇒ consistante dans le cas d'un quantile.

Imputation équilibrée

Principe

Dans le cas d'une imputation aléatoire, la variance d'imputation est du même ordre de grandeur que la variance d'échantillonnage et la variance due à la non-réponse.

La variance d'imputation est annulée si l'équation

$$\sum_{k \in S_{nr}} d_k [\hat{\sigma}_k \epsilon_k^*] = 0 \quad (1)$$

est vérifiée.

Cette contrainte définit une équation d'équilibrage : nous proposons donc de sélectionner les résidus ϵ_k^* à l'aide de la méthode du Cube afin que (1) soit (approximativement) vérifiée.

Consistance de la fonction de répartition estimée

On note $F_N(t) = N^{-1} \sum_{k \in U} 1(y_k \leq t)$ la fonction de répartition de la variable y , et

$$F_I(t) = N^{-1} \left[\sum_{k \in S_r} d_k 1(y_k \leq t) + \sum_{k \in S_{nr}} d_k 1(y_k^* \leq t) \right]$$

son estimateur imputé.

Théorème (CDH, 2010)

Si les résidus ϵ_k^ sont sélectionnés indépendamment et avec remise, alors $\hat{F}_I(t) - F_N(t) \rightarrow_{\mathbb{P}} 0$.*

Théorème (CDH, 2010)

Si les résidus ϵ_k^ sont sélectionnés de façon équilibrée et avec remise, alors $\hat{F}_I(t) - F_N(t) \rightarrow_{\mathbb{P}} 0$.*

Etude par simulations

Cadre

Deux populations de taille 10 000, générées selon le modèle

$$y_i = \beta z_i + \sqrt{z_i} \eta_i,$$

où les z_i sont générés selon une loi gamma et les η_i selon une loi normale. On utilise $R^2 = 0.36$ (pop. 1) et $R^2 = 0.64$ (pop. 2).

Echantillon S de taille $n = 500$ sélectionné par tirage réjectif (probabilités proportionnelles à z_i). La non-réponse est générée selon un mécanisme bernoullien ($p_0 = 0.5$ et $p_0 = 0.75$).

On réalise $R = 1,000$ simulations. On s'intéresse à l'estimation :

- (i) de la moyenne μ_y ,
- (ii) de $F_N(t_\alpha)$, avec $\alpha = 0.25, 0.50$.

Méthode d'imputation

Trois méthodes d'imputation sont comparées :

1. Imputation par le ratio déterministe (DRI),
2. Imputation par le ratio aléatoire (RRI),
3. Imputation par le ratio aléatoire équilibré (BRI).

Pour chacune des trois méthodes, on calcule

- (i) le biais relatif (RB),
- (ii) l'erreur quadratique moyenne (MSE),
- (iii) l'efficacité relative (RE) définie par

$$RE = \frac{MSE\left(\hat{\theta}_I^{(\cdot)}\right)}{MSE\left(\hat{\theta}_I^{(RRI)}\right)}.$$

Table: Biais relatif (%) de l'estimateur imputé de la moyenne

		DRI	RRI	BRI	DRI	RRI	BRI
		$p_0 = 0.5$			$p_0 = 0.75$		
Population 1	RB	-0.01	0.43	0.10	0.40	0.54	0.52
Population 2	RB	0.51	0.72	0.73	0.73	0.80	0.87

DRI, imputation par le ratio déterministe ;

RRI, imputation par le ratio aléatoire ;

BRI, imputation par le ratio aléatoire équilibré.

Table: Efficacité relative de l'estimateur imputé de la moyenne

		DRI	RRI	BRI	DRI	RRI	BRI
		$p_0 = 0.5$			$p_0 = 0.75$		
Population 1	RE	0.78	1	0.82	0.85	1	0.87
Population 2	RE	0.81	1	0.85	0.90	1	0.93

DRI, imputation par le ratio déterministe ;

RRI, imputation par le ratio aléatoire ;

BRI, imputation par le ratio aléatoire équilibré.

Table: Biais relatif (%) de l'estimateur imputé $\hat{F}_I(t_\alpha)$

			DRI	RRI	BRI	DRI	RRI	BRI
α			$p_0 = 0.5$			$p_0 = 0.75$		
Pop. 1	0.25	RB	-50.13	0.00	-2.71	-24.95	0.13	-0.95
	0.50	RB	-3.41	0.21	0.09	-1.49	0.10	0.11
Pop. 2	0.25	RB	-44.57	-0.46	-2.47	-22.27	-0.43	-1.33
	0.50	RB	-2.87	0.23	0.04	-1.82	-0.21	-0.26

DRI, imputation par le ratio déterministe ;

RRI, imputation par le ratio aléatoire ;

BRI, imputation par le ratio aléatoire équilibré.

Table: Efficacité relative de l'estimateur imputé $\hat{F}_I(t_\alpha)$

			DRI	RRI	BRI	DRI	RRI	BRI
α			$p_0 = 0.5$			$p_0 = 0.75$		
Pop.1	0.25	RE	18.41	1	1.00	8.25	1	0.98
	0.50	RE	2.51	1	0.96	1.51	1	0.98
Pop.2	0.25	RE	16.92	1	0.92	6.79	1	1.00
	0.50	RE	1.51	1	0.89	1.26	1	0.95

DRI, imputation par le ratio déterministe ; RRI, imputation par le ratio aléatoire ; BRI, imputation par le ratio aléatoire équilibré

Applications et extensions

Application dans le cas de l'Enquête Patrimoine 2009

⇒ Chaput et al. (2010)

Imputation jointe équilibrée pour préserver le coefficient de corrélation entre deux variables

⇒ Chauvet et Haziza (2010)

Imputation équilibrée dans le cas d'un mélange de variables

⇒ Haziza et al. (2010)

Bibliographie

Chaput, H., Chauvet, G., Haziza, D., Solard, J., Salembier, L. (2010). *Joint Imputation to preserve the relationships between categorical variables*, en cours.

Chauvet, G., Deville, J.-C., and Haziza, D. (2010). *On balanced random imputation in surveys*, en révision dans *Biometrika*.

Chauvet, G., and Haziza, D. (2010). *Fully efficient estimation of coefficients of correlation in the presence of imputed data*, en cours.

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

Haziza, D. (2009). *Imputation and inference in the presence of missing data*, *Handbook of Statistics*, vol.29, chap. 10.

Haziza, D., Nambeu, C-O., Chauvet, G. (2010). *Single imputation for populations containing a large amount of zeroes*, soumis.