Consistance sous un modèle de réponse de la fonction de répartition estimée en présence de données manquantes

Hélène Boistard (Université Toulouse I)
Guillaume Chauvet (Crest, Ensai)
David Haziza (Univ. de Montréal)

Journées de Méthodologie Statistique
Paris, 25/01/2012

Plan de l'exposé

Estimation d'un total

Estimation de la fonction de répartition

Etude par simulations

Estimation d'un total

Notation

On considère une population finie d'individus

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

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

Un échantillon 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 connues et non nulles. Soit $d_k = 1/\pi_k$ le poids de sondage de l'unité k.

Du point de vue de l'échantillonnage, les variables d'intérêt sont fixées et non aléatoires. L'alea provient de la sélection de S.

Estimation d'un total

En situation de réponse complète à la variable y, le total $t_y = \sum_{k \in U} y_k$ peut être estimé sans biais sous le plan de sondage par

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

En situation de non-réponse pour la variable y, une valeur manquante y_k est remplacée par une valeur imputée y_k^* (e.g. Haziza, 2009) :

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

Deux mécanismes aléatoires supplémentaires interviennent :

- le *mécanisme de non-réponse*, qui conduit à l'échantillon S_r 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 $\mathbf{x}_k$ disponible sur l'ensemble de l'échantillon.

$$m \ : \ y_k \ = \ \mathbf{x}_k^\top \boldsymbol{\beta} + \sigma \sqrt{v_k} \epsilon_k.$$

Dans ce modèle :

- $\boldsymbol{\beta}$ et σ^2 sont des paramètres inconnus,
- v_k est une constante connue,
- les résidus ϵ_k sont des variables aléatoires iid, centrées réduites.

Imputation déterministe

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

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

un estimateur du paramètre $\boldsymbol{\beta}$ inconnu, et ω_k un poids d'imputation associé à l'unité k (Haziza, 2009).

Dans ce cas, l'estimateur imputé est égal à

$$\hat{t}_{yI} = \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k \left[\mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_r \right].$$

L'estimateur $\hat{t}_{yI}$ est approximativement mpq non biaisé (approche IM).

Imputation doublement robuste

Afin de se prémunir contre une mauvaise spécification du modèle d'imputation, il est intéressant de disposer d'un mécanisme d'imputation qui conduise à une estimation non biaisée sous un modèle de non-réponse (approche NM).

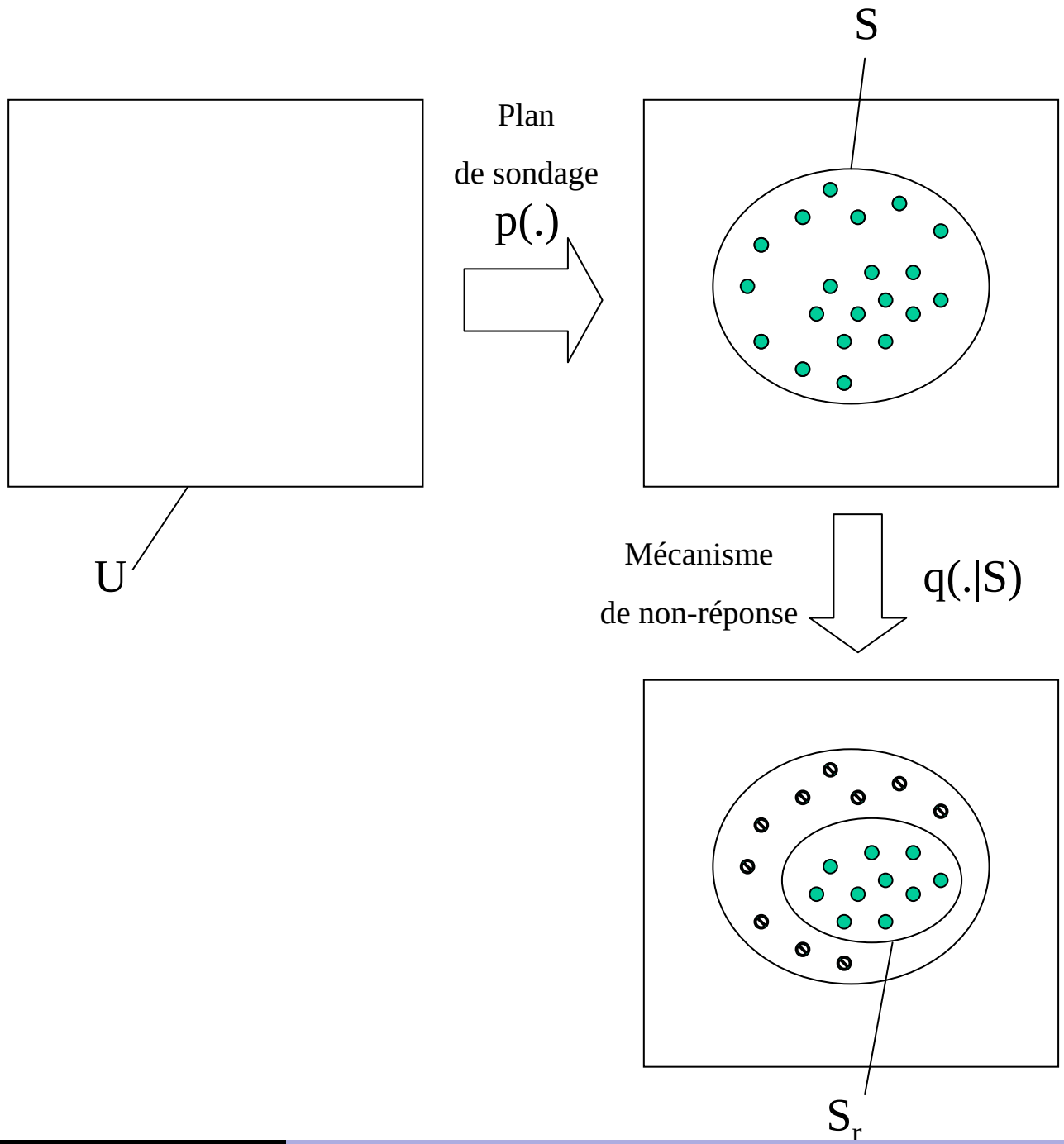
On appelle modèle de non-réponse un jeu d'hypothèses sur le mécanisme (inconnu) de non-réponse. On suppose ici que la probabilité de réponse à la variable y_k suit le modèle logistique

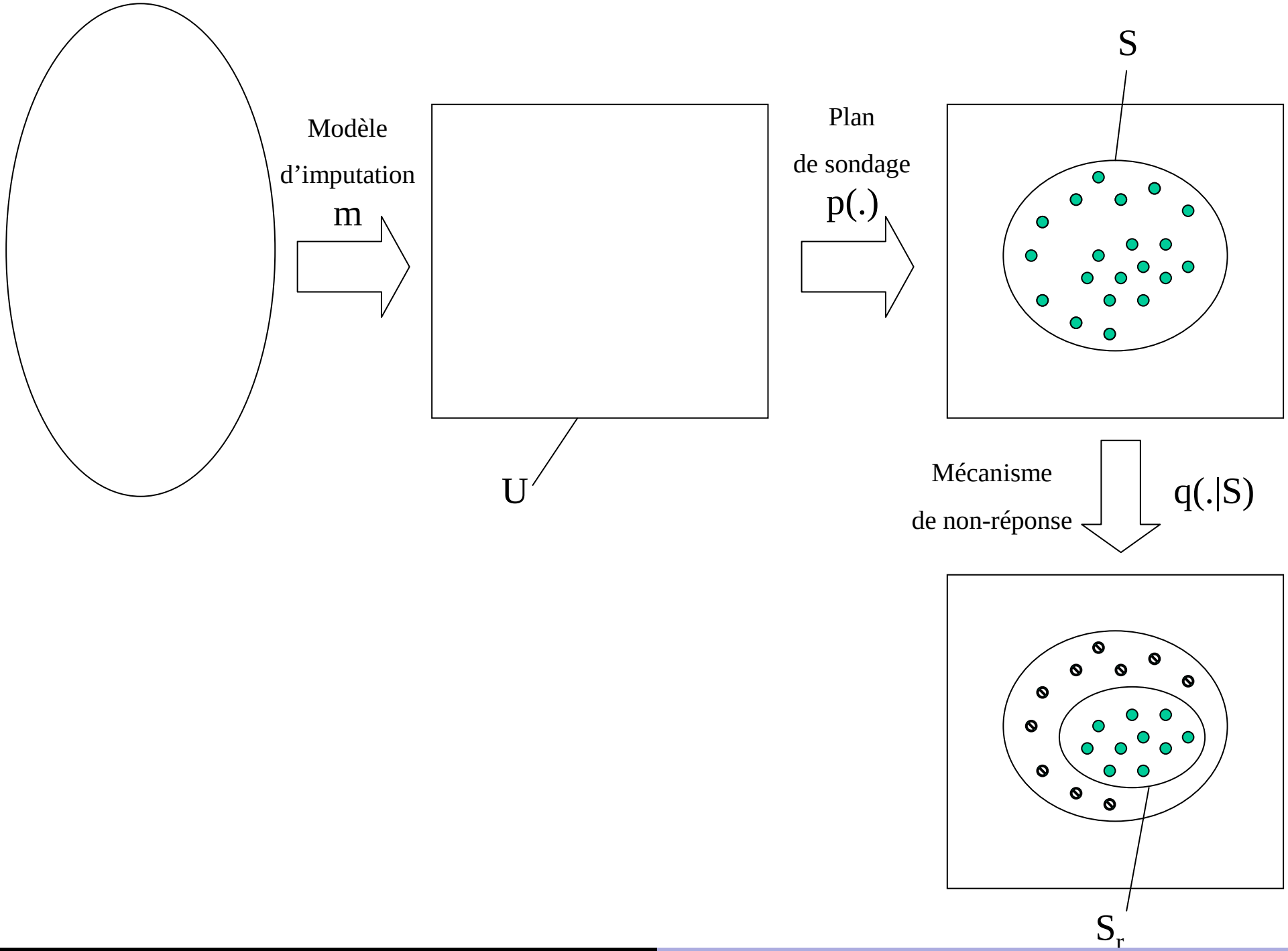
$$p_k \equiv \Pr\left(r_k = 1\right) \quad = \quad \frac{\exp\left(\boldsymbol{\phi}_0^\top \mathbf{x}_k\right)}{1 + \exp\left(\boldsymbol{\phi}_0^\top \mathbf{x}_k\right)}.$$

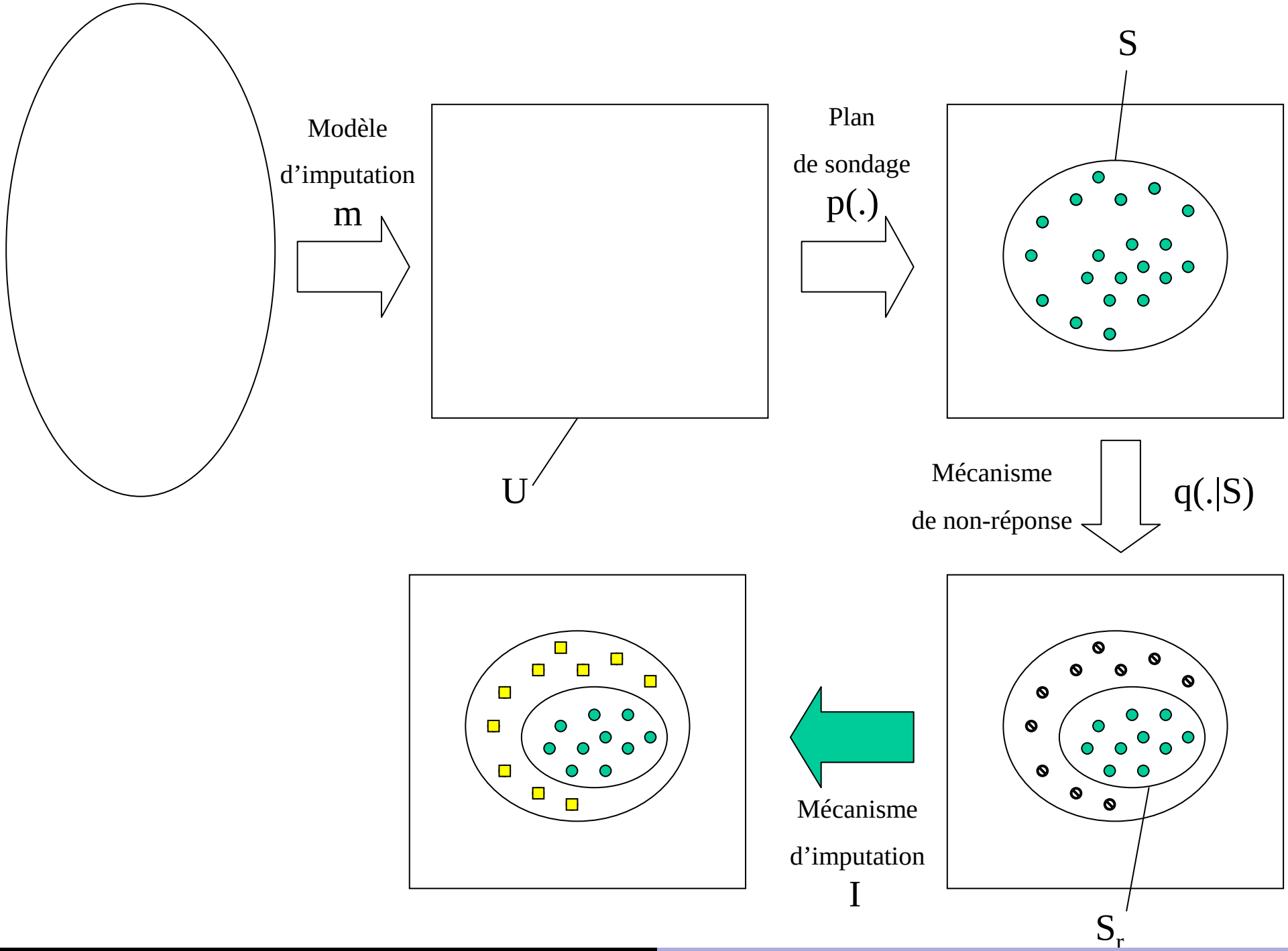
Imputation doublement robuste

L'utilisation du mécanisme d'imputation par la régression déterministe, avec les poids d'imputation $\omega_k = d_k \frac{1-p_k}{p_k}$, conduit à un estimateur imputé du total pq non biaisé (Haziza et Rao, 2006).

On parle d'estimation doublement robuste, voir également Kott (1994), Kim et Park (2006).


S
Plan
de sondage
p(.)
U
Mécanisme
de non-réponse
q(.|S)
S_r


Modèle
d'imputation
m
U
Plan
de sondage
p(.)
S
Mécanisme
de non-réponse
q(.|S)
S_r


Modèle
d'imputation
m
U
Plan
de sondage
p(.)
S
Mécanisme
de non-réponse
q(.|S)
Mécanisme
d'imputation
I
S_r

Estimation de la fonction de répartition

Définition

On s'intéresse à l'estimation de la fonction de répartition

$$F_N(t) = \frac{1}{N}\sum_{k\in U} 1(y_k \leq t).$$

Soit

$$\hat{F}_N(t) = \frac{1}{\hat{N}}\sum_{k\in S} d_k 1(y_k \leq t)$$

son estimateur en situation de réponse complète et

$$\hat{F}_I(t) = \frac{1}{\hat{N}}\left[\sum_{k\in S_r} d_k 1(y_k \leq t) + \sum_{k\in S_m} d_k 1(y_k^* \leq t)\right]$$

son estimateur imputé.

Imputation déterministe

L'utilisation d'une imputation par la régression déterministe conduit à l'estimateur

$$\hat{F}_I(t) = \frac{1}{\hat{N}} \left[\sum_{k \in S_r} d_k 1(y_k \leq t) + \sum_{k \in S_m} d_k 1(\mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_r \leq t) \right].$$

Cependant, cet estimateur est mpq biaisé. Pour résoudre ce problème, Chambers et Dunstan (1986) ont proposé un estimateur corrigé du biais a posteriori.

Une autre solution consiste à utiliser une méthode d'imputation aléatoire.

Imputation aléatoire

L'imputation par la régression aléatoire est obtenue en prenant

$$y_k^* = \mathbf{x}_k^\top \hat{\beta}_r + \hat{\sigma}\sqrt{v_k}\epsilon_k^*,$$

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

Les résidus ϵ_k^* sont tirés au hasard et avec remise, avec des probabilités proportionnelles aux poids d'imputation ω_k, parmi les résidus observés sur les répondants.

On montre que l'estimateur imputé $\hat{t}_{yI}$ est approximativement $mpqI$ non biaisé.

Imputation aléatoire

Sous de faibles hypothèses, Chauvet, Deville et Haziza (2011) montrent que dans le cas d'une imputation par la régression aléatoire, la fonction de répartition imputée est également consistante sous l'approche IM.

En d'autres termes :

$$\hat{F}_I(t) - F_N(t) \quad \rightarrow_{\mathbb{P}} \quad 0.$$

Ce résultat s'étend au cas où les résidus aléatoires sont sélectionnés de façon équilibrée (Deville et Tillé, 2004).

Approche NM

Là encore, pour se prémunir contre une mauvaise spécification du modèle d'imputation, il est intéressant de disposer d'un mécanisme d'imputation qui conduise à une estimation non biaisée sous l'approche NM.

Nous nous restreignons ici au cas du hot-deck aléatoire : une valeur manquante y_k est remplacée en sélectionnant au hasard et avec remise un donneur $y_j \in S_r$, avec des probabilités proportionnelles aux poids d'imputation ω_j.

Résultat obtenu

On suppose que les poids d'imputation sont donnés par $\omega_k = d_k \frac{1-p_k}{p_k}$.

Théorème (BCH, 2012)

Le hot-deck aléatoire donne une estimation consistante de la fonction de répartition sous l'approche NM, i.e.

$$\hat{F}_I(t) - F_N(t) \quad \to_{\mathbb{P}} \quad 0.$$

Etude par simulations

Cadre

Population de taille $N = 10\ 000$, générée selon le modèle

$$y_k \quad = \quad 10 + x_{1i} + x_{2i} + \eta_i,$$

où les x_{1i}, x_{2i} sont générés selon une loi gamma et les η_i selon une loi normale centrée. On utilise $R^2 = 0.70$.

Echantillon S de taille $n = 500$ sélectionné par sondage aléatoire simple. La non-réponse est générée selon un mécanisme poissonien, avec

$$Pr(r_i = 1 | x_{1i}, x_{2i}) \quad = \quad \frac{\exp\left(-1 + 1.6\ x_{1i} + 1.6\ x_{2i}\right)}{1 + \exp\left(-1 + 1.6\ x_{1i} + 1.6\ x_{2i}\right)}.$$

Probabilité de réponse moyenne de 0.60.

Imputation par la régression aléatoire

On réalise $B = 1000$ simulations. On s'intéresse à l'estimation de $F_N(t_\alpha)$, avec $\alpha = 0.05,\ 0.25,\ 0.50,\ 0.75,\ 0.95$.

Pour illustrer les risques d'une mauvaise spécification du modèle d'imputation, on examine les performances de l'imputation par la régression aléatoire non pondérée :

- avec le modèle correct : $\mathbf{x} = (1, x_1, x_2)$,
- avec un modèle incomplet : $\mathbf{x} = (1, x_1)$.

On calcule le biais relatif (RB), et le MSE relatif

$$\text{RMSE}\{\hat{F}_I(t)\} = \frac{\sqrt{\text{MSE}\{\hat{F}_I(t)\}}}{F_N(t)} \times 100.$$

Résultats obtenus

			α				
			0.05	0.25	0.50	0.75	0.95
$\mathbf{x} = (1, x1, x2)$	REGI	RB	0.9	0.8	0.1	-0.1	0.0
		RMSE	23.9	9.3	5.0	2.7	1.1
$\mathbf{x} = (1, x1)$	REGI	RB	-18.8	-12.9	-8.5	-4.6	-1.0
		RMSE	29.4	16.0	10.2	5.6	1.7

Table – Biais relatif et erreur quadratique moyenne de Monte Carlo (en pourcentage) pour la fonction de répartition imputée par la régression aléatoire

Résultats obtenus

			α				
			0.05	0.25	0.50	0.75	0.95
$\mathbf{x} = (1, x1, x2)$	REGI	RB	0.9	0.8	0.1	-0.1	0.0
		RMSE	23.9	9.3	5.0	2.7	1.1
$\mathbf{x} = (1, x1)$	REGI	RB	-18.8	-12.9	-8.5	-4.6	-1.0
		RMSE	29.4	16.0	10.2	5.6	1.7

Table – Biais relatif et erreur quadratique moyenne de Monte Carlo (en pourcentage) pour la fonction de répartition imputée par la régression aléatoire

Imputation par hot-deck

On raisonne maintenant sous l'approche NM. On modélise le mécanisme de non-réponse de façon :

- correcte : $\mathbf{x} = (1, x_1, x_2) \Rightarrow \hat{p}_1$,
- incomplète : $\mathbf{x} = (1, x_1) \Rightarrow \hat{p}_2$.

On examine les performances de l'imputation par hot-deck :

- non-pondéré : $\omega_k = 1$,
- pondéré avec les bonnes probabilités de réponse : $\omega_k = d_k \frac{1-\hat{p}_1}{\hat{p}_1}$,
- pondéré avec les mauvaises probabilités de réponse : $\omega_k = d_k \frac{1-\hat{p}_2}{\hat{p}_2}$.

Résultats obtenus

		α				
		0.05	0.25	0.50	0.75	0.95
RHDI	RB	-29.3	-22.7	-16.2	-9.6	-2.4
	RMSE	37.7	24.7	17.4	10.4	3.0
RHDI-P1	RB	0.6	0.2	0.1	-0.2	0.0
	RMSE	34.2	11.7	6.0	3.1	1.2
RHDI-P2	RB	-17.4	-13.1	-9.0	-5.0	-1.1
	RMSE	33.0	16.7	11.0	6.1	1.8

Table – Biais relatif et erreur quadratique moyenne de Monte Carlo (en pourcentage) pour trois méthodes de hot-deck aléatoire

Résultats obtenus

		α				
		0.05	0.25	0.50	0.75	0.95
RHDI	RB	-29.3	-22.7	-16.2	-9.6	-2.4
	RMSE	37.7	24.7	17.4	10.4	3.0
RHDI-P1	RB	0.6	0.2	0.1	-0.2	0.0
	RMSE	34.2	11.7	6.0	3.1	1.2
RHDI-P2	RB	-17.4	-13.1	-9.0	-5.0	-1.1
	RMSE	33.0	16.7	11.0	6.1	1.8

Table – Biais relatif et erreur quadratique moyenne de Monte Carlo (en pourcentage) pour trois méthodes de hot-deck aléatoire

Résultats obtenus

		α				
		0.05	0.25	0.50	0.75	0.95
RHDI	RB	-29.3	-22.7	-16.2	-9.6	-2.4
	RMSE	37.7	24.7	17.4	10.4	3.0
RHDI-P1	RB	0.6	0.2	0.1	-0.2	0.0
	RMSE	34.2	11.7	6.0	3.1	1.2
RHDI-P2	RB	-17.4	-13.1	-9.0	-5.0	-1.1
	RMSE	33.0	16.7	11.0	6.1	1.8

Table – Biais relatif et erreur quadratique moyenne de Monte Carlo (en pourcentage) pour trois méthodes de hot-deck aléatoire

Bibliographie

Chambers, R.L., Dunstan, R. (1986). *Estimating distribution functions from survey data*, Biometrika, vol 73, pp. 597–604.

Chauvet, G., Deville, J.-C., and Haziza, D. (2010). *On balanced random imputation in surveys*, Biometrika, vol 98, 459-471.

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.

Kim, J.K. and Park, H.A. (2006). *Imputation using response probability*. Canadian Journal Statistics 34, 171-182.

Kott, P.S. (1994). *A note on handling nonresponse in sample surveys*. Journal of the American Statistical Association 89, 693-696.