

# Procédures d'imputation jointes pour des variables catégorielles

Une application à l'enquête Patrimoine 2010

Hélène Chaput, Laurianne Salembier, Julie Solard (Insee)  
Guillaume Chauvet (Crest, Ensai)  
David Haziza (Univ. de Montréal)

Journées de Méthodologie Statistique  
Paris, 25/01/2012

# Plan de l'exposé

Introduction : le cas univarié

Méthodes d'imputation jointes

Application à l'Enquête Patrimoine 2010

# Introduction : le cas univarié

## 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  $x_k$  la valeur prise par une variable d'intérêt  $x$  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 supposées connues et non nulles. Soit  $w_k$  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

Nous nous intéressons d'abord au cas d'une variable  $x$  binaire 0 – 1. Les résultats présentés s'étendent simplement au cas d'une variable catégorielle avec un nombre quelconque de modalités.

En situation de réponse complète à la variable  $x$ , le total  $t_x$  peut être estimé sans biais sous le plan de sondage par

$$\hat{t}_{x\pi} = \sum_{k \in S} w_k x_k.$$

En situation de non-réponse pour la variable  $x$ , une valeur manquante  $x_k$  est remplacée par une valeur imputée  $x_k^*$  (Haziza, 2009). On obtient l'estimateur

$$\hat{t}_{xI} = \sum_{k \in S_r} w_k x_k + \sum_{k \in S_m} w_k x_k^*.$$

## Imputation hot-deck (RHDI)

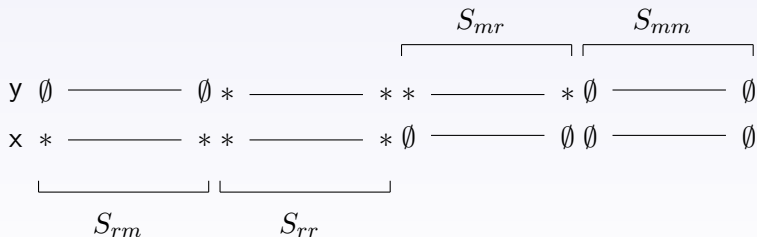
La méthode du hot-deck est couramment utilisée en pratique pour imputer une variable catégorielle. Une valeur manquante  $x_k$  est remplacée par  $x_{(i)}$ , sélectionnée au hasard et avec remise parmi les valeurs observées  $x_i$ ,  $i \in S_r$ .

On note  $q(\cdot)$  le mécanisme de réponse. L'estimateur imputé  $\hat{t}_{xI}$  est approximativement  $pq$ -non biaisé si la variable  $x_k$  et la probabilité  $p_k$  de répondre à cette variable sont non corrélées.

En pratique, l'imputation se fait souvent au sein de classes définies à l'aide de variables explicatives de  $x_k$  et/ou de  $p_k$ , afin que cette condition soit approximativement vérifiée.

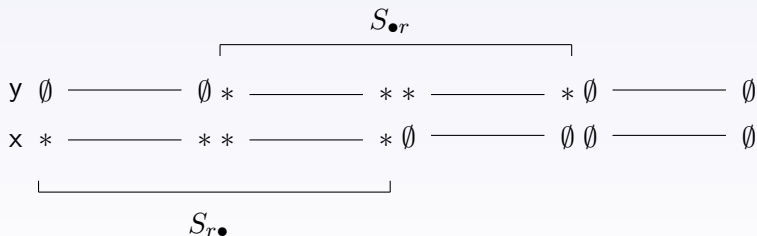
## Cas de deux variables

Considérons maintenant le cas de deux variables  $x_k$  et  $y_k$  affectées par de la non-réponse partielle. La structure obtenue pour les données manquantes est plus complexe.



## Cas de deux variables

Considérons maintenant le cas de deux variables  $x_k$  et  $y_k$  affectées par de la non-réponse partielle. La structure obtenue pour les données manquantes est plus complexe.





## Hypothèses sur le mécanisme de réponse

Nous faisons les hypothèses suivantes sur le mécanisme de réponse :

- 1 les unités répondent indépendamment les unes des autres,
- 2 le mécanisme de réponse aux items est uniforme, au sens où :

$$P(r_{xi} = 1, r_{yi} = 1) \equiv p_{rr},$$

$$P(r_{xi} = 1, r_{yi} = 0) \equiv p_{rm},$$

$$P(r_{xi} = 0, r_{yi} = 1) \equiv p_{mr},$$

$$P(r_{xi} = 0, r_{yi} = 0) \equiv p_{mm}.$$

Là encore, il est possible de définir des classes d'imputation pour que la condition 2 soit approximativement vérifiée.

## Cas de deux variables : imputation marginale (RMI)

Supposons que les deux variables soient imputées indépendamment.

Pour :

$$\begin{array}{ll}
 k \notin S_{r\bullet} & x_k^* = x_{(i)} \quad \text{tiré parmi } x_i, i \in S_{r\bullet}, \\
 l \notin S_{\bullet r} & y_l^* = y_{(j)} \quad \text{tiré parmi } y_j, j \in S_{\bullet r}.
 \end{array}$$

Simulation illustrative : pop artificielle de  $N = 10,000$  individus, coef. de corrélation  $\rho = 0.5$ . On simule  $B = 1,000$  fois un SAS(500) + proba de réponse de 0.55 pour chaque variable.

| $\hat{\rho}_I$    | RMI   |
|-------------------|-------|
| Biais Rel. en %   | -59.9 |
| (Coef Var. en % ) | (8.7) |

Imputer les deux variables séparément conduit à atténuer fortement les relations entre ces variables.

# Méthodes d'imputation jointes

## Procédure d'imputation hot-deck (RHDI)

On utilise un donneur commun si les deux variables sont manquantes.

Pour :

$$\begin{array}{lll}
 k \in S_{mr} & x_k^* = x_{(i)} & \text{tiré parmi } x_i, i \in S_{r\bullet}, \\
 k \in S_{rm} & y_k^* = y_{(j)} & \text{tiré parmi } y_j, j \in S_{\bullet r}, \\
 k \in S_{mm} & (x_k^*, y_k^*) = (x_{(i)}, y_{(i)}) & \text{tiré parmi } (x_i, y_i), i \in S_{rr}.
 \end{array}$$

| $\hat{\rho}_I$    | RMI   | RHDI   |
|-------------------|-------|--------|
| Biais Rel. en %   | -59.9 | -30.3  |
| (Coef Var. en % ) | (8.7) | (10.7) |

Le biais diminue, mais reste très important.

## Procédure d'imputation jointe (JHDI)

L'idée consiste à **tirer un donneur** dans la **distribution conditionnelle de la variable imputée**.

$$\begin{array}{llll}
 k \in S_{mr} & x_k^* = x_{(i)} \text{ tiré parmi} & x_i | y_i = y_k, & i \in S_{rr}, \\
 k \in S_{rm} & y_k^* = y_{(j)} \text{ tiré parmi} & y_j | x_j = x_k, & j \in S_{rr}, \\
 k \in S_{mm} & (x_k^*, y_k^*) = (x_{(i)}, y_{(i)}) & \text{tiré parmi } (x_i, y_i), & i \in S_{rr}.
 \end{array}$$

| $\hat{\rho}_I$    | RMI   | RHDI   | JHDI   |
|-------------------|-------|--------|--------|
| Biais Rel. en %   | -59.9 | -30.3  | -0.2   |
| (Coef Var. en % ) | (8.7) | (10.7) | (13.5) |

La corrélation entre les variables  $x$  et  $y$  est préservée, au prix d'une augmentation de la variabilité.

## Procédure d'imputation jointe équilibrée (BHDI)

On peut limiter la variance d'imputation de la procédure précédente en introduisant des contraintes d'équilibrage dans la sélection des donneurs.

Exemple :

- SAS + 10 individus  $k \in S_{mr}$  tels que  $y_k = 1$  ;
- on observe 80 % de  $x_l = 1$  parmi les  $l \in S_{rr}$  tels que  $y_l = 1$  ;
- on impute exactement 8 valeurs  $x_k^* = 1$  tels que  $y_k = 1$ .

| $\hat{\rho}_I$    | RMI   | RHDI   | JHDI   | BHDI   |
|-------------------|-------|--------|--------|--------|
| Biais Rel. en %   | -59.9 | -30.3  | -0.2   | -0.1   |
| (Coef Var. en % ) | (8.7) | (10.7) | (13.5) | (12.0) |

La corrélation entre les variables  $x$  et  $y$  reste préservée, et la variabilité diminue.

# Application à l'Enquête Patrimoine 2010

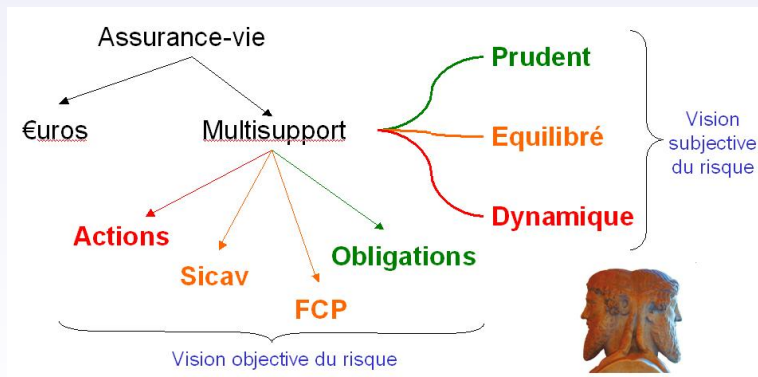
## L'enquête Patrimoine 2010

- Enquête réalisée tous les 6 ans depuis 1986
- Enquête auprès des ménages sur leur patrimoine et les facteurs pouvant l'expliquer
- Plan de sondage complexe, car patrimoine très concentré
- Non réponse : un enjeu crucial.



## L'assurance vie

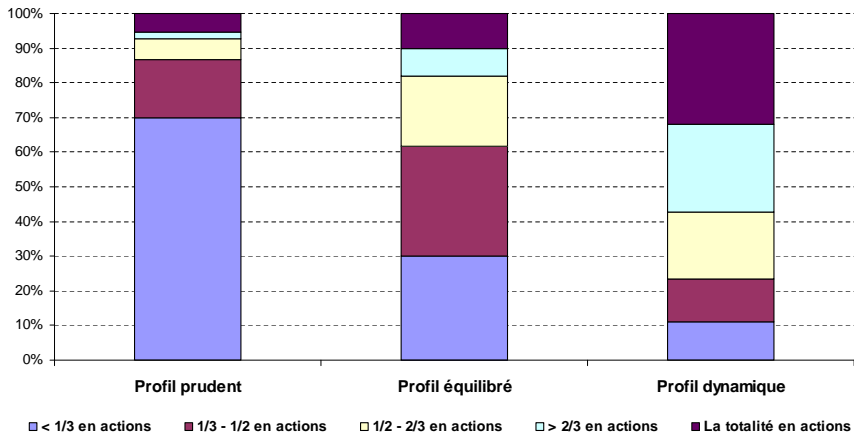
34 % des ménages détiennent au moins une assurance-vie.



## Variables imputées

- Profil de l'assurance-vie :
  - prudent,
  - équilibré,
  - dynamique.
- Part de l'encours investi en actions :
  - moins d'un tiers,
  - entre un tiers et la moitié,
  - entre la moitié et les deux tiers,
  - plus des deux tiers,
  - la totalité.

# Description des données



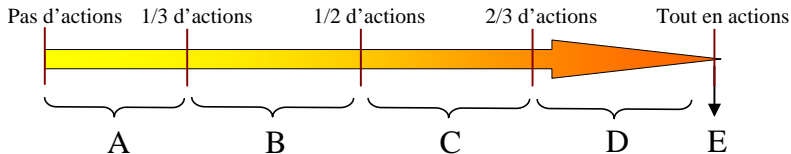
# Description des données

| Part en actions        | Profil de l'assurance-vie |                |                  |                  |
|------------------------|---------------------------|----------------|------------------|------------------|
|                        | ??                        | Profil prudent | Profil équilibré | Profil dynamique |
| ??                     | 510                       | 492            | 234              | 58               |
| < 1/3 en actions       | 45                        | 1308           | 316              | 59               |
| 1/3 - 1/2 en actions   | 17                        | 312            | 331              | 65               |
| 1/2 - 2/3 en actions   | 12                        | 114            | 213              | 101              |
| > 2/3 en actions       | 12                        | 33             | 83               | 133              |
| La totalité en actions | 7                         | 103            | 105              | 169              |

- **Table:** Répartition des assurances-vie selon leur profil et selon leur composition, et coefficient de corrélation des rangs de Spearman entre le profil et la part en actions

|              | Pourcentage d'assurances-vie avec une part en actions |      |      |     |      | Pourcentage d'assurances-vie |              |             | Coefficient de corrélation |
|--------------|---|------|------|-----|------|------------------------------|--------------|-------------|----------------------------|
|              | A   | B    | C    | D   | E    | sans risque                  | risque moyen | risque fort |                            |
| Cas complets | 52.2  | 19.9 | 12.0 | 6.0 | 10.0 | 58.7                         | 26.3         | 15.0        | 0.518                      |
| RHDI         | 51.4  | 20.5 | 11.9 | 6.2 | 10.0 | 60.0                         | 26.4         | 13.6        | 0.434                      |
| JHDI         | 52.1  | 20.5 | 12.0 | 5.7 | 9.8  | 60.0                         | 26.1         | 13.9        | 0.515                      |
| BHDI         | 53.2  | 19.5 | 11.7 | 6.1 | 9.6  | 66.6                         | 25.6         | 13.8        | 0.508                      |

- RHDI = IMPUTATION HOT-DECK
- JHDI = IMPUTATION JOINTE
- BHDI = IMPUTATION JOINTE EQUILIBREE



## Bibliographie

Chaput, H., Chauvet, G., Haziza, D., Salembier, L., Solard, J. (2010). *Joint imputation procedures for categorical variables with application to the French Wealth Survey*, soumis.

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

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

Skinner, C.J., Rao, J.N.K. (2002). *Jackknife variance for multivariate statistics under hot deck imputation from common donors*, *JSPI*, 102, 149-167.