

Données Manquantes dans les Enquêtes

Guillaume Chauvet

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

30 janvier 2012

Panorama du cours

- 1 Introduction et rappels
- 2 Traitement de la non-réponse totale
- 3 Traitement de la non-réponse partielle

Objectifs du cours

- Expliquer le phénomène de non-réponse, et ses conséquences sur l'estimation.
- Décrire les méthodes de correction de la non-réponse totale dans les enquêtes.
- Décrire les méthodes de correction de la non-réponse partielle dans les enquêtes.

Introduction et rappels

Les étapes d'une enquête (Haziza, 2011)

- 1 Planification : objectifs, concepts, champ de l'enquête, ...
- 2 Constitution de la base de sondage
- 3 Conception du questionnaire
- 4 Conception du plan de sondage et tirage de l'échantillon
- 5 Collecte des données
- 6 Traitement des données
- 7 Estimation ponctuelle et estimation de variance

Les étapes d'une enquête (Haziza, 2011)

- 1 Planification : objectifs, concepts, champ de l'enquête, ...
- 2 Constitution de la base de sondage
- 3 Conception du questionnaire
- 4 Conception du plan de sondage et tirage de l'échantillon
- 5 Collecte des données
- 6 Traitement des données
- 7 Estimation ponctuelle et estimation de variance

Rappels sur l'échantillonnage en population finie

Plan de sondage

On se place dans le cadre d'une population finie d'individus, notée U . On s'intéresse à une **variable d'intérêt** y (éventuellement vectorielle), qui prend la valeur y_k sur l'individu k de U .

Les valeurs prises par la variable y sont collectées sur un échantillon S . L'objet de la Théorie des Sondages est d'utiliser cette information afin d'estimer des paramètres définis sur la population entière.

L'échantillon S est sélectionné dans U au moyen d'un **plan de sondage** $p(\cdot)$, i.e. d'une loi de probabilité (supposée connue) sur l'ensemble des parties de U .

Plan de sondage

On suppose en particulier connues les **probabilités d'appartenance** à l'échantillon de chaque unité k :

$$\pi_k = \Pr(k \in S).$$

Si toutes les π_k sont > 0 , le total $t_y = \sum_{k \in U} y_k$ est estimé sans biais par l'**estimateur de Horvitz-Thompson**

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} d_k y_k \quad (1)$$

avec $d_k = 1/\pi_k$ le poids de sondage de l'unité k .

Remarque : les mêmes poids peuvent être utilisés pour toutes les variables d'intérêt.

Plan de sondage

La forme générale de variance est donnée par la formule de Horvitz-Thompson (1953)

$$V_p [\hat{t}_{y\pi}] = \sum_{k,l \in U} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl}$$

avec $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$.

On peut l'estimer sans biais par

$$v_{HT} [\hat{t}_{y\pi}] = \sum_{k,l \in S} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \frac{\Delta_{kl}}{\pi_{kl}}$$

si tous les π_{kl} sont > 0 .

Plan de taille fixe

Si le plan de sondage $p(\cdot)$ est **de taille fixe** égale à n , la variance du π -estimateur peut être alternativement obtenue par la formule de Sen-Yates-Grundy (1954)

$$V_p [\hat{t}_{y\pi}] = -\frac{1}{2} \sum_{k \neq l \in U} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \Delta_{kl}.$$

Un estimateur sans biais est donné par

$$v_{YG} [\hat{t}_{y\pi}] = -\frac{1}{2} \sum_{k \neq l \in S} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \frac{\Delta_{kl}}{\pi_{kl}}.$$

si tous les π_{kl} sont > 0 .

Exemple : le sondage aléatoire simple

Sondage aléatoire simple (SRS) de taille n : plan de taille fixe, où tous les échantillons de taille n ont la même probabilité d'être sélectionnés.

$$\pi_k = \frac{n}{N} \quad \text{et} \quad \pi_{kl} = \frac{n(n-1)}{N(N-1)}.$$

Le π -estimateur peut se réécrire

$$\hat{t}_{y\pi} = \frac{N}{n} \sum_{k \in S} y_k = N\bar{y},$$

et la moyenne dans la population μ_y est estimée par la moyenne dans l'échantillon \bar{y} .

Le sondage aléatoire simple

La variance du π -estimateur est donnée par

$$V_p [\hat{t}_{y\pi}] = N^2(1 - f) \frac{S_y^2}{n}$$

avec $f = n/N$ et $S_y^2 = \frac{1}{N-1} \sum_{k \in U} (y_k - \mu_y)^2$.

Cette variance est estimée sans biais par

$$v_{SRS} [\hat{t}_{y\pi}] = N^2(1 - f) \frac{s_y^2}{n}$$

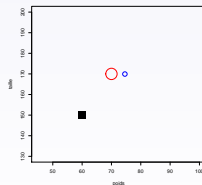
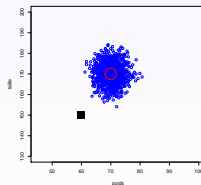
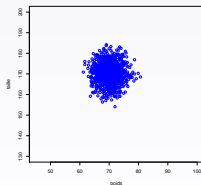
avec $s_y^2 = \frac{1}{n-1} \sum_{k \in s} (y_k - \bar{y})^2$.

Sources d'erreur dans l'estimation

Erreur associée à l'estimateur

Soit $\hat{\theta}$ l'estimateur d'un paramètre θ . La précision de cet estimateur peut être mesurée par :

- son biais : $B[\hat{\theta}] = E[\hat{\theta} - \theta]$
- sa variance : $V[\hat{\theta}] = E[(\hat{\theta} - E(\hat{\theta}))^2]$
- son EQM : $EQM[\hat{\theta}] = B[\hat{\theta}]^2 + V[\hat{\theta}]$.



Sources d'erreur

En pratique, l'erreur totale de l'estimateur, mesurée par

$$\hat{\theta} - \theta,$$

dépend des erreurs réalisées à toutes les étapes de l'enquêtes.

Cela inclut :

- les erreurs de couverture,
- l'erreur d'échantillonnage,
- l'erreur due à la non-réponse,
- les erreurs de mesure.

Erreurs de couverture

Les erreurs de couverture proviennent du fait que la base de sondage et la population-cible ne coïncident pas. On distingue :

- la sous-couverture : des individus de la population-cible sont absents de la base de sondage,
- la sur-couverture : la base de sondage contient des individus qui ne sont pas dans la population-cible.

Exemples de sous-couverture :

- nouvelles entreprises pas encore inscrites dans le répertoire SIRENE,
- enquête téléphonique auprès de ménages, en utilisant une liste d'abonnés à une ligne fixe,
- difficulté de couvrir la population-cible (enquête auprès de SDF).

Erreurs d'échantillonnage et de non-réponse

L'erreur d'échantillonnage provient du fait que l'information n'est collectée que sur une partie de la population : cette erreur est volontaire et planifiée.

L'erreur de non-réponse provient du fait que l'information n'est observée que sur une partie de l'échantillon uniquement : cette erreur est subie et non maîtrisée.

La non-réponse a des conséquences sur le biais et la variance des estimateurs.

Erreurs de mesure

Les erreurs de mesure proviennent du fait que les valeurs obtenues sont différentes des vraies valeurs de la variables d'intérêt.

Parmi les causes des erreurs de mesure :

- questionnaire mal conçu,
- problème d'enquêteur,
- appel à la mémoire des enquêtés,
- erreur de codage.

Dans ce qui suit, on supposera que les erreurs de couverture et de mesure peuvent être négligées. On se focalisera sur l'erreur due à l'échantillonnage et sur l'erreur due à la non-réponse.

Les types de non-réponse

Type de non-réponse

Dans le contexte des enquêtes, on distingue deux types de non-réponse :

- la non-réponse totale ("unit non-response") : aucune information n'est relevée pour une unité,
- la non-réponse partielle ("item non-response") : une partie seulement de l'information est relevée pour une unité.

y_1	y_2	y_3	y_4	y_p
*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*
*	*	*	*	*	*	*	*	*	*
\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset
*	*	\emptyset	*	\emptyset	*	\emptyset	*	*	\emptyset
\emptyset	*	*	*	\emptyset	*	\emptyset	*	*	\emptyset
*	*	*	*	*	*	*	*	\emptyset	\emptyset
\emptyset	\emptyset	\emptyset	*	*	\emptyset	*	*	*	*

⎵ Réponse totale

⎵ Non-réponse totale

⎵ Non-réponse partielle

Effets de la non-réponse

La non-réponse a des conséquences sur le biais et la variance des estimateurs :

- les caractéristiques des non-répondants sont généralement différentes de celles des répondants
⇒ biais de non-réponse
- la non-réponse diminue la taille de l'échantillon effectivement observée
⇒ variance de non-réponse
- l'utilisation d'une imputation aléatoire (voir plus loin) conduit à ajouter une variabilité supplémentaire
⇒ variance d'imputation

Traitement de la non-réponse dans les enquêtes

La non-réponse totale est habituellement traitée par une méthode de repondération :

- on supprime du fichier les non-répondants totaux,
- on augmente les poids des répondants pour compenser de la non-réponse totale.

La non-réponse partielle est habituellement traitée par imputation : une valeur manquante est remplacée par une valeur plausible.

L'objectif prioritaire est de réduire autant que possible le biais de non-réponse : cela passe par une recherche des facteurs explicatifs de la non-réponse.

Quelques facteurs de non-réponse totale (Haziza, 2011)

- Mauvaise qualité de la base de sondage,
- Impossibilité de joindre l'individu,
- Type d'enquête (obligatoire ou volontaire),
- Fardeau de réponse,
- Méthode de collecte (interview, téléphone, courrier, ...),
- Durée de collecte,
- Suivi (et relance) des non-répondants,
- Formation des enquêteurs.

Quelques facteurs de non-réponse partielle (Haziza, 2011)

- Questionnaire mal conçu,
- Fardeau de réponse,
- Questions délicates,
- Formation des enquêteurs,
- Appel à la mémoire des enquêtés.

La prévention (ou la correction) de la non-réponse se fait à toutes les étapes de la collecte des données.

Traitement de la non-réponse totale

Le problème

La non-réponse totale ("unit non-response") survient lorsqu'aucune information (autre que celle de la base de sondage) n'est relevée pour une unité.

On va traiter ce problème par repondération : on fait porter aux répondants le poids des non-répondants. Cette repondération se justifie sous une modélisation du mécanisme de non-réponse.

Cette modélisation permet d'estimer les probabilités de réponse à l'enquête, pour obtenir les poids corrigés de la non-réponse totale.

Les étapes du traitement de la non-réponse totale

- 1 Identification des non-répondants,
- 2 Recherche des facteurs explicatifs de la non-réponse,
- 3 Estimation des probabilités de réponse,
- 4 Calcul des poids corrigés de la non-réponse totale.

Identification des non-répondants

Un point important : la distinction entre individus hors-champ et individus non-répondants. Le **champ** de l'enquête désigne l'ensemble des individus statistiques auxquels on s'intéresse.

Certains individus de l'échantillon sont hors-champ, i.e. ont été sélectionnés mais ne font pas partie du champ de l'enquête. Ils ne sont donc pas pris en compte dans l'estimation.

Exemple : échantillonnage dans les logements Résidences Principales en 1999, pour représenter les logements Résidences Principales en 2006.

Les individus **non-répondants** font partie du champ de l'enquête, mais leur réponse n'est pas observée (refus de répondre, impossible à joindre, perte de questionnaire, ...). Cette non-réponse doit être compensée.

Modélisation du mécanisme de non-réponse

Echantillonnage en deux phases

Dans le cadre d'une enquête, on peut être amené à sélectionner l'échantillon en deux temps :

- On sélectionne tout d'abord un gros sur-échantillon S selon un plan de sondage $p(\cdot)$.
- On tire ensuite dans S un sous-échantillon S_0 selon un plan de sondage $q(\cdot|S)$.

On parle d'échantillonnage en deux phases. Cette méthode est par exemple utilisée pour cibler une population spécifique.

Exemple : Enquête Vie Quotidienne et Santé, utilisée comme filtrage pour l'enquête Handicaps-Incapacités-Dépendances (Joinville, 2002).

Estimateur par expansion

Le total t_y peut théoriquement être estimé sans biais par le π -estimateur, mais les probabilités d'inclusion sont généralement difficiles (voire impossibles) à calculer.

On a recours à un autre estimateur, appelé l'estimateur par expansion.

On note :

- $\pi_k = \Pr(k \in S)$ la probabilité de sélection de l'unité k dans S ,
- $\pi_{0k|S} = \Pr(k \in S_0|S)$ la probabilité de sélection de l'unité k dans S_0 , conditionnellement à S .

Estimateur par expansion

L'estimateur par expansion est défini par

$$\hat{t}_{y,exp} = \sum_{k \in S_0} \frac{y_k}{\pi_k \pi_{0k|S}}.$$

C'est un estimateur sans biais du total t_y :

$$\begin{aligned} E[\hat{t}_{y,exp}] &= E_p E_q [\hat{t}_{y,exp} | S] \\ &= E_p \left[\sum_{k \in S} \frac{y_k}{\pi_k} \right] \\ &= t_y. \end{aligned}$$

Variance de l'estimateur par expansion

Notons que dans le calcul de l'espérance de l'estimateur, deux mécanismes aléatoires interviennent :

- l'un associé au plan $p(\cdot)$ utilisé en 1ère phase,
- l'autre associé au plan $q(\cdot|S)$ utilisé en 2nde phase.

La variance de cet estimateur est donnée par :

$$\begin{aligned}
 V[\hat{t}_{y,exp}] &= V_p E_q [\hat{t}_{y,exp} | S] + E_p V_q [\hat{t}_{y,exp} | S] \\
 &= \underbrace{V_p \left[\sum_{k \in S} \frac{y_k}{\pi_k} \right]}_{\text{Variance Phase 1}} + \underbrace{E_p V_q [\hat{t}_{y,exp} | S]}_{\text{Variance Phase 2}}.
 \end{aligned}$$

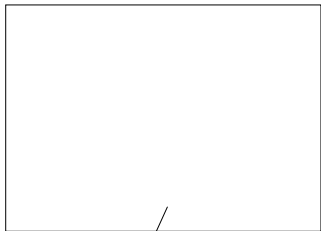
Modélisation du mécanisme de non-réponse

En situation de non-réponse totale :

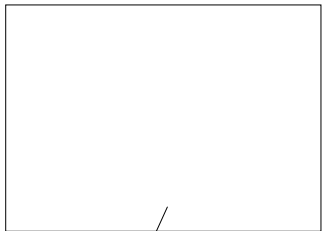
- le mécanisme de sélection de l'échantillon S est connu,
- le mécanisme de non-réponse qui conduit au sous-échantillon de répondants S_r est en revanche inconnu.

On a recours à une modélisation du mécanisme aléatoire conduisant à S_r sous la forme d'un échantillonnage en deux phases :

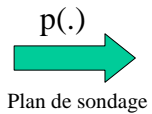
- la 1ère phase correspond à la sélection de l'échantillon S ,
- la 2nde phase correspond à la "sélection" du sous-échantillon de répondants S_r
⇒ mécanisme de non-réponse

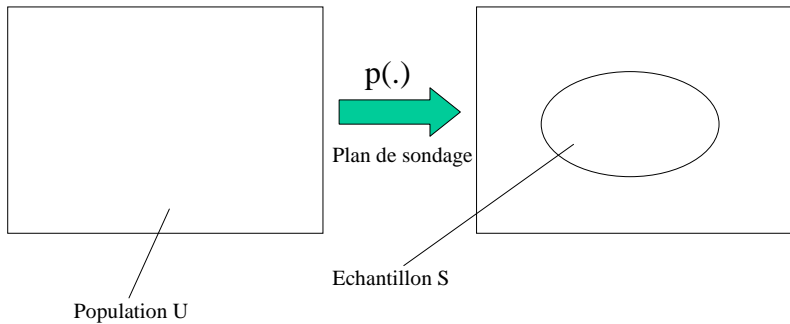


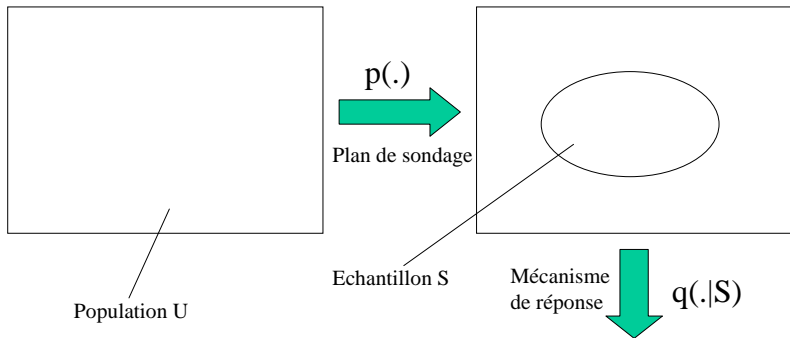
Population U

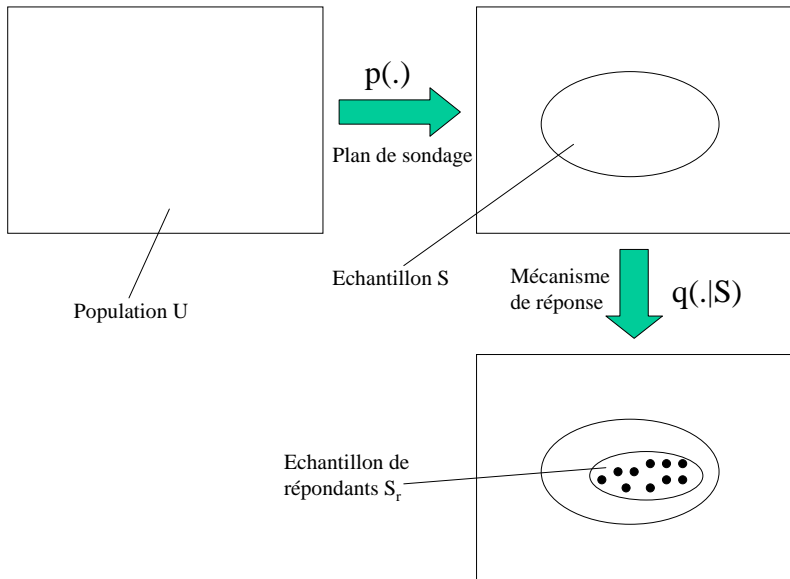


Population U









Mécanisme de non-réponse

On note r_k la variable indicatrice de réponse pour l'individu k , valant 1 si l'individu a répondu à l'enquête et 0 sinon.

On note p_k la probabilité de réponse pour l'unité k :

$$\begin{aligned} p_k &= \Pr(k \in S_r | S) \\ &= \Pr(r_k = 1 | S). \end{aligned}$$

On fait l'hypothèse que :

- toutes les probabilités de réponse vérifient $0 < p_k \leq 1$: pas de non-répondants irréductibles,
- les individus répondent indépendamment les uns des autres :

$$\Pr(k, l \in S_r | S) \equiv p_{kl} = p_k p_l.$$

Cette dernière hypothèse peut être affaiblie (Haziza et Rao, 2003 ; Skinner et D'Arrigo, 2011).

Types de mécanisme

On distingue schématiquement trois types de mécanisme de non-réponse :

- uniforme (ou MCAR),
- ignorable (ou MAR),
- non-ignorable (ou NMAR).

Le mécanisme est dit uniforme (ou Missing Completely At Random) quand $p_k = p$, i.e. quand tous les individus ont la même probabilité de réponse. C'est une hypothèse généralement peu réaliste.

Exemple : non-réponse provenant de la perte de questionnaires.

Types de mécanisme

On parle de mécanisme de non-réponse ignorable (ou Missing At Random) quand les probabilités de réponse peuvent être expliquées à l'aide de l'information auxiliaire disponible :

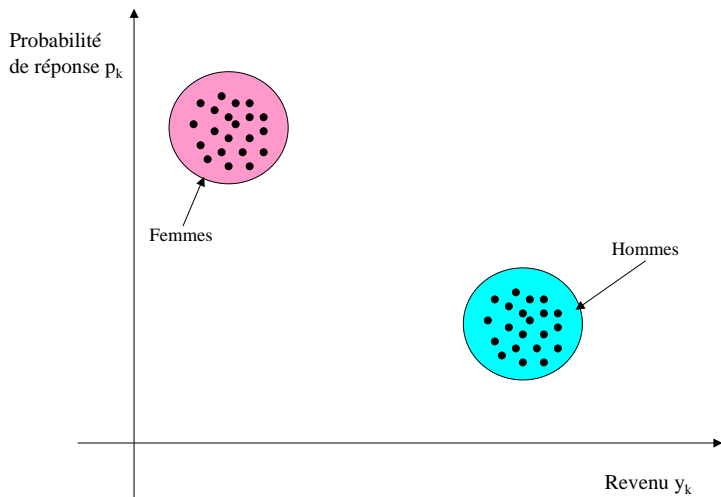
$$\Pr(r_k = 1 | \mathbf{y}_S, \mathbf{z}_S) = \Pr(r_k = 1 | \mathbf{z}_S),$$

avec

- \mathbf{y}_S le vecteur des valeurs prises par la variable y sur les individus de S ,
- \mathbf{z}_S le vecteur des valeurs prises par un vecteur \mathbf{z} de variables auxiliaires sur les individus de S .

Exemple : enquête sur le revenu + non-réponse expliquée par le sexe des individus.

Un exemple de non-réponse MAR



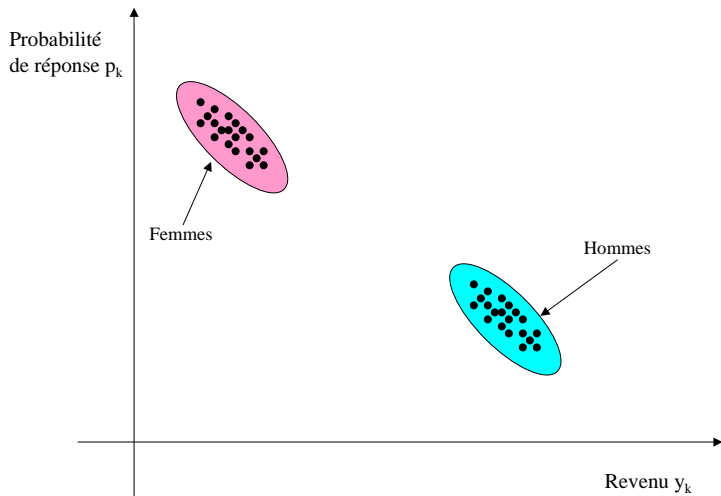
Types de mécanisme

Un mécanisme de non-réponse qui n'est pas ignorable est dit non-ignorable (ou Non Missing At Random). Cela signifie que la non-réponse dépend de la variable d'intérêt, même une fois que l'on a pris en compte les variables auxiliaires.

Il est très difficile de corriger de la non-réponse non ignorable, ou même de la détecter. Dans la suite, nous supposons être dans le cas d'un mécanisme MAR.

Exemple : enquête sur le revenu + non-réponse expliquée par le croisement sexe \times revenu.

Un exemple de non-réponse NMAR



Estimation en présence de non-réponse totale

Estimation par expansion

Si les probabilités de réponse p_k sont connues, on est dans le cas d'un échantillonnage en deux phases. On peut alors estimer le total t_y par l'estimateur par expansion

$$\hat{t}_{y,exp} = \sum_{k \in S_r} \frac{y_k}{\pi_k p_k}.$$

L'écart au vrai total t_y peut se décomposer sous la forme

$$\hat{t}_{y,exp} - t_y = (\hat{t}_{y\pi} - t_y) + (\hat{t}_{y,exp} - \hat{t}_{y\pi}) \quad (2)$$

avec $\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}$ l'estimateur en situation de réponse complète.

Dans le 2nd membre de (2) :

- le 1er terme représente l'erreur d'échantillonnage,
- le 2nd terme représente l'erreur due à la non-réponse.

Estimation par expansion

Cet estimateur est sans biais sous le double alea correspondant au plan de sondage et au mécanisme de non-réponse.

Sous ce modèle, le mécanisme de non-réponse est vu comme un plan poissonien. D'où la variance de l'estimateur par expansion :

$$V[\hat{t}_{y,exp}] = \underbrace{V_p \left[\sum_{k \in S} \frac{y_k}{\pi_k} \right]}_{\text{Variance Echantillonnage}} + \underbrace{E_p \left[\sum_{k \in S} \left(\frac{y_k}{\pi_k} \right)^2 \times \frac{1 - p_k}{p_k} \right]}_{\text{Variance Non Réponse}}.$$

Cette variance est donc toujours plus grande qu'en situation de réponse complète.

Estimation des probabilités de réponse

En pratique, les probabilités de réponse p_k sont inconnues et doivent être estimées. On peut par exemple postuler un modèle paramétrique de la forme

$$p_k = f(\mathbf{z}_k, \beta),$$

avec

- \mathbf{z}_k un vecteur de variables auxiliaires connu sur l'ensemble de l'échantillon S ,
- $f(\cdot, \cdot)$ une fonction connue,
- β un paramètre inconnu.

On parle du **modèle de non-réponse**.

Estimation des probabilités de réponse

Un choix classique en pratique consiste à prendre

$$f(\mathbf{z}_k, \beta) = \frac{\exp(\mathbf{z}_k^\top \beta)}{1 + \exp(\mathbf{z}_k^\top \beta)}.$$

Le modèle de non-réponse peut alors se réécrire à l'aide de la fonction de lien logit :

$$\log \frac{p_k}{1 - p_k} = \mathbf{z}_k^\top \beta,$$

ce qui correspond au modèle de régression logistique.

D'autres fonctions de lien sont possibles. On peut également utiliser une modélisation non paramétrique des probabilités p_k (Da Silva et Opsomer, 2006 et 2009).

Estimateur du total

On peut alors obtenir (par exemple, à l'aide de la PROC LOGISTIC de SAS) un estimateur $\hat{\beta}$ du paramètre β , et des probabilités de réponse estimées

$$\hat{p}_k = f(\mathbf{z}_k, \hat{\beta}).$$

On obtient l'estimateur corrigé de la non-réponse totale

$$\begin{aligned}\hat{t}_{yr} &= \sum_{k \in S_r} \frac{y_k}{\pi_k \hat{p}_k} \\ &= \sum_{k \in S_r} \tilde{d}_k y_k,\end{aligned}$$

avec $\tilde{d}_k = d_k / \hat{p}_k$.

Propriétés de l'estimateur

Si le modèle de non-réponse est correctement spécifié, i.e. :

- la fonction de lien $f(\cdot, \cdot)$ est correcte,
- \mathbf{z}_k inclut toutes les variables explicatives de la non-réponse,

alors les propriétés de \hat{t}_{yr} sont approximativement les mêmes que celles de $\hat{t}_{y,exp}$ utilisant les "vraies" probabilités de réponse.

On obtient donc :

$$E[\hat{t}_{yr}] \simeq t_y$$

$$V[\hat{t}_{yr}] \simeq V_p \left[\sum_{k \in S} \frac{y_k}{\pi_k} \right] + E_p \left[\sum_{k \in S} \left(\frac{y_k}{\pi_k} \right)^2 \times \frac{1 - p_k}{p_k} \right]$$

$$\simeq V_p \left[\sum_{k \in S} \frac{y_k}{\pi_k} \right] + E_p E_q \left[\sum_{k \in S_r} \left(\frac{y_k}{\pi_k} \right)^2 \times \frac{1 - \hat{p}_k}{\hat{p}_k^2} \right].$$

Cas des groupes homogènes de réponse

Un modèle de non-réponse couramment utilisé en pratique consiste à supposer que la probabilité de réponse p_k est constante au sein de groupes S_1, \dots, S_C partitionnant l'échantillon S :

$$\forall k \in S_c \quad p_k = p_c.$$

On les appelle les groupes homogènes de réponse (GHR). Cette modélisation a l'avantage :

- d'être simple à mettre en oeuvre,
- d'offrir une certaine robustesse contre une mauvaise spécification du modèle de non-réponse.

Exemple : enquête sur le revenu + GHR définis en croisant sexe et tranche d'âge.

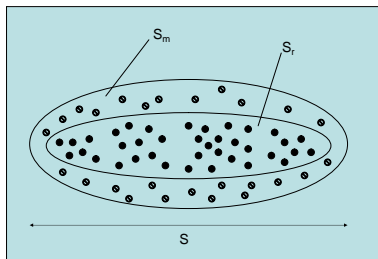
Cas des groupes homogènes de réponse

Au sein de chaque GHR S_c , la probabilité p_c est estimée par

$$\hat{p}_c = \frac{n_{rc}}{n_c},$$

en notant

- n_c le nombre d'individus dans S_c ,
- n_{rc} le nombre de répondants dans S_c .



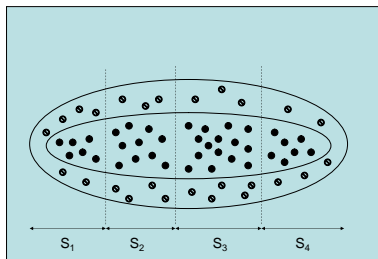
Cas des groupes homogènes de réponse

Au sein de chaque GHR S_c , la probabilité p_c est estimée par

$$\hat{p}_c = \frac{n_{rc}}{n_c},$$

en notant

- n_c le nombre d'individus dans S_c ,
- n_{rc} le nombre de répondants dans S_c .



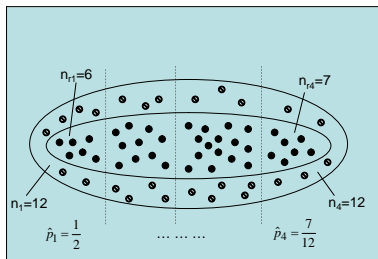
Cas des groupes homogènes de réponse

Au sein de chaque GHR S_c , la probabilité p_c est estimée par

$$\hat{p}_c = \frac{n_{rc}}{n_c},$$

en notant

- n_c le nombre d'individus dans S_c ,
- n_{rc} le nombre de répondants dans S_c .



Estimation

Sous le modèle correspondant aux GHR, on obtient l'estimateur

$$\begin{aligned}\hat{t}_{yr} &= \sum_{k \in S_r} \frac{y_k}{\pi_k \hat{p}_k} \\ &= \sum_{c=1}^C \frac{n_c}{n_{rc}} \sum_{k \in S_{rc}} \frac{y_k}{\pi_k}.\end{aligned}$$

Sa variance est approximativement donnée par

$$V[\hat{t}_{yr}] \simeq V_p \left[\sum_{k \in S} \frac{y_k}{\pi_k} \right] + E_p E_q \left[\sum_{c=1}^C \frac{1 - \hat{p}_c}{\hat{p}_c^2} \sum_{k \in S_{rc}} \left(\frac{y_k}{\pi_k} \right)^2 \right].$$

Détermination des GHR

En pratique, on peut constituer ces groupes de la façon suivante :

- 1 On effectue une régression logistique afin d'expliquer les probabilités de réponse en fonction de l'information auxiliaire disponible.
- 2 On peut ensuite :
 - soit ordonner les individus k selon les \hat{p}_k (méthode des scores), puis diviser l'échantillon en groupes de tailles approximativement égales (méthode des quantiles égaux) ;
 - soit utiliser les variables qui ressortent de façon significative dans la régression logistique, et les croiser pour définir les groupes (méthode par croisement).

En résumé

- 1 Identification des non-répondants
⇒ séparation des individus hors-champ et des non-répondants
- 2 Recherche des facteurs explicatifs de la non-réponse
⇒ e.g., régression logistique pour identifier les \mathbf{z}_k explicatifs
- 3 Estimation des probabilités de réponse
⇒ e.g., méthode des scores ou méthode par croisement pour définir les GHR
- 4 Calcul des poids corrigés de la non-réponse totale.

Calage de l'estimateur

Notons finalement qu'une fois corrigé de la NR totale, l'estimateur peut-être calé sur une information auxiliaire afin de réduire sa variance.

Plus précisément, si des totaux t_x sur la population sont connus, les poids \tilde{d}_k peuvent être modifiés pour reproduire ces totaux. On obtient l'estimateur calé

$$\hat{t}_{yw} = \sum_{k \in S_r} w_k y_k \quad \text{avec} \quad w_k = \tilde{d}_k F(\lambda^\top \mathbf{x}_k).$$

Les deux redressements sont de natures différentes :

- la correction de la NR totale ($d_k \Rightarrow \tilde{d}_k$) vise à réduire le **bias** dû à la non-réponse,
- le calage ($\tilde{d}_k \Rightarrow w_k$) vise à réduire la **variance** de l'estimateur.

Traitement de la non-réponse partielle

Le modèle d'imputation

Le problème

La non-réponse partielle ("item non-response") survient lorsqu'une unité répond à l'enquête, mais renseigne une partie des variables seulement.

On va traiter ce problème par imputation : une valeur manquante est remplacée par une valeur plausible. Cette imputation se justifie sous une modélisation de la variable d'intérêt appelée le modèle d'imputation.

L'imputation permet de recréer un fichier de données complet, ce qui facilite l'analyse. En revanche, elle perturbe les relations entre les variables et peut donner une impression artificielle de précision si l'imputation n'est pas prise en compte dans les calculs de variance.

Estimateur imputé

On se place après l'étape de correction de la NR totale. Pour simplifier les notations, on notera simplement dans la suite S l'échantillon (après prise en compte éventuelle de la NR totale), et d_k le poids d'un individu k (éventuellement redressé).

En l'absence de non-réponse partielle pour la variable y , le total t_y peut donc être estimé par

$$\hat{t}_y = \sum_{k \in S} d_k y_k.$$

Soit

- $S_{ry} \equiv S_r$ le sous-échantillon d'individus ayant renseigné la variable y ,
- $S_{my} \equiv S_m$ le sous-échantillon d'individus n'ayant pas renseigné la variable y .

Estimateur imputé

Pour un individu k , soit y_k^* la valeur imputée pour remplacer y_k , si cette dernière est manquante.

L'estimateur imputé est donné par

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

L'erreur totale $\hat{t}_{yI} - t_y$ peut se décomposer sous la forme

$$\hat{t}_{yI} - t_y = (\hat{t}_{y\pi} - t_y) + (\hat{t}_y - \hat{t}_{y\pi}) + (\hat{t}_{yI} - \hat{t}_y),$$

avec

- $\hat{t}_{y\pi} - t_y \Rightarrow$ erreur d'échantillonnage,
- $\hat{t}_y - \hat{t}_{y\pi} \Rightarrow$ erreur due à la non-réponse totale,
- $\hat{t}_{yI} - \hat{t}_y \Rightarrow$ erreur due à la non-réponse partielle.

Estimateur imputé

Au stade de la correction de la non-réponse partielle, les deux premiers termes d'erreur (dûs à l'échantillonnage et à la non-réponse totale) sont incompressibles.

L'objectif de l'imputation est de limiter au maximum l'erreur due à la non-réponse partielle

$$\hat{t}_{yI} - \hat{t}_y = \sum_{k \in S_m} d_k (y_k^* - y_k).$$

L'erreur d'imputation sera limitée :

- si les valeurs imputées y_k^* sont proches des valeurs réelles y_k ;
- ou si les écarts entre valeurs imputées y_k^* et valeurs réelles y_k se compensent en moyenne.

Les étapes du traitement de la non-réponse partielle

- 1 Identification des valeurs manquantes,
- 2 Choix d'un modèle d'imputation,
- 3 Recherche des facteurs explicatifs de la variable d'intérêt,
- 4 Choix du mécanisme d'imputation,
- 5 Imputation des valeurs manquantes.

Identification des valeurs manquantes

Deux points importants :

- distinguer les non-répondants partiels des non-répondants totaux,
- distinguer la non-réponse partielle des valeurs manquantes dues à la forme du questionnaire.

Point 1 : l'imputation ne concerne que les individus qui ont répondu globalement à l'enquête (répondants totaux), mais pas spécifiquement à la variable d'intérêt y (non-répondant partiel). Les deux mécanismes de non-réponse sont généralement différents.

Point 2 : ne pas traiter par imputation l'absence d'une valeur y_k due à la forme du questionnaire (question filtre).

Modèle d'imputation

Le **mécanisme d'imputation** (i.e., la façon de remplacer les valeurs manquantes) 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.

$$m : y_k = \mathbf{z}_k^\top \boldsymbol{\beta} + \sigma \sqrt{v_k} \epsilon_k \quad \text{pour } k \in S. \quad (3)$$

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.

Modèle d'imputation

Le modèle d'imputation est une formulation mathématique de la question : suis-je capable de bien expliquer la variable y_k à l'aide de l'information \mathbf{z}_k disponible? Le modèle utilisé doit permettre de décrire au mieux la variable d'intérêt y_k .

Le modèle d'imputation utilisé doit s'adapter au type de variable traité. En particulier, le modèle de régression (3) est adapté à l'étude d'une variable quantitative, mais n'est généralement pas adapté à l'étude d'une variable qualitative.

Méthodes d'imputation

Types de méthodes

On peut classer les méthodes d'imputation en deux groupes :

- les **méthodes déterministes** : elles conduisent à la même valeur imputée si le mécanisme d'imputation est répété,
- les **méthodes aléatoires** : la valeur imputée inclut une composante aléatoire, et peut donc changer si le mécanisme d'imputation est répété.

On peut ajouter une troisième famille de méthodes, transversale. Les **méthodes d'imputation par donneur** consistent à piocher un individu parmi les répondants, et à utiliser la valeur observée pour la variable y pour remplacer la valeur manquante.

Mécanisme d'imputation par la régression

L'imputation par la régression déterministe s'appuie sur le modèle (3). Elle est obtenue en prenant $y_k^* = \mathbf{x}_k^\top \hat{\boldsymbol{\beta}}_r$, avec

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

où ω_k désigne un **poids d'imputation** attaché à l'unité k (Haziza, 2009). On utilise généralement $\omega_k = 1$ (imputation non pondérée) ou $\omega_k = d_k$ (imputation pondérée par les poids de sondage).

Dans ce cas, l'estimateur imputé est égal à

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

Imputation par la moyenne

L'imputation par la moyenne est un cas particulier de l'imputation par la régression. Elle s'appuie sur le modèle simplifié

$$m : y_k = \beta + \sigma \epsilon_k \quad \text{pour } k \in S, \quad (4)$$

obtenu avec $\mathbf{z}_k = z_k = 1$ et $v_k = 1$.

On a

$$\hat{\beta}_r = \frac{\sum_{k \in S_r} \omega_k y_k}{\sum_{k \in S_r} \omega_k} \equiv \bar{y}_{\omega r},$$

la moyenne des répondants. L'estimateur imputé est égal à

$$\hat{t}_{yI} = \sum_{k \in S_r} d_k y_k + \sum_{k \in S_m} d_k [\bar{y}_{\omega r}].$$

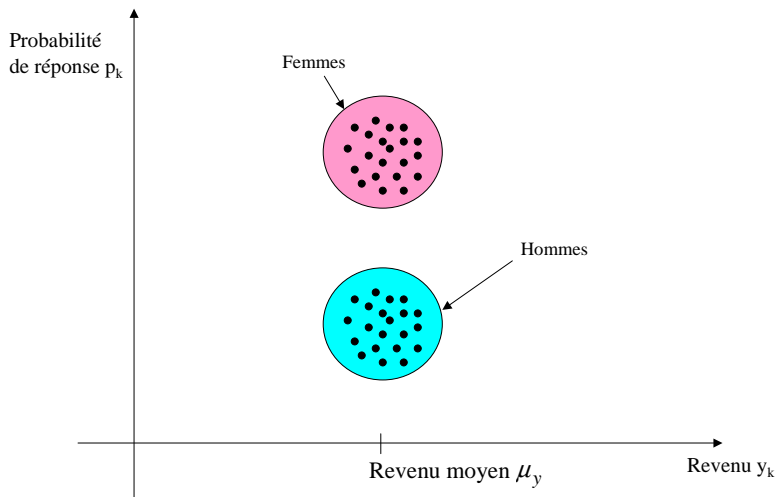
Imputation par la moyenne

Cette méthode d'imputation s'appuie sur le modèle (4). Elle part donc de l'hypothèse que tous les individus de la population ont en moyenne le même comportement pour la variable y .

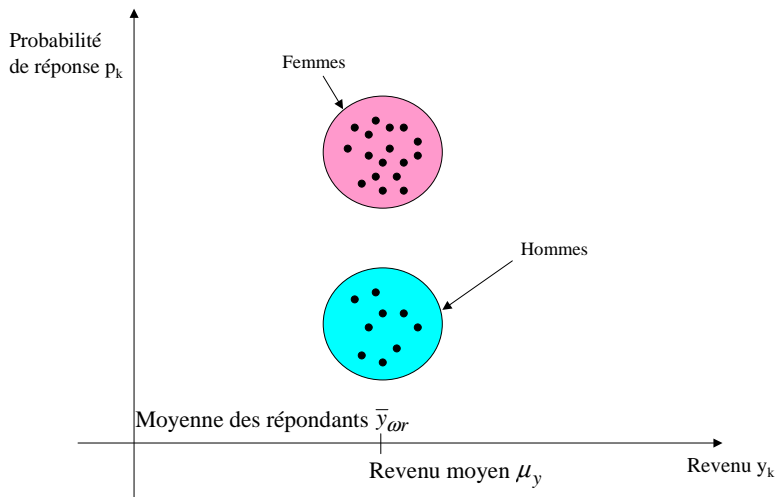
Cette hypothèse est souvent peu réaliste. L'estimateur \hat{t}_{yI} sera également approximativement non biaisé, si le comportement moyen des individus de S_r ne diffère pas du comportement moyen des individus de S , par rapport à la variable y .

Ce sera le cas si le mécanisme de non-réponse partielle ne conduit pas à un échantillon de répondants S_r ayant un comportement particulier relativement à y .

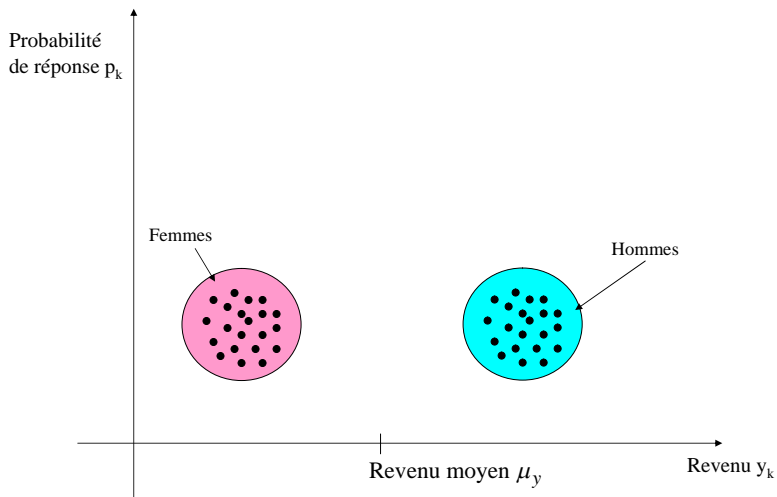
Cas favorable 1



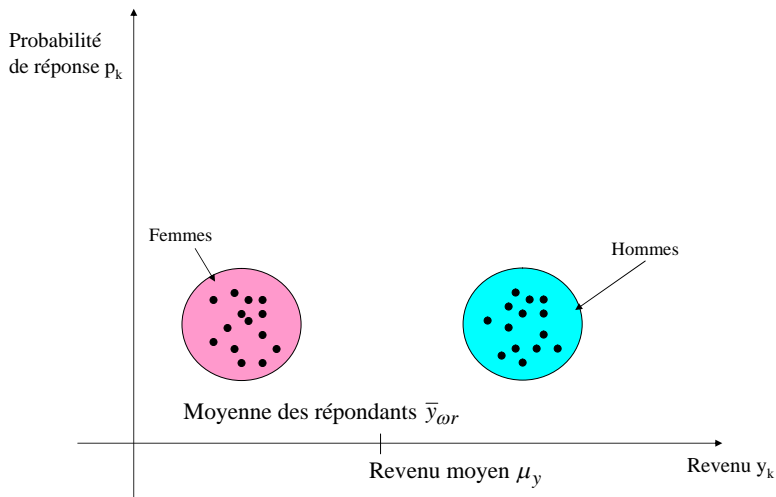
Cas favorable 1 (suite)



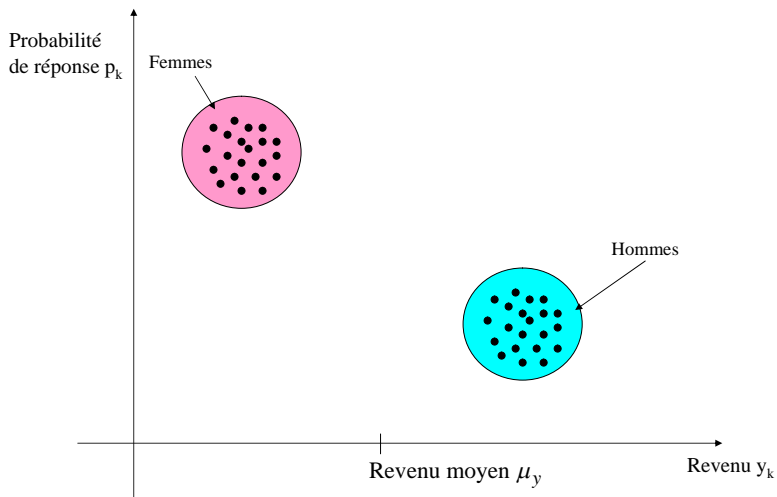
Cas favorable 2



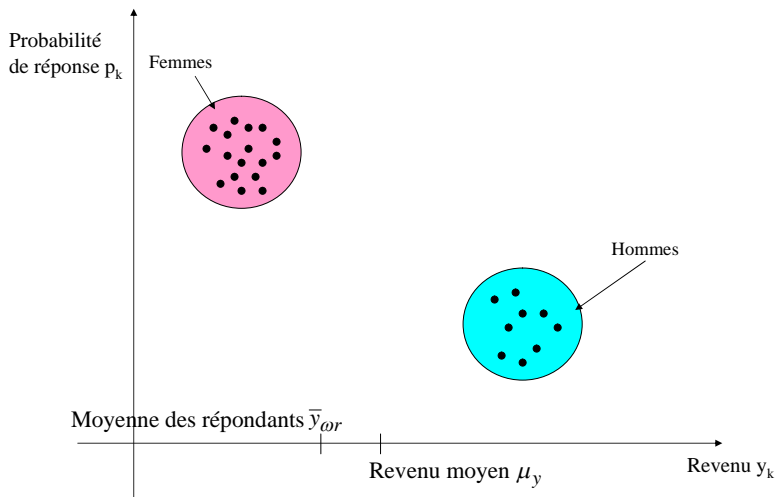
Cas favorable 2 (suite)



Cas défavorable



Cas défavorable (suite)



Imputation par la moyenne

L'imputation par la moyenne conduit donc à une estimation approximativement non biaisée du total :

- soit si tous les individus de l'échantillon sont peu différents par rapport à la variable d'intérêt,
- soit si tous les individus de l'échantillon ont des probabilités de réponse voisines.

En pratique, ces hypothèses sont rarement vérifiées sur l'ensemble de l'échantillon. On peut en revanche essayer de partitionner l'échantillon en classes S_1, \dots, S_H de façon à ce que au sein de chaque classe :

- les individus soient peu différents par rapport à y ;
- ou : les probabilités de réponse soient voisines.

Imputation par la moyenne dans des classes

On parle d'**imputation par la moyenne dans les classes d'imputation**. Cette méthode s'appuie sur le modèle

$$m : y_k = \beta_h + \sigma_h \epsilon_k \quad \text{pour } k \in S_h. \quad (5)$$

Exemple : imputation de la variable revenu par la moyenne, dans des classes définies selon le sexe.

Pour un individu k non-répondant de la classe S_h , on obtient $y_k^* = \hat{\beta}_{rh}$ avec

$$\hat{\beta}_{rh} = \frac{\sum_{k \in S_{rh}} \omega_k y_k}{\sum_{k \in S_{rh}} \omega_k} \equiv \bar{y}_{wrh},$$

en notant $S_{rh} = S_h \cap S_r$.

Construction des classes d'imputation

En pratique, on peut constituer les classes d'imputation de la façon suivante :

- 1 soit en modélisant la variable y :
 - on effectue une régression afin d'obtenir une prédiction \hat{y}_k de y_k , en fonction de l'information auxiliaire disponible.
 - on constitue les classes d'imputation en ordonnant les individus selon les \hat{y}_k , ou en croisant les variables qui ressortent de façon significative.
- 2 soit en modélisant la probabilité de réponse à la variable y :
 - on effectue une régression logistique afin d'obtenir une prédiction des probabilités de réponse \hat{p}_{yk} .
 - on constitue les classes d'imputation en ordonnant les individus selon les \hat{p}_{yk} , ou en croisant les variables qui ressortent de façon significative.

Imputation par hot-deck

La méthode du **hot-deck** consiste à remplacer une valeur manquante y_k en sélectionnant au hasard et avec remise un donneur $y_j \in S_r$, avec des probabilités proportionnelles aux poids d'imputation ω_j .

C'est la version aléatoire de l'imputation par la moyenne. Elle s'appuie sur le même modèle d'imputation : on suppose que les individus de la population ont en moyenne le même comportement par rapport à la variable y .

Le hot-deck a l'avantage d'aller chercher une valeur effectivement observée : en particulier, la méthode est applicable pour une variable catégorielle. En revanche, il s'agit d'une méthode d'**imputation aléatoire** : elle conduit donc à une augmentation de la variance.

Imputation par hot-deck dans des classes

Comme l'imputation par la moyenne, l'imputation par hot-deck est généralement réalisée au sein de classes d'imputation : une valeur manquante y_k est remplacée en sélectionnant au hasard un donneur parmi les répondants de la même classe.

Le modèle d'imputation est le même que pour l'imputation par la moyenne dans des classes. L'estimateur imputé sera approximativement non biaisé :

- si les individus d'une même classe sont peu différents par rapport à y ;
- ou : si les probabilités de réponse sont voisines au sein d'une même classe.

Imputation par donneur

Le hot-deck est un cas particulier des méthodes d'imputation par donneur. On peut également utiliser :

- **l'imputation par la valeur précédente** : une valeur manquante $y_{k,t}$ est remplacée par la valeur observée à une date précédente $y_{k,t-1}$,
⇒ efficace si la variable mesurée évolue peu dans le temps,
- **l'imputation par le plus proche voisin** : une valeur manquante y_k est remplacée en choisissant le donneur le plus proche du non-répondant k , au sens d'une fonction de distance à définir (en fonction des variables auxiliaires disponibles)

Imputation par donneur

Les méthodes par donneurs ont l'avantage

- d'imputer des valeurs effectivement observées,
- de pouvoir être utilisées pour les variables catégorielles,
- de permettre d'imputer plusieurs variables à la fois (aide à préserver le lien entre les variables).

Pour plus de détails sur les méthodes d'imputation possibles, voir Haziza (2009,2011).

Quelle méthode d'imputation utiliser?

Dans le cas que l'on considère ici (estimation d'un total), les méthodes d'imputation déterministes sont préférables car elles ne conduisent pas à une augmentation de la variance. Si le modèle d'imputation est correctement spécifié, l'imputation conduira à une estimation approximativement non biaisée du total.

Dans le cas général, la méthode d'imputation utilisée dépend du type de variable (quanti/quali), et de l'analyse que l'on souhaite faire : estimation d'un total, calcul d'une régression, d'une médiane, ...

Si on s'intéresse à la distribution de la variable imputée, les méthodes d'imputation déterministes ne sont généralement pas adaptées. Par exemple, l'imputation par la moyenne "écrase" de façon artificielle la variable imputée au niveau de sa valeur moyenne.

Problèmes liés à la non-réponse

L'imputation ne crée pas d'information : elle peut donner une fausse impression de précision, car elle conduit à un fichier de données complet, "comme si" on n'observait aucune non-réponse partielle.

L'imputation tend à perturber les relations entre les variables. Si l'objet de l'analyse est par exemple d'étudier une régression entre deux variables, l'imputation des données manquantes doit être réalisée de façon à préserver la relation entre ces variables.

Bibliographie

- Ardilly, P. (2006), *Les Techniques de Sondage*, Technip, Paris.
- Da Silva, D.N., et Opsomer, J.D. (2006). *A kernel smoothing method to adjust for unit nonresponse in sample surveys*. Canadian Journal of Statistics, 34, 563-579.
- Da Silva, D.N., et Opsomer, J.D. (2009). *Nonparametric propensity weighting for survey nonresponse through local polynomial regression*. Survey Methodology, 35, 165-176.
- Haziza, D. (2009). *Imputation and inference in the presence of missing data*, Handbook of Statistics, vol. 29, chap. 10.
- Haziza, D. (2011). *Traitement de la non-réponse totale et partielle dans les enquêtes*. FCDA, Ensaï.
- Haziza, D., et Rao, J.N.K. (2003). *Inference for population means under unweighted imputation for missing survey data*. Survey Methodology, 29, 81-90.
- Joinville, O. (2002). *Mise en oeuvre du logiciel POULPE pour estimer la précision de l'enquête HID*, Rapport de Stage de 2nde année, ISUP.
- Skinner, C.J., et D'Arrigo, J. (2011). *Inverse probability weighting for clustered nonresponse*. Biometrika, 98, 953-966.