Nº 12-001-X au catalogue
ISSN 1712-5685

Techniques d'enquête

Quelques remarques sur un petit exemple de Jean-Claude Deville au sujet de la non-réponse non-ignorable

par Yves Tillé

Date de diffusion : le 20 décembre 2016

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à www.statcan.gc.ca.

Vous pouvez également communiquer avec nous par :

Courriel à STATCAN.infostats-infostats.STATCAN@canada.ca

Téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros sans frais suivants :

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-877-287-4369

Programme des services de dépôt

- Service de renseignements 1-800-635-7943
- Télécopieur 1-800-565-7757

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site www.statcan.gc.ca sous « Contactez-nous » > « Normes de service à la clientèle ».

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population du Canada, les entreprises, les administrations et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques exactes et actuelles.

Signes conventionnels dans les tableaux

Les signes conventionnels suivants sont employés dans les publications de Statistique Canada :

.	indisponible pour toute période de référence
..	indisponible pour une période de référence précise
...	n'ayant pas lieu de figurer
0	zéro absolu ou valeur arrondie à zéro
0^s	valeur arrondie à 0 (zéro) là où il y a une distinction importante entre le zéro absolu et la valeur arrondie
^{p}	provisoire
^{r}	révisé
x	confidentiel en vertu des dispositions de la *Loi sur la statistique*
^{E}	à utiliser avec prudence
F	trop peu fiable pour être publié
*	valeur significativement différente de l'estimation pour la catégorie de référence (p<0,05)

Techniques d'enquête, décembre 2016
Vol. 42, nº 2, pp. 325-335
Statistique Canada, nº 12-001-X au catalogue

Quelques remarques sur un petit exemple de Jean-Claude Deville au sujet de la non-réponse non-ignorable

Yves Tillé[1]

Résumé

Un exemple présenté par Jean-Claude Deville en 2005 est soumis à trois méthodes d'estimation : la méthode des moments, la méthode du maximum de vraisemblance et le calage généralisé. Les trois méthodes donnent exactement les mêmes résultats pour les deux modèles de non-réponse. On discute ensuite de la manière de choisir le modèle le plus adéquat.

Mots-clés : Calage; calage généralisé; méthode des moments; vraisemblance.

1 L'exemple de Deville

Lors d'une conférence à l'Université de Neuchâtel, Jean-Claude Deville (2005) a présenté un exemple simple pour illustrer l'intérêt du calage généralisé pour traiter la non-réponse non-ignorable (au sujet du calage généralisé voir Deville 2000, 2002, 2004; Kott 2006; Chang et Kott 2008; Kott et Chang 2010; Lesage et Haziza 2015). Cet exemple est recopié ci-dessous dans son intégralité.

> *« Les corrections destinées à compenser les effets de la non-réponse demandent une connaissance très précise des facteurs qui la causent. En particulier, si ce que l'on veut mesurer influe directement sur la probabilité de réponse, on est amené à prendre des risques avec les données. Voici un petit exemple fictif : un groupe d'étudiants est interrogé sur sa consommation de drogue. Les résultats de l'enquête sont les suivants :*

Tableau 1.1
Exemple de Deville

	OUI	NON	NON-RÉPONSE	ENSEMBLE
Garçons	40	80	180	300
Filles	20	160	120	300
Ensemble	60	240	300	600

> *Naïvement on dirait que le pourcentage de consommateurs est estimé par 60/(240+60)=25 %. Cette estimation est faite sous l'hypothèse que les non-répondants ont le même comportement que les répondants. Mais on remarque que le taux de réponse des filles est plus important que celui des garçons. Pour corriger cela, on calcule le taux de consommateurs chez les filles, soit 1/9, et chez les garçons soit 3/9, et on conclut que la population étudiante observée est*

1. Yves Tillé, Institut de Statistique, Université de Neuchâtel, Rue de Bellevaux, 51, 2000 Neuchâtel, Suisse. Courriel : yves.tille@unine.ch.

consommatrice à 2/9=22,2 %. Si maintenant on pense que c'est le fait de consommer qui induit la non-réponse, le modèle a deux paramètres p_{oui} *et* p_{non}*, respectivement probabilité de répondre des consommateurs et des non-consommateurs. On trouve que ces probabilités valent respectivement 0,2 et 0,8. Le nombre estimé de consommateurs est donc de 200 chez les garçons et 100 chez les filles et l'estimation du pourcentage global est de 50 %! ».*

L'exemple est a priori très simple et éclaire parfaitement la typologie habituelle des trois mécanismes de non-réponse. Chacune des trois estimations proposées dans l'exemple correspond à l'une de ces trois catégories.

- MCAR (*Missing Completely at Random*) : la probabilité de réponse ne dépend ni de la variable d'intérêt (le fait de consommer de la drogue), ni de la variable auxiliaire (le sexe).
- MAR (*Missing at Random*) : la probabilité de réponse ne dépend pas de la variable d'intérêt y après avoir conditionné sur la variable auxiliaire x (le sexe). Dans le cas présent, la probabilité de réponse ne dépendrait alors que du sexe.
- NMAR (*Not Missing at Random*) : la probabilité de réponse dépend de la variable d'intérêt elle-même (le fait de consommer de la drogue) même si l'on prend en compte la variable auxiliaire x.

L'exemple montre l'intérêt du calage généralisé qui permet de traiter directement le cas NMAR. Jean-Claude Deville traite le problème en considérant que les probabilités p_{oui} et p_{non} sont des paramètres à estimer. Cet exemple peut être traité de multiples façons selon le point de vue que l'on a sur l'inférence.

Dans ce qui suit, nous montrons qu'il existe au moins trois méthodes pour traiter ce problème : la méthode des moments, la méthode du maximum de vraisemblance et le calage. La méthode du maximum de vraisemblance n'a pas été traitée par Jean-Claude Deville. Nous développons complètement les calculs pour les deux premières méthodes d'estimation en considérant les deux modèles. Nous calculons aussi les résultats pour le calage et le calage généralisé.

Nous montrons que les trois résultats obtenus sont identiques. La fonction de vraisemblance estimée pourrait être utilisée pour réaliser un choix entre les deux modèles. Malheureusement, cette fonction prend la même valeur pour les deux modèles, ce qui ne permet pas de choisir le modèle. Nous proposons cependant une piste pour réaliser un choix.

Dans la section 2, on présente la notation utilisée. La section 3 est consacrée à l'estimation par la méthode des moments et la section 4 à l'estimation par la méthode du maximum de vraisemblance. Dans la section 5, on applique les méthodes de calage et de calage généralisé. On termine par une discussion sur les intérêts respectifs des différentes méthodes en section 6.

2 Notation

Le tableau 2.1 contient la notation pour le tableau 1.1.

Tableau 2.1
Notation pour le tableau 1.1

	Drogué	Non Drogué	Manquant	Total
Homme	r_{HD}	r_{HS}	m_H	$n_{H.}$
Femme	r_{FD}	r_{FS}	m_F	$n_{F.}$
Total	$r_{.D}$	$r_{.S}$	m	n

On supposera, par simplicité, que l'on est face à un recensement. Autrement dit, les 600 étudiants n'ont pas été sélectionnés de manière aléatoire. La seule source d'aléa est donc le mécanisme de non-réponse. Cette hypothèse n'est pas tellement restrictive, car elle est équivalente à considérer que l'échantillon est aléatoire mais que les raisonnements qui suivent sont réalisés conditionnellement à l'échantillon aléatoire. L'objectif est d'estimer le tableau 2.2 d'effectifs. Ce tableau est donc supposé ne pas être aléatoire. Il s'agit donc de distribuer les non-répondants m_H et m_F en consommateurs de drogue ou non-consommateurs.

Tableau 2.2
Effectifs à estimer à partir du tableau 1.1

	Drogué	Non Drogué	Total
Homme	n_{HD}	n_{HS}	$n_{H.}$
Femme	n_{FD}	n_{FS}	$n_{F.}$
Total	$n_{.D}$	$n_{.S}$	n

Par ailleurs, on suppose que la non-réponse suit un plan de Poisson, autrement dit chaque individu décide de répondre ou non avec une probabilité fixée indépendamment des autres individus. La probabilité de réponse peut, quant à elle, varier d'un individu à l'autre.

Les deux vecteurs (r_{HD}, r_{HS}, m_H), et (r_{FD}, r_{FS}, m_F) suivent chacun une loi multinomiale dont les paramètres dépendent du modèle utilisé. Le cas MCAR, qui est complètement trivial, ne sera pas étudié. Dans le tableau 2.3 qui représente le cas MAR, la probabilité de réponse ne dépend que du sexe (p_H pour les hommes, p_F pour les femmes). Dans le tableau 2.4 qui représente le cas NMAR, la probabilité de réponse ne dépend que du fait de consommer de la drogue ou pas (q_D pour les drogués, q_S pour les autres).

Tableau 2.3
Cas 1 : Modèle MAR, la non-réponse dépend du sexe

	Drogué	Non Drogué	Manquant	Total
Homme	$\mathrm{E}(r_{HD}) = n_{HD} p_H$	$\mathrm{E}(r_{HS}) = n_{HS} p_H$	$\mathrm{E}(m_H) = n_{H.}(1 - p_H)$	$n_{H.}$
Femme	$\mathrm{E}(r_{FD}) = n_{FD} p_F$	$\mathrm{E}(r_{FS}) = n_{FS} p_F$	$\mathrm{E}(m_F) = n_{F.}(1 - p_F)$	$n_{F.}$
Total	$\mathrm{E}(r_{.D})$	$\mathrm{E}(r_{.S})$	m	n

Tableau 2.4
Cas 2 : Modèle NMAR, la non-réponse dépend du fait de se droguer ou non

	Drogué	Non Drogué	Manquant	Total
Homme	$\mathrm{E}(r_{HD}) = n_{HD}q_D$	$\mathrm{E}(r_{HS}) = n_{HS}q_S$	$\mathrm{E}(m_H) = n_{HD}(1-q_D) + n_{HS}(1-q_S)$	$n_{H.}$
Femme	$\mathrm{E}(r_{FD}) = n_{FD}q_D$	$\mathrm{E}(r_{FS}) = n_{FS}q_S$	$\mathrm{E}(m_F) = n_{FD}(1-q_D) + n_{FS}(1-q_S)$	$n_{F.}$
Total	$\mathrm{E}(r_{.D})$	$\mathrm{E}(r_{.S})$	m	n

3 Estimation par la méthode des moments

3.1 Cas MAR

La méthode des moments permet une estimation rapide des paramètres. Pour le cas MAR, on obtient de la troisième colonne du tableau 2.3 les équations :

$$\mathrm{E}(m_H) = n_{H.}(1-p_H),$$
$$\mathrm{E}(m_F) = n_{F.}(1-p_F),$$

ce qui donne les estimateurs

$$\hat{p}_H = 1 - \frac{m_H}{n_{H.}},$$

$$\hat{p}_F = 1 - \frac{m_F}{n_{F.}},$$

et donc, à partir des deux premières colonnes,

$$\hat{n}_{.D} = \frac{r_{HD}}{\hat{p}_H} + \frac{r_{FD}}{\hat{p}_F} = r_{HD}\frac{n_{H.}}{n_{H.}-m_H} + r_{FD}\frac{n_{F.}}{n_{F.}-m_F},$$

$$\hat{n}_{.S} = \frac{r_{HS}}{\hat{p}_H} + \frac{r_{FS}}{\hat{p}_F} = r_{HS}\frac{n_{H.}}{n_{H.}-m_H} + r_{FS}\frac{n_{F.}}{n_{F.}-m_F}.$$

L'estimation des probabilités de réponse est $\hat{p}_H = 0{,}4$ et $\hat{p}_F = 0{,}6$. On obtient donc l'estimation donnée dans le tableau 3.1.

Tableau 3.1
Estimation : cas MAR

	OUI	NON	ENSEMBLE
Garçons	100,00	200,00	300
Filles	33,33	266,66	300
ENSEMBLE	133,33	466,66	600

3.2 Cas NMAR

Pour le cas NMAR, on obtient du tableau 2.4 les équations :

$$\mathrm{E}(m_H) = E(r_{HD})\frac{1-q_D}{q_D} + \mathrm{E}(r_{HS})\frac{1-q_S}{q_S},$$

$$\mathrm{E}(m_F) = E(r_{FD})\frac{1-q_D}{q_D} + \mathrm{E}(r_{FS})\frac{1-q_S}{q_S}.$$

Après quelques calculs, on obtient les estimateurs suivants pour les probabilités de réponse :

$$\hat{q}_D = \frac{r_{HD}r_{FS} - r_{FD}r_{HS}}{(m_H + r_{HD})r_{FS} - (m_F + r_{FD})r_{HS}},$$

$$\hat{q}_S = \frac{r_{HD}r_{FS} - r_{FD}r_{HS}}{(m_F + r_{FS})r_{HD} - (m_H + r_{HS})r_{FD}}.$$

Finalement, on obtient

$$\hat{n}_{.D} = \frac{r_{.D}}{\hat{q}_D} = r_{.D}\frac{(m_H + r_{HD})r_{FS} - (m_F + r_{FD})r_{HS}}{r_{HD}r_{FS} - r_{FD}r_{HS}} = r_{.D}\frac{n_{H.}r_{FS} - n_{F.}r_{HS}}{r_{HD}r_{FS} - r_{FD}r_{HS}},$$

$$\hat{n}_{.S} = \frac{r_{.S}}{\hat{q}_S} = r_{.S}\frac{(m_F + r_{FS})r_{HD} - (m_H + r_{HS})r_{FD}}{r_{HD}r_{FS} - r_{FD}r_{HS}} = r_{.S}\frac{n_{F.}r_{HD} - n_{H.}r_{FD}}{r_{HD}r_{FS} - r_{FD}r_{HS}}.$$

Comme l'écrit Deville, l'estimation des probabilités de réponse est $\hat{q}_D = 0{,}2$ et $\hat{q}_S = 0{,}8$. On obtient alors l'estimation donnée dans le tableau 3.2.

Tableau 3.2
Estimation : cas NMAR

	OUI	NON	ENSEMBLE
Garçons	200	100	300
Filles	100	200	300
ENSEMBLE	300	300	600

4 Estimation par la méthode du maximum de vraisemblance

4.1 Cas MAR

La distribution de probabilité est multinomiale. Dans le cas MAR, la fonction de vraisemblance vaut :

$$\mathcal{L}(n_{HD}, n_{FD}, p_H, p_F) = \frac{n_{H.}!}{r_{HD}!\, r_{HS}!\, m_H!}\left(\frac{n_{HD}p_H}{n_{H.}}\right)^{r_{HD}}\left(\frac{(n_{H.} - n_{HD})p_H}{n_{H.}}\right)^{r_{HS}}\left(\frac{n_{H.}(1-p_H)}{n_{H.}}\right)^{m_H}$$
$$\times\frac{n_{F.}!}{r_{FD}!\, r_{FS}!\, m_F!}\left(\frac{n_{FD}p_F}{n_{F.}}\right)^{r_{FD}}\left(\frac{(n_{F.} - n_{FD})p_F}{n_{F.}}\right)^{r_{FS}}\left(\frac{n_{F.}(1-p_F)}{n_{F.}}\right)^{m_F}.$$

En annulant les dérivées partielles de la log-vraisemblance par rapport aux paramètres p_H et p_F, on obtient deux équations à deux inconnues. La solution donne les estimateurs

$$\hat{p}_H = 1 - \frac{m_H}{n_{H.}},$$

$$\hat{p}_F = 1 - \frac{m_F}{n_{F.}}.$$

En annulant les dérivées par rapport à n_{HD} et n_{FD}, on obtient les estimateurs

$$\hat{n}_{HD} = \frac{r_{HD}}{\hat{p}_H} \text{ et } \hat{n}_{FD} = \frac{r_{FD}}{\hat{p}_F}.$$

Donc,

$$\hat{n}_{.D} = \hat{n}_{HD} + \hat{n}_{FD} = \frac{r_{HD}}{\hat{p}_H} + \frac{r_{FD}}{\hat{p}_F}.$$

Ces estimateurs sont exactement les mêmes que ceux obtenus par la méthode des moments.

4.2 Cas NMAR

Dans le cas NMAR, la fonction de vraisemblance vaut :

$$\mathcal{L}(n_{HD}, n_{FD}, q_D, p_S) = \frac{n_{H.}!}{r_{HD}!\, r_{HS}!\, m_H!}\left(\frac{n_{HD} q_D}{n_{H.}}\right)^{r_{HD}}\left(\frac{(n_{H.} - n_{HD})\, q_S}{n_{H.}}\right)^{r_{HS}}\left(\frac{n_{HD}(1-q_D) + (n_{H.} - n_{HD})(1-q_S)}{n_{H.}}\right)^{m_H}$$
$$\times \frac{n_{F.}!}{r_{FD}!\, r_{FS}!\, m_F!}\left(\frac{n_{FD} q_D}{n_{F.}}\right)^{r_{FD}}\left(\frac{(n_{F.} - n_{FD})\, q_S}{n_{F.}}\right)^{r_{FS}}\left(\frac{n_{FD}(1-q_D) + (n_{F.} - n_{FD})(1-q_S)}{n_{F.}}\right)^{m_F}.$$

En annulant les dérivées partielles de la log-vraisemblance par rapport aux quatre paramètres q_D, q_S, n_{HD} et n_{FD}, on obtient un système de quatre équations de degré deux assez compliqué à quatre inconnues. Nous avons vérifié au moyen d'un logiciel de calcul symbolique que la solution donnée par la méthode des moments est une solution de ce système d'équations. Évidemment, comme le système est de degré deux, il existe une seconde solution. Cependant cette seconde solution donne, pour l'exemple de Deville, des valeurs négatives qui ne sont pas acceptables pour estimer des probabilités et des effectifs.

5 Estimation par calage et calage généralisé

5.1 Notation

Pour définir le calage, nous allons définir la notation suivante. Soit $U = \{1, \ldots, k, \ldots, N\}$ l'ensemble des personnes interrogées (ici $N = 600$) et $R \subset U$ l'ensemble des répondants à la question concernant la consommation de drogue. On définit également

$$\mathbf{x}_k = \begin{cases} (1 \quad 0)^{\mathrm{T}} & \text{si l'individu } k \text{ est un homme} \\ (0 \quad 1)^{\mathrm{T}} & \text{si l'individu } k \text{ est une femme.} \end{cases}$$

et

$$\mathbf{z}_k = \begin{cases} (1 \quad 0)^{\mathrm{T}} & \text{si l'individu } k \text{ a répondu qu'il consomme de la drogue} \\ (0 \quad 1)^{\mathrm{T}} & \text{si l'individu } k \text{ a répondu qu'il ne consomme pas de la drogue.} \end{cases}$$

En utilisant la notation définie précédemment,

$$\sum_{k \in U} \mathbf{x}_k = \begin{pmatrix} n_{H.} \\ n_{F.} \end{pmatrix}, \quad \sum_{k \in R} \mathbf{x}_k = \begin{pmatrix} n_{H.} - m_H \\ n_{F.} - m_F \end{pmatrix}, \quad \sum_{k \in R} \mathbf{z}_k = \begin{pmatrix} r_{.D} \\ r_{.S} \end{pmatrix},$$

$$\sum_{k \in R} \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}} = \begin{pmatrix} n_{H.} - m_H & 0 \\ 0 & n_{F.} - m_F \end{pmatrix}, \quad \sum_{k \in R} \mathbf{x}_k \mathbf{z}_k^{\mathrm{T}} = \begin{pmatrix} r_{HD} & r_{HS} \\ r_{FD} & r_{FS} \end{pmatrix},$$

et

$$\sum_{k \in R} \mathbf{z}_k \mathbf{z}_k^{\mathrm{T}} = \begin{pmatrix} r_{.D} & 0 \\ 0 & r_{.S} \end{pmatrix}.$$

5.2 Estimation par calage simple

En utilisant le calage simple tel qu'il est décrit dans Deville et Särndal (1992), on cherche un poids qui s'écrit

$$w_k = F\left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}\right),$$

où $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ est un vecteur de paramètres et $F(.)$ est une fonction de calage, c'est-à-dire une fonction strictement croissante, telle que $F(0) = 1$ et dont la dérivée $F'(.)$ est telle que $F'(0) = 1$.

Le vecteur $\boldsymbol{\lambda}$ est identifié en résolvant par la méthode de Newton le système d'équation

$$\sum_{k \in R} F\left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}\right) \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k. \tag{5.1}$$

Finalement, l'estimateur par calage est donné par

$$\begin{pmatrix} \hat{n}_{.D} \\ \hat{n}_{.S} \end{pmatrix} = \sum_{k \in R} w_k \mathbf{z}_k.$$

Dans notre application, l'équation (5.1) devient

$$\sum_{k \in R} F\left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}\right) \mathbf{x}_k = \begin{pmatrix} (n_{H.} - m_H) F(\lambda_1) \\ (n_{F.} - m_F) F(\lambda_2) \end{pmatrix} = \sum_{k \in U} \mathbf{x}_k = \begin{pmatrix} n_{H.} \\ n_{F.} \end{pmatrix}.$$

On obtient directement que

$$w_k = F\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) = \begin{cases} n_{H.}/(n_{H.} - m_H) & \text{si l'individu } k \text{ est un homme} \\ n_{F.}/(n_{F.} - m_F) & \text{si l'individu } k \text{ est une femme.} \end{cases}$$

Donc, les estimateurs calés sont

$$\hat{n}_{.D} = r_{HD}\frac{n_{H.}}{n_{H.} - m_H} + r_{FD}\frac{n_{F.}}{n_{F.} - m_F}$$

$$\hat{n}_{.S} = r_{HS}\frac{n_{H.}}{n_{H.} - m_H} + r_{FS}\frac{n_{F.}}{n_{F.} - m_F},$$

ce qui est exactement le même résultat que ceux donnés par les méthodes des moments et du maximum de vraisemblance. Dans ce cas, la solution ne dépend pas de la fonction de calage utilisée. Évidemment, l'exemple est particulièrement simple. Dans tous les cas plus complexes que la définition de catégories ne se chevauchant pas, le résultat dépend de la fonction de calage utilisée.

5.3 Calage généralisé

Dans le calage généralisé tel qu'il est défini dans (Deville 2000, 2002, 2004; Kott 2006), les poids s'écrivent

$$w_k = F\left(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}\right).$$

Le vecteur $\boldsymbol{\lambda}$ est identifié en résolvant le système d'équation

$$\sum_{k\in R} F\left(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)\mathbf{x}_k = \sum_{k\in U}\mathbf{x}_k. \tag{5.2}$$

Enfin, l'estimateur par calage généralisé est donné par

$$\begin{pmatrix} \hat{n}_{.D} \\ \hat{n}_{.S} \end{pmatrix} = \sum_{k\in R} w_k \mathbf{z}_k.$$

Dans notre application, l'équation (5.2) devient :

$$\sum_{k\in R} F\left(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)\mathbf{x}_k = \begin{pmatrix} r_{HD}F(\lambda_1) + r_{HS}F(\lambda_2) \\ r_{FD}F(\lambda_1) + r_{FS}F(\lambda_2) \end{pmatrix} = \sum_{k\in U}\mathbf{x}_k = \begin{pmatrix} n_{H.} \\ n_{F.} \end{pmatrix},$$

ce qui peut s'écrire de manière matricielle

$$\begin{pmatrix} r_{HD} & r_{HS} \\ r_{FD} & r_{FS} \end{pmatrix}\begin{pmatrix} F(\lambda_1) \\ F(\lambda_2) \end{pmatrix} = \begin{pmatrix} n_{H.} \\ n_{F.} \end{pmatrix}.$$

On résout simplement ce système linéaire

$$\begin{pmatrix} F(\lambda_1) \\ F(\lambda_2) \end{pmatrix} = \begin{pmatrix} r_{HD} & r_{HS} \\ r_{FD} & r_{FS} \end{pmatrix}^{-1} \begin{pmatrix} n_{H.} \\ n_{F.} \end{pmatrix} = \begin{pmatrix} \dfrac{n_{H.}r_{FS} - n_{F.}r_{HS}}{r_{FS}r_{HD} - r_{FD}r_{HS}} \\ \dfrac{n_{H.}r_{FD} - n_{F.}r_{HD}}{r_{FD}r_{HS} - r_{FS}r_{HD}} \end{pmatrix}.$$

Les estimateurs sont donc :

$$\hat{n}_{.D} = r_{.D} \frac{n_{H.}r_{FS} - n_{F.}r_{HS}}{r_{FS}r_{HD} - r_{FD}r_{HS}}$$
$$\hat{n}_{.S} = r_{.S} \frac{n_{H.}r_{FD} - n_{F.}r_{HD}}{r_{FD}r_{HS} - r_{FS}r_{HD}}.$$

À nouveau, la solution ne dépend pas de la fonction de calage utilisée. La solution est identique à la solution obtenue par les méthodes des moments et du maximum de vraisemblance. Ici également cette propriété découle de la simplicité de l'exemple. Dans tous les cas plus complexes, le résultat dépend de la fonction de calage utilisée.

6 Discussion

L'exemple de Deville est particulièrement heureux, car pour les deux modèles, les trois méthodes d'estimation fournissent exactement les mêmes estimateurs. Évidemment, si le modèle est plus compliqué, l'usage de la méthode du maximum de vraisemblance devient laborieux voire impossible. La méthode de calage et de calage généralisé fonctionne dans tous les cas pour autant que le nombre de variables de calage dont les totaux sont connus soit suffisant et que la matrice

$$\sum_{k \in R} \mathbf{x}_k \mathbf{z}_k^{\mathrm{T}}$$

soit inversible. Dans cet exemple, le déterminant de cette matrice apparaît au dénominateur des estimateurs. Un faible déterminant rend donc les estimations particulièrement hasardeuses. Lesage et Haziza (2015) recommandent par ailleurs de vérifier que les corrélations entre les variables $\mathbf{x}_k$ et $\mathbf{z}_k$ soient suffisamment élevées afin d'éviter une possible amplification du biais.

Si les variables sont quantitatives, les solutions vont dépendre de la fonction de calage utilisée $F(.)$. On préconise l'utilisation de la fonction de calage $F\left(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) = 1 + \exp\left(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)$ qui a l'avantage de fournir des poids supérieurs à 1. L'inverse de ces poids peut dès lors être interprété comme une probabilité de réponse estimée au moyen d'un modèle logistique.

La difficulté principale reste évidemment le choix entre les deux modèles proposés. Dans l'exemple de Deville, on pourrait trouver plus « logique » de voir la non-réponse plutôt dépendre du fait de consommer de la drogue que du sexe. Cependant, on se trouve assez démuni pour établir un choix entre les deux modèles. Les valeurs des deux fonctions de vraisemblance pour les paramètres estimés sont exactement égales. Est-il possible d'aller au-delà de l'intime conviction pour choisir le modèle ? Comme suggéré dans Haziza et

Lesage (2016), nous préconisons dans tous les cas de calculer les deux pondérations et de comparer les poids et les estimations obtenues avec chacune d'elles.

Une piste consiste peut-être à calculer un indice de dispersion des probabilités de réponse comme la variance. En effet, si cette variance est élevée, cela signifie que le modèle a permis de calculer des probabilités de réponse plus contrastée d'un individu à l'autre et donc qu'il a pu mieux prendre en compte cette non-réponse. La validation par recherche de poids contrasté est la base de l'identification des groupes de réponse homogènes pour toutes les méthodes de segmentation par exemple avec l'algorithme CHAID (*Chi-square Automatic Interaction Detector*) développé par Kass (1980). En effet, avec cette méthode, à chaque étape, on scinde les groupes de réponses homogènes selon les catégories qui rendent les probabilités de réponse les plus contrastées. En appliquant ce même principe pour réaliser le choix du modèle, on peut choisir le modèle qui fournit les poids les plus contrastés. En effet, si la variance est faible, cela signifie que le modèle de non-réponse n'a pas pu mettre en évidence des différences de probabilités de non-réponse entre les individus. La variance des probabilités de réponse est par ailleurs le carré du R-indicateur défini par Schouten, Cobben et Bethlehem (2009), utilisé ici pour choisir un modèle de non-réponse.

Dans les deux cas, la moyenne des probabilités de réponse vaut 0,5. En effet,

$$\overline{p} = n_{H.}\frac{n_{H.}\hat{p}_H + n_{F.}\hat{p}_F}{n} = \frac{300\times 0{,}4 + 300\times 0{,}6}{600} = 0{,}5$$

et

$$\overline{q} = \hat{n}_{.D}\frac{n_{.D}\hat{q}_D + \hat{n}_{.S}\hat{q}_S}{n} = \frac{300\times 0{,}2 + 300\times 0{,}8}{600} = 0{,}5.$$

Pour le modèle MAR, la variance vaut

$$V_{MAR} = \frac{n_{H.}\left(\hat{p}_H - \overline{p}\right)^2 + n_{F.}\left(\hat{p}_F - \overline{p}\right)^2}{n} = \frac{300\left(0{,}4-0{,}5\right)^2 + 300\left(0{,}6-0{,}5\right)^2}{600} = 0{,}01.$$

Pour le modèle NMAR, la variance vaut

$$V_{NMAR} = \frac{\hat{n}_{.D}\left(\hat{q}_D - \overline{q}\right)^2 + \hat{n}_{.S}\left(\hat{q}_S - \overline{q}\right)^2}{n} = \frac{300\left(0{,}2-0{,}5\right)^2 + 300\left(0{,}8-0{,}5\right)^2}{600} = 0{,}09.$$

La plus grande variance du modèle NMAR plaide en sa faveur. Les probabilités de réponse sont en effet beaucoup plus contrastées.

Remerciements

L'auteur remercie Audrey-Anne Vallée pour sa lecture méticuleuse d'une version précédente de ce texte et un arbitre anonyme pour ses commentaires particulièrement pertinents.

Bibliographie

Chang, T., et Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.

Deville, J.-C. (2000). Generalized calibration and application to weighting for non-response. Dans *Compstat - Proceedings in Computational Statistics:* 14ième Symposium tenu à Utrecht, Pays-Bas, pages 65-76, New York: Springer.

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. Dans les *Actes des Journées de Méthodologie Statistique*, Paris. Insee-Méthodes.

Deville, J.-C. (2004). Calage, calage généralisé et hypercalage. Rapport technique, document interne, INSEE, Paris.

Deville, J.-C. (2005). Calibration, past, present and future? Présentation à la conférence : *Calibration Tools for Survey Statisticians*, Neuchâtel.

Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Haziza, D., et Lesage, E. (2016). A discussion of weighting procedures for unit nonresponse. Va paraitre dans le *Journal of Official Statistics*.

Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 119-127.

Kott, P.S. (2006). Utilisation de la pondération par calage pour la correction de la non-réponse et des erreurs de couverture. *Techniques d'enquête*, 32, 2, 149-160. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2006002/article/9547-fra.pdf.

Kott, P.S., et Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105(491), 1265-1275.

Lesage, E., et Haziza, D. (2015). On the problem of bias and variance amplification of the instrumental calibration estimator in the presence of unit nonresponse. En révision pour le *Journal of Survey Statistics and Methodology*.

Schouten, B., Cobben, F. et Bethlehem, J. (2009). Indicateurs de la représentativité de la réponse aux enquêtes. *Techniques d'enquête*, 35, 1, 107-121. Article accessible à l'adresse http://www.statcan.gc.ca/pub/12-001-x/2009001/article/10887-fra.pdf.