

Méthodes de rééchantillonnage pour l'estimation de variance

A. C. Davison et S. Sardy

*Institut de Mathématiques, Ecole Polytechnique Fédérale de Lausanne,
1015 Lausanne, Suisse*

Anthony.Davison@epfl.ch, Sylvain.Sardy@epfl.ch

4 mai 2006

Resumé

Nous passons en revue des techniques de rééchantillonnage utilisées pour l'estimation de variance en sondage. Les techniques de rééchantillonnage considérées sont basées sur la linéarisation, le jackknife, les répétitions équilibrées répétées, et le bootstrap. Quelques unes des procédures d'estimation de variance utilisent des mécanismes de calibration et d'imputation afin d'améliorer l'estimation. Notre objectif est d'obtenir des conclusions pratiques basées sur des considérations théoriques et des comparaisons empiriques.

Quelques mots clés : bootstrap, calibrage, fonction d'influence empirique, imputation, jackknife, linéarisation, non-réponse, valeurs manquantes.

1 Introduction

Des formules exactes pour la variance des estimateurs, basées sur des données de sondages, sont disponibles seulement pour des classes d'estimateurs limitées, quoique très importantes, de sorte que des approximations sont couramment utilisées dans la pratique. Des approximations classiques sont dérivées à partir de variantes de la méthode delta, pour lesquelles des estimateurs grossiers sont obtenus en utilisant un développement de Taylor. Lorsque l'échantillon est petit ou que l'estimateur est complexe, il est naturel de se poser la question de l'exactitude du développement sur lequel repose la validité des formules résultantes, et de rechercher d'autres approches. Une classe importante d'alternatives est basée sur les procédures de rééchantillonnage, qui impliquent des calculs répétés sur l'échantillon obtenu au départ. La plus importante de celles-ci est le bootstrap, introduit par Efron (1979), mais il y a aussi le jackknife (Quenouille, 1949a; Tukey, 1958), qui fut introduit à la base pour l'estimation du biais dans l'analyse de séries temporelles, et par le semi-échantillonnage (McCarthy, 1969), qui puise ses racines dans des travaux entrepris durant la fin des années 1950 au bureau américain de Census (Hartigan, 1969; Hall, 2003; Wolter, 1985, Section 3.1). Néanmoins, le bootstrap est la procédure la plus flexible et la plus puissante, et pour cette raison se trouve largement utilisé dans les applications.

Le but de cet article est de revoir des travaux publiés sur les applications de ces idées dans le contexte des enquêtes, avec un accent sur les possibilités d'implémentation pratique plutôt que sur la théorie. Le chapitre 2 décrit brièvement les types d'estimateurs considérés et les procédures d'ajustement et d'imputation qui peuvent être utilisées pour les modifier. Le chapitre 3 décrit les approches de linéarisation, et est suivi de descriptions concernant l'application du jackknife, les répétitions équilibrées répétées, et le bootstrap, dans le contexte des sondages. Des comparaisons numériques concernant les principales méthodes de rééchantillonnage sont passées en revue dans le chapitre 7, et nous concluons sur une brève discussion.

2 Préliminaires

2.1 Paramètres et estimateurs

Afin d'établir la notation nous considérons initialement le cas d'une réponse complète pour un schéma d'échantillonnage stratifié de probabilités inégales à un degré sans remise, avec N unités divisées en H strates, à partir duquel n unités sont échantillonnées. Soit n_h le nombre d'unités échantillonnées à partir des N_h unités de population dans la strate h , et soit π_{hi} la probabilité d'inclusion pour l'unité i de cette strate. Dans des enquêtes sur les ménages, cette unité pourrait consister en une grappe d'individus, auquel cas

nous supposons que la réponse de l'unité d'intérêt résume la grappe. Ainsi la taille de la population en nombre d'unités échantillonnées est $N = \sum_{h=1}^H N_h$ et $n = \sum_{h=1}^H n_h$. Soit x_{hi} et y_{hi} les variables qui ont été mesurées sur les unités, où y_{hi} est la réponse d'intérêt scalaire et x_{hi} est un vecteur $q \times 1$ de variables auxiliaires.

Les paramètres de la population finie peuvent être classifiés en deux grands groupes. Le premier groupe, qui est le plus important, contient des quantités qui sont des fonctions lisses de la réponse de la population finie, comme le total

$$\tau = \sum_{h=1}^H \sum_{i=1}^{N_h} y_{hi},$$

le rapport des deux totaux de populations pour deux variables différentes,

$$\psi = \tau_1 / \tau_2, \tag{1}$$

ou le changement dans le rapport entre deux occasions d'échantillonnage différents,

$$\delta(t_1, t_2) = \psi_{t_1} - \psi_{t_2}. \tag{2}$$

L'autre groupe principal de paramètres comprend les fonctions non lisses des réponses de la population finie, comme la médiane,

$$y_{0.5} = \text{median} \{y_{hi} : i = 1, \dots, N_h, h = 1, \dots, H\},$$

d'autres quantiles, et des statistiques qui se basent dessus. Un exemple est $F(ay_b)$, où F est la fonction de distribution du revenu des ménages dans une population; prendre $a = b = 0.5$ donne la proportion de ménages dont le revenu est en-dessous de la moitié de la médiane.

L'estimation des paramètres de la population finie est basée sur les données des n unités échantillonnées et sur leur probabilité d'inclusion dans le plan d'échantillonnage. Le plus important estimateur d'un total est l'estimateur de Horvitz–Thompson

$$\hat{\tau} = \sum_{h=1}^H \sum_{i=1}^{n_h} \omega_{hi} y_{hi} = \omega^T y, \tag{3}$$

où y est le vecteur $n \times 1$ de réponses échantillonnées et ω est le vecteur $n \times 1$ de leurs poids $\omega_{hi} = 1/\pi_{hi}$, les probabilités d'inclusion inverses. La variance exacte de cet estimateur implique les probabilités d'inclusion conjointes pour les différentes unités et est aisément obtenu. Elle peut être utilisée pour développer des formules de variance pour des estimateurs de paramètres tel que (1) et (2) ci-dessus, mais des complications surviennent lorsque les poids eux-mêmes sont aléatoires, ou lorsqu'une partie de la réponse n'est pas disponible.

2.2 Calibration

Dans beaucoup de cas les totaux de la population sont connus pour certaines des variables auxiliaires x , et cette information peut être utilisée pour augmenter la précision de l'estimation. Supposons que q_C marges des q variables auxiliaires soient connues, avec $q_C \leq q$, soit c le vecteur $q_C \times 1$ de marges connues, et soit X_C la matrice de variables auxiliaires $n \times q$ dont les marginales c pour la population entière sont connues. Utilisant l'estimation d'un total pour illustrer l'idée de la calibration, la qualité de l'estimateur de Horvitz–Thompson peut être amélioré en calibrant les poids w_{hi} de sorte à être aussi près que possible des poids originaux ω dans une certaine métrique G , sous les contraintes que les variables auxiliaires pondérées concordent avec les marges (Deville et Särndal, 1992). Une formulation mathématique de ceci est le problème d'optimisation

$$\min_{w_{hi}} \sum_{h=1}^H \sum_{i=1}^{n_h} \omega_{hi} G(w_{hi}/\omega_{hi}), \text{ tel que } \sum_{h=1}^H \sum_{i=1}^{n_h} w_{hi} x_{Chi} = c.$$

Une mesure de distance commune utilisée dans la pratique est la métrique ℓ_2 , $G(x) = (x - 1)^2/2$; dans ce cas les poids calibrés valent

$$w = \omega + \Omega X_C (X_C^T \Omega X_C)^{-1} (c - X_C^T \omega), \quad (4)$$

où Ω désigne la matrice diagonale dont les éléments sont les ω_{hi} . Ainsi l'estimateur calibré de Horvitz–Thompson pour le total s'écrit

$$\begin{aligned} \hat{\tau} &= w^T y = \omega^T y + (c - X_C^T \omega)^T (X_C^T \Omega X_C)^{-1} X_C^T \Omega y \\ &= \omega^T y + (c - X_C^T \omega)^T \hat{\gamma}, \end{aligned} \quad (5)$$

$\hat{\gamma}$ étant l'estimateur de régression lorsqu'on fait une régression de y sur X_C avec poids Ω . Le deuxième terme de $\hat{\tau}$ est un ajustement aux poids ω qui tient compte de la différence entre $X_C^T \omega$ et de la valeur de la population cible c des marges.

Un grand nombre d'autres mesures de distance entre w et ω peut être envisagé. Par exemple, la mesure multiplicative $G(x) = x \log x - x + 1$ (Deville *et al.*, 1993) garantit que les poids calibrés, comme les poids originaux, sont positifs. Cependant les résultats, en utilisant diverses métriques G , sont asymptotiquement équivalents (Deville et Särndal, 1992), et en pratique Deville *et al.* (1993) ont observé que, bien que les poids individuels peuvent varier beaucoup en fonction des méthodes, le choix de G a un impact relativement faible sur l'estimation ponctuelle et sur l'estimation de variance. Bien que les hypothèses à partir desquelles on déduit l'équivalence asymptotique de ces méthodes ne sont souvent pas vérifiées en pratique, l'effet sur la conclusion globale de la méthode de calibration est généralement petit. Etant donné que le calcul de poids pour des mesures plus compliquées, telles que la mesure multiplicative, nécessite plus d'effort de calcul,

nos recherches se concentrent surtout sur les poids calibrés (4) avec la métrique ℓ_2 , ce qui nécessite seulement un calcul de régression. Quelques une des études numériques décrites ci-dessous ont utilisé d’autres formes de calibration, tel que l’ajustement itératif proportionnel, mais les conclusions sont très proches de celles utilisant la métrique ℓ_2 .

2.3 Imputation

Dans la pratique l’application de méthodes d’enquêtes amène rarement à des jeux de données complets. Il existe deux types principaux de non-réponses, la non-réponse de type unité (‘unit non-response’ en anglais), pour laquelle toutes les données pour une unité échantillonnée manquent, et la non-réponse de type article (‘article non-response’), pour laquelle une partie mais pas toutes les données pour l’unité manquent.

La non-réponse unité est souvent traitée par calibration ou toute autre forme d’ajustement de poids, mais la non-réponse article impute les données manquantes en utilisant un modèle d’imputation. Ce modèle est utilisé pour prédire les réponses manquantes pour des unités avec des non-réponse article, et est généralement construit sur la base des données pour ces réponses avec les données complètes. Il sera rarement le même que le mécanisme qui a engendré les non-réponses, et il peut donc y avoir du biais. L’imputation est dite propre (Rubin, 1996) si les estimateurs ponctuels et de variance sont approximativement non biaisés pour leurs analogues avec données complètes.

Les modèles d’imputation et les valeurs imputées sont de deux types : déterministe, ce qui signifie que le modèle impute une fonction déterministe des données observées ; et stochastique, ce qui signifie que le modèle impute une fonction aléatoire des données observées. L’imputation de type ‘hot deck’, où les données provenant d’unités qui ne répondent pas sont remplacées par des valeurs échantillonnées à partir d’unités qui répondent choisie de sorte à donner une bonne concordance avec celles qui manquent — par exemple, à partir de la même strate — est une forme d’imputation stochastique.

Une approche déterministe répandue pour l’imputation de réponses manquantes basée sur les vecteurs correspondants de variables auxiliaires x_{hi} est d’utiliser un modèle linéaire ou l’une de ses généralisations, tel que la régression robuste ou logistique. Tous ces estimateurs sont des M-estimateurs, et les équations normales correspondantes pour une estimation des paramètres θ du modèle d’imputation choisi à travers les strates peuvent s’écrire comme une équation vectorielle

$$\sum_{h=1}^H \sum_{i=1}^{n_h} x_{hi} \psi(y_{hi}, x_{hi}; \theta) = 0, \quad (6)$$

où ψ est la dérivée de la fonction de perte, dont la forme dépend de ce modèle ; pour la

régression logistique les y_{hi} sont des variables indicatrices binaires et

$$\psi(y, x; \theta) = y - \frac{\exp(x^\top \theta)}{1 + \exp(x^\top \theta)}.$$

Posant $\psi(u) = u$ revient à considérer la fonction de perte ℓ_2 . Un autre cas particulier est l'imputation de rapport utilisant un scalaire x , obtenu en posant $\psi(y, x; \theta) = y - \theta x$. Un modèle d'imputation plus robuste est obtenu en utilisant la fonction de perte robustifiée (Huber, 1981)

$$\begin{cases} u^2/2, & |u| \leq \xi, \\ \xi |u| - \xi^2/2, & |u| > \xi, \end{cases}, \quad \xi > 0,$$

dont la dérivée est $\psi(u) = \text{sign}(u) \min(|u|, \xi)$. La quantité ξ contrôle la robustesse de l'ajustement, avec $\xi \rightarrow \infty$ recouvrant le fragile estimateur de moindres carrés, et $\xi \rightarrow 0$ donnant une plus grande robustesse. Le paramètre d'échelle σ peut être estimé en utilisant une équation d'estimation simultanée, ou séparément ; ceci est rarement crucial. Une fois que le M-estimateur du modèle linéaire $\hat{\beta}$ de β a été trouvé, la réponse manquante pour un individu avec variable explicative x peut être prédite par $x^\top \hat{\beta}$. Des explications similaires s'appliquent à la régression logistique, qu'on utiliserait pour prédire la valeur manquante d'une réponse binaire.

Pour les calculs numériques ci-dessous nous supposons ou bien un modèle linéaire commun à travers les strates,

$$y_{hi} = x_{hi}^\top \beta + \epsilon_{hi}, \quad (7)$$

ou bien un modèle linéaire différent pour chaque strate,

$$y_{hi} = x_{hi}^\top \beta_h + \epsilon_{hi}, \quad h = 1, \dots, H. \quad (8)$$

Nous ajustons alors ce modèle à ces individus pour lesquels à la fois x_{hi} et y_{hi} sont disponibles, et utilisons ensuite le modèle linéaire ajusté pour imputer les réponses prédites pour les individus pour lesquels y_{hi} manque. Soit $z_{hi} = I(y_{hi} \neq \text{NA})$ la variable aléatoire indicatrice correspondant à la réponse observée, soit $Z = \text{diag}(z)$ la matrice diagonale $n \times n$ de ces indicatrices, et soit X la matrice $n \times q$ qui contient les variables auxiliaires correspondantes à la fois aux répondants et aux non-répondants. Soit aussi $\hat{y} = X\hat{\beta}$ le vecteur $n \times 1$ de valeurs ajustées par le modèle de régression utilisé pour l'imputation. Alors (4) implique que l'estimateur calibré et imputé de Horvitz–Thompson peut s'écrire comme

$$\begin{aligned} \hat{\tau} &= w^\top \{Zy + (I - Z)\hat{y}\} = \omega^\top Zy + (c - X_C^\top \omega)^\top (X_C^\top \Omega X_C)^{-1} X_C^\top \Omega Zy \\ &\quad + \omega^\top (I - Z)X\hat{\beta} + (c - X_C^\top \omega)^\top (X_C^\top \Omega X_C)^{-1} X_C^\top \Omega (I - Z)X\hat{\beta}. \end{aligned} \quad (9)$$

Etant donné la discussion précédente, seules des modifications mineures seraient nécessaires si à la place on désirait implémenter des modèles d'imputations basés sur des M-estimateurs.

Le calcul de la variance de l'estimateur calibré et imputé de Horvitz–Thompson, comme si les réponses imputées \hat{y} étaient les vraies réponses, peut conduire à une sous-estimation considérable de la vraie variance, de sorte que la technique d'estimation de variance doit refléter l'inflation de variance due à l'imputation.

Ceci est réalisé de manière relativement simple en utilisant une technique de rééchantillonnage, en traitant un estimateur tel que (9) comme n'importe quel estimateur. Une approche plus simple et moins intensive en termes de calculs est d'utiliser une formule de linéarisation standard comme si les valeurs imputées avaient été observées, mais ceci peut conduire à d'importantes sous-estimations de la variance.

Une puissante variante est l'imputation multiple (Rubin, 1987), où les formules standards sont calculées pour différents jeux de données pour lesquels les données manquantes ont été imputées stochastiquement, et sont alors combinées de manière à faire une allocation propre pour l'effet de l'imputation. Cette approche a été considérée comme sujette à controverse pour des enquêtes par certains auteurs; voir par exemple Fay (1996). D'autres discussions récentes à propos de l'imputation multiple peuvent être trouvées dans Zhang (2003) et Nielsen (2003), ainsi que leurs commentaires.

Avec une imputation propre, les estimateurs ponctuels et de variance sont approximativement non biaisés. En pratique, il est impossible de savoir si une approche d'imputation est propre ou pas, parce que le mécanisme des données manquantes est inconnu. Comme c'est généralement le cas en statistique, les modèles dont l'utilisation est décrite ci-dessus sont simplement une approximation brute d'une réalité beaucoup plus complexe. Pour cette raison Rubin (1996) recommande d'inclure autant de variables auxiliaires que possible, l'idée étant d'approximer cette réalité au mieux avec les données disponibles, même au prix d'une augmentation de la variance des estimations éventuelles.

2.4 Discussion

Pour des sondages simples, des livres tels que Cochran (1977) contiennent des formules donnant des estimations de variance d'une grande variété d'estimateurs. Cependant, la complexité des procédures modernes de sondage ne permet souvent pas de dériver des formules simples, et une alternative est de se tourner vers les méthodes de rééchantillonnage telles que le jackknife ou le bootstrap.

L'adaptation des méthodes d'estimation de variance basées sur le rééchantillonnage, à l'ajustement d'enquêtes, nécessite un soin particulier, parce qu'il doit prendre en compte les structures de dépendances complexes induites par le schéma d'échantillonnage ainsi que n'importe quelle procédure de calibration ou d'imputation. Sinon, la variance risque d'être fortement sur- ou sous-estimée.

3 Linéarisation

3.1 Linéarisation de Taylor

Beaucoup d'estimateurs s'expriment comme fonction différentiable $\hat{\theta} = g(\hat{\tau})$ d'un vecteur d'estimateurs linéaires $\hat{\tau}$; un exemple étant le rapport $\hat{\theta} = g(\hat{\tau}_{Y_1}, \hat{\tau}_{Y_2})$ avec $g(x_1, x_2) = x_1/x_2$. Le développement en séries de Taylor linéaire de g autour de la moyenne de population τ de $\hat{\tau}$ donne

$$\hat{\theta} = g(\hat{\tau}) \doteq g(\tau) + \nabla g(\tau)^T(\hat{\tau} - \tau),$$

dont la variance est $\nabla g(\tau)^T \text{var}(\hat{\tau}) \nabla g(\tau)$. L'estimateur de variance est obtenu en remplaçant les inconnues dans cette formule par $\nabla g(\hat{\tau})$ et la matrice de covariance empirique pour $\hat{\tau}$. Lorsqu'il est appliqué à des échantillons stratifiés, l'estimateur de variance résultant est

$$v_L = \nabla g(\hat{\tau})^T \widehat{\text{var}}(\hat{\tau}) \nabla g(\hat{\tau}) = \nabla g(\hat{\tau})^T \left(\sum_{h=1}^H \hat{V}_h \right) \nabla g(\hat{\tau}), \quad (10)$$

où \hat{V}_h est la contribution à la matrice de variance estimée pour $\hat{\tau}$ de la h -ème strate. Si g est linéaire, alors v_L est un estimateur sans biais de la variance de $\hat{\theta}$. Sous des conditions faibles, Krewski et Rao (1981) montre que v_L est un estimateur consistant de la variance asymptotique de $\hat{\theta}$.

Parfois il n'est pas possible d'exprimer la fonction g sous une forme explicite, par exemple lorsque l'estimateur de lieu est défini comme solution d'une équation d'estimation. Rao (1988), Binder (1996) et Deville (1999) discutent de la linéarisation pour de tels paramètres implicites.

3.2 Linéarisation jackknife

Une approche plus générale de construction d'estimateurs de variance par linéarisation pour une statistique différentiable $\hat{\theta}$ est l'utilisation des composantes de la fonction d'influence (Hampel *et al.*, 1986), qui peut être interprétée comme les dérivées de $\hat{\theta}$ par rapport aux changements dans les poids associés aux observations individuelles. La généralité provient de l'utilisation d'un développement en série de von Mises plutôt que de Taylor, permettant l'obtention de formules théoriques pour la variance pour des estimateurs comme la médiane. Voir par exemple Berger et Skinner (2003), qui utilisent des procédures de noyau pour estimer la fonction de densité qui survient dans ces formules.

Dans beaucoup de cas la quantité à estimer θ peut s'écrire comme une fonctionnelle $t(F)$ de la fonction de distribution sous-jacente F qui a engendré les données. Un estimateur simple de $t(F)$ est alors $t(\hat{F})$, où \hat{F} est la fonction de distribution empirique des données. Pour la moyenne, par exemple, $t(F) = \int y dF(y)$ et $t(\hat{F}) = \bar{Y}$ représente

son analogue empirique. Dans le cas de l'échantillonnage aléatoire simple avec remise, et sous quelques hypothèses de différentiabilité pour $t(\cdot)$, l'estimateur $\hat{\theta} = t(\hat{F})$ peut être développé autour de $\theta = t(F)$ en utilisant

$$t(\hat{F}) \doteq t(F) + n^{-1} \sum_{i=1}^n L_t(Y_i; F),$$

où

$$L_t(y; F) = \lim_{\epsilon \rightarrow 0} \frac{t\{(1 - \epsilon)F + \epsilon\delta_y\} - t(F)}{\epsilon}$$

est la *fonction d'influence* pour $t(\hat{F})$, δ_y étant la fonction de distribution mettant une masse ponctuelle en y . Ce développement peut être utilisé pour établir que l'estimateur est asymptotiquement non biaisé et gaussien. Sa variance $v_L(F) = n^{-1}\text{var}\{L_t(Y; F)\}$ peut être estimée par

$$\hat{v}_L = n^{-2} \sum_{i=1}^n l_i^2, \quad (11)$$

où $l_i = L_t(y_i; \hat{F})$ sont les valeurs d'influence empiriques pour la fonctionnelle statistique t évaluée en l'observation y_i et la fonction de distribution empirique \hat{F} .

Ici on peut se représenter l_i comme la dérivée de t en \hat{F} dans la direction de la distribution attribuant plus de masse sur la i -ème observation. Davison et Hinkley (1997, §2.11) donnent plus de détails sur ces développements et des exemples de calculs, et Campbell (1980) décrit leur extension dans les situations d'une population finie.

Pour l'échantillonnage stratifié aléatoire sans remise, (11) pourrait être modifiée en

$$v_L = \sum_{h=1}^H (1 - f_h) \frac{1}{(n_h - 1)n_h} \sum_{i=1}^{n_h} l_{hi}^2, \quad (12)$$

où l_{hi} est la valeur d'influence empirique correspondant à la i -ème observation dans la strate h . Dans des situations simples, (11) et (12) deviennent des formules de linéarisation standard, mais l'utilisation de fonctions d'influence conduit à une approche plus générale pour l'estimation de variance.

3.3 Estimateur de Horvitz–Thompson

Nous considérons maintenant l'estimateur de Horvitz–Thompson et donnons des formules pour sa fonction d'influence empirique pour l'échantillonnage stratifié dans trois situations de complexité croissante :

- L'estimateur de Horvitz–Thompson standard donné dans (3), pour lequel

$$l_{hi} = n_h \omega_{hi} y_{hi} - \omega_h^T y_h;$$

– L'estimateur de Horvitz–Thompson calibré (5), pour lequel

$$l_{hi} = (n_h \omega_{hi} y_{hi} - \omega_h^T y_h) + (X_{C_h}^T \omega_h - n_h \omega_{hi} x_{C_{hi}})^T \hat{\gamma} \\ + n_h \omega_{hi} (c - X_C^T \omega)^T (X_C^T \Omega X_C)^{-1} x_{C_{hi}} (y_{hi} - x_{C_{hi}}^T \hat{\gamma}),$$

où ω_h et y_h sont des vecteurs de poids et de réponses $n_h \times 1$ pour la h -ème strate, X_{C_h} est la matrice de calibration des covariables $n_h \times q_C$ pour la h -ème strate, et $\hat{\gamma} = (X_C^T \Omega X_C)^{-1} X_C^T \Omega y$; et

– l'estimateur calibré de Horvitz–Thompson (9) avec imputation de valeurs manquantes. Soit

$$\hat{\gamma}_M = (X_C^T \Omega X_C)^{-1} X_C^T \Omega (I - Z) \hat{y}$$

correspondant à $\hat{\gamma}$ pour les individus avec réponse manquante, et soit $l_i(\hat{\beta})$ les éléments du vecteur $q \times 1$ de fonctions d'influence pour les coefficients de régression d'imputation, correspondant à la différentiation par rapport à la i -ème unité dans la strate h . Alors pour $i = 1, \dots, n_h$ and $h = 1, \dots, H$,

$$l_{hi} = (n_h \omega_{hi} z_{hi} y_{hi} - \omega_h^T Z_h y_h) + (X_{C_h}^T \omega_h - n_h \omega_{hi} x_{C_{hi}})^T \hat{\gamma} \\ + n_h \omega_{hi} (c - X_C^T \omega)^T (X_C^T \Omega X_C)^{-1} x_{C_{hi}} (z_{hi} y_{hi} - x_{C_{hi}}^T \hat{\gamma}) \\ + \{n_h \omega_{hi} (1 - z_{hi}) \hat{y}_{hi} - \omega_h^T (I_h - Z_h) \hat{y}_h\} + \omega^T (I - Z) X l_i(\hat{\beta}) \\ + (X_{C_h}^T \omega_h - n_h \omega_{hi} x_{C_{hi}})^T \hat{\gamma}_M \\ + n_h \omega_{hi} (c - X_C^T \omega)^T (X_C^T \Omega X_C)^{-1} x_{C_{hi}} \{(1 - z_{hi}) \tilde{y}_{hi} - x_{C_{hi}}^T \hat{\gamma}_M\} \\ + (c - X_C^T \omega)^T (X_C^T \Omega X_C)^{-1} X_{C_h}^T \Omega_h (I_h - Z_h) X_h l_i(\hat{\beta}). \quad (13)$$

Le modèle d'imputation (7) pour la régression des moindres carrés déterministe, par exemple, donne

$$l_i(\hat{\beta}) = n_h z_i (X^T Z X)^{-1} x_i (y_i - x_i^T (X^T Z X)^{-1} X^T Z y), \quad i = 1, \dots, \sum_{h=1}^H n_h,$$

où X est la matrice de régression. Pour le modèle d'imputation (8), dans lequel les coefficients de régression peuvent varier en fonction de la strate, les $l_i(\hat{\beta})$ dans (13) sont

$$l_i(\hat{\beta}_h) = n_h z_i (X_h^T Z_h X_h)^{-1} x_i (y_i - x_i^T (X_h^T Z_h X_h)^{-1} X_h^T Z_h y_h),$$

où X_h , Z_h , et y_h sont la matrice de covariables, la matrice indicatrice pour les réponses observées, et le vecteur de réponses pour la strate h .

Ces formules sont utilisées par la suite dans nos comparaisons numériques, et en particulier, elles forment la base des méthodes d'imputation multiple que nous appliquons dans la section 7.4.

4 Jackknife

4.1 Idées de base

Le jackknife, introduit comme méthode pour l'estimation du biais par Quenouille (1949a,b) et ensuite proposé pour l'estimation de variance par Tukey (1958), implique la suppression systématique de groupes d'unités, le re-calcul de la statistique avec chaque groupe supprimé tour à tour, et ensuite la combinaison de toutes ces statistiques recalculées.

Le jackknife le plus simple consiste à écarter chaque observation d'un échantillon tour à tour, re-calculer la statistique pour obtenir des valeurs t_{-1}, \dots, t_{-n} en plus de sa valeur originale, t , et ensuite à estimer le biais et la variance de l'estimateur par

$$\frac{n-1}{n} \sum_{i=1}^n (t_{-i} - t) = (n-1)(\bar{t} - t), \quad \frac{n-1}{n} \sum_{i=1}^n (t_{-i} - \bar{t})^2.$$

Pour que ces estimateurs soient consistants, la statistique t doit être suffisamment lisse comme fonction des observations, par exemple posséder un développement de Taylor ou similaire. Certains types d'estimateurs non-lisses, comme la médiane et d'autres estimateurs basés sur les quantiles, ne possèdent pas un développement convenable, de sorte que les estimateurs de leur variance de type jackknife simple ne sont pas consistants (Efron, 1982). Bien que ceci est un désavantage majeur du point de vue théorique, une considération pratique plus importante est la performance d'un estimateur pour les types et tailles d'échantillons qu'on rencontre dans les applications, et il se révèle qu'en dépit de son inconsistance, l'estimateur de variance jackknife pour des quantiles d'échantillons basés sur des échantillons d'enquêtes, peut être compétitif par rapport à d'autres estimateurs en terme de moindres carrés moyens (Rao *et al.*, 1992). Shao et Wu (1989) a montré que l'inconsistance peut être réparée en supprimant des groupes de d observations, où $d \rightarrow \infty$ lorsque $n \rightarrow \infty$.

Rao et Wu (1985) ont montré que la linéarisation et le jackknife sont asymptotiquement équivalents au premier ordre, et que dans certains cas ils produisent exactement les mêmes estimations de variance. Quelques simulations tendent aussi à montrer que le jackknife est moins biaisé mais plus variable que l'estimateur de variance de type linéarisation. Berger et Skinner (2005) développent une théorie pour l'estimation de variance jackknife pour une grande classe d'estimateurs sous échantillonnage de probabilités différentes ; dans d'autres travaux pas encore publiés Y. G. Berger et J. N. K. Rao ont développé ceci pour l'imputation.

Une façon de comprendre le jackknife simple est de le voir comme un moyen numérique pour obtenir les dérivées de t par rapport aux observations. Ecarter une observation entraîne une perturbation $O(n^{-1})$ de l'échantillon, et comme la variance de l'échantillonnage

est typiquement de $O(n^{-1/2})$, on peut soupçonner que les perturbations dues au jackknife sont trop locales pour donner une bonne idée de la variance de l'estimateur. L'expérience ainsi que des travaux théoriques suggèrent que les estimateurs de variance jackknife typiquement sous-estiment légèrement la vraie variance, et que cette sous-estimation peut être attribuée à la localité du développement de Taylor.

Miller (1974) propose une excellente revue des premiers travaux sur le jackknife, et une étude récente plus détaillée est donnée par Shao et Tu (1995). Des estimateurs plus généraux de sous-échantillonnage, utiles surtout pour les séries temporelles et d'autres formes de données corrélées, ont été proposées par Politis *et al.* (1999)

4.2 Jackknife modifié

Il y a beaucoup de complications dans le contexte de l'échantillonnage dans les enquêtes. La première est que les plans d'échantillonnage sans remise typiquement utilisés résultent en observations corrélées, tandis qu'une deuxième est que les données sont typiquement stratifiées. Avec des données d'enquêtes stratifiées, le jackknife simple entraîne la suppression tour à tour de chaque unité primaire d'échantillonnage. Une manière de voir ceci est que la suppression de la i -ème unité dans la h -ème strate change le poids pour la j -ème unité dans la k -ème strate en

$$\omega_{kj}^{(hi)} = \begin{cases} \omega_{kj}, & k \neq h, \\ \omega_{hj}n_h/(n_h - 1), & k = h, j \neq i, \\ 0, & k = h, j = i. \end{cases}$$

Dans les cas qui nous intéressent ici le rapport $n_h/(n_h - 1) \doteq 1$, et ainsi l'utilisation de ces poids actualisés va donner les statistiques $\hat{\theta}_{h,-i}$, c'est-à-dire $\hat{\theta}$ calculé sans la i -ème unité de la strate h , pour $i = 1, \dots, n_h$ et $h = 1 \dots, H$. L'estimation de variance par jackknife simple est donnée par

$$v_J = \sum_{h=1}^H \frac{(1 - f_h)(n_h - 1)}{n_h} \sum_{i=1}^{n_h} (\hat{\theta}_{h,-i} - \bar{\theta}_h)^2, \quad \bar{\theta}_h = n_h^{-1} \sum_{i=1}^{n_h} \hat{\theta}_{h,-i}, \quad (14)$$

où $f_h = n_h/N_h$ est la fraction d'échantillonnage dans la h -ème strate, qui doit être incluse pour tenir compte du plan d'échantillonnage. Rao et Wu (1985) montrent que remplacer $\bar{\theta}_h$ dans (14) par $n^{-1} \sum_{h=1}^H \sum_{i=1}^{n_h} \hat{\theta}_{h,-i}$ ou par $H^{-1} \sum_{h=1}^H \bar{\theta}_h$ a peu d'effet sur la performance de v_J dans le sens que les trois versions sont asymptotiquement équivalentes au deuxième ordre; Kovar *et al.* (1988) observent que les trois estimateurs de variance résultants sont également comparables pour des échantillons finis.

Une troisième complication qui arrive en pratique, est que le jackknife simple nécessite un gros effort de calcul parce qu'un total de n suppressions est nécessaire; le fardeau

calculatoire résultant est souvent trop grand pour que ce simple jackknife se révèle utile. La suppression de d éléments d'échantillon à la fois (Shao et Wu, 1989) nécessiterait $\binom{n}{d}$ re-calculs de la statistique, et est ainsi même encore moins faisable, mais il est courant de réduire le nombre de réplifications nécessaires en séparant la strate h en m_h blocs disjoints de d observations, de sorte que $dm_h = n_h$, et d'ensuite systématiquement re-calculer la statistique avec chacun de ces blocs supprimés. Le nombre total de re-calculs est ainsi $m = \sum_{h=1}^H m_h$ plutôt que n ou $\binom{n}{d}$. La suppression d'une partie de la h -ème strate d'un bloc dont les unités ont les indices $i \in \mathcal{S}$ change le poids pour la j -ème unité dans la k -ème strate en

$$\omega_{kj}^{(h\mathcal{S})} = \begin{cases} \omega_{kj}, & k \neq h, \\ \omega_{hj}n_h/(n_h - |\mathcal{S}|), & k = h, j \notin \mathcal{S}, \\ 0, & k = h, j \in \mathcal{S}, \end{cases}$$

où $|\mathcal{S}|$ représente le nombre d'éléments d dans \mathcal{S} . Des variantes de ces procédures jackknife groupées choisissent les membres du bloc aléatoirement lors de chaque re-calcul, avec ou sans remise (Krewski et Rao, 1981; Shao et Tu, 1995, Section 5.2.2). Etant donné que l'échantillonnage avec remise est simple et produit des résultats comparables à l'échantillonnage sans remise, l'échantillonnage avec remise jackknife à d -suppressions semble préférable. Pour chacun de ceux-ci l'estimation de variance est

$$v_{J,-d} = \sum_{h=1}^H \frac{(1 - f_h)(n_h - d)}{n_h} \sum_{i=1}^{m_h} (\hat{\theta}_{h,-i}^d - \bar{\theta}_h^d)^2,$$

où $\hat{\theta}_{h,-i}^d$ est l'estimateur calculé à partir des données avec le i -ème bloc de taille d exclu de la h -ème strate, et $\bar{\theta}_h^d$ est la moyenne pour cette strate. De légères variations des formules ci-dessus sont nécessaires lorsque n_h n'est pas exactement divisible par d , auquel cas un bloc a la taille comprise dans $1, \dots, d$.

Le jackknife à d -suppressions a l'avantage théorique supplémentaire d'étendre la portée des fonctions avec lesquelles le jackknife peut être appliqué. Comme mentionné ci-dessus, il s'avère que si $d, m_h \rightarrow \infty$ en même temps que $n_h \rightarrow \infty$, alors le jackknife de bloc produit des estimateurs de variance consistants (Shao et Wu, 1989). Comme avec d'autres estimateurs de sous-échantillonnage, ce n'est pas clair comment choisir d et m_h lorsque n_h est fini. Moins la statistique d'intérêt est lisse, plus d doit être grand, au moins en théorie, mais une méthode fiable s'appliquant aux petits échantillons pour choisir d n'existe pas encore.

Kim et Sitter (2003) décrivent une approche pour réduire le nombre de re-calculs pour un estimateur de variance jackknife utilisé lors de l'échantillonnage à deux phases, et donnent d'autres références pour ce domaine.

4.3 Imputation

Rao et Shao (1992) décrivent une version consistante de l'estimateur jackknife simple utilisant un mécanisme d'imputation de type 'hot deck' particulier pour prendre en compte les non-réponses ; voir aussi Fay (1996) pour une vision élargie, et Chen et Shao (2001), qui montrent que cette approche échoue pour un autre schéma d'imputation populaire, l'imputation de plus proche voisin. L'idée clé est que, étant donné que la valeur d'une unité qui répond affecte à la fois cette unité et n'importe quelle valeur imputée qui utilise sa valeur, la suppression d'une telle unité doit prendre en compte ceci ; la suppression d'une unité qui ne répond pas affecte cependant l'estimateur de manière usuelle. Ceci peut être étendu à l'échantillonnage stratifié à plusieurs degrés par suppression de chacune des unités d'échantillonnage primaires.

Ci-dessous nous utilisons une légère généralisation de cette idée en autorisant la suppression de blocs, c'est-à-dire, en procédant à une ré-imputation chaque fois qu'un bloc est supprimé ; en utilisant les répondants des données restantes pour estimer le modèle utilisé pour l'imputation déterministe. Etant donné que ceci est simplement une version de l'estimateur basé sur les données répondantes mais tenant compte de l'effet de non-réponse, aucune nouvelle théorie n'est requise lorsque le modèle d'imputation est adéquat.

5 Répliques équilibrées répétées

5.1 Idées de base

Le demi-échantillonnage équilibré (McCarthy, 1969) est la forme la plus simple de réplique équilibrée répétée ('balanced repeated replication'). Il fut à la base développé pour les plans stratifiés à plusieurs degrés avec deux unités d'échantillonnages primaires effectuées avec remise lors du premier degré. Lorsqu'un échantillon consiste en deux observations dans chacune des H strates, un semi-échantillon est formé en prenant une observation pour chaque strate. Le poids ω_{hi} attaché à la i -ème unité dans la strate h est changé en $2\omega_{hi}$ si cette unité est incluse dans le semi-échantillon, et en 0 sinon, et la statistique est re-calculée avec le semi-échantillon et ces nouveaux poids.

Il y a un total de 2^H telles statistiques répliquées, dont les résultats peuvent être combinés aux estimations de variance de la statistique originelle. Ceci est clairement infaisable du point de vue calculatoire à moins que H ne soit petit, mais des idées de design expérimental permettent d'obtenir la même estimation de variance avec moins de répliques, à conditions qu'elles soient équilibrées. Soit $\alpha_{hr} = \pm 1$ selon que la première ou la deuxième des deux unités dans la h -ème strate est retenue dans la r -ème réplique, et soit $\hat{\theta}_r$ la valeur de l'estimateur $\hat{\theta}$ calculé en utilisant le r -ème demi-échantillon et ses

poids ajustés. Alors l'ensemble des R demi-échantillons est équilibré si

$$\sum_{r=1}^R \alpha_{hr} \alpha_{kr} = 0, \quad \text{pour tous } h \neq k, k, h = 1, \dots, H,$$

et dans ce cas l'estimateur de variance de réplication équilibrée répétée

$$v_{\text{BRR}} = R^{-1} \sum_{r=1}^R (\hat{\theta}_r - \hat{\theta})^2 \quad (15)$$

produit la même estimation de variance que par calcul utilisant l'ensemble entier de 2^H demi-échantillons, pour des statistiques linéaires. Pour beaucoup de statistiques, v_{BRR} est égal à la fois à l'estimateur jackknife v_J et à l'estimateur de linéarisation. Si nécessaire, $\hat{\theta}$ en (15) peut être remplacé par $R^{-1} \sum_{r=1}^R \hat{\theta}_r$ sans changer les propriétés asymptotiques de l'estimateur de variance, tandis que d'autres variantes sont aussi possibles.

Lorsqu'il n'y a pas d'observations manquantes, v_{BRR} est consistant pour les variances de fonctions non-linéaires (Krewski et Rao, 1981). Il s'avère que v_{BRR} est stochastiquement proche de l'estimateur de variance de linéarisation, mais pas aussi proche que ne l'est l'estimateur jackknife. La consistance du design de v_{BRR} pour des estimateurs de quantiles et ceux basés sur les quantiles a été établie par Shao et Wu (1992) et Shao et Rao (1994).

Un ensemble équilibré de demi-échantillons peut être obtenu pour tout H en sélectionnant R comme étant le plus petit multiple de quatre qui est plus grand ou égal à H ; c'est-à-dire $H \leq R \leq H + 3$. Des valeurs appropriées pour α_{hr} sont données par les éléments d'une matrice de Hadamard $R \times R$. Par exemple, la matrice de Hadamard utile pour $H = 7$ est la matrice 8×8

$$\alpha = \begin{pmatrix} 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 & -1 & -1 \\ -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \end{pmatrix},$$

à partir de laquelle on identifie les demi-échantillons en prenant α_{hr} comme étant le r -ème élément de la ligne h . Ainsi les strates correspondent aux lignes et les répliques aux colonnes; par exemple la première colonne identifie la première des huit répliques en incluant la première unité à partir des strates 1, 2, 3 et 5, et la deuxième unité à partir des strates 4, 6 et 7.

On répète alors ceci pour chacune des huit colonnes, avec les éléments de matrices donnant le demi-échantillon approprié pour la réplique qui correspond à chaque colonne. Wolter (1985, appendice A) donne la liste de telles matrices jusqu'à $R = 100$ et donne des références pour construire des plus grandes.

La calibration est appliquée à un demi-échantillon en calibrant les poids ré-échellonnés en utilisant les marges de la populations.

5.2 Strate de taille $n_h > 2$

On a proposé deux généralisations principales quand il y a plus de $n_h = 2$ observations par strate. La première (Gurney et Jewett, 1975; Gupta et Nigam, 1987; Wu, 1991; Sitter, 1993) utilise des tableaux orthogonaux mais nécessite un grand nombre de répliques, la rendant peu pratique pour beaucoup d'applications.

La deuxième généralisation, une approche plus simple et plus pragmatique, est de regrouper les unités d'échantillonnages primaires dans chaque strate en deux groupes de tailles $m_h = \lfloor n_h/2 \rfloor$ et $n_h - m_h$, et d'appliquer la réplique équilibrée répétée utilisant les groupes ('grouped balanced repeated replication') plutôt que des unités individuelles (Rao et Shao, 1996; Wolter, 1985, Section 3.7). Dans ces schémas de réplique équilibrée répétée groupée, les poids pour la r -ème réplique sont ajustés en utilisant la formule

$$\omega_{hi}^r = \begin{cases} \omega_{hi} \left[1 + \left\{ \frac{(n_h - m_h)(1 - f_h)}{m_h} \right\}^{1/2} \right], & \alpha_{hr} = 1, \\ \omega_{hi} \left[1 - \left\{ \frac{m_h(1 - f_h)}{n_h - m_h} \right\}^{1/2} \right], & \alpha_{hr} = -1, \end{cases} \quad (16)$$

où le facteur $(1 - f_h)$ est un ajustement pour l'effet de l'échantillonnage sans remise et la condition $m_h \leq n_h/2$ assure que les poids ajustés sont positifs. Le choix $m_h = n_h/2$ conduit à l'ajustement de poids original de $2\omega_{hi}$ ou 0 lorsque f_h peut être négligé. L'estimateur de variance est encore défini par (15), mais l'effet du regroupement n'induit plus l'équivalence avec l'estimateur de variance standard dans le cas linéaire.

5.3 Estimateurs améliorés

L'estimateur de variance basé sur la réplique équilibrée répétée v_{BRR} peut s'avérer très variable, pour deux raisons principales. La première est que la loi de (15) est approximativement $c\chi_{R-1}^2$, où $c > 0$ est une constante. Ainsi le coefficient de variation de (15) est à peu près $(2/R)^{1/2}$; ceci est appréciable lorsque H est petit. Le coefficient de variation peut être réduit en augmentant les degrés de libertés, par exemple en séparant les H vraies données en sous-strates artificielles, auxquelles on applique alors le demi-échantillonnage.

La deuxième difficulté est que certaines statistiques peuvent être fortement sensibles aux perturbations des poids, et si c'est le cas, il peut s'avérer impossible de calculer toutes les estimations répliquées $\widehat{\theta}_1, \dots, \widehat{\theta}_R$. Ceci se produit par exemple avec les estimateurs de rapport, lorsque le dénominateur peut s'annuler pour certaines re-pondérations, et lorsque la suppression de la moitié de l'échantillon résulte en des tailles d'échantillons plus petites que tolérées dans l'enquête originale.

Une solution suggérée par Robert Fay du US Census Bureau (Dippo *et al.*, 1984; Fay, 1989) est d'utiliser un schéma de re-pondérage plus doux, et où un facteur ε est choisi, avec $0 < \varepsilon \leq 1$, et les poids sont perturbés en

$$\omega_{hi}^r = \begin{cases} \omega_{hi} \left[1 + \varepsilon \left\{ \frac{(n_h - m_h)(1 - f_h)}{m_h} \right\}^{1/2} \right], & \alpha_{hr} = 1, \\ \omega_{hi} \left[1 - \varepsilon \left\{ \frac{m_h(1 - f_h)}{n_h - m_h} \right\}^{1/2} \right], & \alpha_{hr} = -1, \end{cases}$$

qui produit des estimations répliquées $\widehat{\theta}_1(\varepsilon), \dots, \widehat{\theta}_R(\varepsilon)$ et un estimateur de variance

$$v_{\text{BRR}}(\varepsilon) = \frac{1}{R\varepsilon^2} \sum_{r=1}^R \left\{ \widehat{\theta}_r(\varepsilon) - \widehat{\theta} \right\}^2.$$

En posant $\varepsilon = 1$ on retrouve les poids (16) et l'estimateur de variance (15), mais d'autres valeurs de ε reviennent à sous-pondérer plutôt que supprimer des groupes d'unités, et réduisent ainsi fortement l'instabilité des estimations répliquées. Rao et Shao (1999) ont étudié les aspects théoriques de ces modifications de poids et ont considéré également leur extensions à l'imputation pour la non-réponse de type article. Leurs conclusions principales sont : le schéma de pondération plus doux produit une estimation de variance consistante sous les mêmes conditions que l'approche originale, c'est-à-dire, avec $\varepsilon = 1$; lorsque $\widehat{\theta}$ est une fonction lisse des totaux $v_{\text{BRR}}(\varepsilon)$, l'estimateur converge vers l'estimateur de variance de la linéarisation de Taylor lorsque $\varepsilon \rightarrow 0$; la différence entre les deux est petite lorsque ε est une fonction de la taille de l'échantillon n ; et que des schémas d'imputation convenablement appliqués produisent une estimation de variance consistante utilisant $v_{\text{BRR}}(\varepsilon)$.

Des travaux numériques (Judkins, 1990; Rao et Shao, 1999) suggèrent que de bons résultats sont obtenus avec $\varepsilon = 0.5$; en particulier la variance de l'estimateur de la médiane en utilisant l'imputation de rapport possède alors des propriétés bien meilleures pour les petits échantillons qu'en posant $\varepsilon = 1$.

5.4 Réplication équilibrée répétée groupée de manière répétitive

Lorsque H est grand, la réplication équilibrée répétée groupée engendre un estimateur de variance consistant, mais ceci peut être très variable si $H \ll n$. De plus elle est

inconsistante lorsque H est fixé et que $n_h \rightarrow \infty$ (Rao et Shao, 1996), suggérant que cela peut mal fonctionner pour H petit. Ceci a été observé empiriquement, et comme mentionné ci-dessus, est due en partie au fait que lorsque H est fixé l'estimateur de variance v_{BRR} a une loi proche de $c\chi_{R-1}^2$.

Pour réduire la variabilité de v_{BRR} et pour enlever l'effet de la séparation aléatoire particulière des unités de la strate en groupes, Rao et Shao (1996) a proposé de répéter la méthode sur des groupes sélectionnés de manière aléatoire pour fournir m estimations de variance, dont la moyenne produira une estimation de variance globalement plus stable.

Cet estimateur de variance basé sur la réplication équilibrée répétée groupée de manière répétitive ('repeatedly grouped balanced repeated replication') s'avère être consistant lorsque $mR \rightarrow \infty$. En particulier, choisir m tel que $mR \doteq n$ engendre un estimateur de variance assez efficace, ce qui implique que m devrait être grand lorsque le nombre de strates H est petit. Lorsque H est grand, la règle suggère de choisir seulement $m = 1$. La conclusion évidente de ceci est que calculer l'estimateur $\hat{\theta}$ pour $O(n)$ répliques est généralement trop encombrant pour l'application pratique, où on doit typiquement prendre $mR \ll n$.

Les simulations suggèrent que l'estimateur de variance basé sur la réplication équilibrée répétée groupée de façon répétitive a une performance similaire à celle du jackknife groupé, et que les deux sont meilleurs que l'estimateur de variance de réplication équilibrée répétée groupée, qui a une grande variance. Des variantes de ces procédures incluent le sous-échantillonnage aléatoire d'unités pour former des groupes, suivi de l'application des formules de variance usuelles (Kovar *et al.*, 1988).

5.5 Imputation

Le travail de Rao et Shao (1999) sur le comportement de $v_{\text{BRR}}(\varepsilon)$ lorsqu'il y a des non-réponses unités a déjà été mentionné. Shao *et al.* (1998) ajuste la réplication équilibrée répétée à la présence de non-réponses, en tenant compte d'un mécanisme d'imputation déterministe ou stochastique. Dans un plan d'échantillonnage stratifié à plusieurs degrés, ils établissent la consistance de l'estimateur de variance ajusté basé sur la réplication équilibrée répétée pour des fonctions de statistiques lisses et non lisses.

6 Bootstrap

6.1 Idées de base

Le bootstrap a été intensément étudié pour des données indépendantes et identiquement distribuées, où l'échantillonnage est fait avec remise, mais son adaptation aux

enquêtes complexes n'est pas directe, parce que le schéma d'échantillonnage de probabilité de l'enquête induit une dépendance. On doit également tenir compte de la calibration et de l'imputation lorsqu'on effectue le bootstrap sur l'échantillon d'enquête.

L'idée du bootstrap est d'imiter la manière dont ont été engendrées les données originales. Comme les méthodes de répliation équilibrée répétée et du jackknife, le bootstrap implique de re-calculer la statistique, utilisant maintenant le rééchantillonnage à partir d'une population estimée \widehat{F} pour obtenir des échantillons bootstrap qui peuvent être représentés par \widehat{F}^* , donnant les statistiques correspondantes $\widehat{\theta}^* = t(\widehat{F}^*)$. En répétant ce processus R fois indépendamment, l'estimation de variance par bootstrap est donnée par

$$v_B = (R - 1)^{-1} \sum_{r=1}^R (\widehat{\theta}_r^* - \overline{\widehat{\theta}^*})^2, \quad \overline{\widehat{\theta}^*} = R^{-1} \sum_{r=1}^R \widehat{\theta}_r^*.$$

Pour des données stratifiées, on effectue le rééchantillonnage indépendamment dans chaque strate. Le bootstrap standard utilise l'échantillonnage avec remise, correspondant à un échantillonnage indépendant à partir d'une population originale, mais ceci ne concorde pas avec l'échantillonnage sans remise généralement utilisé dans le contexte d'enquêtes, et le résultat est qu'il manque la correction d'échantillon fini, ce qui donne un estimateur de variance biaisé.

Le contexte le plus simple pour voir ceci est lorsqu'on estime la moyenne d'une population de taille N par échantillonnage de probabilités égales sans remise. Un estimateur non biaisé de la variance de la moyenne d'échantillon $\bar{y} = n^{-1} \sum_{i \in \mathcal{S}} y_i$, où $\mathcal{S} \subset \{1, \dots, N\}$ contient alors les n indices échantillonnés, est donné par

$$(1 - f) \frac{s^2}{n}, \quad s^2 = \frac{1}{n - 1} \sum_{i \in \mathcal{S}} (y_i - \bar{y})^2,$$

où $f = n/N$ est la fraction d'échantillonnage. Le schéma de bootstrap le plus simple prend des échantillons de taille n avec remise à partir de $\{y_i : i \in \mathcal{S}\}$, et utilise les propriétés d'échantillonnage de la moyenne bootstrap \bar{y}^* pour estimer ceux de \bar{y} sous répétition du design d'échantillonnage original. La variance bootstrap de \bar{y}^* est

$$\text{var}^*(\bar{y}^*) = \frac{n - 1}{n} \frac{s^2}{n},$$

Cependant, et en général, ceci ne vaut pas $\text{var}(\bar{y})$ à moins que n soit grand et que la fraction d'échantillonnage soit petite, ou à moins que $f = 1/n$. Le même problème survient essentiellement avec l'échantillonnage stratifié, où la variance d'un estimateur global est une somme de contribution des strates individuelles; voir (14).

Ceci implique que l'estimateur de variance de bootstrap standard peut s'avérer inconsistant dans des plans d'échantillonnage dans lesquels $H \rightarrow \infty$ mais où les fractions d'échantillonnage f_h sont à une distance bornée de zéro. Un point de vue est que

l'échec asymptotique est sans importance si l'estimateur de variance bootstrap possède un carré moyen de l'erreur faible pour le type d'enquête rencontré en pratique. Cet échec du bootstrap conventionnel a engendré des travaux pour le modifier. Efron (1982) a reconnu le problème d'échellonnage des petites strates, et suggéré la solution simple consistant à prendre des échantillons bootstrap de taille $n_h - 1$, mais ceci n'arrange pas l'effet de la fraction d'échantillonnage non négligeable f_h , pour lesquelles des approches plus élaborées sont requises, et que nous décrivons maintenant.

6.2 Bootstrap sans remise

L'approche la plus simple et la plus satisfaisante dans le cas d'un échantillon non stratifié est de créer une pseudo population utilisant N/n répliques de l'échantillon, et d'ensuite échantillonner à partir de cette population artificielle sans remise (Gross, 1980; Chao et Lo, 1985). Pour des échantillons stratifiés (Bickel et Freedman, 1984; Booth *et al.*, 1994), la h -ème strate est répliquée N_h/n_h fois, et ensuite on applique l'échantillonnage stratifié séparément dans chaque strate.

Bickel et Freedman (1984) généralisent cette méthode à la situation où N_h/n_h n'est pas entier, par randomisation entre deux tailles de populations différentes, et dérivent quelques propriétés théoriques, mais McCarthy et Snowden (1985) décrivent quelques situations où cette approche n'est pas faisable.

Sitter (1992b) modifie la méthode bootstrap pour contourner les problèmes de l'estimateur de variance de Bickel et Freedman (1984) en sélectionnant des échantillons sans remise de taille m_h dans chaque pseudo-population de taille $k_h n_h = N_h$, où

$$m_h = n_h - (1 - f_h) \quad \text{and} \quad k_h = (1 - \frac{1-f_h}{n_h})/f_h, \quad (17)$$

pour obtenir un estimateur de variance non biaisé et consistant, au moins pour une statistique linéaire.

Lorsque la fraction d'échantillonnage inverse N_h/n_h n'est pas entière, toutes ces méthodes nécessitent une randomisation entre les pseudo-populations dont les tailles approchent N_h . Les détails de la randomisation sont importants, et en plus d'être quelque peu encombrant à programmer, des erreurs peuvent conduire à l'échec de ces méthodes (Presnell et Booth, 1994). De plus le besoin d'une pseudo-population pour chaque strate requiert de la mémoire additionnelle, ce qui n'est pas pratique pour de grandes enquêtes.

6.3 Bootstrap avec remise

Sans créer une pseudo-population, McCarthy et Snowden (1985) proposent d'imiter le design d'échantillonnage original en sélectionnant des échantillons de taille m_h avec

remise dans chaque strate avec

$$m_h = (n_h - 1)/(1 - f_h) \quad (18)$$

de sorte à obtenir un estimateur de variance non biaisé et consistant pour un estimateur linéaire. Lorsque m_h n'est pas entier, une randomisation est à nouveau requise.

6.4 Bootstrap rééchellonné

Rao et Wu (1988) proposent aussi de faire un bootstrap avec remise, mais ils atteignent la variance standard en rééchellonnant les échantillons bootstrap de taille m_h suivant

$$y_h^* = \bar{y}_h + \sqrt{m_h(1 - f_h)/(n_h - 1)}(\bar{y}_h^* - \bar{y}_h),$$

où m_h peut être n'importe quel entier positif. Ceci inclut la méthode de Efron (1982) comme cas particulier. Ils montrent que les estimateurs de distribution sont consistants. Pour estimer les moments de troisième ordre, Rao et Wu (1988) suggèrent de prendre

$$m_h = \lfloor (1 - f_h)(n_h - 2)^2 / (1 - 2f_h)(n_h - 1) \rfloor,$$

ce qui évite le problème de m_h non entier comme dans (17) et (18). La méthode a aussi l'avantage de ne pas créer et stocker les pseudo-populations, mais elle a quelques problèmes, notamment que pour $m_h \geq n_h$, elle peut conduire à des valeurs impossibles de la statistique bootstrapée.

6.5 Bootstrap miroir-concordance

Le bootstrap miroir-concordance ('mirror-match bootstrap') (Sitter, 1992a) est un croisement entre le bootstrap avec et sans remise. Dans chaque strate, il produit un échantillon de taille $n_h^* < n_h$ sans remise $k_h = n_h(1 - f_h^*)/\{n_h^*(1 - f_h)\}$ fois. Pour concorder avec les moments de troisième ordre, Sitter (1992a) suggère $n_h^* = f_h n_h$. Les estimateurs de variance bootstrap miroir-concordance sont consistants pour des statistiques linéaires. Il est équivalent au bootstrap avec remise de McCarthy et Snowden (1985) si $n_h^* = 1$. Une conclusion de ceci est que une randomisation entre les entiers englobants est requise lorsque k_h n'est pas un entier (Sitter, 1992a).

6.6 Non-réponse

Lorsqu'il manque des réponses, le mécanisme d'imputation doit être appliqué à chaque ré-échantillon \hat{F}^* (Shao et Sitter, 1996). Ainsi nous devons ré-imputer de manière répétitive les répondants de l'échantillon bootstrappé pour ajuster le modèle d'imputation et ensuite imputer les non-répondants de l'échantillon bootstrap. C'est pourquoi la méthode

est lourde en terme de calcul, mais d'un autre côté elle fournit des estimateurs de variance consistants pour les médianes et d'autres estimateurs basés sur les quantiles.

6.7 Commentaires

Davison et Hinkley (1997, §3.7) contient des commentaires additionnels et des exemples. Sitter (1992b) a utilisé la simulation pour comparer diverses méthodes de bootstrap avec et sans remise, et a trouvé qu'ils se comportent bien et sont comparables. La considérable complexité ajoutée par les méthodes miroir-concordance et de ré-échellonnage nous amène à préférer le bootstrap de rééchantillonnage simple dans les cas où la fraction d'échantillonnage est petite, et sinon d'utiliser l'approche de pseudo-population. Lahiri (2003) et Shao (2003) passent en revue comment les méthodes de bootstrap peuvent être appliquées aux données d'enquêtes complexes, tandis que Rust et Rao (1996) décrivent l'application des méthodes de réplication dans ce contexte. Presnell et Booth (1994) mettent l'accent sur des difficultés théoriques d'un grand nombre d'approches publiées.

7 Comparaisons empiriques

7.1 Cas simples

De nombreuses simulations ont été effectuées pour étudier la performance relative des estimateurs de variance dont nous avons discuté.

Kovar *et al.* (1988) ont considéré l'échantillonnage aléatoire simple à un degré stratifié avec remise basé sur des pseudo-données qui ressemblent aux vraies populations avec $H = 32$ strates à partir desquelles des échantillons de tailles n_h ou bien 2 ou bien 5 sont prélevés. Les paramètres qu'ils considèrent sont le rapport, le coefficient de régression, le coefficient de corrélation, et la médiane. Sur la base d'une mesure de biais et de stabilité, ils ont comparé le jackknife, la réplication équilibrée répétée, le bootstrap avec remise par ré-échellonnage et les méthodes de linéarisation. Ils observent que les estimateurs de variance de linéarisation et de jackknife se comportent de manière équivalente et ont la meilleure performance en termes de biais relatif et de stabilité, que la réplication équilibrée répétée a la deuxième meilleure performance, et que le bootstrap par ré-échellonnage de Rao et Wu (1988), la seule méthode bootstrap qu'ils considèrent, produit la pire performance de toutes. La réplication équilibrée répétée et ce bootstrap tendent à sur-estimer les variances de statistiques non linéaires. Ils considèrent aussi la couverture d'intervalles de confiance, et concluent que les intervalles jackknife et de linéarisation tendent à sous-couvrir, tandis que ceux basés sur la réplication équilibrée répétée tendent à sur-couvrir. Les intervalles de confiance bootstrap Studentisés améliorent la situation, mais tendent également à sur-couvrir.

Rao et Shao (1996) comparent les méthodes de réplification équilibrée répétée de groupe et de réplification équilibrée répétée de manière répétitive de groupe lorsque $n_h = 48$ dans chacun des $H = 5$ strates ; la fraction d'échantillonnage était d'environ 1/10. L'échantillonnage aléatoire simple au premier degré avec remise était utilisé à l'intérieur de chaque strate. Ils ont aussi considéré des paramètres divers, et observé que le jackknife marche le mieux lorsque l'estimateur ponctuel est lisse mais marche mal pour la médiane ; ceci ne nous surprend pas étant donné que l'estimateur de variance jackknife est inconsistant pour un quantile d'échantillon. La méthode de réplification de groupe équilibrée répétée a un petit biais relatif, mais est très instable. Cette instabilité est corrigée par la réplification équilibrée répétée groupée de manière répétitive, mais au prix de plus de calculs.

Sitter (1992b) a comparé diverses méthodes de bootstrap, à savoir le bootstrap avec remise, le bootstrap de ré-échellonnage, une version du bootstrap miroir-concordance et une version du bootstrap sans remise, sous échantillonnage aléatoire simple stratifié à un degré sans remise, et a conclu que les méthodes bootstrap sont comparables pour l'estimation de la variance de la médiane.

7.2 Calibration

7.2.1 Etude néerlandaises

Boonstra et Nieuwenbroek (2003) ont effectué une comparaison des variances de la linéarisation de Taylor et de la réplification équilibrée répétée pour des estimateurs de régression généralisée de totaux et de rapport de totaux à partir de deux enquêtes différentes, l'une étant un design deux-par-strate dessiné à partir d'une population de compagnies, et l'autre un design à deux degrés dessiné à partir d'une population de personnes.

Les formules de linéarisation de Taylor qui furent utilisées étaient de type standard, tandis que différentes variantes de réplification équilibrée répétée furent utilisées, incluant le regroupement d'unités d'échantillonnages primaires en deux groupes pour chaque strate, et l'utilisation de strates artificielles obtenus par division aléatoire de chaque strate d'unités d'échantillonnages primaires en sous-strates, qui sont chacune divisées en deux groupes aléatoires.

Une approche similaire a été appliquée à l'intérieur de chaque strate de première strate, entraînant un schéma de demi-échantillonnage plus complexe amenant aussi des perturbations plus douces des poids originaux à travers un facteur de Fay ε (§5.3).

La première simulation de Boonstra et Nieuwenbroek (2003) englobe 2000 échantillons pris dans 84 niveaux dont la taille N_h se trouve dans le domaine 10–20 : deux observations ont été prises dans chaque strate pour un total de 168 observations. On a effectué

la calibration basée sur les valeurs de deux variables catégoriques en utilisant un estimateur de régression du type décrit dans Section 2.2, et les estimateurs de variance ont été calculés en utilisant la linéarisation de Taylor, la linéarisation de Taylor avec des poids de régression, et la réplication équilibrée répétée avec divers facteurs de Fay. Ces estimateurs de Taylor ne tiennent pas compte de l'aléatoire des poids de régression. Les conclusions générales sont que toutes les méthodes sous-estiment légèrement les variances, et que la méthode de Taylor standard tend à sous-estimer le plus. La seule exception à ceci est la réplication équilibrée répétée standard, qui sur-estime quelque peu les vraies variances. Toutes les méthodes fournissent des intervalles de confiance à 95 % avec une couverture autour de 90–93 %; ceci est une conséquence de la sous-estimation de la variance.

Dans une deuxième étude, Boonstra et Nieuwenbroek (2003) ont créé une population artificielle en utilisant des données à partir de l'enquête néerlandaise à deux degrés de la force de travail et les a utilisées pour comparer diverses formes de réplication équilibrée répétée. La population artificielle comprend 188,216 personnes dans 70 strates contenant chacune de 10 à plusieurs centaines d'unités d'échantillonnage primaires. Chacune de ces unités contient de 15 à 40 ménages. Cinq cents échantillons ont été dessinés suivant un design à deux degrés avec échantillonnage aléatoire simple avec remise à chaque degré, avec des fractions d'échantillonnage de $1/5$ à chaque degré produisant des échantillons généraux de 4 % de la population. Les estimateurs de variance utilisés étaient les estimateurs de Taylor correspondant à ceux mentionnés dans le paragraphe précédent, la réplication équilibrée répétée groupée avec des strates artificielles pour un total de 120 strates et de 120 ré-échantillons. Toutes les estimations de réplication équilibrée répétée ont utilisé un facteur de Fay $\varepsilon = 0.57$.

La Table 1, qui contient quelques résultats typiques de cette étude, montre que la réplication équilibrée répétée est de manière appréciable moins efficace que ne l'est l'approche de linéarisation standard de Taylor, qui est peu affectée par l'utilisation des poids calibrés. L'utilisation de strates artificielles supplémentaires améliore légèrement la réplication équilibrée répétée, mais elle reste beaucoup moins efficace que l'approche standard.

7.2.2 Etudes britanniques et suisses

Deux exercices de simulation majeurs entrepris pour comparer les performances du rééchantillonnage et des méthodes plus standard pour l'estimation de variance dans des enquêtes de force de travail ont utilisé des données simulées basées sur de vraies enquêtes suisses et britanniques (Canty et Davison, 1999). Celles-ci comportaient des vagues de participants se chevauchant, et les statistiques considérées étaient les totaux, les rapports, et différences de totaux et rapports entre deux vagues successives. On a utilisé la calibration par ajustement itératif proportionnel pour ajuster les poids aux

Statistique	Méthodes d'estimation de variance	RB (%)	RMSE relatif (%)
Total	standard	2.2	15.4
	BRR regroupé	-0.1	29.2
	BRR regroupé, avec strates artificielles	2.1	22.5
Rapport	standard	0.1	4.8
	standard, avec poids de régression	-0.4	4.8
	BRR regroupé	-0.9	25.0
	BRR regroupé, avec strates artificielles	0.6	13.7

TAB. 1 – Performances typiques des estimateurs de variance standard de réplication équilibrée répétée, de Boonstra et Nieuwenbroek (2003). RB désigne le biais relatif et RMSE la racine du carré moyen de l'erreur des 500 estimations. BRR désigne la réplication équilibrée répétée.

marges des trois variables catégoriques, terminant après cinq itérations. L'effet de données manquantes n'a pas été considéré.

Nous résumons ici les résultats de l'étude britannique, où la population consistait en environ 60,000 adresses dans la première vague, avec autour de 12,000 nouvelles adresses dans la deuxième vague utilisées pour remplacer 12,000 adresses de la première vague ; soit autour de 20% d'unités changées entre deux vagues.

Il y avait 30 strates utilisées, chacune contenant quelques individus présents seulement dans la première vague, quelques uns présents seulement dans la deuxième vague, et certains présents dans les deux vagues, dans les proportions 1 :1 :3. La fraction d'échantillonnage était à peu près $f = 1/48$, donnant une taille d'échantillon de 1250 dans chaque vague échantillonnée.

Un total de 5000 échantillons a été pris à partir de la population artificielle suivant ce schéma d'échantillonnage, et différentes estimations de variance ont été calculées pour une variété de statistiques lisses, basées sur chacun de ces échantillons.

Comme les conclusions étaient similaires pour toutes les statistiques considérées, la discussion ci-dessous décrit seulement l'estimation de variances pour les totaux et pour les différences de totaux entre deux vagues. L'estimateur utilisé pour le total est

$$\hat{\tau} = \sum_{h,j,k} w_{hjk} y_{hjk},$$

où la somme est sur les strates h , les adresses j , et les personnes vivant à cette adresse k .

Les méthodes d'estimation de variance utilisés étaient :

- une méthode standard basée sur une formule obtenue en traitant $\hat{\tau}$ comme une variable binomiale avec dénominateur $n = \sum_h n_h$ le nombre total d’adresses échantillonnées, c’est-à-dire (Canty et Davison, 1999, equation (1))

$$d^2 \frac{\hat{\tau}(1 - \hat{\tau})}{n} \left(\sum_{h,j,k} w_{hjk} \right),$$

où d est un effet de design ;

- la linéarisation jackknife, où (12) est utilisée avec

$$l_{hj} = n_h \sum_k w_{hjk} e_{hjk} - \sum_i \sum_k w_{hik} e_{hik},$$

avec le résidu $e_{hjk} = y_{hjk} - \hat{y}_{hjk}$ d’une régression de la variable y sur les colonnes de la matrice X dont les lignes contiennent les variables de calibration x_{hjk} utilisant la matrice de poids diagonale Ω comprenant les poids originaux ω_{hjk} pour l’inclusion de l’individu k de l’adresse j de la strate h , et

$$w_{hjk} = \omega_{hjk} \left\{ 1 + (c^T - 1^T \Omega X)(X^T \Omega X)^{-1} x_{hjk}^T \right\}$$

sont les poids recalibrés ;

- le jackknife groupé, avec 10 groupes dans chacune des 30 strates, de sorte que chaque re-calcul de la variances nécessite 300 calculs de la statistique ;
- la réplication groupée équilibrée répétée, avec les 30 strates subdivisées en groupe suivant la présence de l’unité dans la première vague, la deuxième vague, ou les deux vagues, donnant un total de 90 répliques ; et
- le bootstrap standard, approprié étant donnée la faible fraction d’échantillonnage, avec 100 répliques.

L’effet de re-pondération a été étudié pour le jackknife, pour la réplication équilibrée répétée, et pour le bootstrap. Trois ensembles de poids ont été utilisés : ceux pour l’échantillon original, qui sont ‘incorrects’ parce qu’ils ne tiennent pas compte des changements dus au rééchantillonnage, et ceux obtenus après une et cinq itérations de l’algorithme d’ajustement proportionnel itératif utilisé pour la calibration.

La Table 2 présente des résultats typiques pour cette étude. Toutes ces méthodes de rééchantillonnage tiennent compte de la variabilité supplémentaire introduite par calibration, mais la méthode standard ne le fait pas, de sorte qu’elle tend à sous-estimer la vraie variance, bien que sa variance soit petite.

Le bootstrap et la linéarisation jackknife ont un carré moyen de l’erreur plus petit que le jackknife et que la réplication équilibrée répétée. La Figure 1 montre un résumé graphique des résultats correspondants pour l’estimation du taux de chômage, un rapport, pour lequel à la fois le numérateur et le dénominateur sont aléatoires, expliquant la différence dans l’échelle entre cette figure et Table 2.

Statistique	Méthode	Biais	RB (%)	SD	MSE
Total	standard	-35.2	-6.1	25.4	1880
	bootstrap	-14.6	-2.5	55.0	3230
	jackknife	45.3	7.8	96.8	11400
	linéarisation jackknife	-13.6	-2.4	38.4	1650
	BRR	28.0	4.8	79.6	7100
Changement dans le total	standard	-42.7	-7.2	21.6	2290
	bootstrap	-8.2	-1.4	55.5	3140
	jackknife	64.3	10.8	107.0	15600
	linéarisation jackknife	-4.7	-0.8	41.8	1770
	BRR	55.5	9.3	85.5	10400

TAB. 2 – Résumé de statistiques pour les écarts-types du total de chômage et du changement dans le total de chômage (Canty et Davison, 1999). RB désigne le biais relatif, SD l'écart type standard et MSE le carré moyen de l'erreur des 500 estimations. BRR désigne la réplication équilibrée répétée.

La performance relativement faible du jackknife et de la réplication équilibrée répétée est due à une certaine instabilité de ces estimateurs de variance, qui sous-estiment quelques fois de manière significative la vraie variance, bien qu'il semble y avoir un faible biais systématique.

Bien que cela ne soit pas en rapport immédiat ici, la bonne performance de la méthode standard, résulte de sa faible variabilité; elle sous-estime la vraie variance de manière consistante, et toujours à peu près de la même quantité.

L'avantage principal des méthodes de rééchantillonnage est de tenir compte de la variabilité supplémentaire introduite par la calibration, mais l'inconvénient correspondant est que la calibration doit être effectuée pour chaque estimation de rééchantillonnage. Ceci peut représenter un fardeau calculatoire supplémentaire lourd, et lorsque la calibration nécessite des itérations, il est naturel de se demander combien d'étapes itératives sont requises. Cette étude a comparé les effets de l'absence d'utilisation de pas, de l'utilisation d'un pas, et cinq pas dans l'algorithme d'ajustement proportionnel itératif, lorsqu'on calibre les jeux de données rééchantillonnées. Pour le bootstrap et la réplication équilibrée répétée, il semble que bien qu'il y ait une différence entre utiliser aucun pas ou bien un pas, il y a peu de changement après : un pas itératif est suffisant. Le jackknife présente des propriétés plutôt étranges, à savoir que cinq pas itératifs produisent des estimateurs fortement variables.

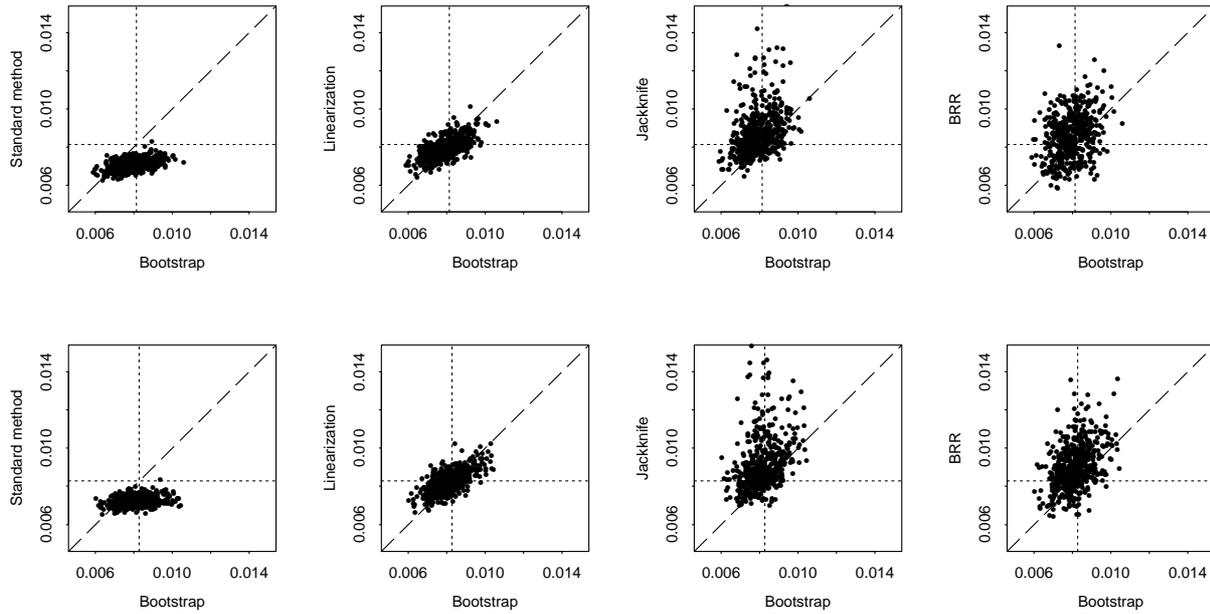


FIG. 1 – Ecart-types pour le taux de chômage (ligne du haut) et changement du taux de chômage (ligne du bas) ; la droite trait-tillée $y = x$, les lignes pointillées sont les ‘vrais’ écart-types d’échantillonnage. De Canty et Davison (1999).

7.3 Imputation

Kovar *et al.* (1988) a effectué une étude de simulation dans le but de tester la méthode de réplication équilibrée répétée ajustée au mécanisme d’imputation. Ils ont considéré l’échantillonnage stratifié de grappes avec $n_h = 2$ pour $H = 32$ strates, et des taux de manque de données de 10 %, 20%, 30%, 40%, et 50%. Pour les statistiques de moyenne et de médiane d’échantillons, ils ont observé que la méthode ajustée de réplication équilibrée répétée marche bien en terme de biais relatif et de coefficient de variation, contrairement à celui non ajusté qui sous-estime la variance des statistiques.

7.4 Imputation et calibration

Les simulations dont nous avons discuté ci-dessus ne traitent pas le cas de données calibrées et imputées. Sur la base d’une simulation réaliste basée sur l’enquête du budget des ménages suisses de l’année 1998 (Renfer, 2001), nous considérons l’estimateur calibré et imputé de Horvitz–Thompson du total des dépenses pour le pain et pour les produits céréaliers, sur la base de données complètes de $N = 9275$ ménages dans $H = 7$ strates de tailles variées.

On dispose également d’un ensemble de 14 variables auxiliaires sur chaque ménage, dont 10 marges de population sont connues. Pour la simulation, nous considérons les

$N = 9275$ ménages comme la population totale, pour laquelle nous connaissons les dépenses totales.

Nous effectuons alors un échantillonnage stratifié sans remise et avec probabilités d'inclusion égales $1/8$ à l'intérieur de 6 strates, et $3/8$ dans les autres strates, donnant une taille d'échantillon de 1332. On applique la non-réponse article pour la variable réponse en utilisant une probabilité de défaut uniforme à travers l'échantillon entier.

Nous calculons les estimations de Horvitz–Thomson calibrées et imputées sur chacun des 500 échantillons simulés, et appliquons diverses techniques d'estimation pour obtenir leur variances.

Pour atteindre la complexité calculatoire du bootstrap, pour lequel une simulation séparée a montré que 100 répliques étaient adéquates, nous utilisons à peu près le même nombre de blocs de suppressions que lorsqu'on applique le jackknife de bloc avec remise. Ceci a été appliqué avec 13 blocs choisis aléatoirement dans chaque strate, donnant environ 91 calculs en tout pour chaque estimation de variance jackknife.

On a appliqué deux formes de réplification équilibrée répétée, la première utilisant une séparation aléatoire de chaque strate en deux moitiés pour chaque réplification ; aucun facteur de Fay n'était utilisé mais les poids pour les observations incluses dans la réplique étaient multipliées par un facteur de deux avant la calibration. La deuxième forme, la réplification équilibrée répétée groupée de manière répétitive, prend la moyenne des estimations de variance de 13 telles séparations.

Le bootstrap a utilisé 100 répliques de l'estimateurs calibré et imputé de Horvitz–Thompson, obtenu par la procédure de Shao et Sitter (1996), c'est-à-dire, avec des réponses manquantes imputées de manière déterministe en utilisant un modèle linéaire ajusté aux répondants pleinement bootstrappés, et avec le jeux de donnée imputé calibré aux poids par régression linéaire.

L'estimateur de linéarisation jackknife est celui donné par (12) et (13).

On a appliqué les formules standard pour l'imputation multiple, en utilisant 30 imputations aléatoires à partir d'un modèles linéaire ajusté aux données complètes ; pour l'imputation paramétrique nous avons utilisé un modèle d'erreur normal homoscedastique, où les valeurs des paramètres de régression et de variance changent aléatoirement et indépendamment suivant les distributions ajustées normales et chi-carrées entre simulations ; tandis que les erreurs d'imputations non paramétriques ont été simulées suivant un résidu bootstrap basé sur un modèle (Davison et Hinkley, 1997, p. 262).

Les Figures 2 et 3 comparent les performances de ces techniques d'estimation de variance pour des taux de défaut de 0%, 20%, 40%, and 60%. Le jackknife de bloc sous-estime les vraies variances, qui sont systématiquement sur-estimées par la réplification équilibrée répétée groupée de manière répétitive.

La réplification équilibrée répétée sans regroupement répété est clairement fortement

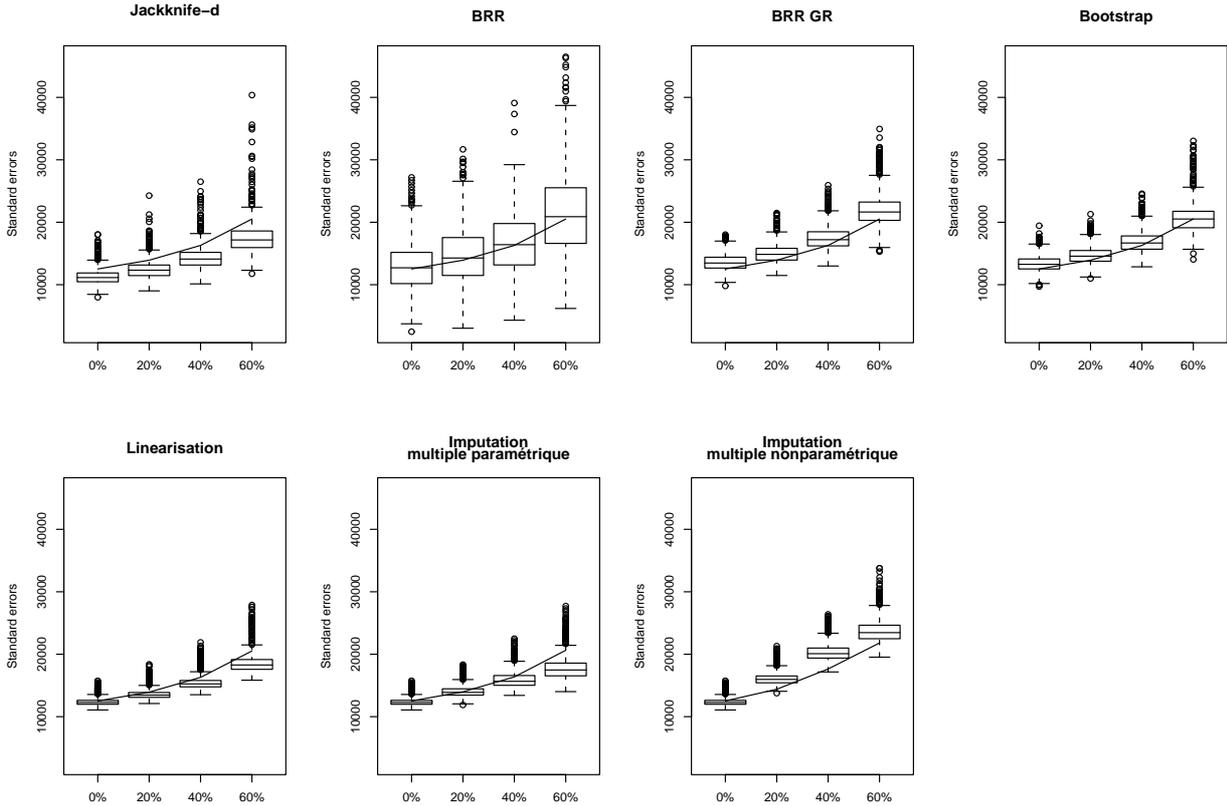


FIG. 2 – Comparaison entre estimateurs de variance de rééchantillonnage en présence de calibration et d'imputation, en tant que fonction de la proportion de données manquantes. Simulation basée sur l'enquêtes du budget des ménages suisses en 1998. La ligne montre les vraies variances, estimée à partir de 10,000 simulations, et les boxplots montrent les écart-types calculés pour 500 échantillons.

variable en comparaison, en accord avec les résultats de Rao et Shao (1996), mais la réplication équilibrée répétée groupée marche beaucoup mieux, bien qu'elle sur-estime la variance lorsqu'il y a un fort taux de défaut de réponses.

La linéarisation jackknife fonctionne bien pour de faible niveaux de défauts, et de manière générale elle produit des variances qui sont plutôt trop basses mais assez stables. Le bootstrap se comporte bien en terme à la fois de biais et de variance, et semble être la meilleure de ces méthodes. Les méthodes d'imputation multiple produisent des erreurs dans des directions opposées, lorsqu'il manque une fraction substantielle des données.

De manière générale l'approche bootstrap de Shao et Sitter (1996) et la méthode de linéarisation de la Section 3.2 semblent les meilleures en termes de biais et de stabilité. Si on se préoccupe du temps calcul, l'avantage va à la linéarisation, qui est jusqu'à cinquante fois plus rapide que les autres méthodes incluses dans cette étude.

Une simulation séparée a été effectuée pour vérifier l'effet de l'utilisation de strates

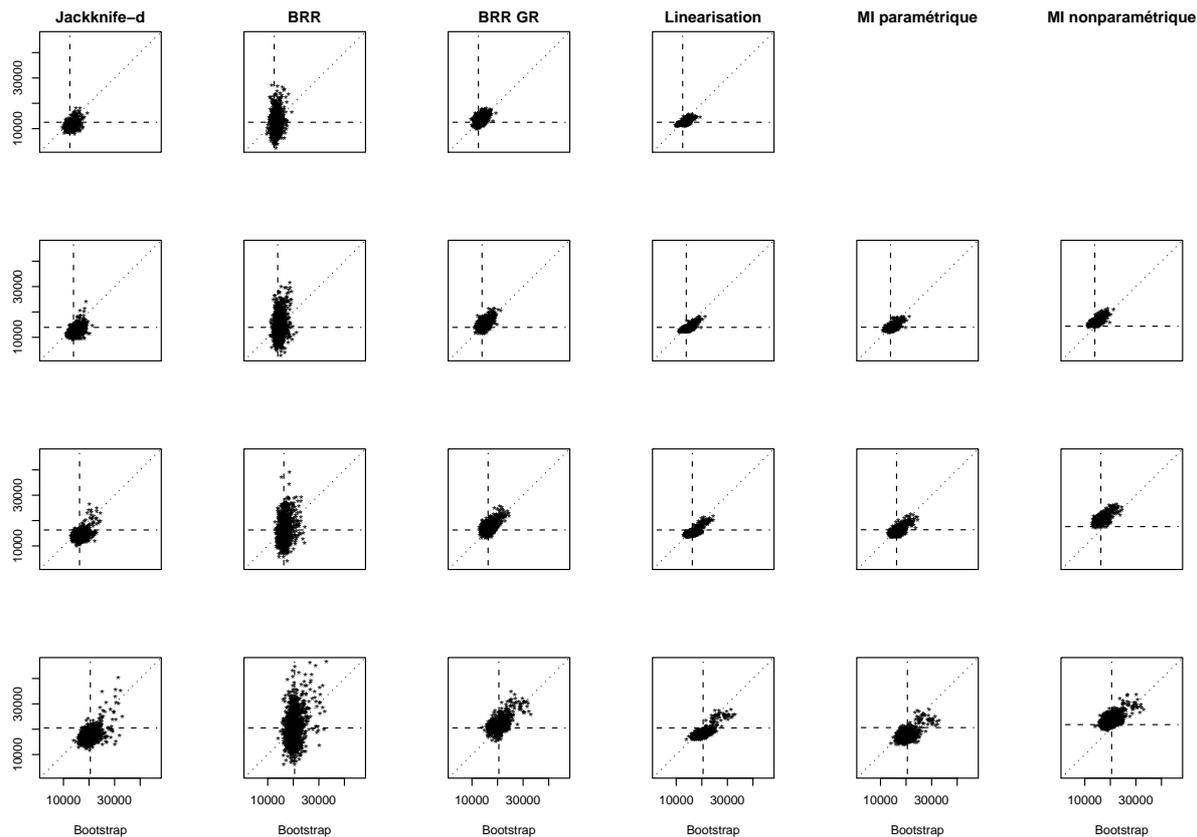


FIG. 3 – Comparaison des erreurs standard de rééchantillonnage en présence de calibration et d'imputation; de haut en bas 0%, 20%, 40%, 60% non-réponses unité. Les traits-tillés sont les 'vrais' écarts-types d'échantillonnage, et la ligne pointillée montre $x = y$. Simulation basée sur l'enquête de budget des ménages suisses de 1998.

artificielles pour la réplication équilibrée répétée. Chacune des $H = 7$ strates a été divisée en quatre parties, et les estimateurs de variances ont été calculés pour 1000 échantillons avec des réponses complètes. Les résultats montrent que l'introduction des strates artificielles améliore beaucoup les estimateurs de variances de réplication équilibrée répétée.

8 Discussion

Il semble de manière générale que les estimateurs de variance bootstrap et de linéarisation sont des plus prometteurs pour l'utilisation dans les enquêtes complexes. Le bootstrap a l'avantage d'être un outil à utilités multiples, d'être applicable tel quel dans beaucoup de situations, et d'être utilisable à la fois pour des statistiques lisses et non lisses. De plus, et contrairement au jackknife ou à la réplication équilibrée répétée, le nombre de re-calculs requis est plutôt contrôlé par l'utilisateur que déterminé par la méthode. Ses inconvénients principaux sont le fardeau calculatoire qu'il nécessite, par-

ticulièrement lorsqu'il est utilisé avec l'imputation, et le fait que de la programmation spéciale est nécessaire s'il est appliqué à des situations avec des grandes fractions d'échantillonnages f_h

La linéarisation jackknife demande des calculs spéciaux des valeurs de la fonction d'influence adapté à des circonstances particulières, mais elle n'entraîne pas de rééchantillonnage et est ainsi beaucoup plus rapide que le bootstrap. Elle ne peut pas être appliquée de manière sûre pour des statistiques non lisses, mais la règle d'enchaînement peut être utilisées pour obtenir des valeurs d'influence pour des estimateurs complexes.

La réplication équilibrée répétée et le jackknife sont presque compétitifs dans certains cas, mais en général ils se comportent moins bien que les autres méthodes, et il semble nécessaire de les ajuster pour obtenir de meilleures performances.

La linéarisation jackknife et le bootstrap peuvent être appliqués pour des estimateurs de changement

$$\hat{\theta} = t(\hat{F}_2, \hat{F}_3) - t(\hat{F}_1, \hat{F}_2),$$

basés sur des enquêtes de panel en séparant les unités en trois parties : celles présentes seulement à la première occasion d'échantillonnage, celles présentes lors des deux occasions, et celles présentes seulement à la deuxième occasion (Canty et Davison, 1999) ; celle-ci sont représentées respectivement par \hat{F}_1 , \hat{F}_3 et \hat{F}_2 . Chacune des H strates originales est séparée en trois parties et l'estimateur est calculé en fonction. Si nous ignorons les corrections pour les fractions d'échantillonnage, l'estimateur de variance de linéarisation jackknife est donné par

$$\sum_{h=1}^H \left\{ \frac{1}{n_{h1}(n_{h1} - 1)} \sum_{j=1}^{n_{h1}} (l_{hj}^1)^2 + \frac{1}{n_{h2}(n_{h2} - 1)} \sum_{j=1}^{n_{h2}} (l_{hj}^2)^2 + \frac{1}{n_{h3}(n_{h3} - 1)} \sum_{j=1}^{n_{h3}} (l_{hj}^3)^2 \right\},$$

où la h -ème strate contient n_{h1} unités présentes à la première occasion, n_{h2} présentes aux deux occasions, et n_{h3} présentes à la deuxième occasion seulement, et où l_{hj}^k sont les influence empiriques correspondantes. Un argument similaire s'applique au bootstrap, pour lequel on applique le rééchantillonnage avec ces mêmes sous-strates, et aux autres méthodes de rééchantillonnage.

Remerciements

Ce travail a été conduit avec le soutien financier de l'Union Européenne, dans le contexte du projet DACSEIS (<http://www.dacseis.de>). Nous remercions les autres membres de l'équipe DACSEIS pour leur collaboration. L'article a été grandement amélioré grâce aux commentaires de Siegfried Gabler et Chris Skinner. Nous remercions également Andreas Kaufmann pour son aide précieuse pour la préparation de ce document.

Références

- Berger, Y. G. et Skinner, C. J. (2003) Variance estimation for a low income proportion. *Applied Statistics* **52**, 457–468.
- Berger, Y. G. et Skinner, C. J. (2005) A jackknife variance estimator for unequal probability sampling.
- Bickel, P. J. et Freedman, D. A. (1984) Asymptotic normality and the bootstrap in stratified sampling. *Annals of Statistics* **12**, 470–482.
- Binder, D. A. (1996) Linearization methods for single phase and two-phase samples : A cookbook approach. *Survey Methodology* **22**, 17–22.
- Boonstra, H. J. H. et Nieuwenbroek, N. (2003) *An empirical comparison of BRR and linearization variance estimators*. DACSEIS.
- Booth, J. G., Butler, R. W. et Hall, P. (1994) Bootstrap methods for finite populations. *Journal of the American Statistical Association* **89**, 1282–1289.
- Campbell, C. (1980) A different view of finite population estimation. In *Proceedings of the Section on Survey Research Methods*, pp. 319–324. Alexandria, Virginia : American Statistical Association.
- Canty, A. J. et Davison, A. C. (1999) Resampling-based variance estimation for labour force surveys. *The Statistician* **48**, 379–391.
- Chao, M. T. et Lo, S. H. (1985) A bootstrap method for finite populations. *Sankhyā A* **47**, 399–405.
- Chen, J. et Shao, J. (2001) Jackknife variance estimation for nearest-neighbor imputation. *Journal of the American Statistical Association* **96**, 260–269.
- Cochran, W. G. (1977) *Sampling Techniques*. Third edition. New York : Wiley.
- Davison, A. C. et Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. Cambridge : Cambridge University Press.
- Deville, J. C. (1999) Variance estimation for complex statistics and estimators : Linearization and residual techniques. *Survey Methodology* **25**, 193–203.
- Deville, J. C. et Särndal, C. E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.
- Deville, J. C., Särndal, C. E. et Sautory, O. (1993) Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* **88**, 1013–1020.

- Dippo, C. S., Fay, R. E. et Morganstein, D. H. (1984) Computing variances from complex samples with replicate weights. In *Proceedings of the Section on Survey Research Methods*, pp. 489–494. Washington DC : American Statistical Association.
- Efron, B. (1979) Bootstrap methods : Another look at the jackknife. *Annals of Statistics* **7**, 1–26.
- Efron, B. (1982) *The Jackknife, the Bootstrap, and other Resampling Plans*. Philadelphia : SIAM.
- Fay, R. E. (1989) Theory and application of replicate weighting for variance calculations. In *Proceedings of the Social Statistics Section*, pp. 212–217. American Statistical Association.
- Fay, R. E. (1996) Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association* **91**, 490–498.
- Gross, S. (1980) Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, pp. 181–184. Alexandria, VA : American Statistical Association.
- Gupta, V. K. et Nigam, A. K. (1987) Mixed orthogonal arrays for variance estimation with unequal numbers of primary selections per stratum. *Biometrika* **74**, 735–742.
- Gurney, M. et Jewett, R. S. (1975) Constructing orthogonal replications for standard errors. *Journal of the American Statistical Association* **70**, 819–821.
- Hall, P. G. (2003) A short prehistory of the bootstrap. *Statistical Science* **18**, 158–167.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. et Stahel, W. A. (1986) *Robust Statistics : The Approach Based on Influence Functions*. New York : Wiley.
- Hartigan, J. A. (1969) Using subsample values as typical values. *Journal of the American Statistical Association* **64**, 1303–1317.
- Huber, P. J. (1981) *Robust Statistics*. New York : Wiley.
- Judkins, D. R. (1990) Fay’s method of variance estimation. *Journal of Official Statistics* **6**, 223–239.
- Kim, J. K. et Sitter, R. R. (2003) Efficient replication variance estimation for two-phase sampling. *Statistica Sinica* **13**, 641–653.
- Kovar, J. G., Rao, J. N. K. et Wu, C. F. J. (1988) Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics* **16**, 25–45.
- Krewski, D. et Rao, J. N. K. (1981) Inference from stratified samples : Properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics* **9**, 1010–1019.

- Lahiri, P. (2003) On the impact of bootstrap in survey sampling and small-area estimation. *Statistical Science* **18**, 199–210.
- McCarthy, P. J. (1969) Pseudo-replication : Half samples. *Review of the Interational Statistics Institute* **37**, 239–264.
- McCarthy, P. J. et Snowden, C. B. (1985) The bootstrap and finite population sampling. *Vital and Health Statistics* **2**, 2–95.
- Miller, R. G. (1974) The jackknife - A review. *Biometrika* **61**, 1–15.
- Nielsen, S. F. (2003) Proper and improper multiple imputation (with discussion). *International Statistical Review* **71**, 593–627.
- Politis, D. N., Romano, J. P. et Wolf, M. (1999) *Subsampling*. New York : Springer-Verlag.
- Presnell, B. et Booth, J. G. (1994) Resampling methods for sample surveys. Technical Report 470, Department of Statistics, University of Florida, Gainesville.
- Quenouille, M. H. (1949a) Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society, Series B* **11**, 68–84.
- Quenouille, M. H. (1949b) Notes on bias in estimation. *Biometrika* **43**, 353–360.
- Rao, J. N. K. (1988) Variance estimation in sample surveys. In *Handbook of Statistics, Volume 6*, ed. P. R. K. amd C. R. Rao, pp. 427–447. Amsterdam : Elsevier Science.
- Rao, J. N. K. et Shao, J. (1992) Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**, 811–822.
- Rao, J. N. K. et Shao, J. (1996) On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association* **91**, 343–348.
- Rao, J. N. K. et Shao, J. (1999) Modified balanced repeated replication for complex survey data. *Biometrika* **86**, 403–415.
- Rao, J. N. K. et Wu, C. F. J. (1985) Inference from stratified samples : Second-order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association* **80**, 620–630.
- Rao, J. N. K. et Wu, C. F. J. (1988) Resampling inference with complex survey data. *Journal of the American Statistical Association* **83**, 231–241.
- Rao, J. N. K., Wu, C. F. J. et Yue, K. (1992) Some recent work on resampling methods for complex surveys. *Survey Methodology* **18**, 209–217.
- Renfer, J.-P. (2001) *Description and process of the Household and Budget Survey of 1998 (HBS 1998)*. Swiss Federal Statistical Office. 1-19.

- Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York : Wiley.
- Rubin, D. B. (1996) Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489. with discussion, 507-515, and rejoinder, 515-517.
- Rust, K. F. et Rao, J. N. K. (1996) Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research* **5**, 283–310.
- Shao, J. (2003) Impact of the bootstrap on sample surveys. *Statistical Science* **18**, 191–198.
- Shao, J., Chen, Y. et Chen, Y. (1998) Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association* **93**, 819–831.
- Shao, J. et Rao, J. N. K. (1994) Standard errors for low income proportions estimated from stratified multistage samples. *Sankhyā B* **55**, 393–414.
- Shao, J. et Sitter, R. R. (1996) Bootstrap for imputed survey data. *Journal of the American Statistical Association* **91**, 1278–1288.
- Shao, J. et Tu, D. (1995) *The Jackknife and Bootstrap*. New York : Springer-Verlag.
- Shao, J. et Wu, C. F. J. (1989) A general theory for jackknife variance estimation. *Annals of Statistics* **17**, 1176–1197.
- Shao, J. et Wu, C. F. J. (1992) Asymptotic properties of the balanced repeated replication method for sample quantiles. *Annals of Statistics* **20**, 1571–1593.
- Sitter, R. R. (1992a) A resampling procedure for complex survey data. *Journal of the American Statistical Association* **87**, 755–765.
- Sitter, R. R. (1992b) Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics* **20**, 135–154.
- Sitter, R. R. (1993) Balanced repeated replications based on orthogonal multi-arrays. *Biometrika* **80**, 211–221.
- Tukey, J. W. (1958) Bias and confidence in not quite large samples (abstract). *Annals of Mathematical Statistics* **29**, 614.
- Wolter, K. M. (1985) *Introduction to Variance Estimation*. New York : Springer-Verlag.
- Wu, C. F. J. (1991) Balanced repeated replications based on mixed orthogonal arrays. *Biometrika* **78**, 181–188.
- Zhang, P. (2003) Multiple imputation : Theory and method. *International Statistical Review* **71**, 581–592.