ESTIMATION DE LA FONCTION DE RÉPARTITION ET DES FRACTILES D'UNE POPULATION FINIE

R. REN

INSEE, Unité Méthodes Statistiques

1. Introduction

L'estimation de la moyenne ou du total d'une population finie et l'estimation de la précision de ces estimateurs sont les deux parties principales de la théorie des sondages, tandis que l'estimation de la fonction de répartition et l'estimation des fractiles d'une population finie sont moins bien reconnues par les statisticiens d'enquête. On ne la trouve pas dans les ouvrages classiques (Kish (1965), Cochran (1977), Sukhatme (1984)) et on en parle peu dans la littérature. Särndal, Swensson et Wretman (1992) ont parlé un peu de ce problème dans le cadre de l'estimation de la médiane. En fait, la fonction de répartition joue un rôle important dans le domaine de la statistique classique. Si on connaît cette fonction, tous les paramètres importants (par exemple, la moyenne, la variance, les fractiles et le mode s'ils existent) deviennent connus. De ce fait, si on peut établir une estimation précise de la fonction de répartition d'une variable, on peut aussi obtenir des estimations précises des paramètres de la loi de cette variable. Par exemple, si $\hat{F}_y$ est un estimateur de la fonction de répartition F_y et $\theta(F_y)$ un paramètre, un estimateur naturel de $\theta(F_y)$ est obtenu par $\hat{\theta}(F_y) = \theta(\hat{F}_y)$. Cet estimateur peut être amélioré lorsqu'on dispose d'informations auxiliaires, par exemple, sur la fonction de répartition d'une variable auxiliaire connue dans la base de sondage.

A la suite des idées de Sedransk et Sedransk (1979), le premier article concernant l'estimation de la fonction de répartition d'une population finie parut dans Chambers et Dunstan (1986), avec, depuis cette date, une dizaine d'articles publiés sur ce sujet : Kuk (1988, 1993), Kuo (1988), Dunstan et Chambers (1989), Rao, Kovar et Mantel (1990), Chambers, Dorfman et Hall (1992), Rao et Liu (1992), Chambers, Dorfman et Wehrly (1993), Chen et Qin (1993), Dorfman (1993), Kuk et Mak (1994), Nascimento Silva et Skinner (1995).

Les méthodes d'estimation de la fonction de répartition proposées par ces articles peuvent être rangées en deux classes, comme pour l'estimation de la moyenne : les méthodes fondées sur le plan de sondage et les méthodes à l'aide d'un modèle de surpopulation. Dorfman (1993) a fait une comparaison de ces deux méthodes. La méthode de post-stratification (Nascimento Silva et Skinner (1995)) et la méthode du noyau (Kuo (1988), Kuk (1993)) sont aussi appliquées. Il y a un point que nous voudrions souligner : ces méthodes et les estimateurs sont obtenus par des démarches plus ou moins indépendantes et il manque une théorie unificatrice. L'objectif de cet article est d'unifier les méthodes proposées dans la littérature, de développer une théorie plus approfondie sur l'estimation de la fonction de répartition en intégrant les estimateurs proposés dans une même théorie, et de trouver de nouveaux estimateurs. Les méthodes de calage développées dans Ren et Deville (1998, 2000a, 2000b) sont en particulier étudiées lorsque l'information auxiliaire a la forme d'une fonction de répartition. Des comparaisons des différentes méthodes sont exposées et validées par des simulations numériques.

2. Les méthodes fondées sur le plan de sondage

Supposons une population finie de taille N, avec la variable d'intérêt Y, l'individu i a la valeur caractéristique Y_i, i=1, 2, ... , N. La fonction de répartition de Y sur cette population finie est définie par :

$$F_y(t) = \frac{1}{N}\sum_{i\in U}\Delta(t - Y_i), \tag{2.1}$$

où $\Delta(t - Y_i) = 1, \text{si } t \geq Y_i$ et $\Delta(t - Y_i) = 0, \text{si } t < Y_i$. L'objectif est d'estimer cette fonction à partir de l'échantillon.

Pour chaque t donné, $F_y(t)$ est une moyenne sur la population d'une variable transformée : $\Delta(t - Y)$. Donc, toutes les méthodes que l'on connaît pour l'estimation de la moyenne ou du total peuvent être adaptées pour l'estimation de la fonction de répartition. Soit P un plan de sondage quelconque, la probabilité d'inclusion de l'unité i dans l'échantillon est notée par π_i , i=1, 2, ... , N. Supposons que les plans de sondage considérés soient des plans avec des probabilités d'inclusion strictement positives : $\pi_i > 0$, pour tout i dans la population. Soit s un échantillon tiré au sein de la population selon le plan de sondage P ; les observations sur l'échantillon de la variable d'intérêt sont notées $Y_i, i \in s$.

2.1 Estimateur de Horvitz-Thompson

Un estimateur naturel de $F_y(t)$ est l'estimateur de Horvitz-Thompson (Kuk (1988)) :

$$\hat{F}_{HTY}(t) = \frac{1}{N}\sum_{i\in s} d_i \Delta(t - Y_i) \ , \tag{2.2}$$

où $d_i = \pi_i^{-1}$ est le poids de sondage associé à l'individu *i*. Pour chaque *t* donné, $\hat{F}_{HTY}(t)$ est un estimateur sans biais de $F_y(t)$, de variance :

$$V\left(\hat{F}_{HTY}(t)\right) = \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U} a_{ij}\Delta(t - Y_i)\Delta(t - Y_j) \ , \tag{2.3}$$

où $a_{ij} = (\pi_{ij} - \pi_i\pi_j) / \pi_i\pi_j$ avec $\pi_{ij} = \pi_i$ si $i = j$. Cette variance peut être estimée sans biais, lorsque $\pi_{ij} > 0$ pour tous *i* et *j*, par :

$$\hat{V}\left(\hat{F}_{HTY}(t)\right) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s} a_{ij}\pi_{ij}^{-1}\Delta(t - Y_i)\Delta(t - Y_j) \ . \tag{2.4}$$

$\hat{F}_{HTY}$ est une fonction constante par morceaux, monotone croissante, mais n'est pas nécessairement une fonction de répartition car $\hat{F}_{HTY}(+\infty) \neq 1$, si $\sum_{i\in s} d_i \neq N$. D'autre part, si *N* est inconnu a priori, $\hat{F}_{HTY}$ n'est plus calculable. En normalisant $\hat{F}_{HTY}$, on obtient :

$$\hat{F}_{NY}(t) = \sum_{i\in s} d_i\Delta(t - Y_i) / \sum_{i\in s} d_i \ , \tag{2.5}$$

fonction monotone croissante et telle que et $\hat{F}_{NY}(+\infty) = 1$. C'est un estimateur par le ratio généralisé ; il est donc asymptotiquement sans biais de $F_y(t)$ avec une erreur quadratique moyenne approchée :

$$MSE\left(\hat{F}_{NY}(t)\right) = \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U} a_{ij}\left(\Delta(t - Y_i) - F_y(t)\right)\left(\Delta(t - Y_j) - F_y(t)\right).$$

Lorsque $\pi_{ij} > 0$ pour tous *i* et *j*, cette erreur quadratique moyenne s'estime par :

$$\hat{MSE}\left(\hat{F}_{NY}(t)\right) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s} a_{ij}\pi_{ij}^{-1}\left(\Delta(t - Y_i) - \hat{F}_{NY}(t)\right)\left(\Delta(t - Y_j) - \hat{F}_{NY}(t)\right).$$

2.2 Estimateur par le ratio

Supposons mainteant des informations auxiliaires, par exemple, la fonction de répartition F_x d'une variable auxiliaire X (univariée) :

$$F_x(t) = \frac{1}{N}\sum_{i\in U}\Delta(t - X_i). \tag{2.6}$$

Les observations de cette variable sur l'échantillon s sont notées X_i, $i \in s$. Les méthodes de l'estimation par le ratio et par la différence peuvent alors être adaptées à l'estimation de la fonction de répartition.

Notons $\hat{F}_{HTX}$ l'estimateur de Horvitz-Thompson de F_x ; nous pouvons introduire l'estimateur par le ratio généralisé :

$$\hat{F}_{RY}(t) = \frac{\sum_{i\in s} d_i\Delta(t - Y_i)}{\sum_{i\in s} d_i\Delta(t - X_i)} F_x(t) = \frac{\hat{F}_{HTY}(t)}{\hat{F}_{HTX}(t)} F_x(t), \tag{2.7}$$

défini sur l'ensemble des valeurs t, où $\hat{F}_{HTX}(t)$ est non nul. Cet estimateur est ici appliqué aux variables transformées $\Delta(t - Y)$ et $\Delta(t - X)$. On notera que les conditions de proportionnalité entre les variables, souhaitables pour l'utilisation d'un tel estimateur, ont peu de chance d'être satisfaites, surtout uniformément par rapport à t.

Nous savons que, quand la taille de l'échantillon n est grande, cet estimateur est asymptotiquement sans biais avec une erreur quadratique moyenne approchée :

$$\begin{aligned} MSE(\hat{F}_{RY}(t)) = \frac{1}{N^2}\sum_{i\in u}\sum_{j\in U} a_{ij}\left[\Delta(t - Y_i) - R(t)\Delta(t - X_i)\right]\times \\ \left[\Delta(t - Y_j) - R(t)\Delta(t - X_j)\right], \end{aligned} \tag{2.8}$$

où a_{ij} est défini dans l'expression (2.3), $R(t)$ est le ratio :

$$R(t) = \frac{F_y(t)}{F_x(t)}, \tag{2.9}$$

défini pour les valeurs t telle que $F_x(t) \neq 0$. Cette erreur peut être estimée asymptotiquement sans biais, lorsque $\pi_{ij} > 0$ pour tous i et j, par :

$$\begin{aligned} \hat{MSE}(\hat{F}_{RY}(t)) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s} a_{ij}\pi_{ij}^{-1}\left[\Delta(t - Y_i) - \hat{R}(t)\Delta(t - X_i)\right]\times \\ \left[\Delta(t - Y_j) - \hat{R}(t)\Delta(t - X_j\right], \end{aligned} \tag{2.10}$$

où $\hat{R}(t) = \hat{F}_{HTY}(t) / \hat{F}_{HTX}(t)$ est un estimateur de $R(t)$ défini pour les valeurs t, où $\hat{F}_{HTX}(t) \neq 0$.

$\hat{F}_{RY}$ est un estimateur consistant au sens de Cochran, c'est-à-dire qu'il vérifie $\hat{F}_{RY}(t) = F_y(t)$ quand $n = N$, pour toute valeur $t \in \mathrm{R}$ telle que $F_x(t) \neq 0$. Cependant, cette propriété n'est plus satisfaite lorsque la variable auxiliaire X est proportionnelle à la variable d'intérêt Y. On peut considérer ce dernier point comme un défaut de l'estimateur. Ceci provient du fait que la proportionnalité entre Y et X n'implique pas la proportionnalité entre les variables transformées $\Delta(t - Y)$ et $\Delta(t - X)$. D'autre part, quand Y est beaucoup plus petit que X, $\hat{F}_{RY}(t)$ n'a pas beaucoup de sens. L'intervalle de définition de $\hat{F}_{RY}(t)$ peut par exemple être vide. Cet inconvénient peut être contourné, lorsque le ratio $R = \sum_{i \in U} Y_i / \sum_{i \in U} X_i$ est connu, en prenant RX comme nouvelle variable auxiliaire, avec comme fonction de répartition, $F_{RX}(t) = F_x(R^{-1}t)$. L'estimateur corrigé est:

$$\widetilde{F}_{RY}(t) = \frac{\sum_{i \in s} d_i \Delta(t - Y_i)}{\sum_{i \in s} d_i \Delta(t - RX_i)} F_x(R^{-1}t) = \frac{\hat{F}_{HTY}(t)}{\hat{F}_{HTX}(R^{-1}t)} F_x(R^{-1}t). \quad (2.11)$$

Alors, $\widetilde{F}_{RY}$ est un estimateur consistant au sens de Cochran et consistant aux informations auxiliaires, c'est-à-dire qu'il vérifie $\widetilde{F}_{RY}(t) = F_y(t)$ lorsque n=N ou X et Y sont proportionnelles. $\widetilde{F}_{RY}(t)$ est aussi un estimateur asymptotiquement sans biais avec une erreur quadratique moyenne approchée :

$$MSE(\widetilde{F}_{RY}(t)) = \frac{1}{N^2} \sum_{i \in U} \sum_{j \in U} a_{ij} \left[\Delta(t - Y_i) - R^*(t)\Delta(t - RX_i)\right] \times \left[\Delta(t - Y_j) - R^*(t)\Delta(t - RX_j)\right], \quad (2.12)$$

où $R^*(t) = F_y(t) / F_x(R^{-1}t)$ est un ratio défini sur l'ensemble des valeurs t, où $F_x(R^{-1}t)$ est non nul. Lorsque $\pi_{ij} > 0$ pour tous i et j, cette erreur est estimée par :

$$\hat{MSE}(\widetilde{F}_{RY}(t)) = \frac{1}{N^2} \sum_{i \in s} \sum_{j \in s} \frac{a_{ij}}{\pi_{ij}} \left[\Delta(t - Y_i) - \hat{R}^*(t)\Delta(t - RX_i)\right] \times \left[\Delta(t - Y_j) - \hat{R}^*(t)\Delta(t - RX_j)\right], \quad (2.13)$$

où $\hat{R}^*(t) = \hat{F}_{HTY}(t) / \hat{F}_{HTX}(R^{-1}t)$ est défini sur l'ensemble des valeurs t, où $\hat{F}_{HTX}(R^{-1}t)$ est non nul.

Comme le ratio R est généralement inconnu, l'estimateur $\widetilde{F}_{RY}$ est en pratique obtenu en remplaçant R par $\hat{R} = \sum_{i \in s} d_i Y_i / \sum_{i \in s} d_i X_i$ dans (2.11), ceci donne l'estimateur de Rao, Kovar et Mantel (1990):

$$\hat{F}_{RY}(t) = \frac{\sum_{i \in s} d_i \Delta(t - Y_i)}{\sum_{i \in s} d_i \Delta(t - \hat{R} X_i)} F_x(\hat{R}^{-1}t) . \qquad (2.14)$$

Dans ce cas, (2.12) désigne encore l'erreur quadratique moyenne approchée de $\hat{F}_{RY}$, dont un estimateur est obtenu en remplaçant R par $\hat{R}$ dans (2.13).

2.3 Estimateur par différence

La même idée que pour la construction de l'estimateur par différence généralisée pour la moyenne de la population, appliquée aux variables transformées $\Delta(t - Y)$ et $\Delta(t - RX)$, donne l'estimateur par différence généralisée :

$$\widetilde{F}_D(t) = \hat{F}_{NY}(t) + \left[F_x(R^{-1}t) - \hat{F}_{NX}(R^{-1}t)\right] , \qquad (2.15)$$

où $\hat{F}_{NX}(R^{-1}t)$ est l'analogue de $\hat{F}_{NY}(t)$ pour la variable RX. $\widetilde{F}_D$ est un estimateur asymptotiquement sans biais de F_y, avec l'erreur quadratique moyenne approchée :

$$MSE\left(\widetilde{F}_D(t)\right) = \sum_{i \in U} \sum_{j \in U} a_{ij} \left(\Delta(t - Y_i) - \Delta(t - RX_i)\right) \times \left(\Delta(t - Y_j) - \Delta(t - RX_j)\right) , \qquad (2.16)$$

Une estimation de cette erreur est directe, lorsque $\pi_{ij} > 0$ pour tous i et j :

$$\hat{MSE}\left(\widetilde{F}_D(t)\right) = \sum_{i \in s} \sum_{j \in s} a_{ij} \pi_{ij}^{-1} \left(\Delta(t - Y_i) - \Delta(t - RX_i)\right) \times \left(\Delta(t - Y_j) - \Delta(t - RX_j)\right) . \qquad (2.17)$$

Comme dans le cas de l'estimateur par le ratio, $\widetilde{\bar{F}}_D$ dépend du ratio inconnu R. En remplaçant R par $\hat{R} = \sum_{i \in s} d_i Y_i / \sum_{i \in s} d_i X_i$, on obtient

$$\hat{F}_D(t) = \hat{F}_{NY}(t) + \left[F_x(\hat{R}^{-1}t) - \hat{F}_{NX}(\hat{R}^{-1}t)\right]. \qquad (2.18)$$

Dans ce cas, (2.16) désigne encore l'erreur quadratique moyenne approchée de $\hat{F}_D$, dont un estimateur est obtenu en remplaçant R par $\hat{R} = \sum_{i \in s} d_i Y_i / \sum_{i \in s} d_i X_i$ dans (2.17). Lorsqu'on utilise les estimateurs de Horvitz-Thompson $\hat{F}_{HTY}(t)$ et $\hat{F}_{HTX}(R^{-1}t)$ à la place de $\hat{F}_{NY}(t)$ et $\hat{F}_{NX}(R^{-1}t)$ dans (2.18), on obtient l'estimateur de Rao, Kovar et Mantel (1990) :

$$\hat{F}_d(t) = \hat{F}_{HTY}(t) + \left[F_x(\hat{R}^{-1}t) - \hat{F}_{HTX}(\hat{R}^{-1}t)\right]. \qquad (2.19)$$

L'inconvénient est que $\hat{F}_d$ n'est pas nécessairement une fonction de répartition de même que $\hat{F}_{HTY}$ dans (2.2).

3. Les méthodes fondées sur des modèles de surpopulation

Comme dans la théorie traditionnelle des sondages, on distingue les méthodes d'estimation fondées sur le plan de sondage et celles utilisant un modèle de surpopulation [voir Basu (1971), Cassel, Särndal et Wretman (1977)]. Les méthodes fondées sur le plan de sondage ne supposent pas d'information préalable concernant la relation entre la variable auxiliaire et la variable d'intérêt. Toutes les inférences statistiques sont établies sur la seule base du plan de sondage, tandis que les méthodes utilisant un modèle supposent un modèle liant la variable auxiliaire à la variable d'intérêt.

3.1 Un modèle de surpopulation

Nous supposons que la variable d'intérêt Y est liée à la variable auxiliaire X connue sur la population finie par un modèle:

$$Y_i = a(X_i, \varepsilon_i, \beta), \quad i=1, 2, \ldots, N, \qquad (3.1)$$

où a est une fonction connue, et β un vecteur des paramètres inconnus. Toutes les variables X_i, ε_i, i=1, 2, ... , N sont supposées indépendantes entre elles et les couples (X_i, ε_i) équidistribués. Les couples (X_i, Y_i), $i \in U$ sont aussi

indépendants et de même loi $F^*_{x,y}$. Notons les lois marginales de X_i, Y_i et ε_i par F^*_x, F^*_y et G, la loi conditionnelle de Y_i sachant X_i par $F^*_{y/x}$. Lorsque l'application a est monotone croissante inversible par rapport à ε_i, nous pouvons réécrire le modèle (3.1) sous la forme :

$$\varepsilon_i = T(Y_i, X_i, \beta),\ i=1, 2, \ldots, N.$$

On a alors :

$$\begin{aligned} F^*_{y/x}(t|X_i,\beta) &= P(Y_i \le t|X_i,\beta) = P(\varepsilon_i \le T(t,X_i,\beta)|X_i,\beta) \\ &= G(T(t,X_i,\beta)),\quad i=1, 2,\cdots,N. \end{aligned} \tag{3.2}$$

On peut déduire de l'expression (3.2) :

$$F^*_y(t) = E_{F^*_x}\left[F^*_{y/x}(t|X_i,\beta)\right] = E_{F^*_x}\left[G(T(t,X_i,\beta)\right].$$

Les lois $F^*_{x,y}$, F^*_x, F^*_y et G sont généralement inconnues, et $F^*_{y/x}$ est elle aussi inconnue.

Comme d'habitude, nous supposons observable la variable auxiliaire X sur la population entière, de sorte que la loi marginale F^*_x est bien approchée par la distribution marginale sur la population $F_x(t) = \sum_{i\in U} \Delta(t - X_i)/N$, dès que la taille de la population N est grande. Si le paramètre β et la fonction de répartition G étaient connus, un estimateur sans biais de $F^*_y(t)$ serait alors $\frac{1}{N}\sum_{i\in U} G[T(t,X_i,\beta)]$ et vraisemblablement très précis, car fondé sur N observations. L'idée est alors d'utiliser une procédure en plusieurs étapes:

Etape 1 : Estimation de β par $\hat{\beta}_n$ à partir des données de l'échantillon s.

Etape 2 : Calcul des résidus sur l'échantillon définis par

$$\hat{\varepsilon}_i = T(Y_i, X_i, \hat{\beta}_n),\ i \in s,$$

et utilisation de ces résidus pour estimer la fonction de répartition G par :

$$\hat{G}_n(t) = \frac{1}{n}\sum_{i\in s} \Delta(t - \hat{\varepsilon}_i) = \frac{1}{n}\sum_{i\in s} \Delta\left(t - T\left(Y_i, X_i, \hat{\beta}_n\right)\right) \tag{3.3}$$

Etape 3 : Finalement un estimateur de $F^*_y(t)$ est :

$$\hat{F}^*_y(t) = \frac{1}{N}\sum_{i\in U} \hat{G}_n\left(T\left(t, X_i, \hat{\beta}_n\right)\right). \tag{3.4}$$

Les propriétés de cet estimateur en trois étapes sont difficiles à exploiter lorsque la taille de l'échantillon n est petite ; elles peuvent l'être de façon approchée par la théorie asymptotique classique lorsque la taille de l'échantillon est assez grande,

même si elle reste petite par rapport à la taille N de la population. Dans ce cas le second estimateur est asymptotiquement sans biais.

L'estimateur (3.4) n'a pas pris en compte le plan de sondage. Un estimateur corrigé en tenant compte du plan de sondage consiste à estimer G par :

$$\hat{G}_n(t) = \frac{1}{N}\sum_{i\in s} d_i \Delta\left(t - \hat{\varepsilon}_i\right) = \frac{1}{N}\sum_{i\in s} d_i \Delta\left(t - T\left(Y_i, X_i, \hat{\beta}_n\right)\right), \qquad (3.5)$$

où $\hat{\beta}_n$ est un estimateur prenant aussi en compte le plan de sondage.

3.2 Approche par prédiction

Dans ce contexte, le paramètre d'intérêt est la loi marginale F_y sur la population (qui diffère peu de la loi F_y^* sur la surpopulation si la taille N est grande). On cherche un estimateur naturel en s'appuyant sur le modèle de surpopulation. Pour construire l'estimateur, l'idée est de décomposer F_y en deux parties [voir Chambers et Dunstan (1986), Rao, Kovar et Mantel (1990) et Kuk (1988, 1993)] :

$$F_y(t) = \frac{1}{N}\left[\sum_{i\in s}\Delta(t - Y_i) + \sum_{j\notin s}\Delta(t - Y_j)\right]. \qquad (3.6)$$

Dans cette décomposition, seule la seconde partie est inconnue. Comme pour l'estimation de la moyenne sur la population [Basu (1971), Cassel, Särndal et Wretman (1977)], l'idée est de prévoir cette partie inconnue $\sum_{j\notin s}\Delta(t - Y_j)$ à l'aide du modèle. Si $j \notin s$, $\Delta\left(t - Y_j\right)$ est prévu par :

$$E_F\left(\Delta\left(t - Y_j\right)\middle| X_j, \beta\right) = G\left(T(t, X_j, \beta)\right),\ j \notin s.$$

Après remplacement de β et G par des approximations calculées à partir de l'échantillon, nous obtenons l'estimateur de Chambers et Dunstan (1986) :

$$\hat{F}^*_{CD}(t) = \frac{1}{N}\left[\sum_{i\in s}\Delta(t - Y_i) + \sum_{j\notin s}\hat{G}_n\left(T\left(t, X_j, \hat{\beta}_n\right)\right)\right], \qquad (3.7)$$

où $\hat{G}_n(t)$ est défini par (3.3). $\hat{F}^*_{CD}$ est un estimateur construit à partir du seul modèle de surpopulation, sans référence au plan de sondage. Nous allons maintenant exploiter l'estimateur précédent pour un modèle de surpopulation particulier.

3.3 Modèle linéaire hétéroscédastique

Considérons un modèle linéaire hétéroscédastique:

$$Y_i = \beta X_i + v(X_i)\varepsilon_i,\ i = 1,\ 2,\ \cdots,\ N, \tag{3.8}$$

où β est une constante inconnue, v est une fonction connue strictement positive. On a alors $\varepsilon_i = \dfrac{Y_i - \beta X_i}{v(X_i)}$, i=1, 2, ... , N. Soient $(X_i,\ Y_i)$, $i \in s$ les observations sur l'échantillon s de taille n ; pour calculer l'estimateur $\hat{F}^*_{CD}$, nous considérons l'estimateur des moindres carrés de β :

$$\hat{\beta}_n = \left(\sum_{i\in s} \frac{X_i^2}{v^2(X_i)}\right)^{-1} \sum_{i\in s} \frac{Y_i X_i}{v^2(X_i)}, \tag{3.9}$$

ainsi que les résidus sur l'échantillon :

$$\hat{\varepsilon}_i = \frac{Y_i - \hat{\beta}_n X_i}{v(X_i)},\ i \in s\ .$$

Alors un estimateur de la loi G à partir de l'échantillon est :

$$\hat{G}_n(t) = \frac{1}{n}\sum_{i\in s} \Delta(t - \hat{\varepsilon}_i) = \frac{1}{n}\sum_{i\in s} \Delta\left(t - \frac{Y_i - \hat{\beta}_n X_i}{v(X_i)}\right). \tag{3.10}$$

Remplaçant $\hat{G}_n$ dans (3.7), nous obtenons l'estimateur de Chambers et Dunstan (1986) :

$$\hat{F}^*_{CD}(t) = N^{-1}\left\{\sum_{i\in s} \Delta(t - Y_i) + n^{-1}\sum_{j\notin s}\sum_{i\in s} \Delta\left\{\frac{(t - \hat{\beta}_n X_j)}{v(X_j)} - \frac{Y_i - \hat{\beta}_n X_i}{v(X_i)}\right\}\right\},$$

qui peut être récrit comme :

$$\hat{F}^*_{CD}(t) = N^{-1}\left\{\sum_{i\in s} \Delta(t - Y_i) + \sum_{j\in U} \hat{G}_{jn}(t) - \sum_{j\in s} \hat{G}_{jn}(t)\right\}, \tag{3.11}$$

avec : $\hat{G}_{jn}(t) = \dfrac{1}{n}\sum_{i\in s} \Delta\left\{\dfrac{(t - \hat{\beta}_n X_j)}{v(X_j)} - \dfrac{Y_i - \hat{\beta}_n X_i}{v(X_i)}\right\}$, j=1, 2, ... , N.

Sous cette forme, $\hat{F}^*_{CD}$ est un estimateur par différence. Il est facile de vérifier que $\hat{F}^*_{CD}$ n'est pas un estimateur asymptotiquement sans biais par rapport au plan de sondage (dont il ne dépend pas), mais qu'il est asymptotiquement sans biais par rapport au modèle de surpopulation. Chambers et Dunstan (1986) ont montré la normalité asymptotique de $\hat{F}^*_{CD}$ sous certaines conditions de régularité.

$\hat{F}^*_{CD}$ est un estimateur fondé sur un modèle de surpopulation et peut être très biaisé lorsque ce modèle est mal spécifié. On peut modifier cet estimateur $\hat{F}^*_{CD}$ pour le rendre à la fois fondé sur le plan de sondage et sur le modèle. Il suffit de remplacer les estimateurs empiriques intervenant dans la formule (3.11) par des estimateurs fondés sur le plan de sondage. Pour cela, considérons le modèle linéaire (3.8) ; un estimateur de β par les moindres carrés généralisés est:

$$\hat{\beta}_n = \left(\sum_{i \in s} \frac{d_i X_i^2}{v^2(X_i)} \right)^{-1} \sum_{i \in s} \frac{d_i Y_i X_i}{v^2(X_i)}. \qquad (3.12)$$

Notons $\widetilde{G}_{jn}(t)$ et $\widetilde{G}_{jnc}(t)$ les estimateurs analogues de $\hat{G}_{jn}(t)$, fondés sur le plan de sondage à l'aide du modèle, définis par :

$$\widetilde{G}_{jn} = \sum_{i \in s} d_i \Delta \left\{ \frac{(t - \hat{\beta}_n X_j)}{v(X_j)} - \frac{Y_i - \hat{\beta}_n X_i}{v(X_i)} \right\} / \sum_{i \in s} d_i \ , \ j \in U \ ,$$

$$\widetilde{G}_{jnc} = \sum_{i \in s} d_j \pi_{ij}^{-1} \Delta \left\{ \frac{(t - \hat{\beta}_n X_j)}{v(X_j)} - \frac{Y_i - \hat{\beta}_n X_i}{v(X_i)} \right\} / \sum_{i \in s} \pi_{ij}^{-1} d_j \ , \ j \in s \ ,$$

où $d_j^{-1} \pi_{ij}$, $i \in s$ est la probabilité de sélectionner l'individu i dans l'échantillon en sachant que l'individu j est sélectionné. Nous obtenons l'estimateur de Rao, Kovar et Mantel (1990) :

$$\hat{F}^*_{RKM}(t) = N^{-1} \Big[\sum_{i \in s} d_i \Delta(t - Y_i) + \sum_{j \in U} \widetilde{G}_{jn} - \sum_{j \in s} d_j \widetilde{G}_{jnc} \Big]. \qquad (3.13)$$

Cet estimateur est asymptotiquement sans biais de F_y par rapport au plan de sondage, et de variance approchée donnée par :

$$V\left(\hat{F}^*_{RKM}(t)\right) = N^{-2} \sum_{i \in U} \sum_{j \in U} a_{ij} \left(\Delta(t - Y_i) - G_i\right)\left(\Delta(t - Y_j) - G_j\right),$$

où a_{ij} est défini par (2.3) et G_i est défini par :

$$G_i(t) = \frac{1}{N} \sum_{k \in U} \Delta \left(\frac{t - \beta X_i}{v(X_i)} - \frac{Y_k - \beta X_k}{v(X_k)} \right), \ i=1, 2, \ldots, N.$$

Comparé à l'estimateur $\hat{F}^*_{CD}$, $\hat{F}^*_{RKM}$ est à la fois asymptotiquement sans biais par rapport au plan de sondage, et asymptotiquement sans biais par rapport au modèle de surpopulation. Par conséquent, $\hat{F}^*_{RKM}$ est beaucoup plus robuste que $\hat{F}^*_{CD}$ par rapport aux erreurs de spécification. Mais l'estimateur $\hat{F}^*_{RKM}$ a

l'inconvénient d'utiliser les probabilités d'inclusion d'ordre deux, qui ne sont pas toujours disponibles.

4. Les méthodes par calage

Les techniques de calage de Deville et Särndal (1992), Deville, Särndal et Sautory (1993) et les méthodes développées dans Ren et Deville (1998, 2000a, 2000b) peuvent aussi être utilisées pour construire des estimateurs d'une fonction de répartition sur une population finie. Réécrivons l'estimateur (2.11) comme :

$$\widetilde{F}_{RY}(t) = \frac{1}{N}\sum_{i\in s} w_i(t)\Delta(t - Y_i)\,,\ t \in D_s^+\,, \tag{4.1}$$

où $$w_i(t) = \frac{d_i \sum_{j\in U}\Delta(t - RX_j)}{\sum_{l\in s} d_l \Delta(t - RX_l)} = d_i F_x(R^{-1}t) F_{HTX}^{-1}(R^{-1}t)\,, \qquad i \in s\,;$$

$D_s^+ = \left\{ t\,;\ \sum_{l\in s} d_l \Delta(t - RX_l) > 0 \right\}$. Comparé à (2.2), l'estimateur est pondéré avec des poids $w_i(t)$ qui dépendent à la fois de la valeur de t et de l'échantillon s. Si on utilise ces poids pour estimer la fonction de répartition $F_x(R^{-1}t)$, on l'estime parfaitement pour chaque échantillon s donné :

$$\frac{1}{N}\sum_{i\in s} w_i(t)\Delta(t - RX_i) = \frac{1}{N}\sum_{j\in U}\Delta(t - RX_j)\,,\ t \in D_s^+\,. \tag{4.2}$$

Ces contraintes :

$$\sum_{i\in s} w_i(t)\Delta(t - RX_i) = \sum_{j\in U}\Delta(t - RX_j),\quad t \in D_s^+,\ \forall\ s\,,$$

définissent une famille d'équations de calage sur marges pour une variable connue (transformée) : $\Delta(t - RX)$ lorsque t, $t \in D_s^+$ est donné. Si on compare les deux estimateurs définis par (2.2) et (4.1), le premier correspond à une fonction en escalier ayant un nombre de sauts inférieur ou égal à n, la taille de l'échantillon. En revanche, le second estimateur correspond à une fonction en escalier, qui peut avoir un nombre de sauts supérieur à n, parce que pour $i \in s$, $w_i(t)$ peut être elle-même une fonction en escalier. Par conséquent, il peut être plus proche de la fonction de répartition à estimer. Dans les paragraphes suivants, nous allons étudier des estimateurs par calage sur la fonction de répartition d'une variable auxiliaire, en utilisant les méthodes du calage sur marges pour une variable transformée, ou les

méthodes du calage sur fonction de répartition développées dans Ren et Deville (1998, 2000a, 2000b)

4.1 Calage sur marge

Soit F_x la fonction de répartition connue de la variable auxiliaire X. Pour un échantillon s donné, soit $\hat{F}_{HTY}(t)$ l'estimateur de F_y défini par (2.2). Notons $D_s'^{+} = \{t;\ \sum_{i\in s} d_i \Delta(t - X_i) > 0\}$. Pour chaque $t \in D_s'^{+}$ donné, comme dans le cas du calage sur marges, on cherche des poids $w_i(t)$, $i \in s$ qui dépendent à la fois de t et de l'échantillon s, tels que la fonction de répartition connue F_x soit estimée sans erreur :

$$\sum_{i\in s} w_i(t)\Delta(t - X_i) = \sum_{j\in U} \Delta(t - X_j),\ \forall t \in D_s'^{+} \text{ fixé}, \tag{4.3}$$

et que $w_i(t)$ soit proche de d_i. Prenons $\Delta(t - X)$ comme variable auxiliaire dont le total $\sum_{j\in U} \Delta(t - X_j)$ est connu a priori ; les poids obtenus après le calage sur marges sont donnés par :

$$w_i(t) = d_i F^{*}\left[q_i(t)\Delta(t - X_i)\lambda(t)\right],\ i \in s,\ t \in D_s'^{+}; \tag{4.4}$$

où F^{*} est une fonction de calage ; $q_i(t)$ le poids que le statisticien a choisi pour l'individu i observé ; $\lambda(t)$ le multiplicateur de Lagrange à déterminer. L'estimateur de la fonction de répartition s'en déduit par :

$$\hat{F}_w(t) = N^{-1}\sum_{i\in s} d_i F^{*}\left[q_i(t)\Delta(t - X_i)\lambda(t)\right]\Delta(t - Y_i),\ t \in D_s'^{+}. \tag{4.5}$$

Lorsque F^{*} est une fonction linéaire $F^{*}(u) = 1 + u$, on aura pour, $i \in s$:

$$\begin{aligned} w_i(t) = d_i + d_i q_i(t)\Delta(t - X_i)\Big\{ \sum_{i\in s} d_i q_i(t)\Delta(t - X_i) \Big\}^{-1} \times \\ \Big\{ \sum_{k\in U} \Delta(t - X_k) - \sum_{i\in s} d_i \Delta(t - X_i) \Big\}, \end{aligned} \tag{4.6}$$

et, éventuellement, l'estimateur de la fonction de répartition par calage est :

$$\hat{F}_w(t) = N^{-1}\sum_{i\in s} d_i\Delta(t-Y_i) + \hat{B}(t)\Big\{N^{-1}\sum_{k\in U}\Delta(t-X_k) - N^{-1}\sum_{i\in s} d_i\Delta(t-X_i)\Big\},\ t\in D_s'^{+}, \tag{4.7}$$

où $\hat{B}(t)$ est défini par :

$$\hat{B}(t) = \Big\{\sum\nolimits_{i\in s} d_i q_i(t)\Delta(t-X_i)\Big\}^{-1}\sum\nolimits_{i\in s} d_i q_i(t)\Delta(t-Y_i)\Delta(t-X_i),\quad t\in D_s'^{+}. \tag{4.8}$$

L'estimateur défini par (4.7) est un estimateur par régression dont les propriétés sont déjà connues. Quand la fonction de calage F^* est autre que la fonction linéaire, la formule (4.5) définit une classe d'estimateurs équivalents à l'estimateur défini par (4.7) au sens qu'ils ont la même variance approchée.

Le même problème que dans la section 2 se pose quand Y est beaucoup plus petit que X ou l'inverse. Pour un échantillon s donné, si on adapte (4.3) en remplaçant X par $\hat{R}X$, on obtient :

$$\sum_{i\in s} w_i(t)\Delta(t-\hat{R}X_i) = \sum_{j\in U}\Delta(t-\hat{R}X_j),\ \forall t\in D_s''^{+}, \tag{4.9}$$

où $\hat{R} = \sum_{i\in s} d_i Y_i / \sum_{i\in s} d_i X_i$, $D_s''^{+} = \Big\{t; \sum_{i\in s} d_i\Delta(t-\hat{R}X_i) > 0\Big\}$. Les poids obtenus du calage sont :

$$w_i(t) = d_i F^*\Big[q_i(t)\Delta(t-\hat{R}X_i)\lambda(t)\Big],\ t\in D_s''^{+}\quad i\in s. \tag{4.10}$$

Ceci donne ensuite un estimateur analogue à l'estimateur défini par (4.7) quand la fonction de calage F^* est linéaire :

$$\hat{F}_w(t) = \frac{1}{N}\sum_{i\in s} d_i\Delta(t-Y_i) + \hat{B}^*(t)\Big\{N^{-1}\sum_{i\in U}\Delta(t-\hat{R}X_i) - N^{-1}\sum_{i\in s} d_i\Delta(t-\hat{R}X_i)\Big\},\ t\in D_s''^{+} \tag{4.11}$$

où $\hat{B}^*(t)$ est défini par :

$$\hat{B}^*(t) = \Big\{\sum\nolimits_{i\in s} d_i q_i(t)\Delta(t-\hat{R}X_i)\Big\}^{-1}\sum\nolimits_{i\in s} d_i q_i(t)\Delta(t-Y_i)\Delta(t-\hat{R}X_i),\quad t\in D_s''^{+};$$

L'estimateur (4.11) doit, en général, être plus performant que l'estimateur défini par (4.7).

4.2 Les méthodes par calage où les poids ne dépendent pas de t

Nous notons qu'il y a une chose assez gênante pour l'obtention des estimateurs définis par (4.5) et (4.11) et leurs variances ou l'estimation de ces dernières, c'est qu'il faut effectuer pour chaque t donné la procédure de calage sur marges. D'autre part, les propriétés des estimateurs pondérés avec des poids dépendant de t sont difficiles à étudier. En tant qu'un estimateur d'une fonction de répartition, Kuk (1993), Nascimento Silva et Skinner (1995) ont proposé des propriétés principales que doivent posséder les estimateurs de la fonction de répartition. Parmi celles-ci, la continuité à droite, la croissance monotone, $\hat{F}(-\infty) = 0$ et $\hat{F}(+\infty) = 1$. Mais, pour les estimateurs qui utilisent des poids dépendant de t, la croissance monotone ne peut plus être garantie. Par exemple, pour les estimateurs $\hat{F}_{RY}(t)$, $\hat{F}_{D}(t)$, $\hat{F}_{d}^{*}(t)$ et $\hat{F}_{RKM}^{*}(t)$, ils ne sont pas en général des fonctions globalement monotones. L'absence de cette propriété cause des difficultés pour le calcul des fractiles traité dans la section 5. Pour surmonter cette difficulté, on peut construire des estimateurs par calage sur une fonction de répartition connue où les poids sont indépendants de t, par exemple, les poids $\{w_i;\ i \in s\}$ obtenus par l'une des méthodes de calage sur la fonction de répartition F_x, étudiées dans Ren et Deville (1998, 2000a, 2000b). Un estimateur de F_y calé sur la fonction de répartition F_x est obtenu par la formule :

$$\hat{F}_w(t) = \sum_{i \in s} w_i \Delta(t - Y_i), \text{ lorsque } \sum_{i \in s} w_i = 1 ;$$

$$\hat{F}_w(t) = \sum_{i \in s} w_i \Delta(t - Y_i) \Big/ \sum_{i \in s} w_i, \text{ lorsque } \sum_{i \in s} w_i \neq 1 .$$

a. Approche calage

Supposons les observations sur échantillon triées par l'ordre croisant selon les valeurs de la variable auxiliaire X : $x_1 \leq x_2 \leq \cdots \leq x_n$. Lorsque la taille de la population est grande, les poids sont donnés par [Ren et Deville (1998)] :

$$\begin{cases} w_1 = \dfrac{1}{2}\left[F_x(x_2) + F_x(x_1)\right], \quad w_n = 1 - \dfrac{1}{2}\left[F_x(x_n) + F_x(x_{n-1})\right] \\ w_i = \dfrac{1}{2}\left[F_x(x_{i+1}) - F_x(x_i)\right], \quad i = 2, 3, \cdots, n-1; \end{cases} \tag{4.12}$$

b. Approche non-paramétrique

Dans l'approche non-paramétrique de Ren et Deville (2000a), les poids sont donnés par :

$$w_i = \frac{1}{N}\sum_{l\in U}\frac{d_i J\left[(X_l - X_i)/h_n\right]}{\sum_{j\in s} d_j J\left[(X_l - X_j)/h_n\right]}\ ,\ i \in s\ ; \tag{4.13}$$

où J est une fonction de noyau, h_n est une constante telle que $h_n \to 0$ quand $n \to \infty$.

c. Calage sur les rangs

Lorsque la variable auxiliaire est catégorielle ou ne prend qu'un petit nombre de valeurs différentes, notons $R_1, R_2, \cdots, R_N$ les rangs (Ren et Deville (2000b)) des unités dans la population ; les poids obtenus par calage sur les rangs sont alors donnés par :

$$\begin{cases} w_1 = \dfrac{1}{2N}\left[r_2 + r_1\right] - \dfrac{1}{2N},\ \ w_n = 1 - \dfrac{1}{2N}\left[r_n + r_{n-1}\right] + \dfrac{1}{2N}, \\ w_i = \dfrac{1}{2N}\left[r_{i+1} - r_{i-1}\right],\ \ i = 2, 3, \ldots, n-1; \end{cases} \tag{4.14}$$

où $r_1, r_2, \cdots, r_n$ sont les rangs des unités observées dans l'échantillon.

d. Calage sur les moments

Soient $\{w_i ; i \in s\}$ les poids obtenus par calage sur les moments, par exemple, calage sur les moments jusqu'à l'ordre m (Ren et Deville (2000b)) :

$$\sum_{i\in s} w_i X_i^k = \sum_{i\in U} X_i^k,\ \ k = 0, 1, 2, \cdots, m\ . \tag{4.15}$$

L'intérêt de cet estimateur est que, si on utilise les mêmes poids pour estimer la fonction de répartition F_x connue :

$$\hat{F}_{w,x}(t) = \frac{1}{N}\sum_{i\in s} w_i\, \Delta(t - X_i)\ ,$$

alors $\hat{F}_{w,x}$ a les mêmes moments que F_x jusqu'à l'ordre m.

5. Estimation des fractiles d'une population finie

Soit F une fonction de répartition continue à droite et monotone croissante. Pour 0<α<1 donné, dans la statistique conventionnelle, le **α-fractile** t_α de F est défini par : $t_\alpha = \inf\{t;\ F(t) \geq \alpha\}$. Une estimation de t_α à partir de l'échantillon est définie par : $\hat{t}_\alpha = \inf\{t;\ \hat{F}(t) \geq \alpha\}$, où $\hat{F}$ est la fonction de répartition empirique calculée sur l'échantillon. Chambers et Dunstan (1986) ont utilisé cette même définition pour définir le α-fractile de la fonction de répartition d'une population finie. Mais Särndal, Swensson et Wretman (1992) ont défini la médiane d'une population finie d'une autre façon un peu différente. Dans ce rapport, on utilise la même définition de Chambers et Dunstan. Soit une population finie de taille *N*. Supposons que la population est triée par valeur croissante de la variable Y :

$$Y_1 \leq Y_2 \leq \cdots \leq Y_N .$$

La fonction de répartition F_y sur la population de la variable Y définie par (2.1) est une fonction continue à droite et monotone croissante. Pour 0<α<1 donné, le α-fractile de F_y est défini comme :

$$t_\alpha = \inf\{t;\ F_y(t) \geq \alpha\} = \begin{cases} Y_i, & \text{si } \alpha = F_y(Y_i); \\ Y_{i+1}, & \text{si } F_y(Y_i) < \alpha < F_y(Y_{i+1}); \end{cases} \tag{5.1}$$

et noté : $t_\alpha = F_y^{-1}(\alpha)$. Quand $\alpha = 1/2$, $t_{1/2}$ est appelé **médiane**. Il est facile de vérifier que cette définition a défini une application F_y^{-1} de]0, 1[dans]a, b[, où $a = \min_{i \in U}(Y_i)$, $b = \max_{i \in U}(Y_i)$. F_y^{-1} est une fonction monotone croissante et continue à gauche, et vérifie les propriétés suivantes :

$$\begin{cases} F_y^{-1}\left(F_y(t\ )\right) \leq t\ ,\ a < t < b; \\ F_y\left(F_y^{-1}(\alpha)\right) \geq \alpha,\ 0 < \alpha < 1; \\ F_y(t) \geq \alpha\,, \text{ si et seulement si } t \geq F_y^{-1}(\alpha). \end{cases} \tag{5.2}$$

Une estimation de t_α à partir d'un échantillon est définie de la même façon. Soit un échantillon de taille *n* avec les observations triées en ordre croissant :

$$y_1 \leq y_2 \leq \ldots \leq y_n .$$

Soit $\hat{F}_y$ un estimateur quelconque de F_y tel que $\hat{F}_y$ est monotone croissante et continue à droite. Pour 0<α<1, le α-fractile de $\hat{F}_y$ est défini de façon analogue :

$$\hat{t}_\alpha = \inf\left\{t;\ \hat{F}_y(t) \geq \alpha\right\} = \begin{cases} y_i, & \text{si } \alpha = \hat{F}_y(y_i); \\ y_{i+1}, & \text{si } \hat{F}_y(y_i) < \alpha < \hat{F}_y(y_{i+1}); \end{cases} \tag{5.3}$$

et noté : $\hat{t}_\alpha = \hat{F}_y^{-1}(\alpha)$. $\hat{t}_\alpha$ est appelé estimateur de t_α. Quand $\alpha = 1/2$, $\hat{t}_{1/2}$ est l'estimateur de la médiane. Comme dans (5.1), la définition (5.3) définit une application $\hat{F}_y^{-1}$ qui a des propriétés analogues à celles de F_y^{-1} présentées dans (5.2).

Dans la statistique conventionnelle, les propriétés de $\hat{t}_\alpha$ sont étudiées et des résultats de convergence de $\hat{t}_\alpha$ ont été établis sous des conditions légères (Serfling (1980)), où la fonction de répartition estimée est la fonction de répartition empirique basée sur des observations indépendantes et de même loi. Dans le cas de l'échantillonnage en population finie, les propriétés de $\hat{t}_\alpha$ deviennent beaucoup plus compliquées parce que les observations ne sont plus indépendantes et qu'on a beaucoup plus de choix sur l'estimation de la fonction de répartition telle que présentée dans les paragraphes précédents. Néanmoins, on peut obtenir une estimation par intervalle de confiance. Dès que $\hat{t}_\alpha$ est calculé, la variance approchée de $\hat{F}_y(\hat{t}_\alpha)$ est estimable selon la méthode que l'on a utilisée pour estimer la fonction de répartition F_y. Par exemple, quand $\hat{F}_y$ est calculé par (2.2), la variance approchée de $\hat{F}_y(\hat{t}_\alpha)$ est donnée par (2.3) en remplaçant t par t_α :

$$V\left(\hat{F}_y(\hat{t}_\alpha)\right) = \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U} a_{ij}\Delta(t_\alpha - Y_i)(t_\alpha - Y_j), \tag{5.4}$$

où a_{ij} est défini par (2.3). Un estimateur de la variance est obtenu, lorsque $\pi_{ij} > 0$ pour tous i et j, par :

$$\hat{V}\left(\hat{F}_y(\hat{t}_\alpha)\right) = \frac{1}{N^2}\sum_{i\in s}\sum_{j\in s} a_{ij}\pi_{ij}^{-1}\Delta(\hat{t}_\alpha - y_i)(\hat{t}_\alpha - y_j). \tag{5.5}$$

Supposons que $\hat{F}_y$ converge vers une variable Gaussienne de variance (5.4). Les bornes de l'intervalle de confiance $\hat{t}_{L\alpha}$, borne inférieure, et $\hat{t}_{U\alpha}$, borne supérieure, du niveau $(1-\gamma)$ $(0 < \gamma < 1)$ sont calculées par [Woodruff (1952)] :

$$\begin{cases} \hat{t}_{L\alpha} = \hat{F}_y^{-1}\left(\alpha - z_{\frac{1}{2}\gamma}\left[\hat{V}\left(\hat{F}_y(\hat{t}_\alpha)\right)\right]^{\frac{1}{2}} \right), \\ \hat{t}_{U\alpha} = \hat{F}_y^{-1}\left(\alpha + z_{\frac{1}{2}\gamma}\left[\hat{V}\left(\hat{F}_y(\hat{t}_\alpha)\right)\right]^{\frac{1}{2}} \right), \end{cases} \tag{5.6}$$

où $z_{\frac{1}{2}\gamma}$ est le $(1-\gamma/2)$-fractile d'une variable Gaussienne centrée réduite, c'est-à-dire : $z_{\frac{1}{2}\gamma} = \Phi^{-1}\left(1-\frac{1}{2}\gamma\right)$; Φ est la fonction de répartition Gaussienne centrée réduite. On a alors :

$$\begin{aligned} &\Pr\left\{\hat{t}_{L\alpha} \le t_\alpha \le \hat{t}_{U\alpha}\right\} = \Pr\left\{t_\alpha \le \hat{t}_{U\alpha}\right\} + \Pr\left\{t_\alpha \ge \hat{t}_{L\alpha}\right\} - 1 \\ &= \Pr\left\{\hat{F}_y(t_\alpha) \le \hat{F}_y(\hat{t}_{U\alpha})\right\} + \Pr\left\{ t_\alpha \ge \hat{F}_y^{-1}\left(\alpha - z_{\frac{1}{2}\gamma}\left[\hat{V}\left(\hat{F}_y(\hat{t}_\alpha)\right)\right]^{1/2} \right)\right\} - 1. \end{aligned}$$

D'après (5.2) et l'hypothèse de la normalité asymptotique de $\hat{F}_y(t)$, on a :

$$\Pr\left\{\hat{F}_y(t_\alpha) \le \hat{F}_y(\hat{t}_{U\alpha})\right\} \ge \Pr\left\{\hat{F}_y(t_\alpha) \le \alpha + z_{\frac{1}{2}\gamma}\left[\hat{V}\left(\hat{F}_y(\hat{t}_\alpha)\right)\right]^{1/2}\right\} \cong 1 - \frac{1}{2}\gamma .$$

$$\begin{aligned} &\Pr\left\{ t_\alpha \ge \hat{F}_y^{-1}\left(\alpha - z_{\frac{1}{2}\gamma}\left[\hat{V}\left(\hat{F}_y(\hat{t}_\alpha)\right)\right]^{1/2} \right)\right\} \\ &= \Pr\left\{\hat{F}_y(t_\alpha) \ge \alpha - z_{\frac{1}{2}\gamma}\left[\hat{V}\left(\hat{F}_y(\hat{t}_\alpha)\right)\right]^{1/2}\right\} \cong 1 - \frac{1}{2}\gamma. \end{aligned}$$

Par conséquent,

$$\Pr\left\{\hat{t}_{L\alpha} \le t_\alpha \le \hat{t}_{U\alpha}\right\} \ge 1 - \gamma .$$

Ce qui montre que $\left[\hat{t}_{L\alpha}, \hat{t}_{U\alpha}\right]$ est un intervalle de confiance de t_α avec le niveau de confiance approché $1-\gamma$.

Soit X une variable auxiliaire connue sur la population entière, de fonction de répartition F_x. Notons le α-fractile connu de celle-ci : $t_\alpha(x)$ et le α-fractile de la variable d'intérêt Y : $t_\alpha(y)$. Notons les estimateurs de ces fractiles calculés à partir d'un échantillon par $\hat{t}_\alpha(x)$ et $\hat{t}_\alpha(y)$. Supposons que Y et X soient liés par le modèle (3.8). Suivant les idées de Rao, Kovar et Mantel (1990), des estimateurs peuvent être construits par les méthodes usuelles utilisées dans le cadre de l'estimation de la moyenne ou du total.

5.1 Estimateur par régression

Soit $\hat{\beta}_n$ l'estimateur de β défini par (3.12) ; un estimateur par régression de $t_\alpha(y)$ est obtenu par :

$$\hat{t}_{\alpha\text{-}reg}(y) = \hat{t}_\alpha(y) + \hat{\beta}_n\left\{t_\alpha(x) - \hat{t}_\alpha(x)\right\}. \tag{5.7}$$

Lorsque $v(X_i) = X_i^{1/2}$ dans le modèle (3.8), on a $\hat{\beta}_n = \hat{R} = \sum_{i\in s} d_i Y_i / \sum_{i\in s} d_i X_i$, l'estimateur $\hat{t}_{\alpha\text{-}reg}(y)$ devient alors celui de l'estimateur par différence de Rao, Kovar et Mantel (1990) :

$$\hat{t}_{\alpha\text{-}reg}(y) = \hat{t}_\alpha(y) + \hat{R}\left\{t_\alpha(x) - \hat{t}_\alpha(x)\right\}.$$

5.2 Estimateur par le ratio

Un estimateur par le ratio de Rao, Kovar et Mantel (1990) est défini par :

$$\hat{t}_{\alpha\text{-ratio}}(y) = \frac{\hat{t}_\alpha(y)}{\hat{t}_\alpha(x)} t_\alpha(x). \tag{5.8}$$

5.3 Estimateur par différence

Un estimateur par différence peut être défini de façon analogue :

$$\hat{t}_{\alpha\text{-}dif}(y) = \hat{t}_\alpha(y) + \left\{t_\alpha(x) - \hat{t}_\alpha(x)\right\}. \tag{5.9}$$

5.4 Estimateur par calage

Un estimateur de $t_\alpha(y)$ par calage sur la fonction de répartition connue F_x consiste à rechercher des poids w_i tels que $\hat{t}_\alpha(x) = \hat{F}_{w,x}^{-1}(\alpha) = t_\alpha(x)$, où $\hat{F}_{w,x}$ est défini par :

$$\hat{F}_{w,x}(t) = \sum_{i\in s} w_i\left(t - X_i\right) / \sum_{i\in s} w_i \,. \tag{5.10}$$

Ou de façon équivalente, on recherche des poids w_i tels que :

$$\sum_{i\in s} w_i \Delta\left(t_\alpha(x) - X_i\right) = \sum_{i\in U} \Delta\left(t_\alpha(x) - X_i\right),$$

en minimisant une certaine mesure de proximité $d(d_i, w_i)$ entre les poids w_i et les poids de sondage d_i . On utilise ces poids pour construire un estimateur de F_y :

$$\hat{F}_{w,y}(t) = \sum_{i \in s} w_i (t - Y_i) / \sum_{i \in s} w_i \,. \qquad (5.11)$$

Un estimateur du α-fractile de F_y, $\hat{t}_\alpha(y)$, est calculé à partir de $\hat{F}_{w,y}$ selon la formule (5.3).

Des estimateurs de $t_\alpha(y)$ peuvent aussi être construits par d'autres estimateurs de la fonction de répartition calés sur la fonction de répartition connue cités dans la section 4. Soient $\{w_i, i \in s\}$ les poids sortants d'une procédure de calage ; l'estimateur de la fonction de répartition $\hat{F}_{w,y}$ est calculé selon la formule (5.11), l'estimateur du α-fractile $\hat{t}_\alpha(y)$ de F_y est calculé à partir de $\hat{F}_{w,y}$. Ces méthodes possèdent l'avantage que la fonction de répartition estimée $\hat{F}_{w,y}$ est une somme pondérée, et elle est monotone croissante lorsque les poids w_i, $i \in s$ sont tous positifs. Des comparaisons numériques seront exposées par simulation sur des populations artificielles dans le paragraphe suivant.

6. Etudes numériques

Dans cette section, des études numériques sont faites sur des données artificielles pour comparer les différents estimateurs de la fonction de répartition et des fractiles. Des difficultés se présentent dans les calculs de certains estimateurs de la fonction de répartition et des fractiles lorsque les calculs comportent une totalisation sur la population entière. C'est le cas, par exemple, de l'estimateur de Chambers et Dunstan, de l'estimateur de Rao, Kovar et Mantel et de l'estimateur par calage sur fonction de répartition dans une approche non-paramétrique. Il est facile d'effectuer le calcul lorsque l'estimateur de la fonction de répartition a la forme (5.11) et que le calcul de w_i ne comprend que des calculs sur l'échantillon. C'est le cas, par exemple, de l'estimateur de Horvitz-Thompson, de l'estimateur par le ratio et par différence, des estimateurs par calage sur marges et par calage sur la fonction de répartition. Quand le calcul de $\hat{F}_{w,y}$ ou de w_i demande une totalisation sur la population entière, ceci nécessite un temps de calcul important si la taille de la population et la taille de l'échantillon sont grandes, surtout pour les estimateurs des fractiles. Pour cette raison nous avons limité la taille de la population simulée, la taille de l'échantillon et le nombre de répétitions des simulations.

Exemple 6.1. Considérons des populations artificielles engendrées par des modèles linéaires hétéroscédastiques :

$$Y_i = \beta X_i + v(X_i)\varepsilon_i, \; i=1, 2, ..., 10000, \tag{6.1}$$

$$Y_i = \beta X_i^2 \exp\left\{X_i^{1,5}/20\right\} + v(X_i)\varepsilon_i, \; i=1, 2, ..., 10000. \tag{6.2}$$

où X_i et ε_i sont des tirages indépendants effectués dans des lois Gaussiennes $N(10, 1,5)$ et $N(0, 1)$, respectivement ; β=1,5 est constante, $v(X_i) = |X_i|^{1/4}$, i=1, 2, ..., 10000. Le plan de sondage utilisé est *ASSR* de taille 500. Les estimateurs de la fonction de répartition sont calculés pour les valeurs de t : t=13,1, t=15,0 et t=16,9 pour le modèle (6.1) et t=473, t=730 et t=1135 pour le modèle (6.2), correspondant aux divers quantiles : $F_y(t)$ =0,25, 0,50 et 0,75 respectivement. Les caractéristiques des estimateurs (moyenne et racine de MSE) sont mesurées par simulation à partir de 100 échantillons indépendants (les valeurs Y_i, X_i et ε_i i=1, 2, ..., 10000 étant elles fixées une fois pour toutes). Les résultats sont donnés dans le Tableau 6.1.

Les estimateurs à comparer sont :

EstHT = « estimateur de Horvitz-Thompson de (2.5) »,
Estratio = « estimateur par le ratio (2.14) »,
Estdifer = « estimateur par différence (2.18) »,
Estcd = « estimateur de Chambers et Dunstan (3.11) »,
Estrkm = « estimateur de Rao , Kovar et Mantel (3.13) »,
Estcalm = « estimateur par calage sur les moments jusqu'à l'ordre deux, poids (4.15) »,
Estcalf = « estimateur par calage sur la fonction de répartition, poids (4.12) ».

Les résultats du Tableau 6.1 montrent que, pour un modèle linéaire (6.1), l'estimateur de Chambers et Dunstan est préférable et qu'il n'y a pas de différence significative entre les autres estimateurs. Pour le modèle non linéaire (6.2), l'estimateur de Chambers et Dunstan est fortement biaisé parce que le modèle de surpopulation est maintenant inapproprié. En revanche, les estimateurs **Estrkm**, **Estcalm** et **Estcalf** sont robustes au sens où ils sont restés précis en dépit de la mauvaise spécification du modèle de surpopulation. Dans ce cas, les estimations ne sont pas toujours disponibles pour l'estimateur **Estratio** lorsque t = 473, car le ratio n'est pas défini pour certains échantillons. L'estimateur **Estcalf** est beaucoup plus précis que les autres.

Tableau 6.1 Estimation de la fonction de répartition d'une population finie.

Modèle (6.1)

Type d'estimateur	$t = 13{,}1$ $F_y(t)=0{,}25$		$t = 15{,}0$ $F_y(t)=0{,}50$		$t = 16{,}9$ $F_y(t)=0{,}75$	
	$\hat{F}_y(t)$	**Racine de MSE**	$\hat{F}_y(t)$	**Racine de MSE**	$\hat{F}_y(t)$	**Racine de MSE**
EstHT	0,25	0,017	0,50	0,021	0,75	0,016
Estratio	0,26	0,020	0,50	0,021	0,75	0,015
Estdifer	0,26	0,017	0,50	0,021	0,75	0,016
Estcd	0,26	0,009	0,50	0,012	0,75	0,010
Estrkm	0,25	0,016	0,50	0,016	0,75	0,014
Estcalm	0,25	0,016	0,50	0,016	0,75	0,014
Estcalf	0,25	0,018	0,50	0,020	0,75	0,018

Modèle (6.2)

Type d'estimateur	$t = 473$ $F_y(t)=0{,}25$		$t = 730$ $F_y(t)=0{,}50$		$t = 1135$ $F_y(t)=0{,}75$	
	$\hat{F}_y(t)$	**Racine de MSE**	$\hat{F}_y(t)$	**Racine de MSE**	$\hat{F}_y(t)$	**Racine de MSE**
EstHT	0,25	0,015	0,50	0,021	0,75	0,020
Estratio	*	*	0,53	0,079	0,75	0,020
Estdifer	0,25	0,015	0,50	0,022	0,75	0,020
Estcd	0,15	0,100	0,48	0,027	0,79	0,043
Estrkm	0,25	0,012	0,50	0,016	0,75	0,018
Estcalm	0,25	0,011	0,50	0,012	0,75	0,012
Estcalf	0,25	0,002	0,50	0,002	0,75	0,001

* Les estimations ne sont pas toujours disponibles.

Exemple 6.2. Dans cet exemple nous étudions l'estimation des fractiles d'une population finie. Les populations artificielles sont engendrées par les mêmes modèles (et les mêmes paramètres) que ceux utilisés dans l'exemple 6.1. Seules les tailles des populations et les tailles d'échantillons sont modifiées. Les tailles de populations 5000 sont identiques pour les deux modèles ; les tailles d'échantillons sont 200, avec le tirage *ASSR*. Pour économiser le calcul, cinq estimateurs de fractile où les estimateurs de la fonction de répartition ont une forme (4.11), ont été calculés pour des valeurs différentes de α : $\alpha = 0{,}25$, $\alpha = 0{,}50$ et $\alpha = 0{,}75$.

Les cinq estimateurs sont :

EstHT = « estimateur calculé à partir de l'estimateur de Horvitz-Thompson de la fonction de répartition (2.5) ».

Estratio = « estimateur par le ratio (5.8) où les estimateurs $\hat{t}_\alpha(y)$ et $\hat{t}_\alpha(x)$ sont calculés à partir des estimateurs de Horvitz-Thompson de la fonction de répartition (2.5) ».

Estreg = « estimateur par régression (5.7) où les estimateurs $\hat{t}_\alpha(y)$ et $\hat{t}_\alpha(x)$ sont les mêmes que pour l'estimateur **Estratio** ».

Estcalf = « estimateur par calage sur la fonction de répartition calculé à partir de l'estimateur de la fonction de répartition en utilisant les poids (4.12) ».

Estcalm = « estimateur par calage sur les moments jusqu'à l'ordre quatre calculé à partir de l'estimateur de la fonction de répartition en utilisant les poids (4.15) ».

Les résultats sont résumés dans le Tableau 6.2 ; les moyennes et les racines de MSE des estimateurs sont calculées sur 100 tirages indépendants. Pour le modèle linéaire (6.1), il n'y a pas de différence significative entre les estimateurs, ce qui était déjà le cas lorsqu'on s'intéressait à la fonction de répartition. Mais pour le modèle non linéaire (6.2), la différence est significative. L'estimateur **EstHT** possède de mauvaises propriétés ; **Estreg** est préférable à **Estratio** pour $\alpha = 0{,}25$ et $\alpha = 0{,}50$; **Estcalf** est le meilleur pour toutes les valeurs de α comme dans le cas de l'estimation de la fonction de répartition, tandis que l'estimateur **Estcalm** calé sur les moments n'est pas le plus performant comme on s'y attendait.

Tableau 6.2 Estimation des fractiles d'une population finie.

Modèle linéaire (6.1)

Type	$\alpha = 0{,}25$ $t_\alpha(y) = 13{,}1$		$\alpha = 0{,}50$ $t_\alpha(y) = 14{,}9$		$\alpha = 0{,}75$ $t_\alpha(y) = 16{,}9$	
D'estimateur	$\hat{t}_\alpha(y)$	**Racine de MSE**	$\hat{t}_\alpha(y)$	**Racine de MSE**	$\hat{t}_\alpha(y)$	**Racine de MSE**
EstHT	13,1	0,25	14,9	0,27	16,9	0,31
Estratio	13,1	0,21	14,9	0,20	16,9	0,28
Estreg	13,1	0,21	14,9	0,20	16,9	0,28
Estcalf	13,0	0,24	14,9	0,22	16,9	0,32
Estcalm	13,1	0,20	14,9	0,16	16,9	0,24

Modèle non-linéaire (6.2)

Type	α=0,25 $t_\alpha(y) = 464$		α=0,50 $t_\alpha(y) = 722$		α=0,75 $t_\alpha(y) = 1118$	
d'estimateur	$\hat{t}_\alpha(y)$	**Racine de MSE**	$\hat{t}_\alpha(y)$	**Racine de MSE**	$\hat{t}_\alpha(y)$	**Racine de MSE**
EstHT	471	32,1	727	46,2	1123	82,1
Estratio	469	24,0	726	35,5	1122	64,7
Estreg	468	17,7	725	32,5	1122	66,2
Estcalf	463	8,3	721	9,8	1114	20,5
Estcalm	470	18,1	721	24,3	1086	59,5

Bibliographie

Chambers, R. L. et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika.* Vol. 73, N° 3 : 597-604.

Chambers, R. L., Dofman, A. H. & Wehrly, T. E. (1993). Bias Robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association.* Vol. 88, N° 421 : 268-277.

Chambers, R. L., Dunstan, R. et Hall, P. (1992). Properties of estimators of the finite population distribution function. *Biometrika.* Vol. 79, N° 3 : 577-582.

Chen, J. et Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika.* Vol. 80, N° 1 : 107-116.

Cochran, W. G. (1977). *Sampling Techniques.* 3ed Edit. John Wiley & Sons. New York.

Deville, J.-C. et Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association.* Vol. 87, N° 418 : 376-382.

Deville, J.-C., Särndal, C. E. et Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association.* Vol. 88, N° 423 : 1013-1020.

Dorfman, A. H. (1993). A comparison of design-based and model-based estimators of the finite population distribution function. *Australian Journal of Statistics.* Vol. 35, N°1 : 29-41.

Dunstan, R. et Chambers, R. L.(1989). Estimating distribution functions from survey data with limited benchmark information. *Australian Journal of Statistics.* Vol. 31, N° 1 : 1-11.

Francisco, C. A. et Fuller, W. A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics.* Vol. 19, N° 1 : 454-469.

Kish, L. (1965). *Survey Sampling.* John-Wiley & Sons, New York.

Kuk, A. Y. C. (1988). Estimation of distribution functions and medians under sampling with unequal probabilities. *Biometrika.* Vol. 75, N° 1 : 97-103.

Kuk, A. Y. C. (1993). A kernel method for estimating finite population distribution functions using auxiliary information. *Biometrika.* Vol. 80, N° 2 : 385-392.

Kuk, A. Y. C. et Mak, T. K. (1994). A functional approach to estimating finite population distribution functions. *Communication in Statistics-Theory and Methods.* Vol. 23, N° 3 : 883-896.

Kuo, L. (1988). Classical and prediction approaches to estimating distribution functions from survey data. *Proceedings of the section on survey research methods, American Statistical Association.* 280-285.

Nascimento Silva, P. L. D. et Skinner, C. J. (1995). Estimating distribution functions with auxiliary information using poststratification. *Journal of Official Statistics.* Vol. 11, N° 3 : 277-294.

Rao, J. N. K. & Liu, J. (1992). On estimating distribution functions from survey data using supplementary information at the estimation stage. In *Nonparametric Statistics and Related Topics.* (A. K. Md.E. Saleh ed.),

Amsterdam : Elsevier Science Publishers, 399-407.
Rao, J. N. K., Kovar, J. G. et Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*. Vol. 77, N° 2 : 365-375.
Ren, R. et Deville, J.-C. (1998). Calage sur fonction de répartition : méthode par calage. Rapport de recherche.
Ren, R. et Deville, J.-C. (2000a). Calage sur fonction de répartition : méthode non-paramétrique. XXXIIèmes Journées de Statistique. Fès, Maroc.
Ren, R. et Deville, J.-C. (2000b). Une généralisation du calage : calage sur les rangs et calage sur les moments. IIème Colloque Francophone sur les Sondages. Bruxelles.
Särndal, C. E., Swensson, B. et Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New-York.
Sedransk, N. et Sedransk, J. (1979). Distinguishing among distributions using data from complex sample designs. *Journal of the American Statistical Association*, Vol. 74, N° 368 : 754-760.
Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.
Sukhatme, P. V. et Sukhatme, B. V. (1984). *Sampling Theory of Surveys with Applications*. 3ed Edit. Iowa State University Press.
Woodruff, R. S. (1952). Confidence intervals for medians and other position measures. *Journal of American Statistical Association*. Vol. 47 : 635-646.