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

Techniques d'enquête

Inférence statistique fondée sur des échantillons poststratifiés par choix raisonné en population finie

par Omer Ozturk

Date de diffusion : le 20 décembre 2016

Statistique Canada Statistics Canada

Canada

Comment obtenir d'autres renseignements

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

Vous pouvez également communiquer avec nous par :

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

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

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

Programme des services de dépôt

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

Normes de service à la clientèle

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

Note de reconnaissance

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

Signes conventionnels dans les tableaux

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

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

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 2016

Tous droits réservés. L'utilisation de la présente publication est assujettie aux modalités de l'entente de licence ouverte de Statistique Canada.

Une version HTML est aussi disponible.

This publication is also available in English.

Techniques d'enquête, décembre 2016
Vol. 42, nº 2, pp. 255-280
Statistique Canada, nº 12-001-X au catalogue

Inférence statistique fondée sur des échantillons poststratifiés par choix raisonné en population finie

Omer Ozturk[1]

Résumé

Le présent article traite de l'inférence statistique de la moyenne d'une population finie fondée sur des échantillons poststratifiés par choix raisonné (PCR). L'échantillon PCR s'obtient en sélectionnant d'abord un échantillon aléatoire simple, puis en stratifiant les unités sélectionnées en H classes créées par choix raisonné en se basant sur les positions relatives (rangs) des unités dans un petit ensemble de taille H. Cela donne un échantillon présentant des tailles d'échantillon aléatoires dans les classes créées par choix raisonné. Le processus de classement peut être effectué en se servant de variables auxiliaires ou par inspection visuelle afin de déterminer les rangs des observations mesurées. L'article décrit l'élaboration d'un estimateur sans biais et la construction d'un intervalle de confiance pour la moyenne de population. Puisque les rangs déterminés par choix raisonné sont des variables aléatoires, en conditionnant sur les observations mesurées, nous construisons des estimateurs Rao-Blackwellisés de la moyenne de population. Nous montrons que les estimateurs Rao-Blackwellisés donnent de meilleurs résultats que les estimateurs PCR habituels. Les estimateurs proposés sont appliqués aux données du recensement de 2012 du *United States Department of Agriculture*.

Mots-clés : Échantillon poststratifié; correction pour échantillon fini; échantillon d'ensembles ordonnés; échantillon stratifié; estimateur Rao-Blackwellisé.

1 Introduction

Dans le cas de nombreuses études par sondage, en plus de la variable d'intérêt, la base de sondage contient des variables auxiliaires en vue d'améliorer le contenu informatif d'un échantillon. Ces variables auxiliaires ont été utilisées fructueusement pour construire de meilleurs estimateurs, comme les estimateurs par le ratio et par la régression. Ces estimateurs requièrent habituellement des hypothèses de modélisation fortes quant à la relation entre la ou les variables auxiliaires et la variable d'intérêt. MacEachern, Stasny et Wolfe (2004) ont introduit un échantillon poststratifié par choix raisonné (PCR) et construit des estimateurs qui nécessitent des hypothèses de modélisation plus faibles que les estimateurs par le ratio et par la régression.

Un échantillon PCR s'obtient en tirant d'abord un échantillon aléatoire simple de taille n dans une population et en mesurant toutes les unités sélectionnées, $X_i;\ \ i = 1,\ldots,n.$ Pour chacune des unités mesurées, le chercheur sélectionne $H-1$ unités additionnelles afin de former un ensemble de taille H. Cet ensemble contient l'unité mesurée X_i et les $H-1$ unités additionnelles sélectionnées. Les unités que contiennent les ensembles ainsi formés sont ensuite classées de la plus petite à la plus grande sans effectuer de mesure pour déterminer le rang de X_i. Les paires $(X_i, R_i); i = 1,\ldots,n$ constituent l'échantillon PCR. Le processus de classement dans ces ensembles peut être effectué par inspection visuelle des unités ou en se servant d'une variable auxiliaire disponible. En cas de recours à l'inspection visuelle, la personne qui effectue le classement ne doit pas connaître les valeurs réelles de X_i pour éviter tout biais. En cas d'utilisation d'une variable auxiliaire, il faut qu'il existe une relation monotone entre la variable X et la

1. Omer Ozturk, The Ohio State University, Department of Statistics, 1958 Neil Avenue, Columbus, OH, 43210, États-Unis. Courriel : omer@stat.osu.edu.

variable auxiliaire. Ces hypothèses sont beaucoup plus faibles que l'hypothèse de linéarité dans les estimateurs par la régression et par le ratio.

L'information sur le classement dans un échantillon PCR est utilisée en vue d'induire une structure parmi les observations mesurées grâce à la création par choix raisonné de H classes d'unités similaires. La classe créée par choix raisonné h, $h = 1,\ldots,H$ contient toutes les observations mesurées possédant le rang par choix raisonné h. Puisque le rang R_i renseigne sur la position relative de X_i parmi H unités dans un ensemble, les observations dans la classe créée par choix raisonné h sont stochastiquement plus grandes que les observations dans la classe créée par choix raisonné h', pour $h' < h$. Cette structure induite accroît l'information contenue dans l'échantillon. On peut également considérer un échantillon PCR comme étant un échantillon stratifié qui contient H strates. Dans ce cas, la plus grande efficacité peut être établie en partant de la théorie classique de l'échantillonnage stratifié dans les plans d'échantillonnage.

Dans un échantillon PCR, les rangs sont déterminés après avoir sélectionné un échantillon aléatoire simple. Donc, le nombre d'observations, M_h, dans une classe créée par choix raisonné h est une variable aléatoire. La distribution conjointe de $\mathbf{M} = (M_1,\ldots,M_H)$ suit une loi multinomiale de paramètres n et vecteur de probabilité de succès $(1/H,\ldots,1/H)$. Puisque $\mathbf{M}$ est une variable aléatoire, il est fort possible que $M_h = 0$ pour une certaine valeur de h quand la taille d'échantillon n est petite. L'inférence statistique doit alors tenir compte de l'effet des strates vides sur les procédures.

Dans le contexte d'une population infinie, l'échantillon PCR a suscité l'intérêt de très nombreux chercheurs. Pour un petit survol de la littérature, le lecteur est invité à consulter Frey et Feeman (2012, 2013), Frey et Ozturk (2011), Stokes, Wang et Chen (2007), Wang, Lim et Stokes (2008), Wang, Stokes, Lim et Chen (2006), Wang, Wang et Lim (2012), Ozturk (2013, 2014a, 2014b, 2015), ainsi que les références mentionnées dans ces articles.

Un moyen d'éviter une taille d'échantillon aléatoire M_h consiste à classer les unités dans chaque ensemble avant de tirer un échantillon aléatoire simple dans la population. Dans ce cas, on parle d'un plan d'échantillonnage d'ensembles ordonnés (EEO). L'échantillonnage d'ensembles ordonnés a été introduit dans McIntyre (1952, 2005) en vue d'estimer la moyenne de population en recherche agricole. Pour construire un échantillon EEO de taille n, le chercheur détermine d'abord les paramètres du plan, la taille des ensembles H et le vecteur des tailles d'échantillon des classes créées par choix raisonné $\mathbf{m} = (m_1,\ldots,m_H)$, où m_h est le nombre requis d'observations qui doivent être sélectionnées dans la classe créée par choix raisonné h. Ensuite, le chercheur sélectionne nH unités au hasard dans la population et les répartit en n ensembles, chacun de taille H. Dans chacun de ces ensembles, les unités sont classées et la h^{e} statistique d'ordre établie par choix raisonné est mesurée dans m_h ensembles de manière que $\sum_{h=1}^{H} m_h = n$. Les observations mesurées $X_{[h]j}$; $j = 1,\ldots,m_h$; $h = 1,\ldots,H$ constituent un échantillon non équilibré d'ensembles ordonnés, où $X_{[h]j}$ est la statistique d'ordre établie par choix raisonné provenant d'un ensemble de taille H. Si les tailles des échantillons des classes créées par choix raisonné sont toutes égales $m_h \equiv n/H$; $h = 1,\ldots,H$, l'échantillon est alors considéré comme un échantillon équilibré d'ensembles ordonnés. S'il n'y a pas d'erreur de classement, la statistique d'ordre établie par choix raisonné devient la statistique d'ordre habituelle provenant d'un échantillon de taille H. Dans ce cas, on utilise la notation de la statistique d'ordre habituelle pour désigner la h^{e} statistique d'ordre, $X_{(h)j}$.

Ces dernières années, les travaux de recherche sur les échantillonnages PCR et EEO dans un contexte de population finie se sont multipliés. Patil, Sinha et Taillie (1995) ont utilisé un échantillon d'ensembles ordonnés pour estimer la moyenne de population pour une population de taille N quand l'échantillon est tiré sans remise. Deshpande, Frey et Ozturk (2006) ont étendu le scénario de sélection sans remise décrit dans Patil et coll. (1995) à trois plans d'échantillonnage distincts, plan 0, plan 1 et plan 2, et construit les intervalles de confiance pour les quantiles de population. Sous le plan 0, l'échantillon est obtenu en remettant toutes les unités dans la population avant de sélectionner l'ensemble suivant. Sous le plan 1, l'échantillon est obtenu en remettant uniquement les unités non mesurées dans la population avant de sélectionner l'ensemble suivant. Sous le plan 2, l'échantillon est construit en ne remettant aucune des unités dans la population, qu'elles aient ou non été mesurées. Al-Saleh et Samawi (2007), Ozdemir et Gokpinar (2007 et 2008), Jafari Jozani et Johnson (2011, 2012), Gokpinar et Ozdemir (2010), Ozturk et Jafari Jozani (2013), et Frey (2011) ont calculé les probabilités d'inclusion et construit les estimateurs de type Horvitz-Thompson pour la moyenne et le total de population pour une variante des échantillons sous le plan 0, le plan 1 et le plan 2 basée sur des échantillons d'ensembles ordonnés.

Ozturk (2014a) a combiné les données de classement provenant de différentes sources dans un échantillon d'ensembles ordonnés, a estimé les probabilités d'inclusion des unités de la population et construit un estimateur pour la moyenne de population. Pour les situations où les valeurs de population des variables auxiliaires sont disponibles, Ozturk (2016) a utilisé les rangs dans la population (données globales de classement) des unités sélectionnées dans l'échantillon pour induire une plus forte structure dans les données afin d'améliorer le contenu informationnel de l'échantillon. Il a montré que les échantillons construits en se basant sur les données de classement globales offrent de plus grands gains d'efficacité. Une revue complète et à jour de la littérature sur la PCR et l'EEO figure dans un article de synthèse récent publié dans Wolfe (2012).

Dans le présent article, nous considérons des plans d'échantillonnage avec et sans remise pour l'échantillonnage PCR en population finie. La section 2 donne une description détaillée de la construction des plans. Pour chaque plan, nous obtenons les fonctions de masse de probabilité, les moyennes, les variances et les covariances des statistiques d'ordre. Les résultats sont utilisés pour construire un estimateur sans biais de la moyenne de population et des estimateurs sans biais de la variance des moyennes de population. La section 3 décrit la construction d'estimateurs Rao-Blackwellisés en conditionnant sur les observations mesurées $X_i;\ i = 1,\ldots,n.$ La section 4 fournit des données empiriques pour les nouveaux estimateurs. La section 5 présente l'application des procédures proposées aux données du recensement de l'agriculture des États-Unis de 2012 (USDA). La section 6 contient certaines conclusions. Les preuves des théorèmes sont présentées en annexe.

2 Plans d'échantillonnage et estimateur

Nous considérons une population finie de taille N, $\mathcal{P} = \{u_1,\ldots,u_N\}$, où u_j est la j^{e} unité dans la population. Soit X la variable d'intérêt. Les valeurs de X sur les unités de la population seront désignées par $x_1,\ldots,x_N$. Sans perte de généralité, nous supposons que les valeurs de population de la variable aléatoire

X sont ordonnées, $x_1 < \ldots < x_N$, de sorte que le rang de population de l'unité u_i par rapport à la variable X soit i, $R\left(x_{u_i}\right) = s_{u_i} = i$, où s_{u_i} est le rang de x_{u_i} parmi les N unités de la population. En plus de la variable d'intérêt X, nous supposons qu'il existe une variable additionnelle Y présentant une relation monotone avec la variable aléatoire X.

Nous considérons deux plans d'échantillonnage, le plan 0 et le plan 2. Sous les deux plans, un échantillon aléatoire simple $U_S = \left\{u_{s_1}, \ldots, u_{s_n}\right\}$ est sélectionné dans la population $\mathcal{P}$. Sans perte de généralité, l'échantillon U_S sera identifié au moyen du vecteur de rangs $S = \{s_1, \ldots, s_n\}$. Sous le plan 0, les unités sont sélectionnées avec remise, mais sous le plan 2, elles le sont sans remise. Toutes les unités sélectionnées dans l'échantillon sont mesurées pour la variable X. Tout au long de l'exposé, nous désignons $\mathbf{X} = (X_1, \ldots, X_n)$ comme étant un échantillon de taille n, où nous utilisons par commodité la notation $X_i = X_{s_i}$. Il est clair que les X_i; $i = 1, \ldots, n$ sont tous indépendants sous le plan 0, mais qu'ils sont corrélés négativement sous le plan 2. Pour chaque unité mesurée u_{s_i} dans l'échantillon, nous sélectionnons aléatoirement $H - 1$ unités additionnelles sans remise parmi les unités de population restantes pour former n ensembles, chacun de taille H,

$$S_{i,H} = \left\{u_{s_i}, u_{t_1}, \ldots, u_{t_{H-1}}\right\};\ s_i \neq t_h, u_{t_h} \in \mathcal{P};\ h = 1, \ldots, H-1;\ i = 1, \ldots, n.$$

Dans chaque ensemble $S_{i,H}$, nous classons les unités en nous basant sur la variable auxiliaire Y et nous déterminons le rang de l'unité mesurée u_{s_i}, R_i, parmi H unités. Notre échantillon poststratifié par choix raisonné (PCR) est alors constitué des paires (X_i, R_i), $i = 1, \ldots, n$. Sous le plan 0, nous remettons dans la population toutes les unités non mesurées dans l'ensemble $S_{i,H}$ avant de construire l'ensemble suivant. Donc, une même unité non mesurée peut figurer dans plus d'un ensemble. Sous le plan 2, aucune des unités non mesurées n'est remise dans la population avant de construire l'ensemble suivant. Donc, tous les ensembles $S_{i,H}$ sont disjoints.

On peut interpréter le vecteur de rangs $\mathbf{R} = \{R_1, \ldots, R_n\}$ comme une covariable qui remplace des unités similaires, c'est-à-dire des unités ayant les mêmes rangs, dans la même classe obtenue par choix raisonné. Un échantillon PCR fournit sur l'unité mesurée u_{s_i}, outre la valeur mesurée x_i, de l'information supplémentaire sous forme de sa position relative (rang R_i) dans l'ensemble $S_{i,H}$. La qualité de l'information dépend de la force de la relation monotone entre les variables X et Y. Il est évident que, si les rangs R_i, $i = 1, \ldots, n$ sont ignorés, l'échantillon se réduit à un échantillon aléatoire simple.

Le scénario de classement est dit cohérent si la même procédure de classement est utilisée dans tous les ensembles. Sous un scénario de classement cohérent, les égalités qui suivent sont vérifiées.

Lemme 1. *Soit (X_i, R_i) un échantillon PCR construit selon un scénario de classement cohérent et la taille d'ensemble H à partir de la population $\mathcal{P}$.*

i. Pour le plan r, $r = 0, 2$, nous avons

$$\sum_{h=1}^{H} P\left(X_{[h]} = x\right) = P\left(X_j = x \middle| R_j = h\right) = \sum_{h=1}^{H} P\left(X_{(h)} = x\right).$$

ii. Pour le plan 2, nous avons

$$\sum_{h=1}^{H}\sum_{h'=1}^{H} P\left(X_{[h]} = x, X_{[h']} = y\right) = \sum_{h=1}^{H}\sum_{h'=1}^{H} P\left(X_j = x, X_t = y \,|\, R_j = h, R_t = h'\right)$$

$$= \sum_{h=1}^{H}\sum_{h'=1}^{H} P\left(X_{(h)} = x, X_{(h')} = y\right), x \neq y.$$

La partie (i) du lemme 1 est donnée dans Presnell et Bohn (1999) dans des conditions de population infinie. Dans le présent article, nous utilisons un scénario de classement par choix raisonné cohérent sauf indication contraire. La moyenne et la variance conditionnelles de X_i sachant $R_j = h$ et la covariance conditionnelle de X_j, X_t sachant que $R_j = h, R_t = h'$ seront désignées par

$$\mu_{[h]} = E\left(X_{[h]}\right) = E\left(X_i \,|\, R_j = h\right),$$

$$\sigma^2_{[h]} = \mathrm{Var}\left(X_{[h]}\right) = \mathrm{var}\left(X_i \,|\, R_j = h\right)$$

et

$$\sigma_{[h,h']} = \mathrm{cov}\left(X_{[h]}, X_{[h']}\right) = \mathrm{cov}\left(X_j, X_t \,|\, R_j = h, R_t = h'\right).$$

Sous un classement parfait, les crochets dans ces expressions seront remplacés par des parenthèses.

Il existe une différence manifeste entre les échantillons PCR sous le plan 0 et le plan 2. Sous le plan 0, les paires, (X_i, R_i); $i = 1, \ldots, n,$ sont mutuellement indépendantes. Sous le plan 2, dans toute paire d'observations mesurées X_i et X_j, les observations sont négativement corrélées même si leurs rangs R_i et R_j sont indépendants. Les rangs sont indépendants parce qu'ils sont déterminés indépendamment dans les différents ensembles. Nous commençons par étudier les propriétés de distribution des variables aléatoires dans un échantillon PCR tiré de la population $\mathcal{P}$.

Lemme 2. *Soit (X_i, R_i) un échantillon PCR sous classement parfait avec taille d'ensemble H tiré de la population $\mathcal{P}$.*

i) La fonction de masse de probabilité conditionnelle de X_j sachant $R_j = h$ est

$$\beta(i;h) = P\left(X_j = x_i \,\middle|\, R_j = h\right) = \frac{\binom{i-1}{h-1}\binom{N-i}{H-h}}{\binom{N}{H}}, x_i \in \mathcal{P}$$

dans les conditions du plan 0 ainsi que du plan 2.

ii) Les fonctions de masse de probabilité conditionnelles de X_r et X_t sachant que $R_r = h$ et $R_t = h'$, $\beta(i, j; h, h') = P\left(X_r = x_i, X_t = x_j \,\middle|\, R_r = h, R_t = h'\right)$ sont

$$\beta_0(i, j; h, h') = \beta(i; h)\beta(j; h'), (x_i, x_j) \in \mathcal{P}$$

pour le plan 0 et

$$\beta_1(i, j; h, h') =$$

$$\sum_{\lambda=0}^{j-i-1} \frac{\binom{i-1}{h-1}\binom{j-i-1}{\lambda}\binom{N-j}{H-\lambda-h}\binom{j-1-h-\lambda}{h'-1}\binom{N-j-H+\lambda+h}{H-h'}}{\binom{N}{H}\binom{N-H}{H}}, \; i < j, (x_i, x_j) \in \mathcal{P}$$

pour le plan 2.

iii) La moyenne et la variance conditionnelles de X_j sachant son rang $R_j = h$ sont

$$\mu_{(h)} = E(X_j \mid R_j = h) = \sum_{i=1}^{N} x_i \beta(i, h)$$

$$\sigma^2_{(h)} = \mathrm{Var}(X_j \mid R_j = h) = \sum_{i=1}^{N} x_i^2 \beta(i, h) - \mu^2_{(h)}$$

pour le plan 0 ainsi que le plan 2.

iv) La covariance conditionnelle de X_r, X_t sachant leurs rangs sont

$$\mathrm{cov}(X_r, X_t \mid R_r = h, R_t = h') = \sigma_{(h,h')} = 0$$

pour le plan 0 et

$$\sigma_{(h,h')} = \sum_{i=1}^{N}\sum_{i\neq j}^{N} x_i x_j \beta_1(i, j, h, h') - \sum_{i=1}^{N} x_i \beta(i, h) \sum_{i=1}^{N} x_i \beta(i, h')$$

pour le plan 2.

Les preuves des parties (i) et (ii) du lemme susmentionné sont données dans Patil et coll. (1995). Les preuves des autres parties sont très faciles et omises ici.

Le processus de classement dans un échantillon PCR aboutit à un vecteur aléatoire multinomial $\mathbf{M} = (M_1, \ldots, M_H)$, où M_h est le nombre d'observations dans la classe (poststrate) créée par choix raisonné h, $h = 1, \ldots, H$. La distribution marginale de M_h suit une loi binomiale de paramètres n et $1/H$. Pour simplifier la notation, nous utilisons $I_h = I(M_h > 0)$ pour désigner l'événement consistant en ce que la classe créée par choix raisonné h est non vide et $d_n = \sum_{h=1}^{H} I_h$ pour définir le nombre de classes créées par choix raisonné non vides dans un échantillon PCR. Dans le lemme qui suit, nous fournissons certains résultats préliminaires utiles sur le vecteur des tailles d'échantillon des classes créées par choix raisonné, dont la preuve peut être consultée dans Dastbaravarde, Arghami et Sarmad (2016) et dans Ozturk (2014b).

Lemme 3. *Soit* $(X_i, R_i); i = 1, \ldots, n,$ *un échantillon PCR construit sous un scénario de classement cohérent avec taille fixée* H *tiré de* $\mathcal{P}$. *Les égalités qui suivent sont vérifiées sous le plan 0 ainsi que le plan 2 :*

(i) $E\left(\frac{I_1}{d_n}\right) = 1/H$.

(ii) $E\left(\frac{I_1^2}{d_n^2}\right) = \frac{1}{H^2}\sum_{k=1}^{H}\left(\frac{k}{H}\right)^{n-1}$.

(iii) $Var\left(\frac{I_1}{d_n}\right) = \frac{1}{H^2}\sum_{k=1}^{H-1}\left(\frac{k}{H}\right)^{n-1}$.

(iv) $cov\left(\frac{I_1}{d_n}, \frac{I_2}{d_n}\right) = -\frac{1}{H-1}Var\left(\frac{I_1}{d_n}\right)$.

(v) $E\left(\frac{I_1^2}{M_1 d_n^2}\right) = \frac{1}{H^n}\left\{\frac{1}{n} + \sum_{k=2}^{H}\sum_{j=1}^{k-1}\sum_{m_h=1}^{n-k+1}\frac{(-1)^{j-1}}{k^2 m_h}\binom{H-1}{k-1}\binom{k-1}{j-1}\binom{n}{m_h}(k-j)^{n-m_h}\right\}$.

Considérons maintenant l'estimation de la moyenne de population. Nous utilisons

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{et} \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

pour désigner la moyenne et la variance de la population $\mathcal{P}$, respectivement. Soit

$$\hat{\mu}_r = \sum_{h=1}^{H}\frac{I_h}{M_h d_n}\sum_{i=1}^{n} X_i I(R_i = h),\ r = 0, 2$$

l'estimateur de la moyenne de population μ fondé sur le plan r, $r = 0, 2$, respectivement. Dans ces estimateurs, I_h, M_h et d_n sont des variables aléatoires. Elles sont utilisées pour apporter une correction en vue d'obtenir un estimateur sans biais pour μ quand certaines classes créées par choix raisonné sont vides. S'il n'est pas tenu compte des rangs dans un échantillon PCR, celui-ci devient un échantillon aléatoire simple fondé sur le plan 0 ou le plan 2. Dans ce cas, la moyenne de population μ est estimée par

$$\bar{X}_r = \frac{1}{n}\sum_{j=1}^{n} X_j, \quad r = 0, 2$$

pour les données du plan 0 ou du plan 2, respectivement.

Théorème 1. *Soit* $(X_i, R_i);\ i = 1, \ldots, n,$ *un échantillon PCR avec taille d'ensemble* H *construit sous un scénario de classement cohérent fondé sur le plan 0 ou sur le plan 2 à partir de la population finie* $\mathcal{P}$. *i) Les estimateurs* $\hat{\mu}_r$ *sont sans biais pour* μ. *ii) Les variances des estimateurs* $\hat{\mu}_r$ *sont*

$$\sigma^2_{\hat{\mu}_0} = \frac{H}{H-1}\mathrm{Var}\left(\frac{I_1}{d_n}\right)\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2 + E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h-1}^{H}\sigma^2_{[h]}$$

pour $r = 0$ *et*

$$\sigma^2_{\hat{\mu}_2} = \frac{H}{H-1}\mathrm{Var}\left(\frac{I_1}{d_n}\right)\sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2 - \frac{1}{H-1}\left(\frac{1}{H}-E\left(I_1/d_n\right)^2\right)\frac{H^2\sigma^2}{N-1}$$
$$+E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h}^{H}\sigma^2_{[h]} + \left(\frac{H}{H-1}E\left(I_1/d_n\right)^2 - E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{1}{H\left(H-1\right)}\right)\sum_{h=1}^{H}\sigma_{[h,h]}$$

pour $r = 2$.

Toutes les valeurs espérées dans $\sigma^2_{\hat{\mu}_0}$ et $\sigma^2_{\hat{\mu}_2}$ sont calculées sur le vecteur des tailles d'échantillon aléatoires $\mathbf{M}$. Ces valeurs espérées peuvent être calculées facilement au moyen du lemme 2 en utilisant de simples fonctions R. Les estimateurs de la moyenne de population fondés sur un échantillon équilibré d'ensembles ordonnés sous le plan 0 et sous le plan 2 sont donnés par

$$\mu^*_r = \frac{1}{mH}\sum_{i=1}^{m}\sum_{h=1}^{H} X_{[h]i}, \quad r = 0, 2,$$

où $m = n/H$ est la taille du cycle. Puisque les observations sous le plan 0 sont indépendantes, la variance de μ^*_0 est la même que la variance de la moyenne d'échantillon EEO dans une population infinie. La variance de μ^*_2 est donnée par l'équation 4.5 dans Patil et coll. (1995). En ce qui concerne notre notation, la variance de μ^*_2 s'écrit

$$\sigma^2_{\mu^*_2} = \frac{(N-1-n)}{n(N-1)}\sigma^2 - \frac{1}{nH}\sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2 - \frac{1}{nH}\sum_{h=1}^{H}\sigma_{[h,h]}. \tag{2.1}$$

Nous exprimons la variance de $\hat{\mu}_2$ sous une forme légèrement différente afin de la comparer à $\sigma^2_{\mu^*_2}$

$$\sigma^2_{\hat{\mu}_2} = \sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2\left\{\frac{H}{H-1}\mathrm{Var}\left(I_1/d_n\right) - E\left(\frac{I_1^2}{d_n^2 M_1}\right)\right\}$$
$$+\frac{H\sigma^2}{(N-1)(H-1)}\left\{(N-1)(H-1)E\left(\frac{I_1^2}{d_n^2 M_1}\right) - 1 + HE\left(\frac{I_1^2}{d_n^2}\right)\right\}$$
$$+\left\{\frac{H}{H-1}E\left(I_1/d_n\right)^2 - E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{1}{H\left(H-1\right)}\right\}\sum_{h=1}^{H}\sigma_{[h,h]}. \tag{2.2}$$

Il est facile de voir l'impact du vecteur des tailles d'échantillon aléatoires $\mathbf{M}$ sur l'estimateur $\hat{\mu}_2$ dans un échantillon PCR en comparant les équations (2.1) et (2.2). Les expressions entre accolades dans l'équation (2.2) apportent les corrections pour les tailles d'échantillon aléatoires dans l'échantillon PCR. Pour de grandes tailles de population et d'échantillon, $\sigma^2_{\mu^*_2}$, $\sigma^2_{\hat{\mu}_2}$ et $\sigma^2_{\hat{\mu}_0}$ se réduisent à des formes simples.

Corollaire 1. *Supposons que* n *et* N *augmentent de telle façon que le ratio de* n/N *s'approche d'une limite à* f, $\lim_{n\to\infty}(n/N) = f$.

i) Si $f > 0$, *les variances* $\sigma^2_{\hat{\mu}_2}$, $\sigma^2_{\mu^*_2}$ *et* $\sigma^2_{\hat{\mu}_0}$ *convergent vers deux formes simples*

$$\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_2} = \lim_{n\to\infty} n\sigma^2_{\mu^*_2} = (1-f)\sigma^2 - \frac{1}{H}\sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2 - \sum_{h=1}^{H}\sigma_{h,h}$$

et

$$\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_0} = \frac{1}{H}\sum_{h=1}^{H}\sigma^2_{[h]} = \sigma^2 - \frac{1}{H}\sum_{h=1}^{H}(\mu_h - \mu)^2.$$

ii) Si $f = 0$, $\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_2} = \lim_{n\to\infty} n\sigma^2_{\mu_2^*} = \lim_{n\to\infty} n\sigma^2_{\hat{\mu}_0} = \frac{1}{H}\sum_{h=1}^{H}\sigma^2_{[h]}$, *ce qui est la même chose que la variance de la moyenne d'échantillon d'un échantillon d'ensembles ordonnés dans des conditions de population infinie.*

iii) Si f *est strictement positive, alors* $\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_2} = \lim_{n\to\infty} n\sigma^2_{\hat{\mu}_{RSS}} < \lim_{n\to\infty} n\sigma^2_{\hat{\mu}_0}$.

La partie (iii) du corollaire indique que, quand les tailles de l'échantillon et de la population augmentent à un certain taux, les variances des moyennes d'échantillon des échantillons PCR et EEO dans des conditions de population finie sont toujours plus petites que la variance du même estimateur dans des conditions de population infinie. Ce gain d'efficacité est dû à la corrélation négative entre X_i et X_j dans les plans d'échantillonnage sans remise.

Nous construisons maintenant des estimateurs sans biais pour $\sigma^2_{\hat{\mu}_0}$, $\sigma^2_{\hat{\mu}_2}$ et $\sigma^2_{\mu_2^*}$. Commençons par réécrire les estimateurs $\sigma^2_{\hat{\mu}_2}$ et $\sigma^2_{\mu_2^*}$ sous une forme légèrement différente

$$\begin{aligned}\sigma^2_{\hat{\mu}_2} &= \left\{\frac{1}{H(H-1)} + E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{H}{H-1}E\left(\frac{I_1^2}{d_n^2}\right)\right\}\left\{\sum_{h=1}^{H}\sigma^2_{[h]} - \sum_{h=1}^{H}\sigma_{h,h}\right\} \\ &\quad + \frac{H^2\sigma^2}{H-1}\left\{\mathrm{Var}\left(\frac{I_1}{d_n}\right) - \frac{1}{N-1}\left\{\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right\}\right\} \\ &= C_1(n,H)\left\{\sum_{h=1}^{H}\sigma^2_{[h]} - \sum_{h=1}^{H}\sigma_{h,h}\right\} + C_2(n,H,N)\frac{H^2\sigma^2}{H-1} \qquad (2.3)\end{aligned}$$

$$\sigma^2_{\mu_2^*} = \frac{1}{Hn}\left\{\sum_{h=1}^{H}\sigma^2_{[h]} - \sum_{h=1}^{H}\sigma_{h,h}\right\} - \frac{\sigma^2}{(N-1)}.$$

Dans l'équation (2.3), il est clair que les coefficients $C_1(n,H)$ et $C_2(n,H,N)$ sont des quantités connues pour les valeurs données de la taille d'échantillon n et de la taille de l'ensemble H. Soit

$$T_1 = \frac{1}{E\left(\frac{I_1 I_2}{d_n^2}\right)}\sum_{h=1}^{H}\sum_{h\neq h'}^{H}\frac{I_h I_{h'}}{M_h M_{h'} d_n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - X_j)^2 I(R_i = h) I(R_j = h'),$$

$$T_2 = \sum_{h=1}^{H}\frac{HI_h^*}{M_h d_n^*(M_h - 1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(X_i - X_j)^2 I(R_i = h) I(R_j = h),$$

$$T_1^* = \frac{1}{2m^2H^2}\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\sum_{i=1}^{m}\sum_{j=1}^{m}\left(X_{[h]i} - X_{[h']j}\right)^2$$

$$T_2^* = \frac{1}{2m(m-1)H^2}\sum_{h=1}^{H}\sum_{i=1}^{m}\sum_{j\neq i}^{m}\left(X_{[h]i} - X_{[h]j}\right)^2$$

et

$$\hat{\sigma}^2_{EAS} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2,$$

où $I_h^* = I(M_h > 1)$, $d_n^* = \sum_{h=1}^{H} I_h^*$ et m est la taille du cycle dans un échantillon équilibré d'ensembles ordonnés. Partant du lemme 3, on peut établir facilement que $E\left(I_1 I_2 / d_n^2\right) = \left(1/H - E\left(I_1/d_n\right)^2\right)/(H-1)$. D'où, T_1 est une statistique qui dépend uniquement des données. Notons que $\hat{\sigma}^2_{EAS}$ est un estimateur sans biais de σ^2. Dans le théorème suivant, nous donnons d'autres estimateurs sans biais $\left(\bar{\sigma}^2\right)$ de σ^2 basés sur les échantillons PCR et EEO sous le plan 0 et le plan 2.

Théorème 2. *Soit* (X_i, R_i), $i = 1,\ldots,n$, *et* $X_{[h]i}$; $h = 1,\ldots,H$; $i = 1,\ldots,m$, *les échantillons PCR et EEO ayant une taille d'ensemble* H, *respectivement, tous deux provenant de la population* $\mathcal{P}$. *Des estimateurs sans biais de* σ^2, $\sigma^2_{\hat{\mu}_0}$, $\sigma^2_{\hat{\mu}_2}$ *et* $\sigma^2_{\mu_2^*}$ *sont donnés, respectivement, par*

$$\bar{\sigma}^2 = \begin{cases} \left(T_1 + T_2\right)/\left(2H^2\right) & \textit{pour le plan } 0 \\ \dfrac{(N-1)(T_1+T_2)}{2NH^2} & \textit{pour le plan } 2 \\ \dfrac{\left(T_1^* + T_2^*\right)(N-1)}{N} & \textit{pour un EEO équilibré sous le plan } 2, \\ T_1^* + T_2^* & \textit{pour un EEO équilibré sous le plan } 0, \end{cases} \tag{2.4}$$

$$\hat{\sigma}^2_{\hat{\mu}_0} = \frac{Var\left(I_1/d_n\right)}{2(H-1)}T_1 + \left\{E\left(\frac{I_1^2}{d_n^2 M_1}\right) - Var\left(\frac{I_1}{d_n}\right)\right\}\frac{T_2}{2}, \tag{2.5}$$

$$\hat{\sigma}^2_{\hat{\mu}_2} = C_1(n,H)T_2/2 + C_2(n,H,N)\frac{H^2\hat{\sigma}^2_{EAS}}{H-1}, \tag{2.6}$$

$$\bar{\sigma}^2_{\hat{\mu}_2} = C_1(n,H)T_2/2 + C_2(n,H,N)\frac{(N-1)(T_1+T_2)}{2N(H-1)}, \tag{2.7}$$

$$\hat{\sigma}^2_{\mu_2^*} = \frac{T_2^*}{m} - \frac{T_1^* + T_2^*}{N}.$$

Notons que $\hat{\sigma}^2_{\hat{\mu}_2}$ et $\bar{\sigma}^2_{\hat{\mu}_2}$ sont tous deux sans biais pour $\sigma^2_{\hat{\mu}_2}$. L'estimateur $\bar{\sigma}^2$ est également sans biais pour la variance de population σ^2 dans les échantillons EEO et PCR sous le plan 0 et le plan 2. Le théorème 2 indique que tous les estimateurs de variance sont sans biais pour toute taille d'échantillon $n > 1$. Nous notons que $E\left(I_1^2/d_n^2\right) \le 1/H^2$. En utilisant cette borne, on peut montrer que $C_1(n,H) \ge 0$. Par ailleurs, le coefficient $C_2(n,H,N)$ peut être négatif pour certaines valeurs de n, N et H. Rarement, cela peut donner une valeur négative pour $\hat{\sigma}^2_{\hat{\mu}_2}$. Si $\hat{\sigma}^2_{\hat{\mu}_2}$ est négatif, nous proposons un estimateur tronqué

$$\tilde{\sigma}^2_{\hat{\mu}_2} = \begin{cases} \hat{\sigma}^2_{\hat{\mu}_2} & \text{si } \hat{\sigma}^2_{\hat{\mu}_2} > 0 \\ C_1(n,H)T/2 & \text{si } \hat{\sigma}^2_{\hat{\mu}_2} \le 0. \end{cases} \tag{2.8}$$

Cet estimateur est toujours positif et son biais est très faible. Les valeurs de $\tilde{\sigma}^2_{\hat{\mu}_2}$ semblent être toujours positives sur la base de notre étude en simulation limitée.

3 Estimateur Rao-Blackwellisé

À la présente section, nous construisons des estimateurs qui améliorent la performance des estimateurs PCR de la section précédente. Pour un échantillon aléatoire simple donné $X_{s_1},\ldots,X_{s_n}$, les ensembles $S_{j,H}$; $j=1,\ldots,n,$ peuvent être construits sur tous les appariements possibles de $n(H-1)$ unités additionnelles avec n unités entièrement mesurées. Chaque construction crée un nouvel ensemble de rangs, donc une nouvelle estimation. Nous combinons toutes ces estimations en utilisant le théorème de Rao-Blackwell. Soit

$$\tilde{\mu}_r = E(\hat{\mu}\,|\,\mathbf{X}) = E\left\{\sum_{h=1}^{H}\frac{I_h}{M_h d_n}\sum_{j=1}^{n}X_j I\left(R_j=h\right)\Big|\,X_1,\ldots,X_n\right\}$$

$$= \sum_{i=1}^{n}X_i E\left[\sum_{h=1}^{H}\frac{I_h I(R_i=h)}{M_h d_n}\,\middle|\,\mathbf{X}\right] = \sum_{i=1}^{n}X_i a_{i|\mathbf{X}}; r=0,2,$$

où

$$a_{i|\mathbf{X}} = E\left[\sum_{h=1}^{H}\frac{I_h I(R_i=h)}{M_h d_n}\,\middle|\,\mathbf{X}\right].$$

L'espérance en $a_{i|\mathbf{X}}$ est prise sur les distributions conditionnelles de $R_j;\ j=1,\ldots,n,\ M_h;\ h=1,\ldots,N,$ et d_n sachant $\mathbf{X}$. Nous notons que les rangs, $R_j;\ j=1,\ldots,n,$ sont attribués indépendamment dans chaque ensemble $S_{j,H}$. Donc, les distributions conjointes de $R_j;\ j=1,\ldots,n,$ sachant les observations mesurées $X_j;\ j=1,\ldots,n,$ sont toutes indépendantes

$$\alpha_{h_1,\ldots,h_n|\mathbf{X}} = P(R_1=h_1,\ldots,R_n=h_n\,|\,\mathbf{X}) = \prod_{j=1}^{n}P\left(R_j=h_j\,|\,\mathbf{X}\right),\ 1\le h_j\le H.$$

Supposons que les rangs dans la population, $s_j;\ j=1,\ldots,n,$ des unités de l'échantillon sont connus. Pour construire la distribution conditionnelle de $R_j=h_j$ sachant $X_j=x_{s_j}$, nous observons d'abord que

$$P\left(R_j=h_j, X_j=x_{s_j}\right) = \beta\left(s_j,h_j\right)/H \quad \text{et} \quad P\left(X_j=x_{s_j}\right)=1/N.$$

En utilisant ces fonctions de masse de probabilité conjointe et marginale, nous écrivons

$$\alpha_{h_j|s_j} = P\left(R_j=h_j\,|\,\mathbf{X}\right) = P\left(R_j=h_j\,\middle|\,X_j=x_{s_j}\right) = \frac{\binom{s_j-1}{h_j-1}\binom{N-s_j}{H-h_j}}{\binom{N-1}{H-1}}, h_j=1,\ldots,H, x_{s_j}\in\mathcal{P}, \tag{3.1}$$

où x_{s_j} est la s_j^{e} unité la plus petite dans la population. L'évaluation de $a_{s_j|\mathbf{X}}$ dans $\tilde{\mu}_r$ requiert de grandes ressources de calcul. Même si les distributions conditionnelles des rangs R_j sachant $\mathbf{X}$ sont indépendantes, elles ne sont pas identiques. D'où, la distribution conditionnelle du vecteur des tailles d'échantillon des classes créées par choix raisonné $\mathbf{M}$ sachant $\mathbf{X}$ ne suit pas une loi multinomiale.

Nous présentons maintenant une approximation pour évaluer $\tilde{\mu}_r$. Nous reconnaissons d'abord que la distribution conditionnelle de $R_j = h$ sachant $X_j = x_{s_j}$ suit une loi hypergéométrique. Donc, nous pouvons générer R_j à partir de cette loi hypergéométrique pour des valeurs données de $\mathbf{X}$.

Algorithme 1.

I. Choisir un entier B. Pour $b = 1,\ldots,B$, générer $\mathbf{R}^{b,*} = \{R_1^{b,*},\ldots,R_n^{b,*}\}$ à partir de $\alpha_{R_j|X_j=x_{s_j}}$ dans l'équation (3.1).

II. En utilisant $\mathbf{R}^{b,*}$, calculer $I_h^{b,*} = I\left(M^{b,*} > 0\right)$, $M_h^{b,*} = \sum_{j=1}^{n} I\left(R_j^{b,*} = h\right)$, $d_n^{b,*} = \sum_{h=1}^{H} I_h^{b,*}$ et

$$a_i^{b,*} = \sum_{h=1}^{H} \frac{I^{b,*}}{M_h^{b,*} d_n^{b,*}} \sum_{h=1}^{H} I\left(R_i^{b,*} = h\right),\ i = 1,\ldots,n.$$

Nous approximons a_i par $\bar{a}_i^* = \sum_{b=1}^{B} a_i^{b,*} / B$, $i = 1,\ldots,n$. De la loi des grands nombres, il découle que $\bar{a}_i^*$ s'approche de a_i à mesure que B devient grand. L'estimateur Rao-Blackwellisé $\tilde{\mu}_r$ est alors approximé par $\tilde{\mu}_r^* = \sum_{i=1}^{n} \bar{a}_i^* X_i$.

Si les rangs de population des unités de l'échantillon ne sont pas disponibles, l'algorithme 1 peut ne pas être utilisable. Dans ce cas, nous nous servons de l'ensemble de toutes les unités non mesurées pour construire les estimateurs Rao-Blackwellisés. Soit $\mathbf{Y}^{\mathrm{T}} = \left(Y_1,\ldots,Y_{n(H-1)}\right)$ les variables auxiliaires sur les variables aléatoires non mesurées. Nous utilisons l'algorithme qui suit pour obtenir une approximation des estimateurs Rao-Blackwellisés.

Algorithme 2. *Pour $b = 1,\ldots,B$, répéter les étapes I à IV.*

I. Effectuer une permutation aléatoire sur les entrées du vecteur $\mathbf{Y}$ pour obtenir $\mathbf{Y}_b = permuter(\mathbf{Y})$.

II. Répartir les entrées de $\mathbf{Y}_b$ en n ensembles, chacun de taille $H-1$.

III. Apparier ces n ensembles de taille $H-1$ avec n valeurs Y des unités mesurées pour former n ensembles, chacun de taille H. Obtenir le rang de la mesure X à partir du rang de la valeur Y correspondante dans chaque ensemble, $\mathbf{R}_b^* = \left(R_{b,1}^*,\ldots,R_{b,n}^*\right)$.

IV. En utilisant $\mathbf{R}_b^*$, calculer $I_h^{b,*} = I\left(M^{b,*} > 0\right)$, $M_h^{b,*} = \sum_{j=1}^{n} I\left(R_{b,j}^* = h\right)$, $d_n^{b,*} = \sum_{h=1}^{H} I_h^{b,*}$ et

$$a_i^{b,*} = \sum_{h=1}^{H} \frac{I^{b,*}}{M_h^{b,*} d_n^{b,*}} \sum_{h=1}^{H} I\left(R_i^{b,*} = h\right),\ i = 1,\ldots,n.$$

V. Calculer l'estimateur Rao-Blackwellisé $\tilde{\mu}_r = \sum_{i=1}^{n} \bar{a}_i^* X_i$, $r = 0, 2$, où $\bar{a}_i^* = \sum_{b=1}^{B} a_i^{b,*} / B$.

Même si de grandes valeurs de B fournissent une meilleure approximation des estimateurs Rao-Blackwellisés, elles pourraient nécessiter un effort de calcul supplémentaire et ne pas être faisables en pratique. Par ailleurs, même de petites valeurs de B, telles que $B = 5$, pourraient donner lieu à une amélioration importante.

Considérons maintenant la construction d'estimateurs de la variance des estimateurs Rao-Blackwellisés. L'obtention d'expressions analytiques pour les variances de $\tilde{\mu}_r$ est un défi. La difficulté tient au fait qu'il n'existe pas d'expressions analytiques pour le calcul de $a_{i|\mathbf{X}};\ i = 1,\ldots,n.$ Nous faisons donc appel à une procédure bootstrap pour calculer la variance de $\tilde{\mu}_r$. Les estimateurs bootstrap peuvent être construits par une méthode d'insertion de valeurs. Soit θ une fonctionnelle statistique $\theta = T(\mathcal{P})$, où $\mathcal{P}$ est la population finie. L'estimation bootstrap de θ peut s'obtenir au moyen de $\hat{\theta} = T\left(\hat{\mathcal{P}}\right)$, où $\hat{\mathcal{P}}$ est la population empirique. En population finie, la construction de la population empirique joue un rôle important dans le respect des modalités de tirage sans remise des échantillons bootstrap. Soit K la partie entière de N/n et $k = N - Kn.$ Nous construisons $\hat{\mathcal{P}}$ avec

$$\hat{\mathcal{P}} = \left\{ \underbrace{\mathbf{X},\ldots,\mathbf{X}}_{K \text{ fois}}, X_{t_1},\ldots,X_{t_k} \right\},$$

où les $X_{t_j};\ j = 1,\ldots,k,$ sont tirés au hasard de $\mathbf{X} = \left\{X_{s_1},\ldots,X_{s_n}\right\}$. Dans le cas de cette construction, la taille de population, $N,$ est la même dans $\hat{\mathcal{P}}$ et $\mathcal{P}.$ Nous générons des répliques bootstrap des échantillons $\mathbf{X}^{*,1} = \left\{X_{s_1}^{*,1},\ldots,X_{s_n}^{*,1}\right\}$ à partir de $\hat{\mathcal{P}}$ avec remise pour le plan 0 et sans remise pour le plan 2. Pour construire la distribution bootstrap de l'estimateur $\tilde{\mu}_r,$ nous générons des répliques d'échantillon $\mathbf{X}^{*,c}, c = 1,\ldots,C$ et calculons

$$\tilde{\mu}_r^{*,c} = \sum_{i=1}^{n} \overline{a}_i^{*} X_{s_i}^{*,c}, \quad c = 1,\ldots,C$$

au moyen de l'algorithme 1 ou 2. L'estimation bootstrap de la variance de $\tilde{\mu}_r$ s'obtient alors à partir de

$$\hat{\sigma}^2_{\tilde{\mu}_r} = \frac{1}{C-1}\sum_{c=1}^{C}\left\{\tilde{\mu}_r^{*,c} - \tilde{\mu}_r^{*}\right\}^2, \tag{3.2}$$

où $\tilde{\mu}_r^{*}$ est la moyenne de $\tilde{\mu}_r^{*,c}, c = 1,\ldots,C.$

Nous construisons pour μ un intervalle de confiance bootstrap à $100(1-\gamma)\,\%$ par la méthode des centiles en prenant $\left(L_r^{\gamma/2}, L_r^{1-\gamma/2}\right)$, où L_r^a est le a^{e} quantile de $\tilde{\mu}_r^{*}$ satisfaisant $P\left(\tilde{\mu}_r^{*} \le L_r^a \,\middle|\, \hat{\mathcal{P}}\right)$ pour $0 < a < 1.$

4 Résultats empiriques

À la présente section, nous examinons les propriétés en échantillon fini des estimateurs au moyen d'une petite étude en simulation sous une gamme étendue de paramètres de simulation. Les ensembles de données sont générés à partir de populations suivant une loi normale discrète et une loi exponentielle décalée discrète pour une taille de population donnée $N.$ Les populations discrètes sont construites en partant de la fonction quantile

$$x_i = F^{-1}\left(\frac{i}{N+1}\right); \quad i = 1, \ldots, N, \tag{4.1}$$

où F est la fonction de répartition de la loi normale ou de la loi exponentielle. Pour la population normale discrète, nous avons utilisé un paramètre de position de 10 et un paramètre d'échelle de 4. Pour la population exponentielle discrète décalée, nous avons utilisé la fonction de répartition de la loi exponentielle classique pour générer x_i dans l'équation (4.1), puis nous avons décalé chaque x_i en ajoutant 10. La taille de population est fixée à $N = 150$.

Nous avons utilisé des tailles d'échantillon (n) et d'ensemble (H) donnant des valeurs entières pour n/H afin de pouvoir créer un échantillon équilibré d'ensembles ordonnés de taille n. Les combinaisons de tailles d'échantillon et d'ensemble (n, H) sont $(10, 2)$, $(15, 3)$, $(20, 4)$, $(25, 5)$. Pour contrôler la qualité de l'information sur les rangs, nous avons utilisé la variable auxiliaire Y, où $\rho = \text{cor}(X, Y)$ avec $\rho = 1, 0,75$. La valeur de $\rho = 1$ donne un classement parfait et la valeur de $\rho = 0,75$ crée des erreurs dans le classement. La taille de la simulation est fixée à 3 000. Les estimateurs Rao-Blackwellisés sont calculés au moyen de l'algorithme 1 avec $B = 50$ et une taille de réplique bootstrap de 200.

La première partie de la simulation a pour objet d'étudier les gains d'efficacité des estimateurs et la probabilité de couverture des intervalles de confiance de la moyenne de population. Tous les estimateurs sont comparés aux estimateurs Rao-Blackwellisés sous le plan 2 $(\tilde{\mu}_2)$. Soit $D(\mathbf{X})$ n'importe lequel des estimateurs présentés aux sections 2 et 3. L'efficacité relative de $D(.)$ par rapport à $\tilde{\mu}_2$ est donnée par

$$R(D) = \frac{EQM\,(D)}{EQM\,(\tilde{\mu}_2)}, \tag{4.2}$$

où $EQM\,(D)$ est l'erreur quadratique moyenne estimée de l'estimateur D. Dans l'équation (4.2), la valeur $R(D) > 1$ indique que l'estimateur $\tilde{\mu}_2$ est plus efficace que l'estimateur D.

Nous considérons deux types d'intervalles de confiance pour la moyenne de population. L'intervalle de confiance par la méthode des centiles fondée sur une distribution bootstrap est donné à la section 3. Les probabilités de couverture de ces intervalles seront notées $C^a\,(\tilde{\mu}_0)$ pour le plan 0 et $C^a\,(\tilde{\mu}_2)$ pour le plan 2. Un deuxième type d'intervalle de confiance approximatif peut être construit en s'appuyant sur la théorie classique. Notons que nous avons des estimateurs sans biais, $\hat{\sigma}_r$, pour les variances de $\hat{\mu}_r$; $r = 0, 2$. Un intervalle de confiance à $100(1-\gamma)\,\%$ pour μ est donné par

$$\hat{\mu}_r \pm t_{n-1,1-\gamma/2}\hat{\sigma}_r; \quad r = 0, 2,$$

où $t_{n-1,1-a}$ est le a^{e} quantile supérieur de la distribution t à $n-1$ degrés de liberté. Les probabilités de couverture de ces intervalles de confiance seront notées $C^b\,(\hat{\mu}_0)$ pour le plan 0 et $C^b\,(\hat{\mu}_2)$ pour le plan 2. Ahn, Lim et Wang (2014) ont proposé d'utiliser $n - H$ degrés de liberté pour l'approximation t. Ce choix peut également fonctionner pour l'échantillonnage PCR en population finie avec un certain accroissement de la variation due au caractère non équilibré d'un échantillon PCR. Toutefois, cette piste n'est pas poursuivie ici faute d'espace.

Le tableau 4.1 donne les efficacités relatives des estimateurs et les probabilités de couverture des intervalles de confiance pour les populations normales discrètes. Il est clair que l'estimateur sous le plan 2

Rao-Blackwellisé $(\tilde{\mu}_r)$ surpasse tous les autres estimateurs, y compris les estimateurs EEO. En général, les estimateurs EEO sont plus efficaces que les estimateurs PCR en raison du vecteur $\mathbf{M}$ des tailles aléatoires d'échantillon des classes créées par choix raisonné. Cela peut se constater au tableau 4.1 en examinant le ratio

$$\frac{R(\hat{\mu}_r)}{R(\mu_r^*)} = \frac{EQM(\hat{\mu}_r)}{EQM(\mu_r^*)},\ r = 0, 2.$$

Pour $r = 0,\ \rho = 1$ et les combinaisons de tailles échantillon-ensemble $(n, H), (10, 2), (15, 3), (20, 4), (25, 5)$, ces ratios sont de $1{,}267 (1{,}698/1{,}340)$, $1{,}491 (2{,}117/1{,}419)$, $1{,}815 (2{,}985/1{,}644)$, $2{,}391 (3{,}479/1{,}455)$, respectivement. Il est évident que l'estimateur pour l'échantillon d'ensembles ordonnés μ_0^* est plus efficace que l'estimateur PCR $\hat{\mu}_0$. Cela peut s'expliquer par le fait que l'échantillon EEO utilise un vecteur de tailles d'échantillon constantes (non aléatoires) $\mathbf{m} = (n_1, \ldots, n_H)$. D'où, il n'existe pas de variation supplémentaire comparable à celle due au caractère aléatoire de $\mathbf{M}$ dans l'échantillon PCR, ce qui donne une plus petite variance pour l'estimateur.

Le tableau 4.1 (entrées dans les colonnes $R(\mu_0^*)$ et $R(\mu_2^*)$) montre que les estimateurs PCR Rao-Blackwellisés sont meilleurs que les estimateurs EEO. Dans ce cas, on constate une différence nette entre les estimateurs PCR Rao-Blackwellisés et les estimateurs EEO. En ce qui concerne l'échantillon EEO, même si $\mathbf{m}$ est constant, l'information sur le classement (ou le rang R_i appartenant à chaque X_i) s'obtient à partir d'une construction particulière de n ensembles, chacun de taille H. Par ailleurs, les estimateurs PCR Rao-Blackwellisés tiennent compte de toutes les constructions possibles de n ensembles, chacun de taille H. Donc, le contenu de l'information de classement est plus riche dans un échantillon PCR Rao-Blackwellisé que dans un échantillon EEO. Cette information accrue sur le classement rend les estimateurs Rao-Blackwellisés supérieurs aux estimateurs EEO.

Le tableau 4.1 donne aussi les probabilités de couverture des intervalles de confiance. Les probabilités de couverture des intervalles de confiance bootstrap par la méthode des centiles sont légèrement inférieures à la valeur nominale de 0,95. Celles des intervalles de confiance basés sur la distribution t sont raisonnablement proches de la probabilité de couverture nominale de 0,95.

Tableau 4.1
Efficacités relatives des estimateurs et probabilités de couverture d'un intervalle de confiance à 95 % de la moyenne de population. Les ensembles de données sont générés à partir d'une population normale discrète de moyenne $\mu = 10$ et d'échelle $\sigma = 4$

			Efficacités relatives, $R(\bar{X}_0) = \mathrm{Var}(\bar{X}_0)/\mathrm{Var}(\tilde{\mu}_2)$							Probabilités de couverture			
n	H	ρ	$R(\bar{X}_0)$	$R(\bar{X}_2)$	$R(\hat{\mu}_0)$	$R(\hat{\mu}_2)$	$R(\mu_0^*)$	$R(\mu_2^*)$	$R(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_2)$	$C^b(\hat{\mu}_0)$	$C^b(\hat{\mu}_2)$
10	2	1,00	2,182	2,050	1,698	1,571	1,340	1,470	1,147	0,880	0,885	0,943	0,947
15	3	1,00	3,393	3,074	2,117	1,809	1,419	1,732	1,049	0,902	0,896	0,940	0,929
20	4	1,00	5,739	5,008	2,985	2,277	1,644	2,363	1,238	0,907	0,916	0,944	0,924
25	5	1,00	7,791	6,536	3,479	2,262	1,455	2,689	1,283	0,908	0,924	0,937	0,903
10	2	0,75	2,322	2,057	2,236	1,941	1,945	1,761	1,137	0,886	0,890	0,942	0,941
15	3	0,75	3,726	3,282	3,338	2,829	2,641	2,351	1,129	0,901	0,908	0,946	0,937
20	4	0,75	5,383	4,562	4,458	3,922	3,451	2,881	1,139	0,910	0,903	0,946	0,930
25	5	0,75	7,339	6,413	6,054	4,805	4,493	3,527	1,197	0,905	0,904	0,944	0,924

a : Les probabilités de couverture sont calculées en partant d'un intervalle de confiance bootstrap par la méthode des centiles.
b : Les probabilités de couverture sont calculées en partant de $\hat{\mu}_r \pm t_{n-1,\,0{,}975}\hat{\sigma}_{\hat{\mu}_r}$, $r = 0, 2$.

Le tableau 4.2 donne les estimations de la variance des estimateurs de la moyenne provenant de la simulation et des estimateurs donnés par les équations (2.5), (2.6), (2.8) et (3.2) aux sections 2 et 3. Nous avons déjà prouvé que les estimateurs $\hat{\sigma}^2_{\hat{\mu}_r}$, $r = 0, 2$ sont sans biais. Les entrées pour ces estimateurs de variance sont très proches des valeurs correspondantes fondées sur les estimations de variance simulées. L'estimateur de variance tronqué est presque identique à l'estimateur sans biais non tronqué. Cela montre que les valeurs négatives sont rares et qu'il y a peu de différence entre les estimateurs de variance tronqué et non tronqué. Les estimations de variance bootstrap des estimateurs Rao-Blackwellisés sont également très proches des estimations de variance simulées. Des tendances similaires à celles observées dans les tableaux 4.1 et 4.2 se dégagent aussi des tableaux 4.3 et 4.4 pour la population suivant une loi exponentielle décalée.

Tableau 4.2
Estimation de la variance des estimateurs. Les ensembles de données sont générés à partir d'une population normale discrète de moyenne $\mu = 10$ et d'échelle $\sigma = 4$

			Estimations d'après les équations (2.5), (2.6), (2.8), (3.2)					Estimations d'après la simulation			
n	H	ρ	$\hat{\sigma}^2_{\hat{\mu}_0}$	$\hat{\sigma}^2_{\hat{\mu}_2}$	$\tilde{\sigma}^2_{\tilde{\mu}_2}$	$\hat{\sigma}^2_{\tilde{\mu}_0}$	$\hat{\sigma}^2_{\tilde{\mu}_2}$	$V^a(\hat{\mu}_0)$	$V^a(\hat{\mu}_2)$	$V^a(\tilde{\mu}_0)$	$V^a(\tilde{\mu}_2)$
10	2	1,00	1,177	1,078	1,078	0,694	0,646	1,175	1,087	0,794	0,692
15	3	1,00	0,632	0,534	0,534	0,305	0,275	0,628	0,537	0,311	0,297
20	4	1,00	0,392	0,300	0,300	0,169	0,146	0,393	0,299	0,163	0,132
25	5	1,00	0,268	0,175	0,175	0,106	0,087	0,270	0,175	0,099	0,078
10	2	0,75	1,431	1,335	1,335	0,692	0,645	1,463	1,270	0,744	0,654
15	3	0,75	0,896	0,802	0,802	0,306	0,276	0,901	0,763	0,305	0,270
20	4	0,75	0,631	0,531	0,531	0,169	0,145	0,627	0,552	0,160	0,141
25	5	0,75	0,485	0,386	0,386	0,106	0,089	0,506	0,401	0,100	0,083

a : Ces estimations de variance sont obtenues de la simulation,

Tableau 4.3
Efficacités relatives des estimateurs et probabilités de couverture d'un intervalle de confiance à 95 % d'une moyenne de population. Les ensembles de données sont générés à partir d'une population exponentielle décalée discrète de paramètre d'échelle $\sigma = 1$ et de paramètre de décalage égal à 10

			Efficacités relatives, $R(\bar{X}_0) = \mathrm{Var}(\bar{X}_0)/\mathrm{Var}(\tilde{\mu}_2)$							Probabilités de couverture			
n	H	ρ	$R(\bar{X}_0)$	$R(\bar{X}_2)$	$R(\hat{\mu}_0)$	$R(\hat{\mu}_2)$	$R(\mu^*_0)$	$R(\mu^*_2)$	$R(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_2)$	$C^b(\hat{\mu}_0)$	$C^b(\hat{\mu}_2)$
10	2	1,00	1,770	1,663	1,495	1,394	1,193	1,297	1,103	0,838	0,833	0,894	0,950
15	3	1,00	2,472	2,239	1,757	1,538	1,222	1,446	1,027	0,855	0,842	0,905	0,931
20	4	1,00	3,839	3,349	2,353	1,889	1,406	1,879	1,212	0,871	0,884	0,915	0,931
25	5	1,00	4,639	3,892	2,503	1,792	1,235	1,958	1,182	0,865	0,881	0,916	0,915
10	2	0,75	1,900	1,690	1,941	1,690	1,667	1,520	1,128	0,839	0,857	0,898	0,949
15	3	0,75	2,708	2,440	2,626	2,233	2,132	1,815	1,117	0,859	0,870	0,914	0,947
20	4	0,75	3,484	2,996	3,059	2,704	2,430	2,103	1,104	0,869	0,871	0,922	0,938
25	5	0,75	4,758	4,127	4,156	3,298	3,106	2,402	1,245	0,866	0,877	0,913	0,932

a : Les probabilités de couverture sont calculées en partant d'un intervalle de confiance bootstrap par la méthode des centiles.
b : Les probabilités de couverture sont calculées en partant de $\hat{\mu}_r \pm t_{n-1,\,0,975}\hat{\sigma}_{\hat{\mu}_r}$, $r = 0, 1$.

Tableau 4.4
Estimation de la variance des estimateurs. Les ensembles de données sont générés à partir d'une population exponentielle décalée discrète de paramètre d'échelle $\sigma = 1$ et de paramètre de décalage égal à 10

			Estimations d'après les équations (2.5), (2.6), (2.8), (3.2)					Estimations d'après la simulation			
n	H	ρ	$\hat{\sigma}^2_{\hat{\mu}_0}$	$\hat{\sigma}^2_{\hat{\mu}_2}$	$\tilde{\sigma}^2_{\hat{\mu}_2}$	$\hat{\sigma}^2_{\tilde{\mu}_0}$	$\hat{\sigma}^2_{\tilde{\mu}_2}$	$V^a(\hat{\mu}_0)$	$V^a(\hat{\mu}_2)$	$V^a(\tilde{\mu}_0)$	$V^a(\tilde{\mu}_2)$
10	2	1,00	0,077	0,069	0,069	0,046	0,042	0,075	0,070	0,055	0,050
15	3	1,00	0,042	0,036	0,036	0,022	0,020	0,042	0,037	0,025	0,024
20	4	1,00	0,027	0,022	0,022	0,013	0,012	0,027	0,022	0,014	0,012
25	5	1,00	0,019	0,014	0,014	0,009	0,007	0,019	0,014	0,009	0,008
10	2	0,75	0,089	0,083	0,083	0,046	0,043	0,090	0,078	0,052	0,046
15	3	0,75	0,055	0,051	0,051	0,022	0,020	0,057	0,048	0,024	0,022
20	4	0,75	0,039	0,033	0,033	0,013	0,011	0,039	0,034	0,014	0,013
25	5	0,75	0,031	0,025	0,025	0,009	0,008	0,032	0,025	0,009	0,008

a : Ces estimations de variance sont obtenues d'après la simulation.

5 Exemple

À la présente section, nous appliquons les estimateurs proposés pour estimer la production de maïs en Ohio en se basant sur les données du Recensement de 2012 de l'*United States Department of Agriculture* (USDA). La population est constituée de $N = 87$ comtés de l'Ohio (un des comtés est exclu de la population car les données de recensement ne contiennent aucune entrée pour celui-ci). La variable d'intérêt est la production totale de maïs (X) exprimée en boisseaux. Nous utilisons la production de maïs selon le recensement de l'USDA de 2007 (Y) comme variable auxiliaire. La moyenne et l'écart-type de la production de maïs en 2012 sont $\mu_X = 5\,021\,061$ et $\sigma_X = 3\,983\,560$ boisseaux, respectivement. Le coefficient de corrélation entre X et Y est de 0,963. En utilisant cette population, nous avons exécuté une autre étude en simulation pour estimer la production de maïs et avons construit les intervalles de confiance pour la moyenne de population. Des échantillons ont été générés pour les combinaisons de tailles d'échantillon et d'ensemble $(n, H) = (10, 2), (15, 3), (20, 4)$. Les tailles de simulation et des répliques bootstrap étaient fixées à 3 000 et 200, respectivement. Les estimateurs Rao-Blackwellisés sont calculés en se basant sur 50 répliques.

Les efficacités relatives des estimateurs par rapport à $\tilde{\mu}_2$ et les probabilités de couverture des intervalles de confiance sont données au tableau 5.1. Celui-ci indique que les estimateurs sous le plan 2 Rao-Blackwellisé surpassent en performance tous les autres estimateurs pris en considération. Les probabilités de couverture semblent être légèrement plus faibles que le niveau nominal de 0,95.

Tableau 5.1
Efficacités relatives des estimateurs et probabilités de couverture d'un intervalle de confiance à 95 % de la moyenne de population. La population est constituée de 87 comtés de l'Ohio. La variable d'intérêt est la production de maïs (X) en 2012. La variable auxiliaire est la production de maïs (Y) en 2007, $\mu_X = 5\,021\,061$, $\sigma_X = 3\,983\,560$, $\text{cor}(X, Y) = 0,963$ et $N = 87$

		Efficacités relatives, $R(\bar{X}_0) = \text{Var}(\bar{X}_0)/\text{Var}(\tilde{\mu}_2)$							Probabilités de couverture			
n	H	$R(\bar{X}_0)$	$R(\bar{X}_2)$	$R(\hat{\mu}_0)$	$R(\hat{\mu}_2)$	$R(\mu_0^*)$	$R(\mu_2^*)$	$R(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_0)$	$C^a(\tilde{\mu}_2)$	$C^b(\hat{\mu}_0)$	$C^b(\hat{\mu}_2)$
10	2	2,301	1,981	1,829	1,448	1,468	1,280	1,181	0,883	0,896	0,924	0,925
15	3	3,745	3,188	2,353	1,612	1,994	1,454	1,200	0,907	0,919	0,940	0,907
20	4	5,707	4,402	2,901	1,624	2,476	1,143	1,341	0,920	0,920	0,946	0,873

a : Les probabilités de couverture sont calculées en partant d'un intervalle de confiance bootstrap par la méthode des centiles.
b : Les probabilités de couverture sont calculées en partant de $\hat{\mu}_r \pm t_{n-1,\,0,975}\hat{\sigma}_{\hat{\mu}_r}$, $r = 0, 2$.

Le tableau 5.2 donne les estimations de l'écart-type des estimateurs de la moyenne de population d'après les simulations et d'après l'expression analytique dans les équations (2.5), (2.6), (2.8), (3.2). Il est de nouveau évident que les estimations des erreurs-types sont raisonnablement proches des estimations d'après les simulations. Les estimations des écarts-types des estimateurs du total de population sont obtenues en multipliant les entrées du tableau 5.2 par la taille de population $N = 87$.

Tableau 5.2
Estimation de l'écart-type des estimateurs d'après le recensement de l'USDA de 2012. La population est constituée de 87 comtés de l'Ohio. La variable d'intérêt est la production de maïs (X) en 2012. La variable auxiliaire est la production de maïs (Y) en 2007, $\mu_X = 5\,021\,061$, $\sigma_X = 3\,983\,560$, $\mathrm{cor}(X,Y) = 0{,}963$ et $N = 87$

		Estimations d'après les équations (2.5), (2.6), (2.8), (3.2)					Estimations d'après la simulation			
n	H	$\hat{\sigma}_{\hat{\mu}_0}$	$\hat{\sigma}_{\hat{\mu}_2}$	$\tilde{\sigma}_{\tilde{\mu}_2}$	$\hat{\sigma}_{\tilde{\mu}_0}$	$\hat{\sigma}_{\tilde{\mu}_2}$	$\sqrt{V^a(\hat{\mu}_0)}$	$\sqrt{V^a(\hat{\mu}_2)}$	$\sqrt{V^a(\tilde{\mu}_0)}$	$\sqrt{V^a(\tilde{\mu}_2)}$
10	2	1 108 818,7	1 027 289,0	1 027 717,4	883 847,8	833 711,4	1 156 300,5	1 028 629,9	929 090,9	854 940,3
15	3	815 371,3	687 605,0	689 118,9	602 682,4	545 000,2	810 156,1	670 521,9	578 608,4	528 146,5
20	4	652 734,4	472 231,5	477 888,6	454 368,3	392 990,3	638 755,1	478 007,6	434 365,0	375 040,7

a : Ces estimations de variance sont obtenues d'après la simulation.

6 Conclusion

Nous avons élaboré deux plans d'échantillonnage pour l'obtention d'échantillons poststratifiés par choix raisonné en population finie. Les plans présentent deux niveaux de modalités de tirage sans remise, à savoir le plan 0 et le plan 2. Sous le plan 0, toutes les unités mesurées et non mesurées dans un ensemble sont remises dans la population avant de sélectionner les unités suivantes. Sous le plan 2, une modalité de tirage sans remise est appliquée à toutes les unités, qu'elles soient mesurées ou non. Donc, sous le plan 0, les variables aléatoires sont indépendantes, mais sous le plan 2, elles sont négativement corrélées. Dans ces plans, les observations mesurées sont classées et stratifiées en H classes créées par choix raisonné après tirage d'un échantillon aléatoire simple. À l'aide de l'information sur le classement par rangs, nous construisons des estimateurs sans biais de la moyenne de population et de la variance de ces estimateurs sans biais de la moyenne de population. Nous avons démontré que les nouveaux estimateurs basés sur le plan 2 donnent de meilleurs résultats que les estimateurs de la moyenne d'un échantillon aléatoire simple, de la moyenne d'un échantillon PCR sous le plan 0 et de la moyenne d'un échantillon d'ensembles ordonnés. Le présent article porte principalement sur l'estimation de la moyenne de population, mais les résultats s'appliquent également à l'estimation du total de population avec un petit ajustement de la notation.

La poststratification crée des rangs aléatoires et un nombre aléatoire d'observations dans les classes créées par choix raisonné. En conditionnant sur les valeurs mesurées dans l'échantillon, nous construisons les estimateurs Rao-Blackwellisés en calculant la valeur espérée conditionnelle des estimateurs sur toutes les valeurs possibles des rangs aléatoires. Les estimateurs Rao-Blackwellisés sont sans biais et plus efficaces que les estimateurs non conditionnels. Nous construisons une inférence bootstrap en échantillon fini pour la moyenne de population fondée sur tous les estimateurs proposés. Les nouveaux plans d'échantillonnage et estimateurs sont appliqués aux données du recensement de l'USDA de 2012 pour montrer qu'il s'agit d'estimateurs et de plans d'échantillonnage viables dans les études par sondage.

Dans l'un de nos projets courants, nous étendons le plan 0 et le plan 2 à l'échantillonnage à deux degrés où les unités d'échantillonnage primaires et secondaires constituent deux populations finies. Dans ce cas, nous nous attendons à ce que se pose un problème d'optimisation intéressant lié à la sélection des tailles d'échantillon, ainsi que du plan 0 et du plan 2 dans l'échantillonnage de degré I et de degré II.

Annexe

Preuve du lemme 1 : La preuve de la partie (i) est donnée dans Presnell et Bohn (1999) dans des conditions de population infinie. La preuve est essentiellement la même dans des conditions de population finie.

Pour prouver la partie (ii), nous considérons la fonction de masse de probabilité conjointe de X_j et X_t sachant leurs rangs déterminés par choix raisonné $R_j = h$ et $R_t = h'$

$$P\left(X_j = x, X_t = y \,|\, R_j = h, R_t = h'\right) = P\left(X_{[h]} = x, X_{[h']} = y\right);\ x \neq y, (x,y) \in \mathcal{P}.$$

Pour $h = 1,\ldots,H,$ soit $J_h = i$ l'événement pour lequel la i^{e} statistique d'ordre dans un ensemble de taille H est jugée être la h^{e} statistique d'ordre par choix raisonné. En nous servant de cette notation, nous écrivons

$$P\left(X_{[h]} = x, X_{[h']} = y\right) = \sum_{i=1}^{H}\sum_{k=1}^{H} P\left(X_{(i)} = x, X_{(k)} = y, J_h = i, J_{h'} = k\right);\ x \neq y, (x,y) \in \mathcal{P}. \qquad \text{(A.1)}$$

Pour chaque $i,$ les événements $\{J_h = i\};\ h = 1,\ldots,H,$ partitionnent l'espace d'échantillon de la variable aléatoire R_j parce que la i^{e} statistique d'ordre est assignée à un, et uniquement un, rang par choix raisonné. Puisque le rang R_t est obtenu indépendamment à partir d'un autre ensemble disjoint en utilisant la même procédure de classement, les événements $\{J_{h'} = k\};\ h' = 1,\ldots,H,$ partitionnent aussi l'espace d'échantillon. Donc, nous écrivons

$$\sum_{h=1}^{H}\sum_{h'=1}^{H} P\left(X_{(i)} = x, X_{(k)} = y, J_h = i, J_{h'} = k\right) = P\left(X_{(i)} = x, X_{(k)} = y\right);\ x \neq y, (x,y) \in \mathcal{P}. \qquad \text{(A.2)}$$

En combinant les équations (A.1) et (A.2), nous obtenons

$$\begin{aligned}\sum_{h=1}^{H}\sum_{h'=1}^{H} P\left(X_{[h]} = x, X_{[h']} = y\right) &= \sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{i=1}^{H}\sum_{k=1}^{H} P\left(X_{(i)} = x, X_{(k)} = y;\ J_h = i, J_{h'} = k\right) \\ &= \sum_{i=1}^{H}\sum_{k=1}^{H} P\left(X_{(i)} = x, X_{(k)} = y\right);\ x \neq y, (x,y) \in \mathcal{P}.\end{aligned}$$

Cela achève la preuve.

Preuve du théorème 1 :

i) Nous utilisons l'espérance conditionnelle pour écrire

$$E\left(\hat{\mu}_r\right) = \sum_{h=1}^{H} E\left(\frac{I_h}{d_n M_h}\right)\sum_{i=1}^{n} E\left(X_i I\left(R_i = h\right)\,|\,R_i = h\right) = \sum_{h=1}^{H} E\left(\frac{I_h M_h}{d_n M_h}\right) EX_{[h]}.$$

Les variables aléatoires $I_h / d_n;\ h = 1, \ldots, H,$ sont de même loi. En utilisant la partie (i) du lemme 3, l'espérance susmentionnée peut s'écrire

$$E(\hat{\mu}_r) = E\left(\frac{I_1}{d_n}\right)\sum_{h=1}^{n} EX_{[h]} = \frac{1}{H}\sum_{h=1}^{H}\mu_{[h]} = \frac{1}{H}\sum_{h=1}^{H}\mu_{(h)} = \mu, r = 0, 2,$$

ce qui achève la preuve. Les deux dernières égalités dans l'équation qui précède découlent de la partie (i) du lemme 1.

ii) Pour la preuve de $\mathrm{Var}(\hat{\mu}_0),$ nous utilisons la variance conditionnelle

$$\mathrm{Var}(\hat{\mu}_0) = E\{\mathrm{Var}(\hat{\mu}_0 \,|\mathbf{R})\} + \mathrm{Var}\{E(\hat{\mu}_0 \,|\mathbf{R})\}.$$

Puisque les X_j sont sélectionnés avec remise, ils sont tous indépendants. Donc, le premier terme dans l'équation ci-dessus donne

$$E\{\mathrm{Var}(\hat{\mu}_0 \,|\mathbf{R})\} = \sum_{h=1}^{H} E\left(\frac{I_h^2 M_h}{d_n^2 M_h^2}\right)\sigma_{[h]}^2 = E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h=1}^{H}\sigma_{[h]}^2.$$

Nous considérons maintenant

$$\begin{aligned}
\mathrm{Var}\{E(\hat{\mu}_0 \,|\mathbf{R})\} &= \sum_{h=1}^{H}\mathrm{Var}\left(\frac{I_h}{d_n}\right)\mu_{[h]}^2 + \sum_{h=1}^{H}\sum_{h'\neq h}^{H}\mathrm{Cov}\left(\frac{I_h}{d_n}, \frac{I_{h'}}{d_n}\right)\mu_{[h]}\mu_{[h']} \\
&= \mathrm{Var}\left(\frac{I_1}{d_n}\right)\sum_{h=1}^{H}\mu_{[h]}^2 + \mathrm{Cov}\left(\frac{I_1}{d_n}, \frac{I_2}{d_n}\right)\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\mu_{[h]}\mu_{[h']} \\
&= \frac{H}{H-1}\mathrm{Var}\left(\frac{I_1}{d_n}\right)\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2
\end{aligned}$$

ce qui achève la preuve de $\mathrm{Var}(\hat{\mu}_0).$ La dernière égalité est obtenue en utilisant la partie (iv) du lemme 3.

Pour la preuve de $\mathrm{Var}(\hat{\mu}_2),$ nous considérons de nouveau la variance conditionnelle de $\hat{\mu}_2$ sachant les rangs

$$\mathrm{Var}(\hat{\mu}_2) = E(\mathrm{Var}(\hat{\mu}_2 \,|\mathbf{R})) + \mathrm{Var}(E(\hat{\mu}_2 \,|\mathbf{R})).$$

Le deuxième terme du deuxième membre de l'équation ci-dessus peut s'écrire

$$\begin{aligned}
\mathrm{Var}(E(\hat{\mu}_2 \,|\mathbf{R})) &= \mathrm{Var}\left(\sum_{h=1}^{h}\frac{\mu_{[h]} I_h}{d_n}\right) \\
&= \sum_{h=1}^{H}\mathrm{Var}\left(\frac{I_h}{d_n}\right)\mu_{[h]}^2 + \sum_{h=1}^{H}\sum_{h'\neq h}^{H}\mathrm{Cov}\left(\frac{I_h}{d_n}, \frac{I_{h'}}{d_n}\right)\mu_{[h]}\mu_{[h']} \\
&= \mathrm{Var}\left(\frac{I_1}{d_n}\right)\sum_{h=1}^{H}\mu_{[h]}^2 + \mathrm{Cov}\left(\frac{I_1}{d_n}, \frac{I_2}{d_n}\right)\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\mu_{[h]}\mu_{[h']}.
\end{aligned}$$

Notons que $\mathrm{Cov}(I_1/d_n, I_2/d_n) = -\mathrm{Var}(I_1/d_n)/(H-1)$ dans la partie (iv) du lemme 3. En utilisant cette égalité dans l'équation qui précède, nous obtenons

$$\mathrm{Var}(E(\hat{\mu}_2|\mathbf{R})) = \frac{H}{H-1}\mathrm{Var}\left(\frac{I_1}{d_n}\right)\sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2. \tag{A.3}$$

Nous considérons maintenant $\mathrm{Var}(\hat{\mu}_2|\mathbf{R})$

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_2|\mathbf{R}) &= \mathrm{Var}\left\{\sum_{h=1}^{H}\frac{I_h}{d_n M_h}\sum_{j=1}^{n}X_j I(R_j=h)\Big|\mathbf{R}\right\}\\ &= \sum_{h=1}^{H}\frac{I_h^2}{d_n^2 M_h^2}\left(M_h\sigma_{[h]}^2 + M_h(M_h-1)\sigma_{[h,h]}\right) + \sum_{h}\sum_{h'\neq h}^{H}\frac{I_h I_{h'}}{d_n^2 M_h M_{h'}}M_h M_{h'}\sigma_{[h,h']}\\ &= \sum_{h=1}^{H}\frac{I_h^2}{d_n^2 M_h}\sigma_{[h]}^2 + \sum_{h=1}^{H}\frac{I_h^2(M_h-1)}{d_n^2 M_h}\sigma_{[h,h]} + \sum_{h}\sum_{h'\neq h}^{H}\frac{I_h I_{h'}}{d_n^2}\sigma_{[h,h']}.\end{aligned}$$

Il est clair que $I_h/(d_n M_h)$ et $I_h I_{h'}/d_n^2$, $h=1,\ldots,H$, sont de même loi. En utilisant les égalités ci-dessous

$$E\left(\frac{(M_1-1)I_1^2}{d_n^2 M_1}\right) = E\left(\frac{I_1^2}{d_n^2}\right) - E\left(\frac{I_1^2}{d_n^2 M_1}\right)$$

$$E\left(\frac{I_1 I_2}{d_n^2}\right) = \frac{1}{H-1}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)$$

nous écrivons

$$\begin{aligned}E\{\mathrm{Var}(\hat{\mu}_2|\mathbf{R})\} &= E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h=1}^{H}\sigma_{[h]}^2 + E\left(\frac{I_1^2}{d_n^2}\right)\sum_{h=1}^{H}\sigma_{[h,h]} - E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h=1}^{H}\sigma_{[h,h]}\\ &\quad + \frac{1}{H-1}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\sigma_{[h,h']}\\ &= E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h=1}^{H}\sigma_{[h]}^2 + \frac{1}{H-1}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)\sum_{h=1}^{H}\sum_{h=1}^{H}\sigma_{[h,h']}\\ &\quad + \left\{E\left(\frac{I_1^2}{d_n^2}\right) - E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{1}{H-1}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)\right\}\sum_{h=1}^{H}\sigma_{[h,h]}.\end{aligned} \tag{A.4}$$

Nous montrons maintenant que $\sum_{h=1}^{H}\sum_{h'=1}^{H}\sigma_{[h,h']} = H^2\sigma^2/(N-1)$. En utilisant la partie (ii) du lemme 1, nous observons d'abord que

$$\sum_{h=1}^{H}\sum_{h'=1}^{H} E\left(X_{[h]}X_{[h']}\right) = \sum_{h=1}^{H}\sum_{h'=1}^{H} E\left(X_{(h)}X_{(h')}\right)$$

$$= 2\sum_{i=1}\sum_{i<j} x_i x_j \sum_{h=1}\sum_{h'=1} \beta\left(i,j,h,h'\right)$$

$$= 2\sum_{i=1}\sum_{j>i} x_i x_j \left\{ \sum_{h=1}^{H}\sum_{\lambda=0}^{j-i-1} \frac{\binom{i-1}{h-1}\binom{j-i-1}{\lambda}\binom{N-j}{H-\lambda-h}}{\binom{N}{H}} \right.$$

$$\left. \times \sum_{h'=1}^{H} \frac{\binom{j-1-h-\lambda}{h'-1}\binom{N-j-H+\lambda+h}{H-h'}}{\binom{N-H}{H}} \right\}.$$

Dans la dernière somme de l'équation susmentionnée, nous posons que $y = h' - 1$. Après certaines simplifications, nous écrivons

$$\sum_{h=1}^{H}\sum_{h'=1}^{H} E\left(X_{[h]}X_{[h']}\right) = 2\sum_{i=1}\sum_{i<j} x_i x_j \sum_{h=1}^{H}\sum_{\lambda=0}^{j-i-1} \frac{\binom{i-1}{h-1}\binom{j-i-1}{\lambda}\binom{N-j}{H-\lambda-h}}{\binom{N}{H}} \frac{H}{N-H}$$

$$= \frac{H}{N-H}\sum_{i=1}\sum_{j\neq i} x_i x_j \sum_{h=1}^{H} \frac{\binom{i-1}{h-1}\binom{N-i-1}{H-h}}{\binom{N}{H}}.$$

Soit $z = h - 1$. L'expression ci-dessus se réduit à

$$\sum_{h=1}^{H}\sum_{h'=1}^{H} E\left(X_{[h]}X_{[h']}\right) = \frac{H}{N-H}\sum_{i=1}\sum_{j\neq i} x_i x_j \sum_{z=0}^{H-1} \frac{\binom{i-1}{z}\binom{N-i-1}{H-1-z}}{\frac{N(N-1)}{H(H-N)}\binom{N-1}{H-1}}$$

$$= \frac{H^2}{N(N-1)}\sum_{i=1}\sum_{j\neq i} x_i x_j = \frac{H^2}{N(N-1)}\left\{\left(\sum_{i=1}^{N} x_i\right)^2 - \sum_{i}^{N} x_i^2\right\}.$$

En utilisant l'équation qui précède, nous concluons que

$$\sum_{h=1}^{H}\sum_{h'=1}^{H}\sigma_{[h,h']} = \sum_{h=1}^{H}\sum_{h'=1}^{H}E\left(X_{[h]}X_{[h']}\right) - \sum_{h=1}^{H}\sum_{h'=1}^{H}E\left(X_{[h]}EX_{[h']}\right)$$

$$= \frac{H^2}{N(N-1)}\left\{\left(\sum_{i=1}^{N}x_i\right)^2 - \sum_{i}^{N}x_i^2\right\} - \frac{H^2}{N^2}\left(\sum_{i=1}^{N}x_i\right)^2$$

$$= -\frac{H^2}{N-1}\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 = \frac{-H^2}{N-1}\sigma^2. \qquad \text{(A.5)}$$

En insérant l'expression susmentionnée dans l'équation (A.4), nous écrivons

$$E\{\mathrm{Var}(\hat{\mu}_2)|\mathbf{R}\} = E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h=1}^{H}\sigma_{[h]}^2 - \frac{H^2\sigma^2}{(H-1)(N-1)}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)$$

$$+\left\{\frac{H}{H-1}E\left(\frac{I_1^2}{d_n^2}\right) - E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{1}{H(H-1)}\right\}\sum_{h=1}^{H}\sigma_{[h,h]}. \qquad \text{(A.6)}$$

Nous achevons la preuve en combinant les équations (A.3) et (A.6)

$$\mathrm{Var}(\hat{\mu}_2) = \frac{H}{H-1}\mathrm{Var}(I_1/d_n)\sum_{h=1}^{H}\left(\mu_{[h]}-\mu\right)^2 + E\left(\frac{I_1^2}{d_n^2 M_1}\right)\sum_{h=1}^{H}\sigma_{[h]}^2$$

$$-\frac{H^2\sigma^2}{(H-1)(N-1)}\left(\frac{1}{H} - E\left(\frac{I_1^2}{d_n^2}\right)\right)$$

$$+\left\{\frac{H}{H-1}E\left(\frac{I_1^2}{d_n^2}\right) - E\left(\frac{I_1^2}{d_n^2 M_1}\right) - \frac{1}{(H-1)H}\right\}\sum_{h=1}^{H}\sigma_{[h,h]}.$$

Preuve du corollaire 1 : Les preuves de (i) et (ii) sont très simples. Pour la preuve de (iii), nous réécrivons $\lim_{n\to\infty} n\sigma_{\hat{\mu}_2}^2$ sous la forme

$$\lim_{n\to\infty} n\sigma_{\hat{\mu}_2}^2 = \frac{1}{H}\left(\sum_{h=1}^{H}\sigma_{[h]}^2 - \sum_{h=1}^{H}\sigma_{[h,h]}\right) - f\sigma^2.$$

En nous servant de l'égalité (A.5), nous écrivons

$$-\frac{1}{H}\sum_{h=1}^{H}\sigma_{h,h} - f\sigma^2 \approx -\frac{NH}{NH^2}\sum_{h=1}^{H}\sigma_{[h,h]} + \frac{n(N-1)}{NH^2}\sum_{h=1}^{H}\sum_{h'=1}^{H}\sigma_{[h,h']}$$

$$= \sum_{h=1}^{H}\sigma_{[h,h]}\frac{N(n-H)-n}{NH^2} + \frac{n(N-1)}{NH^2}\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\sigma_{[h,h']}$$

$$\leq \frac{n(N-1)}{NH^2}\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\sigma_{[h,h']} \leq 0,$$

où $\approx$ est utilisé pour indiquer une égalité approximative puisque nous remplaçons la limite f par sa valeur finie pour une certaine valeur de n. Cette inégalité donne que

$$\lim_{n\to\infty} n\sigma^2_{\hat{\mu}_2} \le \frac{1}{H}\sum_{h=1}^{H}\sigma^2_{[h]} = \operatorname{var}\left(\sqrt{n}\,\hat{\mu}_0\right)$$

ce qui complète la preuve.

Preuve du théorème 2 : Nous commençons par examiner la valeur espérée de T_1 et T_2 sous le plan 0 et le plan 2. En utilisant l'espérance conditionnelle, la valeur espérée de T_1 se réduit à

$$E(T_1) = \frac{1}{E\left(\frac{I_1 I_2}{d_n^2}\right)}\sum_{h=1}^{H}\sum_{h\neq h'}^{H} E\left(\frac{I_h I_{h'} M_h M_{h'}}{M_h M_{h'} d_n^2}\right) E\left(X_{[h]} - X_{[h']}\right)^2$$

$$= \begin{cases} 2(H-1)\sum_{h=1}^{H}\sigma^2_{[h]} + 2H\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2 & \text{pour le plan 0} \\ 2(H-1)\sum_{h=1}^{H}\sigma^2_{[h]} + 2H\sum_{h=1}^{H}\left(\mu_{[h]} - \mu\right)^2 - 2\sum_{h=1}^{H}\sum_{h'\neq h}^{H}\sigma_{[h,h']} & \text{pour le plan 2.} \end{cases}$$

De façon similaire, on peut montrer que

$$E(T_2) = \sum_{h=1}^{H} E\left(\frac{HI_h^* M_h(M_h-1)}{M_h(Mh-1)d_n^*}\right) 2E\left\{\left(X_{[h]1} - \mu_{[h]}\right) - \left(\mu_{[h]} - X_{[h]2}\right)\right\}^2$$

$$= \begin{cases} 2HE\left(\frac{I_1^*}{d_n^*}\right)\sum_{h=1}^{H}\sigma^2_{[h]} & \text{pour le plan 0} \\ 2HE\left(\frac{I_1^*}{d_n^*}\right)\left\{\sigma^2_{[h]} - \sum_{h}^{H}\sigma_{[h,h]}\right\} & \text{pour le plan 2.} \end{cases}$$

Puisque I_h^*/d_n^*, $h = 1,\ldots,H$, sont de même loi, nous avons $E\left(I_1^*/d_n^*\right) = 1/H$. Il s'ensuit que $E(T_2) = 2\sum_{h=1}^{H}\sigma^2_{[h]}$ pour le plan 0 et $E(T_2) = 2\sum_{h=1}^{H}\sigma^2_{[h]} - 2\sum_{h}^{H}\sigma_{[h,h]}$ pour le plan 2.

Pour la preuve de $\bar{\sigma}^2$, nous commençons par observer que $E(T_1) + E(T_2) = 2H^2\sigma^2$ pour le plan 0 et $E(T_1) + E(T_2) = 2NH^2\sigma^2/(N-1)$ pour le plan 2. Les preuves découlent alors de ces égalités.

Nous achevons la preuve de l'absence de biais de $\hat{\sigma}^2_{\hat{\mu}_0}$ et $\hat{\sigma}^2_{\hat{\mu}_2}$ et $\bar{\sigma}^2_{\hat{\mu}_2}$ en insérant $E(T_1)$ et $E(T_2)$ dans les équations (2.5), (2.6) et (2.7).

Les preuves pour les estimateurs EEO étant similaires, elles sont omises ici.

Bibliographie

Ahn, S., Lim, J. et Wang, X. (2014). The student's t approximation to distributions of pivotal statistics from ranked set samples. *Journal of Korean Statistical Society*, 43, 643-652.

Al-Saleh, M.F., et Samawi, H.M. (2007). A note on inclusion probability in ranked set sampling and some of its variations. *Test*, 16, 198-209.

Dastbaravarde, A., Arghami, N.R. et Sarmad, M. (2016). Some theoretical results concerning non-parametric estimation by using a judgment poststratication sample. *Communications in Statistics - Theory and Methods*, 45, 8, 2181-2203.

Deshpande, J.V., Frey, J. et Ozturk, O. (2006). Nonparametric ranked-set sampling confidence intervals for a finite population. *Environmental and Ecological Statistics*, 13, 25-40.

Frey, J. (2011). Recursive computation of inclusion probabilities in ranked set sampling. *Journal of Statistical Planning and Inference*, 141, 3632-3639.

Frey, J., et Feeman, T.G. (2012). An improved mean estimator for judgment post-stratification. *Computational Statistics and Data Analysis*, 56, 418-426.

Frey, J., et Feeman, T.G. (2013). Variance estimation using judgment post-stratification. *Annals of the Institute of Statistical Mathematics*, 65, 551-569.

Frey, J., et Ozturk, O. (2011). Constrained estimation using judgment post-stratification. *Annals of the Institute of Statistical Mathematics*, 63, 769-789.

Gokpinar, F., et Ozdemir, Y.A. (2010). Generalization of inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 39, 89-95.

Jafari Jozani, M., et Johnson, B.C. (2011). Design based estimation for ranked set sampling in finite population. *Environmental and Ecological Statistics*, 18, 663-685.

Jafari Jozani, M., et Johnson, B.C. (2012). Randomized nomination sampling in finite populations. *Journal of Statistical Planning and Inference*, 142, 2103-2115.

MacEachern, S.N., Stasny, E.A. et Wolfe, D.A. (2004). Judgment post-stratification with imprecise ranking. *Biometrics*, 60, 207-215.

McIntyre, G.A. (1952). A method for unbiased selective sampling using ranked-sets. *Australian Journal of Agricultural Research*, 3, 385-390.

McIntyre, G.A. (2005). A method of unbiased selective sampling using ranked-sets. *The American Statistician*, 59, 230-232.

Ozdemir, Y.A., et Gokpinar, F. (2007). A generalized formula for inclusion probabilities in ranked set sampling. *Hacettepe Journal of Mathematics and Statistics*, 36, 89-99.

Ozdemir, Y.A., et Gokpinar, F. (2008). A new formula for inclusion probablities in median ranked set sampling. *Communications in Statisitcs - Theory and Methods*, 37, 2022-2033.

Ozturk, O. (2013). Combining multi-obsever information in partially rank-ordered judgment post-stratifed and ranked set samples. *The Canadian Journal of Statistics*, 41, 304-324.

Ozturk, O. (2014a). Estimation of population mean and total in finite population setting using multiple auxiliary variables. *Journal of Agricultural, Biological and Environmental Statistics*, 19, 161-184.

Ozturk, O. (2014b). Statistical inference for population quantiles and variance in judgment post-stratified samples. *Computational Statistics and Data Analysis*, 77, 188-205.

Ozturk, O. (2015). Distribution free two-sample methods for judgment-post stratified data. *Statistica Sinica*, 25, 1691-1712.

Ozturk, O. (2016). Estimation of a finite population mean and total using population ranks of sample units. *Journal of Agricultural, Biological and Environmental Statistics*, 21, 1, 181-202.

Ozturk, O., et Jafari Jozani, M. (2013). Inclusion probabilities in partially rank ordered set sampling. *Computational Statistics and Data anlysis*, 69, 122-132.

Patil, G.P., Sinha, A.K. et Taillie, C. (1995). Finite population correction for ranked set sampling. *Annals of the Institute of Statistical Mathematics*, 47, 621-636.

Presnell, B., et Bohn, L.L. (1999). *U-Statistics and Imperfect Ranking in Ranked Set Sampling*, 10, 111-126.

Stokes, S.L., Wang, X. et Chen, M. (2007). Judgment post stratification with multiple rankers. *Journal of Statistical Theory and Applications*, 6, 344-359.

Wang, X., Lim, J. et Stokes, S.L. (2008). A nonparametric mean estimator for judgment post-stratified data. *Biometrics*, 64, 355-363.

Wang, X., Wang, K. et Lim, J. (2012). Isotonized CDF estimation from judgment post-stratification data with empty strata. *Biometrics*, 68, 194-202.

Wang, X., Stokes, L., Lim, J. et Chen, M. (2006). Concomitant of multivariate order statistics with application to judgment poststratification. *Journal of American Statistical Association*, 101(476), 1693-1704.

Wolfe, D.A. (2012). Ranked set sampling: Its relevance and impact on statistical inference. *ISRN Probability and Statistics*, doi:10.5402/2012/568385.