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

Techniques d’enquête

Améliorer la couverture des intervalles de confiance au niveau des degrés de liberté : application au recensement canadien

par Marie-Hélène Toupin et Vincent Martin

Date de diffusion : le 29 juin 2026

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 infostats@statcan.gc.ca

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

- Service de renseignements statistiques 1-800-263-1136
- Service national d'appareils de télécommunications pour les malentendants 1-800-363-7629
- Télécopieur 1-514-283-9350

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.

Techniques d'enquête, juin 2026
Vol. 52, nº 1, pp. 209-233
Statistique Canada, nº 12-001-X au catalogue

Améliorer la couverture des intervalles de confiance au niveau des degrés de liberté : application au recensement canadien

Marie-Hélène Toupin et Vincent Martin[1]

Résumé

La construction d'intervalles de confiance s'appuie très souvent sur une loi de probabilité qui utilise un certain nombre de degrés de liberté comme paramètre. C'est le cas des intervalles de confiance de Student et de Wilson modifié, dont il sera question dans cet article, qui font appel aux quantiles de la distribution de Student dont le nombre de degrés de liberté est généralement inconnu. Pour que la longueur d'un intervalle de confiance soit représentative de la fiabilité d'une estimation, il est essentiel que le taux de couverture réel corresponde au taux nominal. À cette fin, le nombre de degrés de liberté de la distribution de probabilité utilisé en pratique pour calculer l'intervalle de confiance doit être estimé avec la plus grande précision possible. Une règle approximative est souvent employée, bien qu'elle tende à surestimer le nombre réel de degrés de liberté. Dans cet article, une version plus précise des degrés de liberté, issue de l'approximation de Satterthwaite, est obtenue dans le contexte du recensement canadien de la population. Le plan de sondage est équivalent à un plan aléatoire simple sans remise, stratifié en grappes, et la méthode d'estimation de la variance est une adaptation de la méthode des demi-échantillons équilibrés. Une expression explicite des degrés de liberté est obtenue dans ces conditions, permettant d'identifier les facteurs qui les influencent. À titre de comparaison, la formule des degrés de liberté est également établie pour l'estimateur classique de la variance. Une étude par simulations montre que l'utilisation de cette version des degrés de liberté permet de corriger le problème de sous-couverture observé avec la règle approximative, soulignant l'importance de l'évaluation précise de ce nombre.

Mots-clés : Couverture; degrés de liberté; estimation pour un petit domaine; intervalles de confiance; méthode des demi-échantillons équilibrés; recensement canadien de la population.

1. Introduction

Au Canada, le recensement de la population a lieu tous les cinq ans. Le recensement comprend deux questionnaires : un court, auquel toute la population répond, et un détaillé, adressé à un échantillon de ménages, qui porte sur des sujets plus précis tels que les activités quotidiennes, les renseignements socioculturels, la scolarité et les activités sur le marché du travail. Lors du Recensement de 2021, un quart des ménages a été sélectionné aléatoirement pour remplir la version détaillée du questionnaire.

Pour la diffusion du Recensement de 2021, Statistique Canada a utilisé des intervalles de confiance de niveau 95 % comme indicateurs de la fiabilité des estimations issues du questionnaire détaillé. Afin que la fiabilité soit fidèlement reflétée par ces intervalles, il est crucial que le niveau de confiance nominal soit respecté. Cela signifie que le taux de couverture réel des intervalles de confiance devrait être égal ou supérieur au niveau de confiance choisi.

Le respect de ce niveau de confiance nominal dépend notamment du type d'intervalle de confiance utilisé, qui doit être adapté à la nature du paramètre estimé. Dans cet article, deux types d'intervalles de confiance sont étudiés. Premièrement, l'intervalle de confiance de Student, développé pour des variables

1. Marie-Hélène Toupin, Méthodologiste principale, Statistique Canada, Édifice R.H. Coats, 100 promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6. Courriel : marie-helene.toupin@statcan.gc.ca; Vincent Martin, Méthodologiste principal, Statistique Canada, Édifice R.H. Coats, 100 promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6. Courriel : vincent.martin@statcan.gc.ca.

d'intérêt continues, peut s'avérer inadapté pour certains types de statistiques, notamment les proportions lorsqu'elles sont très faibles ou très élevées. Deuxièmement, l'intervalle de confiance de Wilson modifié, moins connu, a été spécialement conçu pour l'estimation de proportions ou d'effectifs. Ce type d'intervalle a été employé pour la diffusion des estimations d'effectifs issues du questionnaire détaillé du Recensement canadien de la population de 2021. Dans les deux cas, le calcul de l'intervalle de confiance fait intervenir un paramètre généralement inconnu : le nombre de degrés de liberté.

La méthode d'estimation de la variance et les degrés de liberté influencent tous deux la couverture des intervalles de confiance. Cet article se concentre spécifiquement sur l'impact des degrés de liberté. Heeringa, West et Berglund (2010, traduction libre) expliquent le rôle joué par les degrés de liberté : « Les distributions de probabilité telles que Student, khi-carré et Fisher jouent un rôle critique dans la construction d'intervalles de confiance pour des valeurs de la population ou comme distributions de référence pour les tests d'hypothèses formels concernant des paramètres de la population. Parmi les quantités qui influencent la forme de ces distributions se trouvent les degrés de liberté. Les degrés de liberté sont les indices du niveau de précision avec lequel la vraie variance de la distribution de référence a été estimée à partir du plan d'échantillonnage. Les plans d'échantillonnage avec un grand nombre de degrés de liberté pour l'estimation de la variance permettent une estimation plus précise de la vraie variance de la distribution de référence. […] Déterminer précisément les degrés de liberté pour une estimation de la variance sous des plans d'échantillonnage complexes tels qu'utilisés en pratique est difficile. »

Une pratique courante pour contourner le fait que le nombre de degrés de liberté est inconnu consiste à supposer que ce nombre est très grand. Dans ce cas, le quantile de la distribution de Student est remplacé par le quantile de la loi normale standardisée. Une autre technique couramment employée consiste à déterminer le nombre de degrés de liberté selon une règle approximative (ou règle du pouce). Par exemple, dans un plan stratifié, une règle fréquemment utilisée est le nombre d'unités primaires d'échantillonnage moins le nombre de strates. Ces deux pratiques tendent à surestimer le nombre de degrés de liberté, ce qui peut entraîner une sous-couverture significative.

L'approche proposée dans cet article consiste à utiliser l'approximation de Satterthwaite (1946) afin d'obtenir une estimation des degrés de liberté plus précise que celle obtenue avec la règle approximative. La formule générale, valable pour un plan de sondage et un estimateur de variance quelconques, est rappelée. À partir de celle-ci, des expressions spécifiques sont dérivées pour le questionnaire détaillé du recensement canadien, fondé sur un plan stratifié avec tirage aléatoire simple sans remise de grappes. Les formules, obtenues pour l'estimation d'un total sur un domaine, se généralisent à tout estimateur linéaire.

Plus l'estimateur de la variance est imprécis, plus le nombre de degrés de liberté sera petit. Dans cet article, deux méthodes d'estimation de la variance seront examinées, à savoir l'estimateur de la variance classique et une variante de la méthode par rééchantillonnage des demi-échantillons équilibrés décrite par Devin et Verret (2016).

La formule spécifique résultante de l'approximation de Satterthwaite offre plusieurs avantages. Premièrement, l'expression algébrique obtenue permet une compréhension plus approfondie des facteurs qui influencent les degrés de liberté. Deuxièmement, une étude de simulations montre qu'elle permet de corriger les problèmes de sous-couverture observés avec la règle approximative pour de petits domaines. Cela revêt un intérêt particulier, car en s'efforçant de répondre au mieux aux besoins des utilisateurs, Statistique Canada vise à diffuser de l'information pour des domaines aussi petits que possible.

Cet article est structuré comme suit. La section 2 présente en détail l'approximation de Satterthwaite, qui fournit une formule générale pour le calcul des degrés de liberté. La notation employée est introduite à la section 3. La section 4 définit les intervalles de confiance de Student et de Wilson modifié. La méthode d'estimation de la variance utilisée pour le recensement est décrite à la section 5. La section 6 établit la formule théorique du nombre de degrés de liberté pour deux méthodes spécifiques d'estimation de la variance, et propose un estimateur de ce nombre. Une étude de simulations complète est ensuite présentée à la section 7. Enfin, la section 8 conclut l'article.

2. Approximation de Satterthwaite

Dans cette section, une formule pour les degrés de liberté dans un contexte général est présentée. Cette formule a été initialement dérivée par Satterthwaite (1946) et proposée par Rust (1986) comme moyen d'obtenir une expression pour les degrés de liberté. Plus de détails concernant l'approximation de Satterthwaite sont fournis dans Valliant et Rust (2010) et Kott (2020).

Soient $\hat{\theta}$ un estimateur sans biais d'un paramètre de la population θ et $v(\hat{\theta})$ un certain estimateur de la variance $\text{Var}_P(\hat{\theta})$ sous le plan de sondage P. Supposons que $v(\hat{\theta})$ dépend également d'un certain processus aléatoire noté *. Par exemple, pour une méthode d'estimation de la variance par répliques, * représenterait alors le processus de création des répliques. Soit $\text{Var}_{P,*}(\hat{\theta})$ la variance de $\hat{\theta}$ sous le plan de sondage P et sous le processus aléatoire *.

Soit $v(\hat{\theta})$, un estimateur sans biais de la variance $\text{Var}_P(\hat{\theta})$. Si $dl \times v(\hat{\theta}) / \text{Var}_P(\hat{\theta})$ est de distribution khi-carré, alors le nombre de degrés de liberté associé à $v(\hat{\theta})$ est donné par

$$dl = \frac{2\left\{\text{Var}_P(\hat{\theta})\right\}^2}{\text{Var}_{P,*}\left\{v(\hat{\theta})\right\}}. \tag{2.1}$$

L'approximation de Satterthwaite, décrite à l'équation (2.1), constitue une expression générique, applicable à divers plans de sondage, estimateurs et méthodes d'estimation de la variance. L'hypothèse classique selon laquelle $dl \times v(\hat{\theta}) / \text{Var}_P(\hat{\theta})$ est de distribution khi-carré et qui est à l'origine de cette équation, peut ne pas être vérifiée dans un contexte de données d'enquête complexe. La valeur de dl donnée à l'équation (2.1) sera appelée le *nombre théorique de degrés de liberté* dans la suite de l'article. L'expression algébrique de dl est dérivée à la section 6 pour deux méthodes d'estimation de la variance, dans le contexte spécifique décrit dans la section suivante.

3. Cadre de référence et notations

Dans le reste de l'article, on suppose un plan d'échantillonnage semblable à celui utilisé pour le questionnaire détaillé du Recensement de la population de 2021. D'abord, les ménages, les unités primaires d'échantillonnage, sont regroupés en un certain nombre de strates. Ensuite, un ménage sur quatre est sélectionné aléatoirement sans remise dans chaque strate. Finalement, tous les individus des ménages sélectionnés font partie de l'échantillon.

Soit U, une certaine population stratifiée regroupant H strates. Supposons que cette population soit composée de N ménages regroupant M individus. Soient N_h, le nombre de ménages de la strate h, M_{hi}, le nombre d'individus du ménage i de la strate h, et y_{hij}, la variable d'intérêt pour l'individu j, du ménage i de la strate h, $h = 1,\ldots,H, i = 1,\ldots,N_h$, $j = 1,\ldots,M_{hi}$. On définit

$$N = \sum_{h=1}^{H} N_h \quad \text{et} \quad M = \sum_{h=1}^{H} M_h,$$

où $M_h = \sum_{i=1}^{N_h} M_{hi}$. Soit $D \subseteq U$, un certain domaine et z_{Dhij}, l'indicateur d'appartenance de l'individu j au domaine. On note $M_D \leq M$, le nombre d'individus de la population faisant partie de ce domaine, c'est-à-dire

$$M_D = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} z_{Dhij}.$$

Le total des y_{hij} sur le domaine D de tous les individus du ménage i de la strate h est noté

$$t_{Dhi} = \sum_{j=1}^{M_{hi}} y_{hij} z_{Dhij}. \tag{3.1}$$

Le paramètre d'intérêt est le total Y_D sur le domaine D, c'est-à-dire

$$Y_D = \sum_{h=1}^{H} \sum_{i=1}^{N_h} t_{Dhi}.$$

Un échantillon aléatoire simple sans remise s_h, composé de n_h ménages regroupant m_h individus, est tiré dans la strate h. Soit I_{hi} l'indicateur d'appartenance à l'échantillon du ménage i de la strate h. Le nombre d'individus de l'échantillon s sur le domaine D, dont la valeur est aléatoire, est noté

$$m_D = \sum_{h=1}^{H} \sum_{i=1}^{N_h} I_{hi} m_{Dhi},$$

où $m_{Dhi} = \sum_{j=1}^{M_{hi}} z_{Dhij}$. Pour la suite, uniquement le cas où $m_D \geq 1$ est considéré. Soit également n_D, le nombre de ménages dont au moins un individu appartient au domaine D dans l'échantillon s, c'est-à-dire

$$n_D = \sum_{h=1}^{H} \sum_{i=1}^{N_h} I_{hi} J_{Dhi}, \tag{3.2}$$

où J_{Dhi} vaut 1 si $m_{Dhi} > 0$, et zéro sinon. L'estimateur de type Narain-Horvitz-Thompson de Y_D est

$$\hat{Y}_D = \sum_{h=1}^{H} w_h \sum_{i=1}^{N_h} I_{hi} t_{Dhi}, \tag{3.3}$$

où $w_h = N_h / n_h$. Pour des raisons logistiques, ce type d'estimateur est utilisé pour les estimations du recensement de la population. Soient σ_{Dh}^2 et m_{Dh4}, respectivement la variance et le moment centré d'ordre 4 des totaux t_{Dhi} dans la strate h, qui sont fixes par rapport au plan de sondage, c'est-à-dire

$$\sigma_{Dh}^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (t_{Dhi} - \mu_{Dh})^2 \quad \text{et} \quad m_{Dh4} = \frac{1}{N_h} \sum_{i=1}^{N_h} (t_{Dhi} - \mu_{Dh})^4, \tag{3.4}$$

où

$$\mu_{Dh} = \frac{1}{N_h} \sum_{i=1}^{N_h} t_{Dhi}.$$

La forme exacte de la variance de $\hat{Y}_D$ par rapport au plan de sondage P est

$$\mathrm{Var}_P(\hat{Y}_D) = \sum_{h=1}^{H} \frac{N_h^2 (N_h - n_h)}{n_h (N_h - 1)} \sigma_{Dh}^2. \tag{3.5}$$

Puisque les variances σ_{Dh}^2, $h = 1, \ldots, H$, sont généralement inconnues, $\mathrm{Var}_P(\hat{Y}_D)$ doit être estimée. L'estimateur sans biais classique de cette variance est

$$v_{cl}(\hat{Y}_D) = \sum_{h=1}^{H} N_h \left(\frac{N_h - n_h}{n_h} \right) S_{Dh}^2, \quad \text{où} \quad S_{Dh}^2 = \frac{1}{n_h - 1} \left\{ \sum_{i=1}^{n_h} t_{Dhi}^2 - \frac{1}{n_h} \left(\sum_{i=1}^{n_h} t_{Dhi} \right)^2 \right\}. \tag{3.6}$$

4. Construction d'un intervalle de confiance

Il y a plusieurs façons de construire un intervalle de confiance. L'intervalle de confiance de Student repose sur la distribution de la statistique pivotale

$$t_{\text{pivot}} = \frac{\hat{Y}_D - Y_D}{\sqrt{v(\hat{Y}_D)}}, \tag{4.1}$$

où $v(\hat{Y}_D)$ est un estimateur sans biais de $\mathrm{Var}_P(\hat{Y}_D)$. En supposant que t_{pivot} suit une loi de Student avec un certain nombre de degrés de liberté $dl > 0$, un intervalle de confiance de niveau $1 - \alpha$ s'obtient alors en prenant deux valeurs de la distribution de t_{pivot} telles que $P(t_1 < t_{\text{pivot}} < t_2) = 1 - \alpha$. Les deux bornes, inférieure BI et supérieure BS, de l'intervalle de confiance sont ensuite déduites algébriquement de telle sorte que $P\left(BI < Y_D < BS\right) = 1 - \alpha$. Choisir t_1 et t_2 de façon symétrique conduit à l'intervalle de confiance de Student de niveau nominal $1 - \alpha$ qui prend la forme

$$\hat{Y}_D \pm t_{dl,1-\alpha/2} \sqrt{v(\hat{Y}_D)}, \tag{4.2}$$

où $t_{dl,1-\alpha/2}$ est le quantile Student à dl degrés de liberté. Comme cas particulier, l'intervalle de confiance de Wald, largement utilisé, s'obtient lorsque le nombre de degrés de liberté est très grand. En présence d'une variable d'intérêt dichotomique, l'intervalle de confiance de Student peut s'avérer inadapté. Il peut

notamment générer des bornes qui sortent des limites cohérentes pour un effectif. Par exemple, lorsque $\hat{Y}_D$ est près de zéro, la borne inférieure donnée à l'équation (4.2) peut devenir négative, entraînant une sous-couverture. Ce problème est particulièrement marqué lorsque la taille de l'échantillon est petite.

L'intervalle de confiance de Wilson modifié a été initialement proposé par Wilson (1927) pour l'estimation d'une proportion dans le cadre d'un échantillonnage aléatoire simple avec remise. Kott et Carr (1997) ont ensuite adapté cette méthode afin qu'elle soit valide pour l'estimation de proportions issues de données d'enquêtes complexes, tels que pour les plans d'échantillonnage sans remise. Plus récemment, Neusy, Savard, Hidiroglou et Martin (2021) ont étendu cette approche afin de l'appliquer à l'estimation d'effectifs. Contrairement à l'intervalle de confiance de Student, les bornes de l'intervalle de confiance de Wilson modifié sont généralement asymétriques autour de l'estimation $\hat{Y}_D$, et la méthode garantit un intervalle toujours à l'intérieur des limites cohérentes. Cette asymétrie résulte de l'utilisation d'un pivot différent de celui présenté à l'équation (4.1); pour plus de détails, voir Neusy et coll. (2021). Par ailleurs, leurs simulations ont démontré que ce type d'intervalle améliore souvent la couverture par rapport à d'autres méthodes, notamment, l'intervalle de confiance de Student.

Soit $\mathcal{H}_D \subseteq \{1, \ldots, H\}$, l'ensemble des indices des strates non vides, c'est-à-dire celles dont au moins un individu fait partie du domaine. On définit $M_I \leq M$ et $m_I \leq m$ de telle sorte que

$$M_I = \sum_{h \in \mathcal{H}_D} M_h \quad \text{et} \quad m_I = \sum_{h \in \mathcal{H}_D} m_h.$$

Soit aussi $\widehat{EP}(\hat{Y}_D)$ l'effet de plan estimé par rapport au plan aléatoire simple avec remise, c'est-à-dire $\widehat{EP}(\hat{Y}_D) = v(\hat{Y}_D)\, m_I \big/ \hat{Y}_D(M_I - \hat{Y}_D)$. L'intervalle de confiance de Wilson modifié de niveau nominal $1-\alpha$ est donné par

$$\frac{\hat{Y}_D + M_I a_D / 2}{1 + a_D} \pm \frac{\sqrt{\hat{Y}_d (M_I - \hat{Y}_D)\, a_D + M_I^2 a_D^2 / 4}}{1 + a_D}, \tag{4.3}$$

où $a_D = t^2_{dl,1-\alpha/2} / m_e$ et $m_e = m_I / \widehat{EP}(\hat{Y}_D)$ est la taille d'échantillon effective. À noter que l'effet de plan estimé $\widehat{EP}(\hat{Y}_D)$ n'est pas défini pour les cas extrêmes $\hat{Y}_D = 0$ et $\hat{Y}_D = M_I$. Ces cas extrêmes sont simplement écartés lors des simulations de la section 7.

Dans le cas de l'intervalle de confiance de Student, supposer que t_{pivot} suit une loi de Student provient du fait de supposer que $dl \times v(\hat{\theta}) / \text{Var}_P(\hat{\theta})$ suit une distribution khi-carré. Par conséquent, l'approximation de Satterthwaite fournit la valeur appropriée des degrés de liberté à utiliser dans ce contexte. Toutefois, cette correspondance n'est pas nécessairement assurée pour l'intervalle de confiance de Wilson modifié, dont les hypothèses sur la statistique pivotale diffèrent, bien que l'on utilise généralement cette même valeur de dl dans ce cas.

Puisque l'estimation de la variance influence les degrés de liberté, il est crucial de porter une attention particulière à la méthode employée pour obtenir cette estimation. La méthode d'estimation de la variance utilisée pour l'estimation du questionnaire détaillé du recensement est décrite en détail dans la section suivante.

5. Méthode des demi-échantillons partiellement équilibrés epsilon (DEPE-ε)

La méthode des demi-échantillons équilibrés (DEE), formalisée par McCarthy (1966), est une technique d'estimation de la variance par répliques, initialement conçue pour un plan d'échantillonnage aléatoire stratifié avec remise comportant exactement deux unités primaires par strate. Pour plus de détails, voir Wolter (2007). La méthode utilisée pour l'estimation du questionnaire détaillé du recensement, appelée DEPE-ε et décrite par Devin et Verret (2016), est une adaptation de la méthode des DEE qui est valide pour des plans d'échantillonnage complexes tels que celui utilisé pour le recensement. Cette technique d'estimation de la variance est présentée ici.

5.1 Réorganisation des ménages de l'échantillon

Puisque le nombre de ménages par strate n_h est généralement supérieur à 2 dans le cas du recensement, les ménages de l'échantillon doivent être réorganisés afin d'appliquer la méthode des DEPE-ε. Wolter (2007) propose deux solutions pour le cas où $n_h > 2$: diviser chaque strate en deux grappes et en choisir une par réplique ou créer plusieurs strates artificielles de deux ménages et en sélectionner un par réplique. La première solution est avantageuse sur le plan computationnel mais coûteuse en terme de perte d'information, tandis que la seconde réduit la perte d'information mais est plus coûteuse en calcul. Devin et Verret (2016) adoptent un compromis entre ces deux approches, illustré aux étapes 1 et 2 de la figure 5.1. La constitution des groupes d'équilibrage (étape 3 de la figure 5.1, en rouge) sera détaillée à la section 5.3.

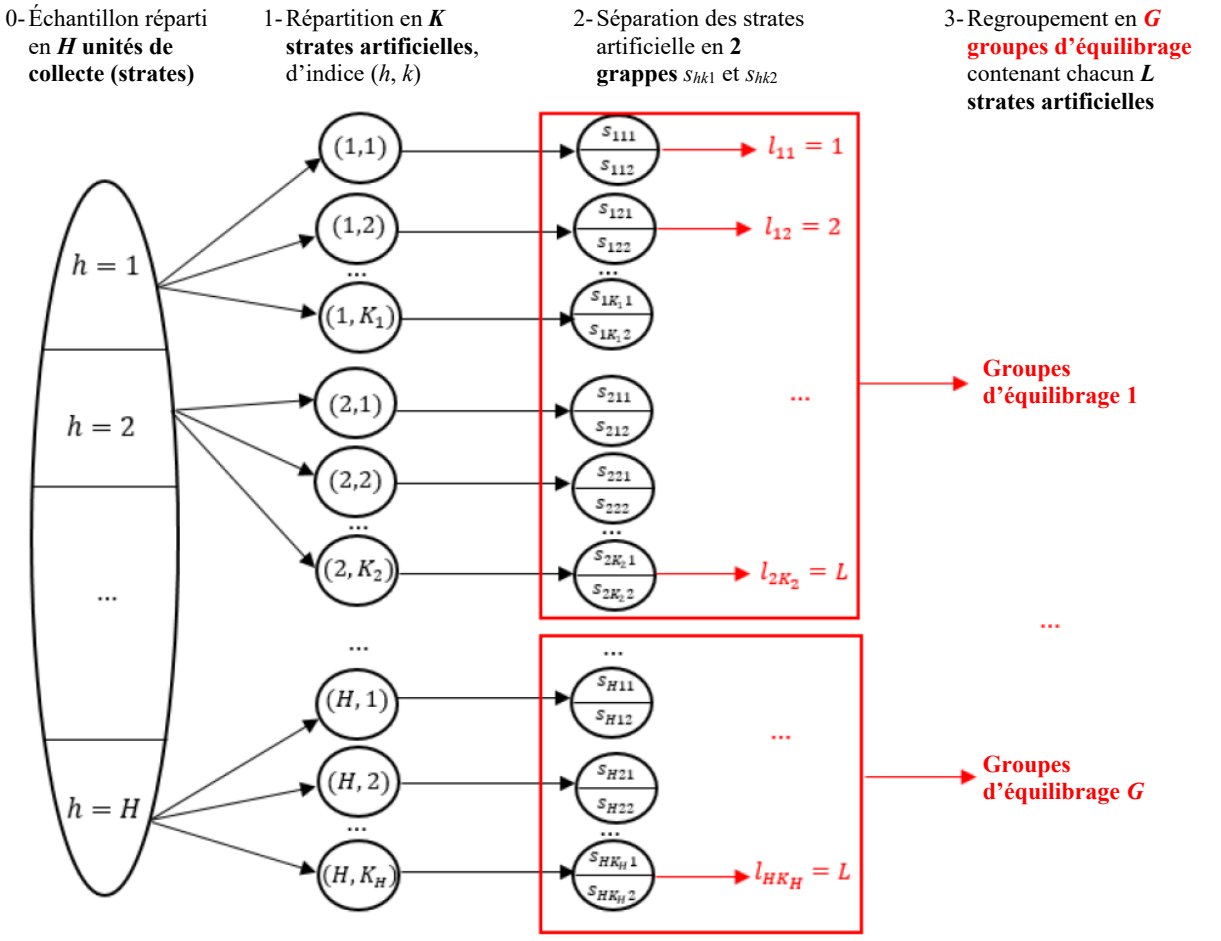
Dans un premier temps, les n_h ménages de chaque strate h sont répartis aléatoirement en K_h groupes appelés « strates artificielles ». Chaque strate artificielle (h,k), avec $k=1,\ldots,K_h$, contient n_{hk} ménages. Ainsi, l'ensemble des n ménages de l'échantillon est réparti en $K=\sum_{h=1}^{H}K_h$ strates artificielles. Cette répartition est réalisée de sorte que les tailles n_{hk} des strates artificielles soient toutes approximativement égales. Dans un second temps, les ménages de chaque strate artificielle (h,k) sont répartis aléatoirement en deux groupes, appelés « grappes », de tailles approximativement égales. Soient s_{hk1} et s_{hk2} l'ensemble des ménages appartenant respectivement aux grappes 1 et 2 de la strate artificielle (h,k). Ainsi, $s_h = \bigcup_{k=1}^{K_h}(s_{hk1}\cup s_{hk2})$.

Soient $t_{Dhk1}=\sum_{i\in s_{hk1}}t_{Dhi}$ et $t_{Dhk2}=\sum_{i\in s_{hk2}}t_{Dhi}$ les totaux sur le domaine D pour les grappes 1 et 2 de la strate artificielle (h,k), où t_{Dhi} est défini à l'équation (3.1). Chaque réplique r, $r=1,\ldots,R$, consiste à sélectionner une seule des deux grappes (s_{hk1} ou s_{hk2}) pour chacune des strates artificielles (h,k), puis à calculer $\hat{Y}_D^{(r)}$, une estimation du total Y_D, via

$$\hat{Y}_D^{(r)} = \sum_{h=1}^{H}\sum_{k=1}^{K_h}\left(w_{hk1}^{(r)}t_{Dhk1}+w_{hk2}^{(r)}t_{Dhk2}\right), \tag{5.1}$$

où $w_{hk1}^{(r)}$ et $w_{hk2}^{(r)}$ sont appelés poids de réplique et sont détaillés ci-dessous.

Figure 5.1 Étapes de création des strates artificielles, grappes et groupes d'équilibrage pour la méthode des DEPE-ε



Note : Demi-échantillons partiellement équilibrés (DEPE).

5.2 Définition des poids de réplique

Dans le cas classique, les poids de réplique valent soit 0, soit $2w_h$ selon la grappe sélectionnée. Ainsi, seulement la moitié de l'échantillon contribue à l'estimation du total à chaque réplique, l'autre moitié recevant un poids nul. Une définition pratique des poids de réplique est alors $(1+\delta_{hk}^{(r)})\, w_h$ ou $(1-\delta_{hk}^{(r)})\, w_h$, avec

$$\delta_{hk}^{(r)} = \begin{cases} 1 \text{ si } s_{hk1} \text{ est sélectionnée à la r}^{\text{e}} \text{ réplique} \\ -1 \text{ si } s_{hk2} \text{ est sélectionnée à la r}^{\text{e}} \text{ réplique.} \end{cases}$$

Comme le soulignent Rao et Shao (1999), la forte perturbation des poids de réplique par rapport aux poids originaux w_h peut poser certains problèmes, notamment lors d'un calage des poids de réplique. Pour atténuer cet effet, un paramètre de perturbation $\varepsilon \in (0,1)$ est introduit et les poids de réplique deviennent $(1-\varepsilon\delta_{hk}^{(r)})\, w_h$ ou $(1+\varepsilon\delta_{hk}^{(r)})\, w_h$. Cette variante, appelée méthode des DEE de Fay ou DEE-ε, est détaillée

par Judkins (1990) et permet la participation de tous les ménages de l'échantillon à chacune des répliques. Devin et Verret (2016) suggèrent de prendre $\varepsilon = \sqrt{1/2}$; cette valeur, employée pour l'estimation du questionnaire détaillé du recensement, sera utilisée pour la suite.

Deux modifications supplémentaires sont appliquées aux poids de réplique. Premièrement, comme proposé par Rao et Shao (1999), pour éviter un biais lié aux différences éventuelles de taille entre les grappes s_{hk1} et s_{hk2}, des facteurs d'ajustement $F_{hk1} = n_{hk2} / n_{hk1}$ et $F_{hk2} = n_{hk1} / n_{hk2}$ sont introduits, où n_{hk1} et n_{hk2} sont les tailles respectives des grappes s_{hk1} et s_{hk2}.

Deuxièmement, dans notre contexte de forte fraction de sondage, le facteur de correction pour population finie $(1 - n_h / N_h) = (1 - 1/w_h)$, tel que suggéré par Wolter (2007), est également appliqué. Ainsi, les poids de réplique utilisés dans la méthode des DEPE-ε s'écrivent

$$w_{hk1}^{(r)} = \left(1 + \delta_{hk}^{(r)} \varepsilon \sqrt{F_{hk1}\left(1 - \frac{1}{w_h}\right)}\right) w_h \quad \text{et} \quad w_{hk2}^{(r)} = \left(1 - \delta_{hk}^{(r)} \varepsilon \sqrt{F_{hk2}\left(1 - \frac{1}{w_h}\right)}\right) w_h.$$

Il est important de souligner que les poids de réplique $w_{hk1}^{(r)}$ et $w_{hk2}^{(r)}$ ainsi définis respectent la cohérence suivante sur la somme totale des poids, c'est-à-dire

$$\sum_{k=1}^{K_h} (n_{hk1} w_{hk1}^{(r)} + n_{hk2} w_{hk2}^{(r)}) = N_h.$$

5.3 Réduction du nombre de répliques

La méthode classique des demi-échantillons considère tous les $R = 2^K$ demi-échantillons possibles, ce qui est très coûteux sur le plan computationnel. La méthode dite « équilibrée » réduit ce nombre en sélectionnant les répliques de manière structurée, de sorte que $\sum_{r=1}^{R} \delta_{hk}^{(r)} = 0$ et $\sum_{r=1}^{R} \delta_{hk}^{(r)} \delta_{h'k'}^{(r)} = 0$ pour toutes les paires distinctes de strates artificielles $(h,k) \neq (h',k')$. En pratique, les valeurs de $\delta_{hk}^{(r)} \in \{-1, +1\}$ sont déterminées à l'aide d'une matrice de Hadamard de dimension $R \times R$. Se référer à l'article de Hedayat et Wallis (1978) pour plus de détails sur les matrices de Hadamard. Selon Wolter (2007), la méthode équilibrée fixe R au plus petit multiple de 4 supérieur à K, réduisant ainsi le nombre de répliques tout en garantissant un estimateur de variance équivalent à l'estimateur classique pour des statistiques linéaires. Toutefois, lorsque K est grand, la mise en œuvre reste complexe sur le plan computationnel.

La méthode des demi-échantillons partiellement équilibrés (DEPE), définie par Wolter (2007), permet de réduire à nouveau le nombre de répliques. Utilisée pour l'estimation du questionnaire détaillé du recensement (Devin et Verret, 2016), elle consiste à former aléatoirement $G \geq 2$ groupes dits « groupes d'équilibrage », regroupant les strates artificielles de plusieurs strates, de manière à obtenir un nombre égal $L = K / G$ de strates artificielles par groupe. Chaque strate artificielle (h,k) reçoit un indice $l_{hk} \in \{1, \ldots, L\}$ aléatoire au sein de son groupe. La formation des groupes d'équilibrage ainsi que la numérotation des strates artificielles de 1 à L sont illustrées à la figure 5.1 (étape 3, en rouge). Le nombre de répliques R est le plus petit multiple de 4 supérieur à L. Les répliques sont structurées comme dans la méthode DEE, mais à l'intérieur de chaque groupe d'équilibrage, en respectant les trois conditions suivantes :

1) $\sum_{r=1}^{R} \delta_{hk}^{(r)} = 0$ pour toutes strates artificielles (h,k),
2) $\sum_{r=1}^{R} \delta_{hk}^{(r)} \delta_{h'k'}^{(r)} = 0$ pour toutes paires de strates artificielles $(h,k) \neq (h',k')$ telles que $l_{hk} \neq l_{h'k'}$,
3) $\delta_{hk}^{(r)} = \delta_{h'k'}^{(r)}$ pour toutes paires de strates artificielles $(h,k) \neq (h',k')$ telles que $l_{hk} = l_{h'k'}$ et tout $r = 1, \ldots, R$.

Les deux premières conditions garantissent que les répliques soient équilibrées pour chaque groupe, tandis que la troisième impose une organisation identique des répliques à l'intérieur d'un même groupe d'équilibrage. Finalement, l'estimateur de la variance de la méthode des DEPE-ε est donné par

$$v_{DEPE}(\hat{Y}_D) = \frac{1}{R\varepsilon^2} \sum_{r=1}^{R} (\hat{Y}_D^{(r)} - \hat{Y}_D)^2, \tag{5.2}$$

où $\hat{Y}_D^{(r)}$ est défini à l'équation (5.1). L'estimateur $v_{DEPE}(\hat{Y}_D)$ n'est pas équivalent à l'estimateur classique v_{cl} donné à l'équation (3.6). Cependant, tel que démontré à l'Annexe A, $E_*\{v_{DEPE}(\hat{Y}_D)\} = v_{cl}$, où E_* désigne l'espérance sous le processus aléatoire de création des répliques. Ce résultat implique que l'estimateur $v_{DEPE}(\hat{Y}_D)$ est sans biais pour $\mathrm{Var}_P(\hat{Y}_D)$, puisque v_{cl} l'est également.

6. Degrés de liberté

Dans cette section, la formule du nombre théorique de degrés de liberté est obtenue pour l'estimateur classique de la variance $v_{cl}(\hat{Y}_D)$ ainsi que pour $v_{DEPE}(\hat{Y}_D)$, dans le cadre d'un plan aléatoire simple stratifié sans remise de grappes (ménages). Les expressions algébriques obtenues aux propositions 6.1 et 6.2 découlent directement de l'équation (2.1). Elles sont obtenues pour l'estimateur du total $\hat{Y}_D$, mais se généralisent à tout estimateur linéaire. Des versions estimées et approximatives du nombre de degrés de liberté sont également présentées.

6.1 Degrés de liberté liés à la méthode classique d'estimation de la variance

La proposition suivante donne le nombre théorique de degrés de liberté associé à $v_{cl}(\hat{Y}_D)$, l'estimateur classique de la variance défini à l'équation (3.6).

Proposition 6.1 : Soient $V_1 = \mathrm{Var}_P(\hat{Y}_D)$ telle que définie à l'équation (3.5) et

$$V_{2,cl} = \sum_{h=1}^{H} \frac{N_h^3 (N_h - n_h)^3}{n_h^3 (n_h - 1)(N_h - 1)^2 (N_h - 2)(N_h - 3)} \times \left[m_{Dh4} (N_h - 1)\{N_h (n_h - 1) - (n_h + 1)\} - \sigma_{Dh}^4 \{N_h^2 (n_h - 3) + 6N_h - 3(n_h + 1)\} \right].$$

Le nombre théorique de degrés de liberté associé à $v_{cl}(\hat{Y}_D)$ est $dl_{cl} = 2V_1^2 / V_{2,cl}$.

Le résultat de la proposition 6.1, dont la preuve se trouve à l'Annexe B, peut également être appliqué à d'autres types d'estimateurs de la variance qui sont équivalents à $v_{cl}(\hat{Y}_D)$. C'est notamment le cas du jackknife et du bootstrap lorsque le nombre de répliques est très grand.

6.2 Degrés de liberté liés à la méthode des DEPE-ε

L'expression algébrique du nombre théorique de degrés de liberté est maintenant dérivée dans le cas où l'estimateur de la variance est $v_{DEPE}(\hat{Y}_D),$ tel que fourni à l'équation (5.2).

Proposition 6.2 : Soient $V_1 = \mathrm{Var}_P(\hat{Y}_D)$ telle que définie à l'équation (3.5) et

$$\begin{aligned} V_{2,DEPE} \;=\; & \sum_{h=1}^{H} \frac{N_h^3 (N_h - n_h)^2}{n_h^4 (N_h - 1)(N_h - 2)(N_h - 3)} \left\{ N_h (N_h + 1) B_h - (N_h - 1)(n_h^2 + 2A_h) \right\} m_{Dh4} \\ & + \sum_{h=1}^{H} \frac{N_h^3 (N_h - n_h)^2}{n_h^4 (N_h - 1)(N_h - 2)(N_h - 3)} \left\{ 2A_h (N_h^2 - 3N_h + 3) + \frac{n_h^2 (N_h^2 - 3)}{N_h - 1} - 3N_h (N_h - 1) B_h \right\} \sigma_{Dh}^4 \\ & + \frac{G-1}{K-1} \sum_{h \neq h'} \frac{N_h^2 N_{h'}^2 (N_h - n_h)(N_{h'} - n_{h'})}{n_h n_{h'} (N_h - 1)(N_{h'} - 1)} \sigma_{Dh}^2 \sigma_{Dh'}^2, \end{aligned}$$

où

$$A_h = \sum_{k=1}^{K_h} n_{hk}^2 \quad \text{et} \quad B_h = \sum_{k=1}^{K_h} \left(\frac{n_{hk1}^2}{n_{hk2}} + \frac{n_{hk2}^2}{n_{hk1}} \right). \tag{6.1}$$

Le nombre théorique de degrés de liberté associé à $v_{DEPE}(\hat{Y}_D)$ est

$$dl_{DEPE} \;=\; \frac{2V_1^2}{V_{2,DEPE}}. \tag{6.2}$$

La preuve de la proposition 6.2 est présentée à l'Annexe C. Rappelons que G et K représentent respectivement le nombre de groupes d'équilibrage et le nombre de strates artificielles liés à la mise en œuvre de la méthode des DEPE-ε. Puisque $v_{DEPE}(\hat{Y}_D),$ tel que défini à l'équation (5.2), s'exprime comme une somme de carrés, il est raisonnable de penser que cet estimateur est distribué selon une loi khi-carré lorsque la taille de l'échantillon est assez grande.

Tel que démontré à l'Annexe A, $E_*\{v_{DEPE}(\hat{Y}_D)\} = v_{cl}(\hat{Y}_D).$ Conséquemment, la variance $V_{2,DEPE}$ au dénominateur de dl_{DEPE} est telle que

$$V_{2,DEPE} \;=\; \mathrm{Var}_P\{v_{cl}(\hat{Y}_D)\} + E_P\left[\mathrm{Var}_*\{v_{DEPE}(\hat{Y}_D)\}\right] \;\geq\; \mathrm{Var}_P\{v_{cl}(\hat{Y}_D)\} \;=\; V_{2,cl},$$

où $V_{2,cl}$ est telle que définie à la proposition 6.1. Il en résulte que $dl_{DEPE} \leq dl_{cl}.$ Ainsi, dl_{cl} constitue une borne supérieure pour $dl_{DEPE},$ et la différence entre les deux représente la perte en termes de degrés de liberté due à l'utilisation de la méthode des DEPE-ε par rapport à la méthode classique.

L'équation (6.2) ne permet pas d'identifier facilement les facteurs qui influencent le nombre de degrés de liberté. Afin d'en faciliter l'interprétation, supposons que toutes les quantités à l'intérieur de chacune des strates sont égales, c'est-à-dire $N_h = N_{h'},$ $n_h = n_{h'},$ $\sigma_{Dh}^2 = \sigma_{Dh'}^2,$ $m_{Dh4} = m_{Dh'4},$ $B_h = B_{h'}$ et $A_h = A_{h'},$ pour tout $h \neq h' \in \{1, \ldots H\}.$ Supposons également que la taille des strates soit très grande, c'est-à-dire $N_h \to \infty.$ Sous ces conditions, la formule de dl_{DEPE} se réduit à

$$dl_{DEPE} \;\approx\; \frac{2H}{B_h \kappa_{2D} + 2A_h / n_h^2 \;+\; (G-1)(H-1)/K - 1},$$

où $\kappa_{2D} = m_{Dh4} / \sigma_{Dh}^4 - 3 \geq -2$ est le coefficient d'aplatissement normalisé (« *Excess Kurtosis* ») des totaux t_{Dhi}. Cette formule simplifiée révèle, d'une part, qu'une grande taille d'échantillon n_h contribue à augmenter dl_{DEPE}. D'autre part, dl_{DEPE} diminue à mesure que κ_{2D} augmente. Un coefficient d'aplatissement élevé indique que la distribution des totaux t_{Dhi} présente un pic prononcé autour de la moyenne et des queues de distribution épaisses. Lorsque la taille du domaine est petite, la probabilité d'observer des totaux nuls augmente, ce qui contribue à accroître κ_{2D} et, par conséquent, à réduire dl_{DEPE}.

Cette formule simplifiée permet également d'identifier des pistes pour optimiser la méthode des DEPE-ε. D'abord, imposer des tailles de strates artificielles n_{hk} aussi égales que possible, ainsi que choisir des tailles de grappes n_{hk1} et n_{hk2} les plus similaires possible au sein de ces strates, permet de limiter la perte de degrés de liberté en minimisant les valeurs de A_h et de B_h. Ensuite, maintenir un petit nombre de groupes d'équilibrage G et un grand nombre de strates artificielles K contribue également à réduire cette perte. Choisir G faible et K élevé implique cependant un nombre accru de répliques R. Ainsi, comme attendu, un grand nombre de répliques conduit à un nombre de degrés de liberté plus élevé.

6.3 Estimation de dl_{DEPE}

Le calcul de dl_{DEPE} requiert la connaissance des paramètres σ_{Dh}^2 et m_{Dh4}, tels que définis à l'équation (3.4), pour chacune des strates $h = 1, \ldots, H$. Ces paramètres sont généralement inconnus. Soit $\bar{t}_{Dhi} = \sum_{i \in s_h} t_{Dhi} / n_h$. Des estimateurs non biaisés de σ_{Dh}^2 et m_{Dh4}, tels que rapportés par Gerlovina et Hubbard (2019), sont donnés respectivement par

$$\hat{\sigma}_{Dh}^2 = \frac{1}{n_h - 1} \sum_{i \in s_h} (t_{Dhi} - \bar{t}_{Dhi})^2$$

et

$$\hat{m}_{Dh4} = \frac{1}{(n_h - 1)(n_h - 2)(n_h - 3)} \left\{ (n_h^2 - 2n_h + 3) \sum_{i \in s_h} (t_{Dhi} - \bar{t}_{Dhi})^4 - \frac{3(2n_h - 3)(n_h - 1)}{n_h} \hat{\sigma}_{Dh}^2 \right\}.$$

Ainsi, un estimateur de dl_{DEPE}, noté $\widehat{dl}_{DEPE}$, consiste en la formule du nombre théorique de degrés de liberté dl_{DEPE}, donnée à l'équation (6.2), pour laquelle on remplace simplement σ_{Dh}^2 et m_{Dh4} par $\hat{\sigma}_{Dh}^2$ et $\hat{m}_{Dh4}$ respectivement. Une version empirique de dl_{cl}, dont l'expression requiert aussi la connaissance de σ_{Dh}^2 et m_{Dh4}, pourrait s'obtenir de la même façon.

6.4 Règles approximatives

En pratique, la valeur des degrés de liberté est souvent déterminée à l'aide d'une règle approximative. La règle standard, généralement utilisée, consiste à soustraire le nombre de strates du nombre d'unités primaires d'échantillonnage. Comme le rapportent Valliant et Rust (2010), cette règle constitue un cas particulier du nombre théorique de degrés de liberté obtenu à la proposition 6.1, dont la précision repose sur

plusieurs hypothèses. Parmi celles-ci, il est supposé que les fractions de sondage sont négligeables dans chaque strate, que la variable d'intérêt suit une distribution normale et que les variances de chaque strate sont équivalentes. Lorsque ces hypothèses ne sont pas respectées, la règle standard peut entraîner une surestimation substantielle du nombre réel de degrés de liberté. Ce problème est d'autant plus probable lorsque la taille de l'échantillon est petite et peut s'avérer particulièrement grave pour les plans d'échantillonnage à un seul degré de liberté.

À cause de contraintes opérationnelles, une règle approximative légèrement différente est utilisée pour l'estimation du questionnaire détaillé du recensement de 2021. Celle-ci consiste en

$$dl_{App} = \min\left(n_D, R\right), \tag{6.3}$$

où R est le nombre de répliques pour la méthode d'estimation de la variance des DEPE-ε et n_d est tel que défini à l'équation (3.2).

Dans la section suivante, quatre versions des degrés de liberté sont comparées dans une *étude* par simulation, c'est-à-dire dl_{DEPE}, $\widehat{dl}_{DEPE}$, dl_{cl} et dl_{App}.

7. Simulations

Dans cette section, les résultats d'une étude par simulation sont présentés dans le but d'évaluer le gain potentiel à utiliser le nombre théorique de degrés de liberté dl_{DEPE} tel que défini à l'équation (6.2) par rapport à la version approximative dl_{App} donnée à l'équation (6.3). L'efficacité de l'estimateur $\widehat{dl}_{DEPE}$ et la réduction du nombre de degrés de liberté due au fait d'utiliser v_{DEPE} comparativement à v_{cl} sont également évaluées. Les scénarios des simulations sont construits afin d'être les plus réalistes possible par rapport aux estimations du questionnaire détaillé du recensement.

7.1 Description des scénarios de simulations

Une population fictive a d'abord été générée. Cette population, formée de $H = 22$ strates de tailles variables, est composée de $N = 4\,030$ ménages regroupant $M = 11\,227$ individus au total. Pour chacun des scénarios décrits dans ce qui suit, 3 000 échantillons indépendants ont été sélectionnés à partir de cette population fictive. Ces échantillons ont été tirés selon un plan aléatoire simple sans remise avec une fraction de sondage de 1/4 dans chacune des 22 strates.

Afin de mettre en œuvre la méthode des DEPE-ε, un total de 198 strates artificielles a été constitué en répartissant, pour chacune des 22 strates, les ménages sélectionnés dans 9 strates artificielles de taille approximativement égale; $G = 2$ groupes d'équilibrage ont été formés, chacun contenant $L = 99$ strates artificielles. Le nombre de répliques a été établi à $R = 100$. Cette réorganisation en 99 strates artificielles et l'utilisation de 100 répliques coïncident avec la procédure utilisée pour l'estimation de la variance lors du Recensement de 2021. La valeur de ε a été fixée à $\sqrt{1/2}$.

Le domaine D a été généré au niveau du ménage, c'est-à-dire que l'indicateur d'appartenance à D est identique pour tous les membres d'un même ménage. Autrement dit $z_{Dhij} = z_{Dhi} \in \{0, 1\}$. Trois domaines de tailles différentes ont été générés. Ainsi, z_{Dhi} a été généré aléatoirement de façon indépendante pour chaque ménage de la population avec des probabilités d'inclusion de 3 %, 5 % ou 10 %. Le tableau 7.1 montre les tailles de ces trois domaines.

Tableau 7.1
Nombre d'individus et de ménages de la population fictive appartenant au domaine *D*

Domaine	Nombre d'individus	Nombre de ménages
3 %	338	121
5 %	570	204
10 %	1 136	428

À noter que des simulations supplémentaires ont également été réalisées pour un domaine dont z_{Dhij} était généré de façon indépendante pour chacun des individus d'un même ménage. Les résultats de ces simulations s'apparentent à ceux obtenus avec les domaines présentés au tableau 7.1, c'est pourquoi ils ne sont pas présentés ici.

Pour le recensement de la population, il est fréquent d'observer une dépendance intra-ménage, c'est-à-dire que les variables d'intérêt pour les membres d'un même ménage sont dépendantes. C'est le cas, par exemple, pour des variables relatives à la langue. Ainsi, tous les scénarios de simulation ont été mis en œuvre d'abord sans, puis avec dépendance intra-ménage. Premièrement, pour les scénarios sous l'indépendance intra-ménage, la variable y_{hij} est générée de façon indépendante de $y_{hij'}$, pour tout $j \neq j'$. Deuxièmement, pour les scénarios avec dépendance intra-ménage, la variable d'intérêt (dichotomique ou continue) a été générée aléatoirement en utilisant une copule normale et en considérant une dépendance telle que le rho de Spearman entre y_{hij} et $y_{hij'}$, pour tous $j \neq j'$, est de 0,95, ce qui correspond à un coefficient de corrélation de Pearson de 0,954. Pour plus de détails au sujet des copules, se référer à Nelsen (2006). Aucune dépendance entre les membres de ménages distincts n'a été considérée.

Pour chacun des 3 000 échantillons, un calage a d'abord été effectué au niveau de chacune des strates. Les poids de calage $\tilde{w}_{hi}$, $h = 1, \ldots, H, i = 1, \ldots, N_h$, ont été obtenus afin de minimiser la somme des différences relatives sous les contraintes

$$\sum_{i=1}^{N_h} \tilde{w}_{hi} I_{hi} = N_h \quad \text{et} \quad \sum_{i=1}^{N_h} \sum_{j=1}^{M_h} \tilde{w}_{hi} I_{hi} = M_h.$$

Ces deux contraintes assurent que la somme des poids au niveau des ménages soit égale au nombre réel de ménages N_h dans la strate h, ainsi que la somme des poids au niveau des individus soit égale au nombre réel d'individus M_h dans la strate h. L'estimateur $\tilde{Y}_D$ du total Y_D utilisé pour les simulations est obtenu en remplaçant les poids originaux par les poids de calage dans l'équation (3.3), c'est-à-dire

$$\tilde{Y}_D = \sum_{h=1}^{H}\sum_{i=1}^{N_h} \tilde{w}_{hi} I_{hi} t_{Dhi}.$$

La variance de $\tilde{Y}_D$ a été estimée à l'aide de la méthode des DEPE-ε, telle que décrite à la section 5. Un calage a également été réalisé sur tous les poids de réplique afin de représenter le plus fidèlement possible les procédures appliquées lors du recensement. Il est difficile d'évaluer l'impact du calage sur le nombre de degrés de liberté, car les estimateurs par calage ne sont pas équivalents à l'estimateur linéaire de Horvitz-Thompson, même asymptotiquement. Bien qu'aucune garantie n'assure que cet impact soit négligeable, aucun ajustement n'a été effectué dans les formules théoriques des degrés de liberté pour tenir compte du calage dans les simulations. Cet exercice est laissé pour des considérations futures. Le calage a néanmoins été inclus dans les simulations afin de reproduire fidèlement les procédures opérationnelles du recensement.

Pour chaque scénario, les intervalles de confiance de niveau 95 % ont été calculés en utilisant les quatre versions des degrés de liberté, c'est-à-dire dl_{DEPE}, $\widehat{dl}_{DEPE}$, dl_{cl} et dl_{App}. La couverture des intervalles de confiance est estimée par la proportion des 3 000 échantillons dont l'intervalle de confiance contient la vraie valeur du total de la population Y_D.

Il convient de souligner que la valeur de dl_{DEPE}, de même que celle de dl_{cl}, est fixe pour tous les 3 000 échantillons d'un même scénario. En effet, les paramètres σ^2_{Dh} et m_{Dh4} qui interviennent dans le calcul de dl_{DEPE} s'obtiennent au niveau de la population, tandis que les autres éléments, tels que n_h et N_h, dépendent du plan de sondage qui est fixe pour les 3 000 répliques. Au contraire, les valeurs de $\widehat{dl}_{DEPE}$ et de dl_{App} varient d'un échantillon à l'autre car elles dépendent de quantités calculées au niveau de l'échantillon; le nombre de ménages dans le domaine n_D dans le cas de dl_{App} et $\hat{\sigma}^2_{Dh}$ et $\hat{m}_{Dh4}$ dans le cas de $\widehat{dl}_{DEPE}$.

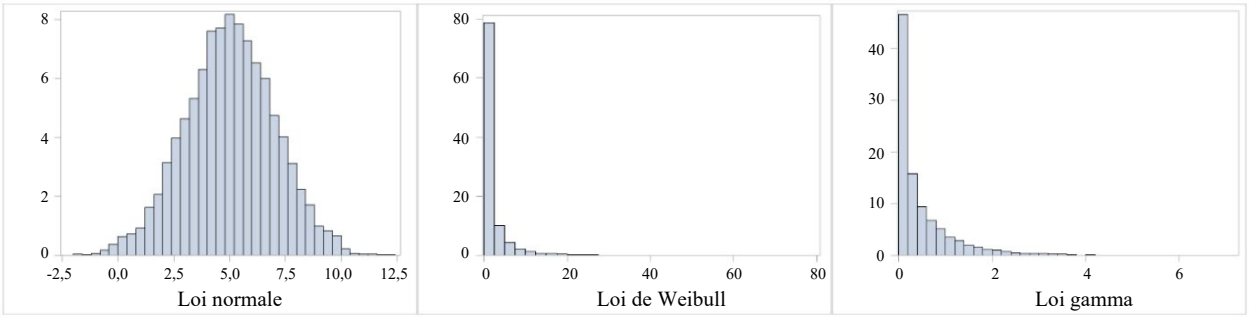
Dans certains cas où dl_{DEPE} était très petit, la couverture était fortement supérieure au taux nominal (près de 100 %). Une explication possible est que l'hypothèse selon laquelle $dl \times v(\hat{\theta}) / \mathrm{Var}_P(\hat{\theta})$ est de distribution khi-carré n'est peut-être pas valide pour certains cas limites. Ainsi, dl_{DEPE}, $\widehat{dl}_{DEPE}$ et dl_{cl} ont été bornés inférieurement à 2,5 degrés de liberté pour l'étude de simulations.

Différents types de variables d'intérêt ont été générés; des variables continues et dichotomiques. La section 7.2 présente les résultats pour l'estimation d'un total de variables continues, tandis que la section 7.3 présente les résultats pour l'estimation d'un effectif.

7.2 Résultats de simulations pour les variables continues

On a considéré les trois distributions de probabilité continues suivantes : La loi normale de moyenne 5 et de variance 4, la loi de Weibull avec paramètre de forme 0,5 et paramètre d'échelle 1 et la loi Gamma avec paramètre de forme 0,5 et paramètre d'échelle 1. La distribution de ces trois variables au niveau de la population est illustrée à la figure 7.1 pour le cas d'indépendance intra-ménage. Des simulations supplémentaires ont également été réalisées pour des variables de type exponentielle, Fisher et Student. Les résultats obtenus s'apparentent à ceux rapportés dans cette section et ne sont pas présentés ici.

Figure 7.1 Distribution des trois variables continues pour les 11 227 individus de la population pour les scénarios sans dépendance intra-ménage



Les tableaux 7.2 (indépendance intra-ménage) et 7.3 (dépendance intra-ménage) présentent dl_{DEPE} et dl_{cl} (calculés au niveau de la population) ainsi que les moyennes $\overline{\widehat{dl}}_{DEPE}$ et $\overline{dl}_{App}$ obtenues à partir des 3 000 échantillons de $\widehat{dl}_{DEPE}$ et dl_{App} respectivement, pour chacun des scénarios. Ces tableaux comparent également la couverture estimée des intervalles de confiance de Student, décrits à l'équation (4.2), ainsi que les longueurs relatives moyennes exprimées en pourcentage de ces intervalles, c'est-à-dire la longueur divisée par la valeur absolue de $\hat{Y}_D$, obtenues pour chacun des scénarios.

Tel qu'attendu, dl_{DEPE} augmente lorsque la taille du domaine augmente. Également, la présence de dépendance intra-ménage semble réduire le nombre de degrés de liberté. En effet, dl_{DEPE} est plus petit sous les scénarios avec dépendance intra-ménage (tableau 7.3) comparativement aux scénarios sous l'indépendance intra-ménage (tableau 7.2). De surcroît, dl_{DEPE} est plus grand en présence d'une variable normale. Ceci concorde avec le fait que la valeur du coefficient d'aplatissement normalisé κ_{2D} est plus petite pour la variable normale que pour les variables Weibull et Gamma.

Tableau 7.2
Couverture estimée par 3 000 répétitions d'intervalles de confiance de Student sur un total de variables continues sous l'indépendance intra-ménage

Type de variable	Taille du domaine	Degrés de liberté				Longueur relative moyenne des intervalles des confiance (en %)				Couverture estimée (en %)			
		dl_{DEPE}	$\overline{\widehat{dl}}_{DEPE}$	dl_{cl}	$\overline{dl}_{App}$	dl_{DEPE}	$\widehat{dl}_{DEPE}$	dl_{cl}	dl_{App}	dl_{DEPE}	$\widehat{dl}_{DEPE}$	dl_{cl}	dl_{App}
Normale	3 %	28,2	26,3	39,2	30,3	72,7	73,1	71,8	72,6	93,9	94,1	93,7	94,0
	5 %	44,1	40,4	68,4	50,7	54,4	54,6	53,9	54,2	93,7	93,7	93,4	93,7
	10 %	67,8	63,2	129,1	99,0	37,2	37,3	36,9	37	95,2	95,2	95,0	95,0
Weibull	3 %	5,8	11,3	6,4	30,3	121,9	113,9	119,1	100,9	94,4	93,1	94,1	91,6
	5 %	9,5	12,2	11,1	50,7	90	89,7	88,2	80,6	93,4	92,8	93,0	91,3
	10 %	11,0	16,6	12,8	99,0	64,5	63,8	63,4	58,2	93,7	92,6	93,5	91,3
Gamma	3 %	19,3	18,4	23,9	30,3	87,1	87,8	85,9	85,1	93,9	94,0	93,6	93,5
	5 %	21,3	21,2	27,6	50,7	68,6	69,0	67,7	66,3	93,8	93,8	93,4	93,2
	10 %	32,3	32,6	43,9	99,0	47,3	47,4	46,8	46,1	94,5	94,5	94,3	94,1

Note : Demi-échantillons partiellement équilibrés (DEPE).

Tableau 7.3
Couverture estimée par 3 000 répétitions d'intervalles de confiance de Student sur un total de variables continues en présence de dépendance intra-ménage

Type de variable	Taille du domaine	Degrés de liberté				Longueur relative moyenne des intervalles des confiance (en %)				Couverture estimée (en %)			
		dl_{DEPE}	$\overline{\widehat{dl}}_{DEPE}$	dl_{cl}	$\overline{dl}_{App}$	dl_{DEPE}	$\widehat{dl}_{DEPE}$	dl_{cl}	dl_{App}	dl_{DEPE}	$\widehat{dl}_{DEPE}$	dl_{cl}	dl_{App}
Normale	3 %	20,1	20,2	26,2	30,3	76,4	76,8	75,3	74,9	94,5	94,5	94,1	94,1
	5 %	35,5	33,6	51,9	50,7	57,6	57,8	57,0	57,0	93,4	93,5	93,1	93,2
	10 %	57,5	53,8	100,5	99,0	38,8	38,8	38,4	38,4	94,7	94,7	94,4	94,4
Weibull	3 %	2,5	6,1	2,5	30,3	238,1	181,8	238,1	136,2	93,6	88,1	93,6	83,8
	5 %	2,5	7,0	2,6	50,7	203,8	151,2	199,9	114,5	95,2	88,1	95,0	84,3
	10 %	3,0	10,6	3,4	99,0	137,7	112,2	128,2	86	95,6	87,7	94,6	85,2
Gamma	3 %	3,5	8,0	4,0	30,3	163,2	138,4	154,1	113,6	94,4	90,2	93,4	87,6
	5 %	5,1	10,2	6,1	50,7	113,8	105,3	108,8	89,6	94,1	91,6	93,0	89,3
	10 %	9,9	16,5	11,7	99,0	69,9	68,7	68,5	62,1	94,0	93,2	93,6	91,8

Note : Demi-échantillons partiellement équilibrés (DEPE).

La version empirique du nombre de degrés de liberté théorique $\widehat{dl}_{DEPE}$ tend à être assez proche en moyenne de dl_{DEPE} sous la normalité. En revanche, elle donne de moins bons résultats pour les variables Weibull et Gamma particulièrement en présence de dépendance intra-ménage. La version empirique $\widehat{dl}_{DEPE}$ semble surévaluer la vraie valeur de dl_{DEPE} dans ces cas.

Rappelons que la différence entre dl_{cl} et dl_{DEPE} illustre la perte en termes de degrés de liberté liée à l'utilisation de la méthode d'estimation de la variance des DEPE-ε par rapport à la méthode classique. Cette perte est particulièrement marquée lorsque le nombre de degrés de liberté est élevé. Par exemple, pour la distribution normale en présence de dépendance intra-ménage et pour le domaine de 10 %, les degrés de liberté passent de 100,5 à 57,5. Cependant, lorsque le nombre de degrés de liberté est grand (disons supérieur à 30), cette perte a un impact limité sur la couverture (de 94,4 % à 94,7 % dans l'exemple précédent). En revanche, lorsque le nombre de degrés de liberté est petit, l'écart absolu entre dl_{cl} et dl_{DEPE} est moins prononcé, mais il affecte davantage la couverture. Par exemple, pour la distribution gamma avec dépendance intra-ménage et le domaine de 3 %, une diminution des degrés de liberté de 4 à 3,5 entraîne une augmentation de la couverture estimée de 1 point de pourcentage, passant de 93,4 % à 94,4 %.

Du côté des degrés de liberté approximatifs, le type de variable et la présence de dépendance intra-ménage n'exercent aucune influence sur dl_{App}. C'est uniquement la taille du domaine qui influence la valeur de dl_{App}. Dans tous les cas, dl_{App} est supérieur en moyenne à dl_{DEPE}. Conséquemment, les longueurs moyennes relatives des intervalles de confiance sont presque toujours plus petites avec dl_{App} comparativement à dl_{DEPE}. Cela implique également que la couverture estimée pour dl_{DEPE} est généralement égale ou supérieure à celle estimée pour dl_{App}. La couverture estimée avec dl_{DEPE} est généralement près de 95 % (entre 93,4 % et 95,5 %) tandis que la couverture estimée avec dl_{App} est plus variable et est inférieure ou égale à 95 % (entre 83,8 % et 95,0 %).

Dans le cas du plus grand domaine (10 %), la couverture observée est globalement près du seuil nominal autant pour dl_{DEPE} que pour dl_{App}. C'est surtout dans les cas des petits domaines que s'observent les problèmes de sous-couverture avec dl_{App}. La performance de dl_{App} est la plupart du temps assez bonne sous l'indépendance intra-ménage (tableau 7.2). Cependant, en présence de dépendance intra-ménage (tableau 7.3), dl_{App} montre une sous-couverture importante pour les variables Weibull et Gamma. Dans ces cas, la couverture estimée est nettement améliorée avec l'utilisation de dl_{DEPE} comparativement à dl_{App}. Par exemple, pour la variable Gamma, pour le domaine de 3 % avec dépendance intra-ménage, la couverture estimée passe de 87,6 % avec dl_{App} à 94,4 % avec dl_{DEPE} et la longueur relative moyenne des intervalles de confiance passe de 113,6 à 163,2 %, soit une augmentation de 49,6 points de pourcentage. Cet exemple précis illustre bien le danger d'une surestimation du nombre de degrés de liberté. Cela laisse croire que les estimations sont plus précises qu'elles ne le sont réellement étant donné des intervalles de confiance trop courts. L'objectif d'utiliser une borne inférieure de 2,5 degrés de liberté pour le nombre théorique de degrés de liberté a été atteint puisqu'aucune sur-couverture importante n'a été observée. Finalement, rien n'indique que le calage effectué ait eu un impact significatif sur les degrés de liberté puisque la couverture estimée demeure près de 95 % dans tous les scénarios avec la formule théorique des degrés de liberté. Ceci suggère que les expressions théoriques développées restent appropriées même en présence de calage, du moins dans le contexte étudié.

7.3 Résultats de simulations pour l'estimation d'un effectif

La majorité des estimations produites à partir des réponses du questionnaire détaillé du recensement estiment un effectif. Dans cette section, une étude de simulations pour les estimations d'effectifs est présentée. Les valeurs possibles de la variable d'intérêt y_{hij} sont 0 et 1 et celles-ci ont été générées selon une loi de Bernoulli avec $P(y_{hij}=1 \mid z_{hij}=1)=p$ pour tous les individus de la population. Pour chacun des trois domaines possibles, 99 variables dichotomiques correspondantes à $p=1,\ldots,$ 99 % ont été générées.

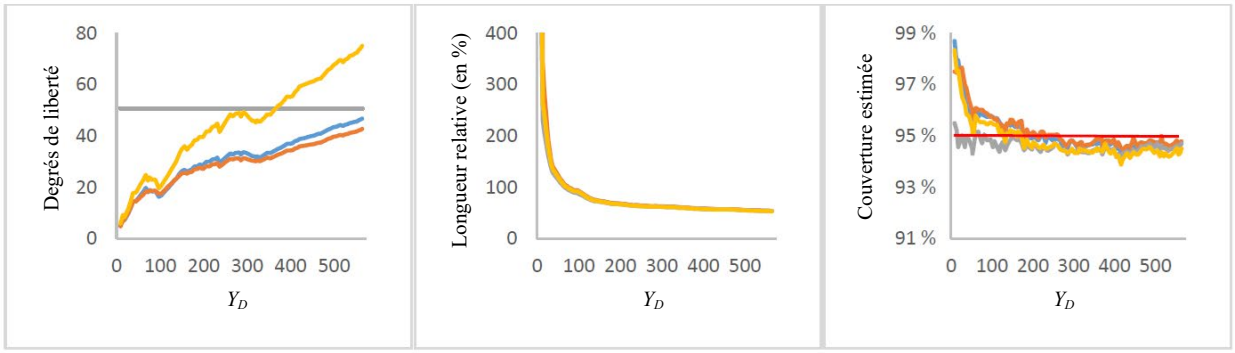
Des intervalles de confiance de type Wilson modifié, tels que donné à l'équation (4.3), ont été calculés en utilisant dl_{DEPE}, $\widehat{dl}_{DEPE}$, dl_{cl} et dl_{App}. Ici, toutes les strates contiennent au moins un individu du domaine, et ce pour les trois domaines générés. Ainsi, l'intervalle de confiance de Wilson modifié est calculé avec $M_I = M$ et $m_I = m.$ À noter que pour certains échantillons simulés (lorsque p est très petite et la taille du domaine est petite), l'effectif estimé était nul. Dans ces cas, l'estimateur de la variance $v_{DEPE}(\hat{Y}_D)$ est aussi nul. Par exemple, pour le domaine de 5 %, ceci a été observé pour environ 0,3 % de tous les échantillons simulés. Pour la simplicité, ces cas ont été retirés de l'étude de simulations et la couverture a été estimée à partir des autres échantillons.

La figure 7.2 présente les résultats obtenus pour le domaine de 5 %; les résultats des domaines de 3 % et de 10 % étant similaires, ils ne sont pas présentés ici. La figure 7.2 illustre les nombres théoriques de degrés de liberté dl_{DEPE} et dl_{cl} ainsi que $\overline{\widehat{dl}}_{DEPE}$ et $\overline{dl}_{App}$, c'est-à-dire les valeurs moyennes de $\widehat{dl}_{DEPE}$ et dl_{App} pour les 3 000 échantillons. Il compare aussi la longueur relative moyenne et la couverture estimée des intervalles de confiance en fonction de l'effectif Y_D.

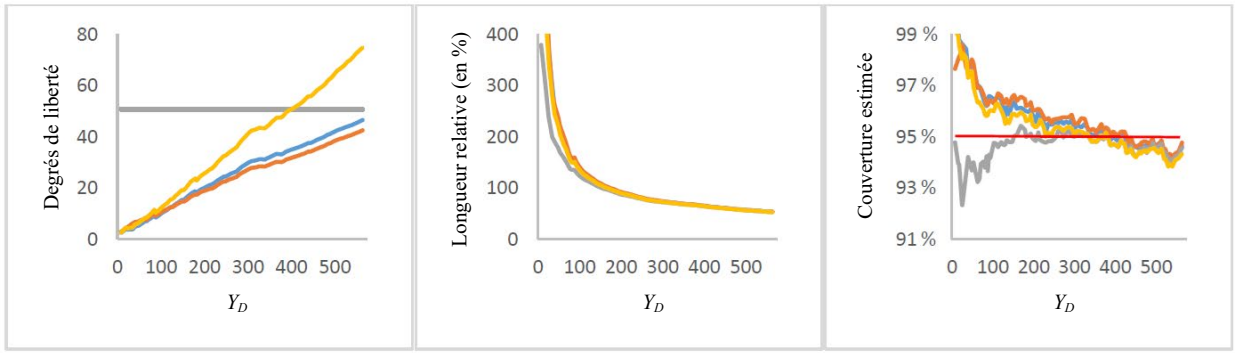
Figure 7.2 Degrés de liberté dl_{DEPE}, $\widehat{dl}_{DEPE}$, dl_{cl} et $\overline{dl}_{App}$ (à gauche), longueur relative moyenne (au milieu) et couverture estimée (à droite) des intervalles de confiance de Wilson modifié pour l'estimation d'un effectif Y_D pour le domaine de 5 %;

dl_{DEPE} (—), $\widehat{dl}_{DEPE}$ (—), dl_{cl} (—), et dl_{App} (—)

(a) sous l'indépendance intra-ménage



(b) en présence de dépendance intra-ménage



Note : Demi-échantillons partiellement équilibrés (DEPE).

Du côté du nombre de degrés de liberté (figure 7.2, à gauche), on peut voir que dl_{DEPE} augmente à mesure que l'effectif Y_D augmente et la valeur de dl_{DEPE} est plus petite en présence de dépendance intra-ménage que sous l'indépendance pour une même valeur de Y_D. Du côté du nombre de degrés de liberté approximatif dl_{App}, celui-ci n'est ni influencé par la taille de l'effectif au niveau de la population, ni influencé par la présence de dépendance intra-ménage. Ces premières observations sont similaires à ce qui a été observé à la section 7.2 pour les variables continues. Ces comportements ne dépendent donc pas de la nature de la variable d'intérêt.

Dans tous les cas, l'estimateur $\widehat{dl}_{DEPE}$ semble estimer assez efficacement la valeur théorique dl_{DEPE}. En effet, le nombre moyen de degrés de liberté estimé $\widehat{dl}_{DEPE}$ est près de la valeur théorique dl_{DEPE} pour tous les scénarios. La différence entre dl_{cl} et dl_{DEPE} s'accentue lorsque l'effectif Y_D augmente. Ainsi, la perte de degrés de liberté liée à l'utilisation de la méthode des DEPE-ε, comparativement à la méthode classique, est plus importante lorsque les effectifs sont élevés.

Du côté de la couverture estimée (figure 7.2, à droite), on note que la couverture obtenue avec les degrés de liberté estimés $\widehat{dl}_{DEPE}$ est assez semblable à celle observée avec les degrés de liberté théoriques dl_{DEPE}.

Sous l'indépendance intra-ménage et lorsque l'effectif Y_D est grand, la couverture estimée avec les quatre versions des degrés de liberté est semblable, généralement assez près du taux nominal de 95 %. En revanche, en présence de dépendance intra-ménage, la couverture estimée avec dl_{App} chute sous les 95 % lorsque l'effectif est petit. Inversement, dl_{DEPE} et $\widehat{dl}_{DEPE}$ ne montrent aucune sous-couverture importante dans tous les scénarios considérés. Lorsque l'effectif est faible, tant dans les situations de dépendance que d'indépendance intra-ménage, une sur-couverture est observée du côté de dl_{DEPE} et de $\widehat{dl}_{DEPE}$, ce qui renforce l'intérêt de leur utilisation pour les estimations sur petits domaines. Cette sur-couverture pourrait s'expliquer par le fait que le calage n'a pas été pris en compte dans la formule théorique du nombre de degrés de liberté.

Sous l'indépendance intra-ménage, les longueurs relatives moyennes (figure 7.2, au milieu) des intervalles de confiance sont très semblables pour les quatre versions des degrés de liberté. Cependant, lorsque Y_D est petit et en présence de dépendance intra-ménage, les longueurs relatives des intervalles de confiance sont légèrement plus petites pour dl_{App} que pour les trois autres alternatives ce qui concorde avec les cas où dl_{App} engendre une sous-couverture.

8. Conclusion

Pour le questionnaire détaillé du Recensement de 2021, Statistique Canada a publié des intervalles de confiance afin de refléter la fiabilité des estimations. La longueur de ces intervalles est un bon indicateur de la fiabilité, à condition que leur couverture respecte le taux nominal. Cet article propose d'utiliser une formule plus précise des degrés de liberté, en recourant à l'approximation de Satterthwaite, afin de corriger des problèmes de sous-couverture observés avec la règle approximative. La formule générale décrite à l'équation (2.1), adaptable à différents estimateurs, plans de sondage et méthodes d'estimation de la variance, montre que plus l'estimateur de la variance est instable, plus le nombre de degrés de liberté diminue. Cependant, obtenir une formule explicite peut être laborieux, notamment pour des estimateurs complexes tels que les quantiles.

Des expressions explicites du nombre théorique de degrés de liberté sont obtenues, sous un plan stratifié aléatoire simple sans remise de grappes, pour deux méthodes d'estimation de la variance : la méthode classique et la méthode des DEPE-ε. Les expressions algébriques obtenues, dérivées pour l'estimation d'un total sur un domaine, peuvent être généralisées à tout estimateur linéaire. La formule des degrés de liberté a été obtenue pour la méthode classique. Elle s'applique également à toute méthode d'estimation de la variance par répliques identique à l'estimateur de variance classique dans le cas linéaire. C'est notamment le cas de la méthode du bootstrap, qui est algébriquement équivalente à l'estimateur classique lorsque le nombre de répliques est infini.

Une version simplifiée de la formule liée à la méthode des DEPE-ε met en évidence les facteurs influençant les degrés de liberté. Pour limiter la perte de degrés de liberté, il est notamment recommandé de privilégier un grand nombre de strates artificielles et un petit nombre de groupes d'équilibrage, ce qui implique un nombre accru de répliques.

Les simulations montrent que l'utilisation du nombre théorique de degrés de liberté corrige la sous-couverture causée par la règle approximative dans plusieurs situations, notamment pour des domaines de petite taille et en présence de dépendance intra-ménage. Dans certains cas limites, une sur-couverture peut apparaître avec les degrés de liberté théoriques, qui s'expliquerait par une violation des hypothèses sous-jacentes à l'approximation de Satterthwaite. Une sur-couverture étant moins problématique qu'une sous-couverture, cela soutient l'usage du nombre théorique de degrés de liberté plutôt que celui de la règle approximative. Une version empirique des degrés de liberté théoriques qui consiste à remplacer les moments centrés d'ordre 2 et 4 par des estimations donne de bons résultats pour l'estimation d'un effectif, mais tend à surestimer dans certains cas la vraie valeur du nombre de degrés de liberté pour les totaux de variables continues.

Finalement, bien que l'utilisation du nombre théorique de degrés de liberté soit avantageuse, leur calcul est beaucoup plus complexe que celui de la règle approximative. Actuellement, le système de diffusion du recensement ne permet pas son implémentation. Celle-ci nécessiterait une modification importante de ce système.

Remerciements

Nous tenons à remercier Claude Girard, François Verret et Jean-François Beaumont de Statistique Canada ainsi que les réviseurs de la revue Technique d'enquête pour leur relecture attentive et leurs précieux commentaires ayant permis d'améliorer la présente version de cet article. Merci également au fond commun de recherche en méthodologie de Statistique Canada pour le soutien financier ayant permis de réaliser ce projet.

Annexe

A. Preuve que $E_*\left\{v_{DEPE}(\hat{Y}_D)\right\} = v_{cl}$

Dans un premier temps, une forme alternative de $v_{DEPE}(\hat{Y}_D)$ qui facilitera les calculs pour la suite est obtenue. Soient $\tilde{w}_h = w_h(w_h - 1)$ et $\Delta_{hk} = \sqrt{n_{hk1}n_{hk2}}(\bar{t}_{Dhk1} - \bar{t}_{Dhk2})$, où $\bar{t}_{Dhk1} = t_{Dhk1}/n_{hk1}$ et $\bar{t}_{Dhk2} = t_{Dhk2}/n_{hk2}$. Ainsi, on obtient

$$v_{DEPE}(\hat{Y}_D) = \sum_{h=1}^{H}\sum_{h'=1}^{H}\sum_{k=1}^{K_h}\sum_{k'=1}^{K_{h'}}\sqrt{\tilde{w}_h\tilde{w}_{h'}}\Delta_{hk}\Delta_{h'k'}\frac{1}{R}\sum_{r=1}^{R}\delta_{hk}^{(r)}\delta_{h'k'}^{(r)}.$$

Soit $l_{hk} \in \{1, \ldots, L\}$ l'indice de la strate artificielle (h,k) dans son groupe d'équilibrage tel qu'illustré à la figure 5.1. À cause des conditions 2 et 3 décrites à la section 5.3, on a que

$$\frac{1}{R}\sum_{r=1}^{R}\delta_{hk}^{(r)}\delta_{h'k'}^{(r)} = \begin{cases} 1 & \text{lorsque } l_{hk} = l_{h'k'} \\ 0 & \text{sinon} \end{cases}.$$

Il s'ensuit que $\sum_{r=1}^{R} \delta_{hk}^{(r)} \delta_{h'k'}^{(r)} / R = \mathrm{I}(l_{hk} = l_{h'k'})$, où $\mathrm{I}(l_{hk} = l_{h'k'}) = 1$ lorsque $l_{hk} = l_{h'k'}$ et $\mathrm{I}(l_{hk} = l_{h'k'}) = 0$ lorsque $l_{hk} \neq l_{h'k'}$. Ainsi,

$$v_{DEPE}(\hat{Y}_D) = \sum_{h=1}^{H} \sum_{h'=1}^{H} \sum_{k=1}^{K_h} \sum_{k'=1}^{K_{h'}} \sqrt{\tilde{w}_h \tilde{w}_{h'}}\ \Delta_{hk} \Delta_{h'k'} \mathrm{I}(l_{hk} = l_{h'k'}).$$

À noter que cette somme contient exactement $K + 2L \times \binom{G}{2} = GK$ éléments non nuls, où G est le nombre de groupes d'équilibrage, K est le nombre de strates artificielles au total et $L = K / G$. Lorsque $h = h'$, on a que $\mathrm{I}(l_{hk} = l_{h'k'}) = 1$ lorsque $k = k'$ tandis que $\mathrm{I}(l_{hk} = l_{h'k'}) = 0$ lorsque $k \neq k'$. Ainsi

$$v_{DEPE}(\hat{Y}_D) = \sum_{h=1}^{H} \sum_{k=1}^{K_h} \tilde{w}_h \Delta_{hk}^2 + \sum_{h \neq h'=1}^{H} \sum_{k=1}^{K_h} \sum_{k'=1}^{K_{h'}} \sqrt{\tilde{w}_h \tilde{w}_{h'}}\ \Delta_{hk} \Delta_{h'k'} \mathrm{I}(l_{hk} = l_{h'k'}). \qquad \text{(A.1)}$$

Le processus aléatoire * peut être séparé en deux processus aléatoires indépendants :

1- la séparation des n_h unités de l'échantillon de la strate h en K_h strates artificielles et la distribution des n_{hk} unités de la strate artificielle (h, k) en deux grappes s_{hk1} et s_{hk2} et
2- la création des groupes d'équilibrage avec la numérotation des $l_{hk} \in \{1, \ldots, L\}$ et le choix des grappes via $\delta_{hk}^{(r)}$.

On notera $E_* = E_{12}$ pour représenter ces deux processus. En utilisant le résultat de l'équation (A.1), on a

$$E_*\left\{v_{DEPE}(\hat{Y}_D)\right\} = \sum_{h=1}^{H} \sum_{k=1}^{K_h} \tilde{w}_h E_1(\Delta_{hk}^2) + \sum_{h \neq h'=1}^{H} \sum_{k=1}^{K_h} \sum_{k'=1}^{K_{h'}} \sqrt{\tilde{w}_h \tilde{w}_{h'}}\ E_1(\Delta_{hk} \Delta_{h'k'})\, E_2\left\{\mathrm{I}(l_{hk} = l_{h'k'})\right\}.$$

Premièrement, $E_1(\Delta_{hk}\Delta_{h'k'}) = E_1(\Delta_{hk})\, E_1(\Delta_{h'k'})$ pour $h \neq h'$ et $E_1(\Delta_{hk}) = E_1(\Delta_{h'k'}) = 0$. Ainsi, $E_1(\Delta_{hk}\Delta_{h'k'}) = 0$. Deuxièmement, un peu d'algèbre nous conduit à

$$E_1(\Delta_{hk}^2) = \frac{n_{hk}}{n_h(n_h - 1)} \left(n_h \sum_{i=1}^{n_h} t_{Dhi}^2 - \left(\sum_{i=1}^{n_h} t_{Dhi} \right)^2 \right) = n_{hk} S_{Dh}^2.$$

Par conséquent,

$$E_*\left\{v_{DEPE}(\hat{Y}_D)\right\} = \sum_{h=1}^{H} \sum_{k=1}^{K_h} \tilde{w}_h n_{hk} S_{Dh}^2 = \sum_{h=1}^{H} N_h \left(\frac{N_h - n_h}{n_h} \right) S_{Dh}^2 = v_{cl}.$$

B. Preuve de la proposition 6.1

Le résultat est une conséquence directe de l'équation (2.1). On a que

$$V_{2,cl} = \sum_{h=1}^{H} N_h^2 \left(\frac{N_h - n_h}{n_h} \right)^2 \mathrm{Var}_P(S_{Dh}^2).$$

La forme de $\mathrm{Var}_P(S_{Dh}^2)$ peut être trouvée notamment dans Cho, Cho and Eltinge (2005).

C. Preuve de la proposition 6.2

À partir du résultat de l'équation (2.1), il suffit de développer algébriquement $V_{2,DEPE} = \mathrm{Var}_{P,*}\left\{v_{DEPE}(\hat{Y}_D)\right\}$. Puisque $v_{DEPE}(\hat{Y}_D)$ est sans biais pour $\mathrm{Var}_P(\hat{Y}_D)$ donné à l'équation (3.5), on a

$$V_{2,DEPE} = E_{P,*}\left\{(v_{DEPE}(\hat{Y}_D))^2\right\} - \left(\sum_{h=1}^{H} \frac{N_h n_h \tilde{w}_h}{(N_h - 1)} \sigma_{Dh}^2\right)^2.$$

En utilisant le résultat de l'équation (A.1), on obtient $E_{P,*}\left\{(v_{DEPE}(\hat{Y}_D))^2\right\} = S_1 + 2S_2 + S_3,$ où

$$S_1 = \sum_{h,h'=1}^{H} \tilde{w}_h \tilde{w}_{h'} \sum_{k=1}^{K_h} \sum_{k'=1}^{K_{h'}} E_{P,1}(\Delta_{hk}^2 \Delta_{h'k'}^2)$$

$$S_2 = \sum_{h_1=1}^{H} \sum_{h_2 \neq h_2'=1}^{H} \tilde{w}_{h_1} \sqrt{\tilde{w}_{h_2} \tilde{w}_{h_2'}} \sum_{k_1=1}^{K_{h_1}} \sum_{k_2=1}^{K_{h_2}} \sum_{h_2'=1}^{K_{h_2'}} E_{P,1}(\Delta_{h_1 k_1}^2 \Delta_{h_2 k_2} \Delta_{h_2' k_2'}) \, E_2\left\{\mathrm{I}(l_{h_2 k_2} = l_{h_2' k_2'})\right\}$$

et

$$S_3 = \sum_{h_1 \neq h_1'} \sum_{h_2 \neq h_2'} \sqrt{\tilde{w}_{h_1} \tilde{w}_{h_1'} \tilde{w}_{h_2} \tilde{w}_{h_2'}} \times \sum_{k_1=1}^{K_{h_1}} \sum_{k_1'=1}^{K_{h_1'}} \sum_{k_2=1}^{K_{h_2}} \sum_{k_2'=1}^{K_{h_2'}} E_{P,1}(\Delta_{h_1 k_1} \Delta_{h_1' k_1'} \Delta_{h_2 k_2} \Delta_{h_2' k_2'}) \, E_2\left\{\mathrm{I}(l_{h_1 k_1} = l_{h_1' k_1'}, l_{h_2 k_2} = l_{h_2' k_2'})\right\}.$$

De la même façon qu'à l'Annexe A, le processus aléatoire * a été séparé en deux processus indépendants; les indices 1 et 2 réfèrent respectivement aux processus aléatoires de création des strates artificielles et de création des groupes d'équilibrage.

Premièrement, notons que $E_{P,1}(\Delta_{hk}) = 0$ et que Δ_{hk} et $\Delta_{h'k'}$ sont indépendants dès que $h \neq h'$. Deuxièmement, $E_{P,1}(\Delta_{hk_1} \Delta_{hk_2}) = 0$ aussitôt que $k_1 \neq k_2,$ puisque

$$E_{P,1}(\overline{t}_{Dhk_1 1} \overline{t}_{Dhk_2 1}) = E_{P,1}(\overline{t}_{Dhk_1 1} \overline{t}_{Dhk_2 2}) = E_{P,1}(\overline{t}_{Dhk_1 2} \overline{t}_{Dhk_2 1}) = E_{P,1}(\overline{t}_{Dhk_1 2} \overline{t}_{Dhk_2 2}).$$

Ces arguments seront utilisés dans les développements qui suivent.

Développement de S_1 : Puisque Δ_{hk} et $\Delta_{h'k'}$ sont indépendants dès que $h \neq h',$ on a

$$S_1 = \sum_{h=1}^{H} \tilde{w}_h^2 \left\{\sum_{k=1}^{K_h} E_{P,1}(\Delta_{hk}^4) + \sum_{k \neq k'} E_{P,1}(\Delta_{hk}^2 \Delta_{hk'}^2)\right\} + \sum_{h \neq h'} \tilde{w}_h \tilde{w}_{h'} \sum_{k=1}^{K_h} \sum_{k'=1}^{K_{h'}} E_{P,1}(\Delta_{hk}^2) \, E_{P,1}(\Delta_{h'k'}^2).$$

Développement de S_2 : Puisque $\Delta_{h_2 k_2}$ et $\Delta_{h_2' k_2'}$ sont indépendants pour $h_2 \neq h_2'$ et que $E_{P,1}(\Delta_{h_2 k_2}) = E_{P,1}(\Delta_{h_2' k_2'}) = 0,$ on a que $E_{P,1}(\Delta_{h_1 k_1}^2 \Delta_{h_2 k_2} \Delta_{h_2' k_2'}) = 0$ dans tous les cas et ainsi, $S_2 = 0.$

Développement de S_3 : Pour $h_1 \neq h_1'$ et $h_2 \neq h_2',$ le seul cas où $E_{P,1}(\Delta_{h_1 k_1} \Delta_{h_1' k_1'} \Delta_{h_2 k_2} \Delta_{h_2' k_2'})$ n'est pas nulle est lorsque $h_1 = h_2,$ $h_1' = h_2',$ $k_1 = k_2$ et $k_1' = k_2'.$ Puisque $E_2\left\{\mathrm{I}(l_{hk} = l_{h'k'})\right\} = (G-1)/(K-1),$ on obtient donc

$$S_3 = \frac{G-1}{K-1} \sum_{h \neq h'} \tilde{w}_h \tilde{w}_{h'} \sum_{k=1}^{K_h} \sum_{k'=1}^{K_{h'}} E_{P,1}(\Delta_{hk}^2) \, E_{P,1}(\Delta_{h'k'}^2).$$

Un long développement algébrique conduit à $E_{P,1}(\Delta_{hk}^2 \Delta_{hk'}^2) = n_{hk} n_{hk'} g_2(h),$

$$E_{P,1}(\Delta_{hk}^2) = \frac{n_{hk} N_h}{N_h - 1}\sigma_{Dh}^2 \quad \text{et} \quad E_{P,1}(\Delta_{hk}^4) = \left(\frac{n_{hk2}^2}{n_{hk1}} + \frac{n_{hk1}^2}{n_{hk2}}\right) g_1(h) + 3n_{hk}^2 g_2(h),$$

où

$$g_1(h) = N_h^2 \frac{(N_h + 1)\, m_{Dh4} - 3(N_h - 1)\,\sigma_{Dh}^4}{(N_h - 1)(N_h - 2)(N_h - 3)} \quad \text{et} \quad g_2(h) = N_h \frac{(N_h^2 - 3N_h + 3)\,\sigma_{Dh}^4 - (N_h - 1)\, m_{Dh4}}{(N_h - 1)(N_h - 2)(N_h - 3)}.$$

Par conséquent, on obtient

$$S_1 = \sum_{h=1}^{H} \tilde{w}_h^2 \left\{ \left\{ g_1(h)\, B_h + g_2(h)\,(2A_h + n_h^2) \right\} \right\} + \sum_{h \neq h'} \frac{\tilde{w}_h \tilde{w}_{h'} n_h n_{h'} N_h N_{h'}}{(N_h - 1)(N_{h'} - 1)} \sigma_{Dh}^2 \sigma_{Dh'}^2$$

et

$$S_3 = \frac{G - 1}{K - 1} \sum_{h \neq h'} \frac{\tilde{w}_h \tilde{w}_{h'} n_h n_{h'} N_h N_{h'}}{(N_h - 1)(N_{h'} - 1)} \sigma_{Dh}^2 \sigma_{Dh'}^2$$

où A_h et B_h sont tels que définis à l'équation (6.1). Conséquemment,

$$V_{2,DEPE} = \sum_{h=1}^{H} \tilde{w}_h^2 \left\{ \left\{ g_1(h)\, B_h + g_2(h)\,(2A_h + n_h^2) \right\} - \frac{N_h^2 n_h^2}{(N_h - 1)^2} \sigma_{Dh}^4 \right\} + \frac{G - 1}{K - 1} \sum_{h \neq h'} \frac{\tilde{w}_h \tilde{w}_{h'} n_h n_{h'} N_h N_{h'}}{(N_h - 1)(N_{h'} - 1)} \sigma_{Dh}^2 \sigma_{Dh'}^2.$$

Finalement, un développement algébrique permet d'obtenir le résultat recherché.

Bibliographie

Cho, E., Cho, M.J. et Eltinge, J. (2005). The variance of the variance of samples from a finite population. *International Journal of Pure and Applied Mathematics*, 21(1).

Devin, N. et Verret, F. (2016). The development of a variance estimation methodology for large-scale dissemination of quality indicators for the 2016 Canadian census long form sample. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Chicago, États-Unis.

Gerlovina, I. et Hubbard, A.E. (2019). Computer algebra and algorithms for unbiased moment estimation of arbitrary order. *Cogent Mathematics & Statistics*, 6(1).

Hedayat, A. et Wallis, W. (1978). Hadamard matrices and their applications. *The Annals of Statistics*, 6(6), 1184-1238.

Heeringa, S.G., West, B.T. et Berglund, P.A. (2010). *Applied Survey Data Analysis*. Chapman and hall/CRC.

Judkins, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, Statistics Sweden, 6(3), 223-239.

Kott, P.S. (2020). *The Degrees of Freedom of a Variance Estimator in a Probability Sample*. RTI Press.

Kott, P.S. et Carr, D.A. (1997). Developing an estimation strategy for a pesticide data program. *Journal of Official Statistics*, 13(4), 367-383.

McCarthy, P.J. (1966). Replication: An approach to the analysis of data from complex surveys. *Vital and Health Statistics, Data Evaluation and Methods Research*, Series 2, 14, 1-38.

Nelsen, R.B. (2006). *An Introduction to Copulas*. Springer Series in Statistics. New York: Springer, 2nd ed.

Neusy, E., Savard, S.-A., Hidiroglou, M.A. et Martin, V. (2021). Modified Wilson intervals for estimated counts with application to Census 2021 long form estimation. Présentation au Comité consultatif sur les méthodes statistiques, mai 2021. Document interne, Statistique Canada.

Rao, J.N.K. et Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, Oxford University Press, 86(2), 403-415.

Rust, K.F. (1986). Efficient replicated variance estimation. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 81-87.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6), 110-114.

Valliant, R. et Rust, K.F. (2010). Degrees of freedom approximations and rules-of-thumb. *Journal of Official Statistics*, 26(4), 585-602.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209-212.

Wolter, K. (2007). *Introduction to Variance Estimation*. Springer Science & Business Media.