Estimation de la variance pour l'enquête longitudinale Histoire de vie et Patrimoine : une comparaison des approches analytique et par *bootstrap*

Document de travail

N°M2025-07-Octobre 2025

Insee
Mesurer pour comprendre

Khaled LARBI
Jean RUBIN

Institut national de la statistique et des études économiques

2025/07

Estimation de la variance pour l'enquête longitudinale Histoire de vie et Patrimoine : une comparaison des approches analytique et par *bootstrap*

Khaled LARBI
Jean RUBIN

Octobre 2025

Remerciements :

Les auteurs souhaitent remercier Emmanuel Gros, Éric Lesage, Marine Guillerm, Pauline Givord et Corinne Prost pour leur relecture attentive de ce document de travail ainsi que Guillaume Chauvet et Olivier Guin pour les différents échanges autour des méthodes de bootstrap en sondage.

Direction de la méthodologie et de la coordination statistique et internationale
Timbre L001
88 Avenue Verdier - CS 70058 - 92541 Montrouge Cedex - France -
Tél. : 33 (1) 87 69 55 00 - E-mail : DG75-L001@insee.fr - Site Web Insee : http://www.insee.fr

Résumé

L'enquête Histoire de vie et Patrimoine (HVP) fait partie d'un ensemble d'enquêtes au niveau européen (Household Finance and Consumption Survey) ayant notamment pour but d'étudier l'évolution du patrimoine des ménages. La diffusion de ces données dans le cadre de comparaisons internationales rend d'autant plus nécessaire de fournir des moyens d'estimer la précision des indicateurs construits. Cette enquête intègre en effet de nombreux mécanismes d'échantillonnages, pouvant rendre difficile pour un utilisateur de quantifier les incertitudes sous-jacentes sans l'aide du producteur. On peut par exemple citer sa structure en panel rotatif qui permet de réaliser des analyses à la fois longitudinales et transversales des données, mais qui implique que l'observation d'un individu à une date donnée puisse être le fruit d'un tirage sur des périodes précédentes. Ce document propose ainsi deux approches d'estimation de la variance : la première s'appuie sur des formules analytiques d'approximation de variance pour chaque mécanisme du plan de sondage et la seconde s'appuie sur une approche par rééchantillonnage des unités primaires. Les méthodes sont ensuite appliquées et comparées sur l'enquête HVP **2020.**

Mots clés : Théorie des sondages après collecte, Estimation de précision, Méthode Analytique, Bootstrap

Classification JEL : C13, C83

Table des matières

1 Introduction

La mesure de la précision des résultats issus d'enquêtes statistiques est un enjeu crucial, tant pour la diffusion que pour l'analyse des données. En effet, la précision des estimations est essentielle pour évaluer la qualité du processus de production statistique, permettant de s'assurer que les conclusions tirées des enquêtes sont fiables. Cette exigence se manifeste dans les rapports communiqués périodiquement à Eurostat ainsi que dans les règlements européens tels que le règlement Integrated European Social Statistics (IESS). De plus, ces indicateurs de précision sont de plus en plus utilisés pour guider l'interprétation des résultats d'enquête, notamment lorsque les données sont ventilées selon des critères (géographique, par exemple).

L'imprécision dans les résultats d'enquête découle de diverses sources d'erreur qui peuvent être classées en deux grandes catégories : les erreurs d'échantillonnage et les erreurs non liées à l'échantillonnage. Les erreurs d'échantillonnage sont dues au fait que les estimations sont basées sur un échantillon de la population plutôt que sur la population entière. Elles se manifestent principalement sous forme de variance dans les résultats, mais n'introduisent pas, sous certaines hypothèses, de biais systématique, ce qui signifie que, en moyenne, les estimations tendent à se rapprocher des vrais paramètres de la population. Pour l'enquête Histoire de vie et Patrimoine 2020, le calcul de la précision des estimations prend en compte principalement les erreurs d'échantillonnage et la non-réponse totale, tout en incorporant les ajustements statistiques comme le calage sur marges pour améliorer l'efficacité des estimateurs.

Ce document propose une analyse comparative entre deux méthodes d'estimation de la variance : l'approche analytique classique et l'approche par bootstrap. L'objectif est de fournir une évaluation complète de la précision des résultats de l'enquête Histoire de vie et Patrimoine 2020, tout en tenant compte des différentes sources d'erreur qui peuvent affecter les estimations. Cette comparaison a pour but d'apporter un éclairage supplémentaire sur les méthodes d'estimation de la variance utilisées, afin de mieux comprendre leurs avantages et leurs limites dans le contexte de l'enquête Histoire de vie et Patrimoine.

Cet exercice s'inscrit dans la continuité d'évaluations réalisées pour l'enquête Patrimoine 2010 (Lamarche et Salembier 2015) et pour l'enquête Patrimoine 2014 (Chevalier 2016).

2 L'enquête Histoire de vie et Patrimoine 2020

L'enquête Histoire de vie et Patrimoine (HVP) 2020 a pour objectif de fournir une analyse approfondie des actifs financiers, immobiliers et professionnels des ménages français, ainsi que de leur endettement associé. Elle permet de suivre dans le temps la distribution du patrimoine au sein des ménages et les taux de détention des différents actifs patrimoniaux. En outre, l'enquête offre des informations détaillées sur les facteurs expliquant la formation du patrimoine, incluant la biographie familiale et professionnelle des répondants, les héritages et donations reçus, les revenus et la situation financière des ménages, ainsi que les motifs de détention ou de non-détention des divers actifs.

La collecte des données pour l'enquête Histoire de vie et Patrimoine 2020 a eu lieu entre septembre 2020 et décembre 2020. Un total de 15 820 logements a été tiré au cours de cette période, permettant de disposer d'un échantillon *représentatif* à l'échelle nationale. L'enquête se concentre sur le ménage comme unité d'enquête principale, tandis que le suivi longitudinal se fait au niveau des individus. Le plan de sondage repose sur trois échantillons distincts : un échantillon issu de l'enquête Patrimoine 2014-2015, un autre de l'enquête 2017-2018, et un échantillon entrant, tiré des fichiers fiscaux pour l'année 2020. Cette approche combinée permet non seulement de suivre l'évolution du patrimoine des ménages sur une période prolongée, mais aussi d'intégrer de nouveaux participants pour enrichir la base de données et maintenir la représentativité transversale de l'échantillon.

L'enquête Histoire de vie et Patrimoine 2020 est réalisée en partenariat étroit avec la Banque de France, dans le cadre d'un projet européen plus large. Depuis 1986, cette enquête a été menée tous les six ans, avant de passer à une fréquence triennale, avec une dimension longitudinale, à partir de 2014. Cette cadence accrue permet de suivre de manière plus précise l'évolution des patrimoines des ménages et des taux de détention des actifs patrimoniaux. L'enquête de 2020, comme les vagues précédentes, constitue une référence unique pour la mesure détaillée du patrimoine des ménages en France, notamment en ce qui concerne le patrimoine professionnel et financier. Elle est la seule enquête à fournir une description exhaustive de ces éléments.

En complément des entretiens face-à-face, l'enquête inclut un dispositif de suivi entre chaque entretien. Ce suivi se fait sous la forme d'un questionnaire auto-administré, disponible en version papier ou en ligne, qui permet de maintenir le contact avec les enquêtés, de mettre à jour leurs coordonnées et d'introduire un module thématique supplémentaire. Ce dispositif contribue à limiter le taux d'attrition et à améliorer la qualité des données longitudinales.

Les données produites par l'enquête Histoire de vie et Patrimoine 2020 sont également utilisées pour alimenter la partie française du Household Finance and Consumption Survey (HFCS),

un dispositif d'harmonisation européen des enquêtes sur le patrimoine piloté par la Banque centrale européenne. Grâce à ce partenariat, les résultats de l'enquête HVP 2020 s'inscrivent également dans une perspective européenne, facilitant ainsi les analyses comparatives entre pays et l'élaboration de politiques économiques adaptées.

2.1 Description du plan de sondage

L'enquête HVP 2020 permet d'obtenir de l'information de deux natures différentes : longitudinale et transversale.

L'inférence longitudinale nécessite le suivi d'individus[1] au cours du temps. Ces individus, appelés **individus-panels**, sont tirés et réinterrogés tous les trois ans à trois reprises. Il est donc possible de calculer pour chaque individu-panel entré (et répondant à l'enquête) en $t-9$ des évolutions entre $t-9$ et $t-6$, entre $t-9$ et $t-3$, … Lorsqu'un individu-panel ne répond pas (ou n'est plus dans le champ de l'enquête) une année donnée, il quitte le dispositif et il n'est pas réinterrogé aux éditions suivantes. Afin de limiter le fardeau de réponse, l'échantillon est rotatif : une partie des individus-panels sont retirés du dispositif à chaque édition (ceux ayant été réinterrogés trois fois) et sont remplacés par de nouveaux individus-panels : il est donc nécessaire de tirer un nouvel échantillon d'individus-panels.

L'inférence transversale sur les ménages en t nécessite l'utilisation d'un échantillon représentatif de logements[2] en t.

Ces deux objectifs conduiraient à tirer deux échantillons : un échantillon d'individus-panels pour combler le retrait des individus-panels ayant déjà été réinterrogés trois fois et un échantillon de logements pour l'inférence transversale.

Néanmoins, il est possible de ne tirer qu'un seul échantillon pour atteindre ces deux objectifs :

- l'échantillon représentatif S^t_{log} de logements en t peut permettre de construire un échantillon des individus-panels entrants dans le dispositif en t : les individus entrants dans le panel sont tous les individus vivant dans un logement de l'échantillon S^t_{log}.
- étant donné que tous les individus-panels vont être réinterrogés en t pour l'inférence longitudinale, il est envisageable de leur poser des questions permettant l'inférence transversale. Ainsi, il est possible d'obtenir de l'information transversale en interrogeant les ménages de S^t_{log} tirés en t mais également les ménages contenant au moins un individu-panel entrés en $t-9$, $t-6$ ou en $t-3$ et répondants en t. Les ménages interrogés via ce suivi d'individus-panels ne sont pas nécessairement les mêmes que ceux tirés aux périodes précédentes, en raison de possibles séparations et unions ayant eu lieu entre temps. Cette

[1] Les indicateurs portent le plus souvent sur les ménages. Pour autant, il est plus simple de suivre les individus au cours du temps et de les interroger sur le patrimoine de leur ménage.

[2] La base de sondage dans laquelle les échantillons de l'enquête HVP sont tirés est une base de logements.

démarche permet d'accroître la taille de l'échantillon et ainsi d'améliorer les estimations transversales réalisées.

Le tirage d'un échantillon représentatif S^t_{log} de logements en t permet donc d'assurer l'estimation longitudinale en actualisant le panel (avec tous les individus vivant dans des logements de S^t_{log}) et d'effectuer des estimations transversales (en se basant sur l'échantillon S^t_{log} complété des ménages contenant au moins un individu-panel en réinterrogation en t)[3].

Dans la suite, nous décrivons le plan de sondage associé à l'échantillon S^t_{log} des logements en t. À partir de ce plan de sondage, il est possible de déduire l'ensemble des individus répondants $S^{t\to t}_{\text{ind,rep}}$ des individus-panels entrants dans le dispositif en t et l'ensemble des ménages répondants en t S^t_{men}.

Description de la modélisation de l'échantillon S^t_{log}

Le tirage de l'échantillon $S^{t\to t}_{\text{ind,rep}}$ des individus-panels entrants dans le dispositif HVP en t est réalisé en considérant tous les individus des logements tirés dans l'échantillon S^t_{log}. Cet échantillon est sélectionné selon un tirage à deux degrés. Cependant, nous introduisons un degré supplémentaire permettant de modéliser la non-réponse des ménages.

Le tirage des logements S^t_{log} est obtenu en tirant :

- un échantillon de regroupements de communes : afin de limiter le coût de collecte et de faciliter l'enquête, les communes de France métropolitaine sont regroupées en plus de 5 000 unités contiguës. Un échantillon de 541 unités primaires S_{UP} est tiré selon un plan stratifié régionalement et équilibré spatialement (Chevalier et al. 2022). La probabilité d'inclusion d'ordre 1 de l'unité primaire u est notée π^{UP}_u
- un échantillon de logements : dans chaque unité primaire, un échantillon de logements est tiré selon un plan stratifié en utilisant une typologie ad-hoc liée aux patrimoines[4]. Dans chaque strate, un échantillon de logements est tiré selon un tirage systématique conditionnellement aux unités primaires. La probabilité d'inclusion d'ordre 1 du logement l conditionnellement aux unités primaires tirées est notée $\pi^{\text{log}|\text{UP}}_l$.
- un échantillon de ménages répondants : il est possible que certains ménages refusent de répondre - ce phénomène est habituellement modélisé comme un plan poissonien conditionnellement à l'échantillon des logements tirés. Les probabilités d'inclusion de ce plan sont modélisées en utilisant des informations auxiliaires sur les logements de l'échantillon. La probabilité de réponse estimée du logement l est notée $\hat{p}_l$.

[3] Il est à noter que si un individu en réinterrogation en t (donc entré dans le dispositif en $t-9$, $t-6$ ou $t-3$) déménage alors seul son ménage en t est considéré dans l'échantillon (le ménage occupant le logement initial n'est pas retenu).

[4] Une typologie permettant d'identifier les ménages contenant de potentiels individus à hauts patrimoines est construite en mobilisant des informations issues de sources fiscales.

Le plan de sondage de S_{log}^{2020} diffère légèrement de celui de S_{log}^{2014} et S_{log}^{2017} : l'échantillon des unités primaires est tiré selon un autre plan de sondage (Gros et Moussallam 2015).

Description du plan de sondage de l'échantillon $S_{\text{ind,rep}}^{t\to t}$

Dans chaque logement répondant, tous les individus répondent à l'enquête (y compris par proxy) lors de la première interrogation. Les individus sont donc tirés selon un plan par grappe conditionnellement aux logements répondants : leur poids conditionnel vaut 1.

Description du plan de sondage des échantillons $S_{\text{ind,rep}}^{t-3\to t}$, $S_{\text{ind,rep}}^{t-6\to t}$ et $S_{\text{ind,rep}}^{t-9\to t}$

De la même manière que pour l'échantillon $S_{\text{ind,rep}}^{t\to t}$, l'échantillon $S_{\text{ind,rep}}^{t-3\to t}$ des individus-panels entrés en $t-3$ et répondants en t est le fruit d'un tirage à plusieurs degrés. Un échantillon de ménages S_{log}^{t-3} représentatif des ménages en $t-3$ est tiré en $t-3$. L'échantillon $S_{\text{ind,rep}}^{t-3\to t-3}$ des individus-panels entrants dans le dispositif en $t-3$ est obtenu en considérant tous les individus de l'échantillon des ménages répondants S_{log}^{t-3}.

Cependant, certains individus-panels entrés en $t-3$ ne répondent pas en t : ce mécanisme d'attrition est modélisé à l'aide d'un plan poissonien conditionnellement à l'échantillon des individus-panels. La probabilité de réponse estimée de l'individu-panel k à l'année t sachant qu'il a répondu à sa première interrogation est notée $\hat{\rho}_k$. En pratique, la probabilité de réponse est estimée entre chaque réinterrogation conditionnellement aux interrogations précédentes.

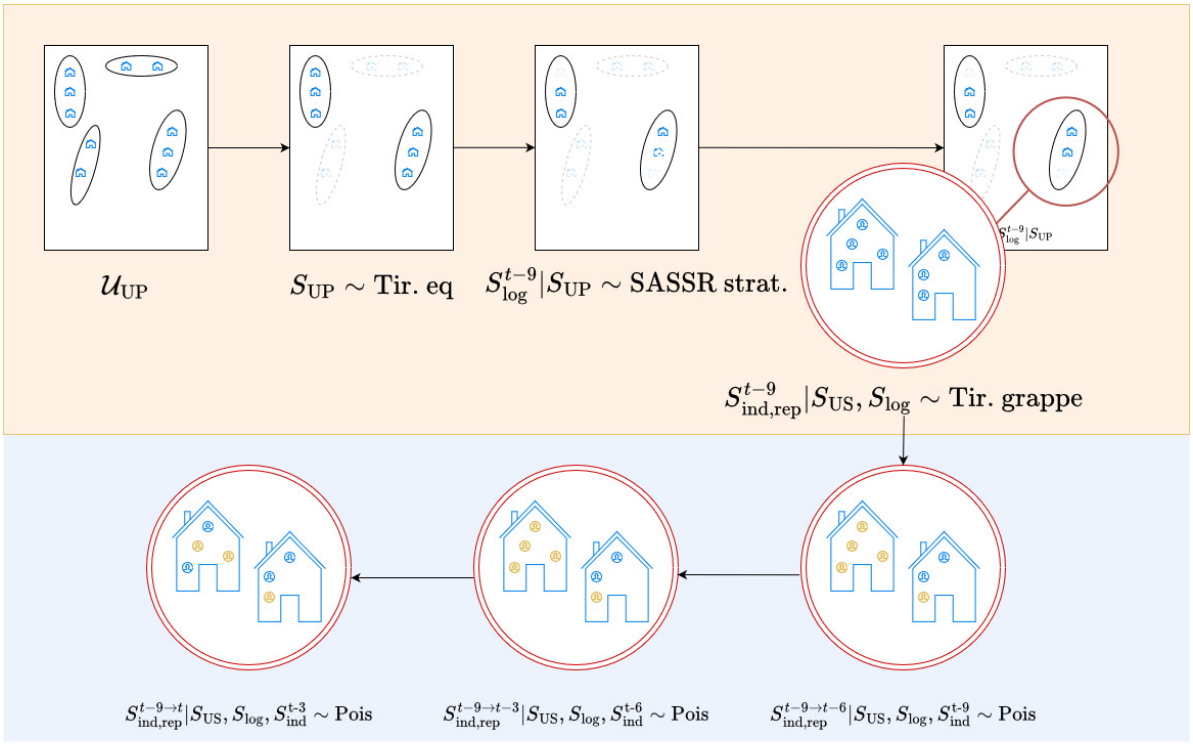


Figure 2.1: Suivi des individus-panels entrés en $t-9$. En orange, il s'agit des étapes de sélection en $t-9$: tirage d'unités primaires, de logements répondants et des individus. En bleu, il s'agit des étapes d'attrition. Les personnes jaunes dans les étapes d'attrition correspondent aux répondants.

$S_{\text{ind}}^{t-6\to t}$ et $S_{\text{ind}}^{t-9\to t}$ sont construits de manière analogue.

Description du plan de sondage de l'échantillon transversal des ménages S_{men}^t

L'échantillon final utilisé pour l'inférence transversale S_{men}^t est obtenu par sondage indirect à partir de l'échantillon $S_{\text{ind,rep}}^t$ des individus-panels entrés en $t-9$, $t-6$, $t-3$ ou t et répondants en t : les ménages de S_{men}^t sont les ménages contenant au moins un individu-panel de $S_{\text{ind,rep}}^t$.

2.2 Redressements et partage des poids

La pondération associée à un individu k de $S^t_{\text{ind,rep}}$ avant redressement est $w^{\text{ind}}_k = \frac{1}{\pi^{\text{UP}}_{u(k)} \pi^{\log|\text{UP}}_{l(k)} \hat{p}_l \hat{\rho}_k}$ où $u(k)$ désigne l'unité primaire dans laquelle le logement initial $l(k)$ de l'individu k se trouve. Afin d'assurer la cohérence avec d'autres sources et d'améliorer la précision des statistiques estimées, un calage sur marges est réalisé avant un partage des poids : la pondération des individus de $S^t_{\text{ind,rep}}$ est calée à l'aide d'informations socio-démographiques et fiscales au niveau individu[5]. La pondération après calage est notée $w^{\text{ind,cal}}_k$

Table 2.1: Variables de calage pour les indicateurs transversaux de l'enquête HVP 2020

Catégorie	Source
Sexe x âge (6 tranches)	Marges EEC 2020, niveau individus
Nombre total de ménages dans la population	Marges EEC 2020, niveau ménages
Âge de la personne de référence (5 tranches)	Marges EEC 2020, niveau ménages
Taille d'unité urbaine (5 tranches)	Marges EEC 2020, niveau ménages
Diplôme de la personne de référence (4 modalités)	Marges EEC 2020, niveau ménages
Zone d'études et d'aménagement du territoire (ZEAT, 3 modalités)	Marges EEC 2020, niveau ménages
Type de ménage (6 modalités)	Marges EEC 2020, niveau ménages
Catégorie Socioprofessionnelle de la personne de référence (6 modalités)	Marges EEC 2020, niveau ménages
Nombre de ménages propriétaire de leur résidence principale	Estimation annuelle du parc de logements
Actif net (marge Ifi)	Marges POTE 2020 (métro + DOM)
Nombre de foyers redevables de l'Ifi	Marges POTE 2020 (métro + DOM)
Revenus d'activité (séparés entre métro et DOM)	Marges POTE 2020 (métro + DOM)
Revenus du patrimoine (séparés entre métro et DOM)	Marges POTE 2020 (métro + DOM)

L'échantillon S^t_{men} peut être considéré comme le fruit de deux étapes de tirage : un tirage d'individus-panels répondants $S^t_{\text{ind,rep}}$ et un tirage indirect à partir de $S^t_{\text{ind,rep}}$.

Dans la suite, l'ensemble des individus-panels dans la population au temps t est noté U^t_{ind}. On supposera que les éléments de U^t_{ind} sont des couples de la forme (k,t) où k est un identifiant de l'individu k n'évoluant pas au cours du temps[6] et t décrit l'année. Un individu k dans la population au temps $t-3$ et t correspond aux éléments $(k,t-3)$ dans U^{t-3}_{ind} et (k,t) dans U^t_{ind}.

[5] Certaines variables sont disponibles au niveau ménage mais sont transformées en variables au niveau individu en vertu du principe de dualité décrit par Lavallée (2009).

[6] Exemple : le numéro d'inscription au répertoire.

La pondération finale des ménages S^t_{men} (i.e contenant au moins un individu-panel répondant de $S^t_{\text{ind,rep}}$) se déduit de la pondération calée sur l'échantillon $S^t_{\text{ind,rep}}$ des individus-panels répondants en t : pour tout ménage $m \in S^t_{\text{men}}$, le poids après redressement et calage au niveau individu w^{men}_m est tel que $w^{\text{men}}_m = \sum_{l \in S^t_{\text{ind,rep}}} \frac{w^{\text{cal}}_l L_{lm}}{L_{\bullet m}}$ où $L_{\bullet m} = \sum_{l \in U^{t-9 \to t}_{\text{ind}}} L_{lm}$ où $U^{t-9 \to t}_{\text{ind}} = U^{t-9}_{\text{ind}} \cup U^{t-6}_{\text{ind}} \cup U^{t-3}_{\text{ind}} \cup U^{t}_{\text{ind}}$ décrit une population intertemporelle et L_{lm} est une variable valant 1 si l'individu-panel $l = (k, \tau),\ \tau \in \{t-9, t-6, t-3, t\}$ est tel que l'individu k appartient au ménage m en t et 0 sinon (voir Lavallée 2009).

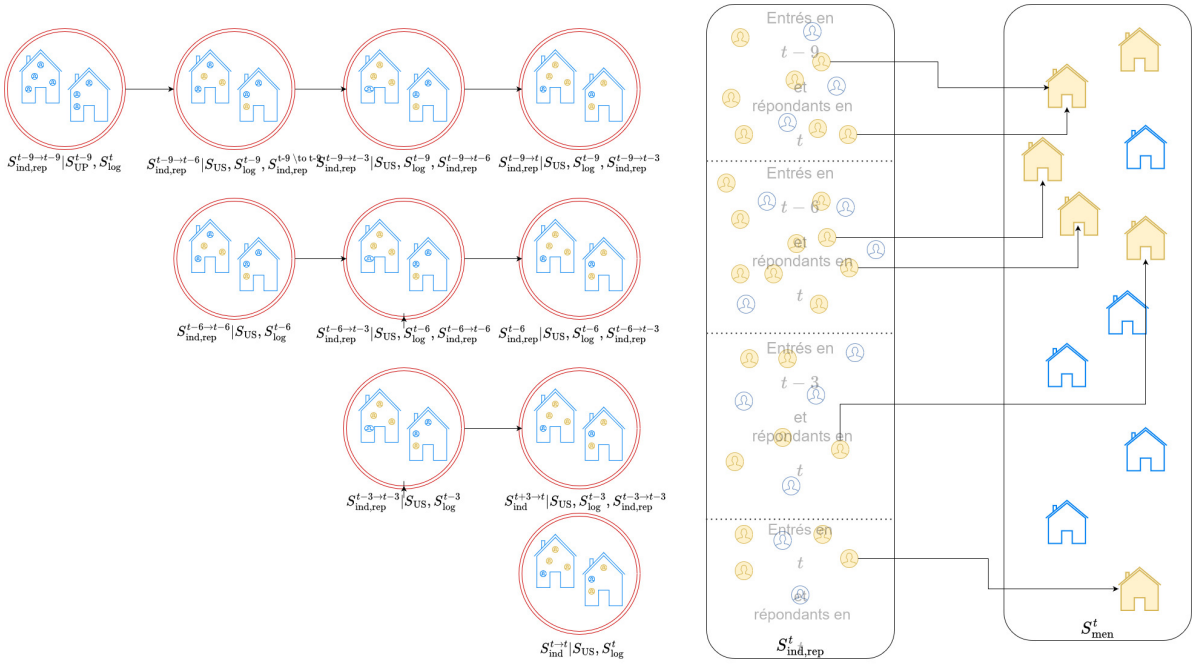
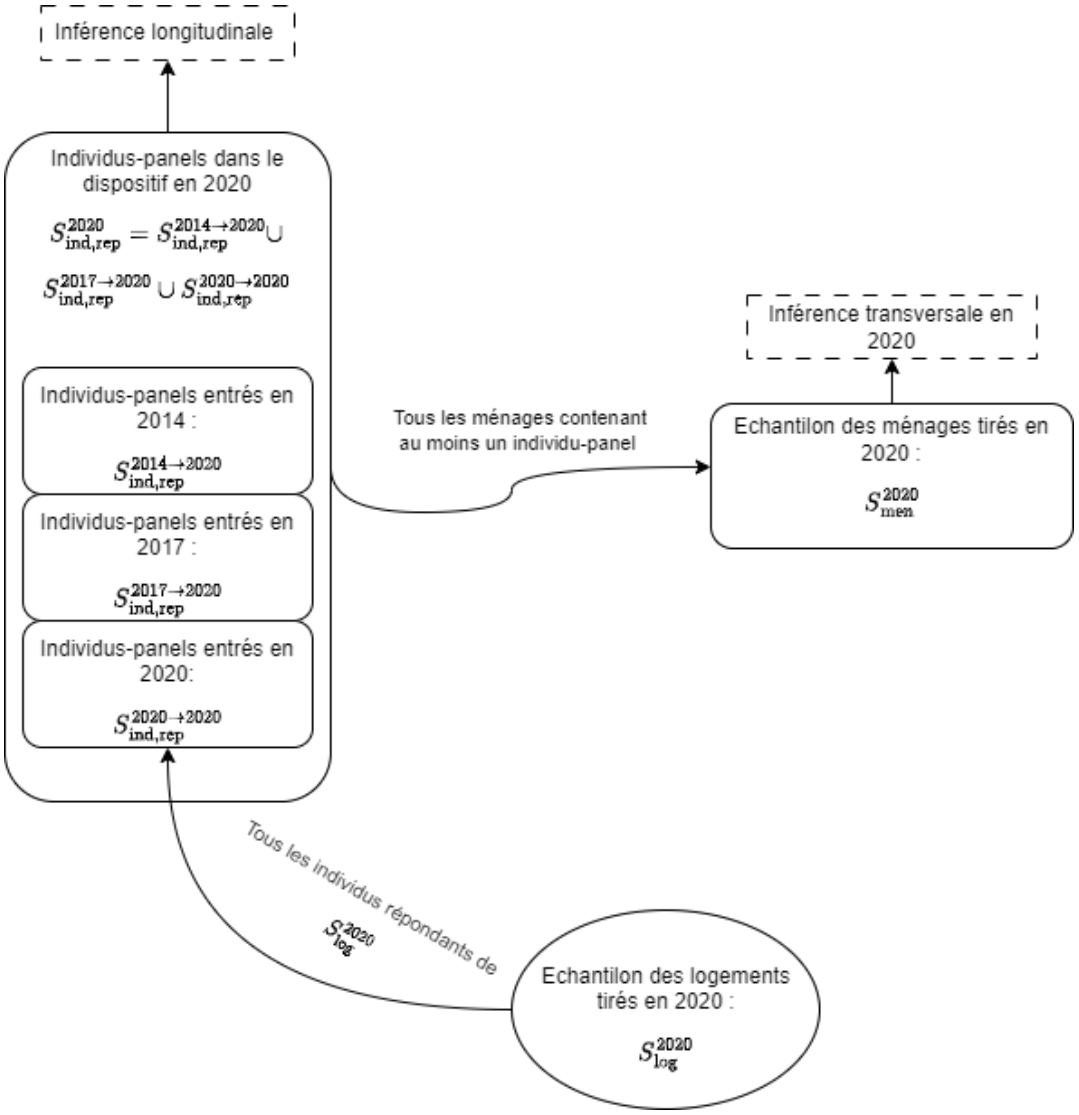


Figure 2.2: L'échantillon transversal des ménages S^t_{men} correspond à l'ensemble des ménages en t contenant au moins un individu-panel entré en $t-9$, $t-6$, $t-3$ ou t et répondants en t. La pondération des ménages de S^t_{men} est obtenue par partage des poids.

Il est à noter que $L_{\bullet m}$ nécessite de connaître tous les individus du ménage m en t et de savoir s'ils étaient dans la population en $t-9$, $t-6$, $t-3$ et t.



L'échantillon des individus-panels entrés en t et répondants $S^{t\to t}_{\text{ind,rep}}$ est obtenu en considérant tous les individus répondants de l'échantillon représentatif S^{t}_{log} de logements en t.
L'échantillon final de ménages utilisés pour l'inférence transversale est S^{t}_{men} obtenu en considérant tous les ménages en 2020 contenant au moins un individu-panel dans le dispositif et répondants en t.
Note de lecture : le document porte sur l'enquête de 2020. Le dispositif longitudinal ayant débuté en 2014, il n'y avait que deux réinterrogations.

3 Estimation de la variance par approche analytique

Dans cette partie, nous détaillons la construction d'un estimateur $\widehat{\mathbb{V}^{\text{analyt}}}$ de la variance sous le plan de l'enquête HVP (et en prenant en compte les redressements de l'enquête) pour des estimateurs définis au niveau de l'échantillon des ménages S^t_{men}[1].

Afin de construire cet estimateur, nous allons proposer un estimateur de la variance $\widehat{\mathbb{V}^{\text{ind}}}$ d'un estimateur du total non calé défini sur l'échantillon $S^t_{\text{ind,rep}}$. Un estimateur de la variance $\widehat{\mathbb{V}^{\text{ind-cal}}}$ prenant en compte le calage réalisé au niveau individu-panel est proposé en se basant sur $\widehat{\mathbb{V}^{\text{ind}}}$ et sur les résidus de la régression des variables d'intérêt sur les variables de calage.

Les estimateurs pour lesquels nous souhaitons obtenir une estimation de la variance utilisent l'échantillon S^t_{men} : l'estimateur final de la variance $\widehat{\mathbb{V}^{\text{analyt}}}$ est obtenu en utilisant l'estimateur $\widehat{\mathbb{V}^{\text{ind-cal}}}$ et des résultats sur le partage de poids.

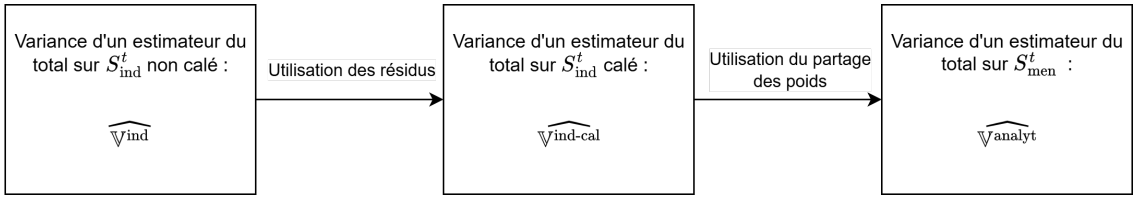


Figure 3.1: Étapes intermédiaires permettant la construction de l'estimateur $\widehat{\mathbb{V}^{\text{analyt}}}$

3.1 Prise en compte du tirage des individus-panels $\widehat{\mathbb{V}^{\text{ind}}}$

La formule de Rao (J. Rao 1975) permet de proposer un estimateur de la variance d'un estimateur linéaire du total sous un plan à deux degrés en se basant sur une décomposition de la variance.

Sous certaines hypothèses décrites ci-après et à l'aide d'un estimateur de la variance d'un estimateur sous le premier degré (en raisonnant conditionnellement aux éléments tirés au deuxième degré - intuitivement, cela revient à considérer que le total au sein d'une unité

[1] Comme l'estimateur du total d'une variable d'intérêt $\{y_k\}_{k\in S^t_{\text{men}}}$, s'écrivant $\sum_{k\in S^t_{\text{men}}} w^{\text{men}}_k y_k$.

primaire est déterministe) et d'un estimateur de la variance d'un estimateur sous le deuxième degré (en raisonnant conditionnellement aux éléments tirés au premier degré - intuitivement, cela revient à considérer qu'il n'y a pas de premier degré), il est possible de proposer un estimateur de la variance sous le plan réunissant le premier et deuxième degré.

Soit p un plan de sondage à deux degrés où p_1 désigne le plan de sondage des unités primaires et $p_{2|1}$ le plan de sondage des unités secondaires conditionnellement aux unités primaires. On notera respectivement S_{UP} et S_{US}, les échantillons d'unités primaires et secondaires obtenus.

On note :

- $\pi_u^{(1)}$, les probabilités d'inclusion d'ordre 1 d'une unité primaire $u \in U_{\text{UP}}$;
- $\pi_l^{(2|1)}$, les probabilités d'inclusion d'ordre 1 d'une unité secondaire $l \in U_{\text{UP}}$; conditionnellement à l'échantillon des UP.
- $\pi_l^{(1,2)}$, les probabilités d'inclusion d'ordre 1 d'une unité secondaire $l \in U_{\text{UP}}$.

Soit $\hat{t}_y$, l'estimateur d'Horvitz-Thompson d'une variable d'intérêt $\{y_k\}$:

$$\hat{t}_y = \sum_{k \in S_{\text{US}}} \frac{y_k}{\pi_k^{(1,2)}} = \sum_{u \in S_{\text{UP}}} \frac{\hat{t}_{y,(u)}}{\pi_u^{(1)}} \text{ avec } \hat{t}_{y,(u)} = \sum_{k \in u \cap S_{\text{US}}} \frac{y_k}{\pi_k^{(2|1)}}$$

Supposons :

- que les tirages des unités secondaires au sein des unités primaires sont indépendants d'une unité primaire à l'autre;
- l'existence d'un estimateur $\widehat{\mathbb{V}^{(1)}}(\hat{t}_y)$ sans biais de la variance lié au premier degré d'échantillonnage $\mathbb{V}^{(1)}(\hat{t}_y)$ pouvant s'écrire sous la forme $\widehat{\mathbb{V}^{(1)}}(\hat{t}_y) = Q((t_{y,(u)})_{u \in S_{\text{UP}}}) = \sum_{i \in S_{\text{UP}}} \sum_{\substack{j \in S_{\text{UP}} \\ j \neq i}} q_{ij} t_{y,(i)} t_{y,(j)} + \sum_{i \in S_{\text{UP}}} q_i t_{y,(i)}^2$ où pour tout $u \in S_{\text{UP}}$, $t_{y,(u)} = \sum_{k \in u} y_k$;
- pour tout $u \in U_{\text{UP}}$, l'existence d'un estimateur sans biais $\hat{t}_{y,(u)}$ (sous le deuxième degré) de $t_{y,(u)}$;
- pour tout $u \in U_{\text{UP}}$, l'existence d'un estimateur sans biais $\widehat{\mathbb{V}^{(2|1)}}(\hat{t}_{y,(u)})$ (sous le deuxième degré) de la variance conditionnelle de l'estimateur $\hat{t}_{y,(u)}$.

D'après la formule de Rao, sous ces hypothèses, la variance totale de l'estimateur d'Horvitz-Thompson sous le plan p peut être estimée sans biais par

$$\widehat{\mathbb{V}^{(1,2)}}(\hat{t}_y) = Q((\hat{t}_{y,(u)})_{u \in S_{\text{UP}}}) + \sum_{u \in S_{\text{UP}}} \left(\frac{1}{{\pi_u^{(1)}}^2} - q_u \right) \widehat{\mathbb{V}^{(2|1)}}(\hat{t}_{y,(u)})$$

L'estimation de la variance se base sur l'écriture du plan de sondage de HVP comme d'un plan à cinq degrés (tirage des unités primaires, des logements, des logements répondants, des individus-panels et des individus-panels répondants). Un estimateur de variance d'un estimateur de la forme $\hat{t}_{y,\text{ind}} = \sum_{k \in S^t_{\text{ind,rep}}} w^{\text{ind}}_k y_k$ est donné en utilisant la formule de Rao de manière itérative.

Il en résulte que :

$$
\begin{aligned}
\widehat{\mathbb{V}^{\text{ind}}}(\hat{t}_{y,\text{ind}}) = \quad & (3.1)\\
& \underbrace{\widehat{\mathbb{V}^{(\text{DT})}_{\mathbf{x}^{\text{eq}}}}\left(\sum_{u \in S_{\text{UP}}}\left(\sum_{l \in u}\sum_{k \in l}\frac{y_k \mathbb{1}_{l \in S_{\text{log,rep}}}\mathbb{1}_{k \in S_{\text{ind,rep}}}}{\hat{p}_l \pi^{\log|\text{UP}}_l \hat{\rho}_k}\right)\frac{1}{\pi^{\text{UP}}_u}\right)}_{\widehat{\mathbb{V}^{(\text{DT})}_{\mathbf{x}^{\text{eq}}}}\left(\sum_{u \in S_{\text{UP}}}\frac{\hat{y}^{\text{UP}}_u}{\pi^{\text{UP}}_u}\right)} + \\
& \sum_{u \in S_{\text{UP}}}\left(\frac{1}{\left(\pi^{\text{UP}}_u\right)^2} - q^{(\text{UP})}_u\right)\underbrace{\widehat{\mathbb{V}^{\text{SRS-strat}}_u}\left(\sum_{l \in u}\left(\sum_{k \in l}\frac{y_k \mathbb{1}_{l \in S_{\text{log,rep}}}\mathbb{1}_{k \in S_{\text{ind,rep}}}}{\hat{p}_l \hat{\rho}_k}\right)\frac{1}{\pi^{\log|\text{UP}}_l}\right)}_{=\widehat{\mathbb{V}^{\text{SRS-strat}}_u}\left(\sum_{l \in u}\frac{\hat{y}^{\log}_l}{\pi^{\log|\text{UP}}_l}\right)} + \\
& \sum_{l \in S^{t-9}_{\log} \cup S^{t-6}_{\log} \cup S^{t-3}_{\log} \cup S^t_{\log}}\left(\frac{1}{\left(\pi^{\log}_l\right)^2} - q^{(\text{UP,log})}_l\right)\underbrace{\left(\sum_{k \in l}\frac{y_k \mathbb{1}_{l \in S_{\text{log,rep}}}\mathbb{1}_{k \in S_{\text{ind,rep}}}}{\hat{\rho}_k}\right)^2 \frac{(1-\hat{p}_l)}{\hat{p}_l^2}}_{=\widehat{\mathbb{V}^{\text{Poisson}}}\left(\sum_{k \in l}\frac{\hat{y}_k}{\hat{p}_k}\right)} + \\
& \sum_{k \in S^{t-9}_{\text{ind}} \cup S^{t-6}_{\text{ind}} \cup S^{t-3}_{\text{ind}} \cup S^t_{\text{ind}}}\left(\frac{1}{\left(\pi^{\text{ind}}_k\right)^2} - q^{(\text{UP,log, log-NR, ind})}_k\right)\underbrace{\left(y_k \mathbb{1}_{k \in S^t_{\text{ind,rep}}}\right)^2 \frac{1-\hat{\rho}_k}{\hat{\rho}_k^2}}_{=\widehat{\mathbb{V}^{\text{Poisson}}}\left(\frac{\hat{y}_k \mathbb{1}_{k \in S^t_{\text{ind,rep}}}}{\hat{\rho}_k}\right)}
\end{aligned}
$$

où

- $\pi^{\log}_l = \pi^{\text{UP}}_{u(l)}\pi^{\log|\text{UP}}_l$ désigne la probabilité d'inclusion d'un logement non conditionnelle où $u(l)$ est l'unité primaire initiale du logement l;
- $\pi^{\text{ind}}_k = \pi^{\text{UP}}_{u(l)}\pi^{\log|\text{UP}}_{l(k)}\hat{p}_{l(k)}$ désigne la probabilité d'inclusion d'un individu non conditionnelle où $u(k)$ est l'unité primaire initiale du logement $l(k)$ de l'individu k;
- $\widehat{\mathbb{V}^{(\text{DT})}_{\mathbf{x}^{\text{eq}}}}$ désigne l'estimateur de Deville-Tillé appliqué avec les variables d'équilibrage $\mathbf{x}^{\text{eq}}$ liées à l'échantillon-maître Nautile (Delta et Paliod 2022). Dans un cadre général, si S est un échantillon tiré selon un plan équilibré sur les variables $\{\mathbf{x}_k\}_k = \{(x_{1,k}, ..., x_{d,k})\}_k$, l'estimateur de Deville-Tillé (Deville et Tillé 2005) appliqué à l'estimation de $\sum_{k \in S}\frac{y_k}{\pi_k}$ est

donné par

$$\widehat{\mathbb{V}_{\mathbf{x}}^{\mathrm{DT}}}\left(\sum_{k\in S}\frac{y_k}{\pi_k}\right)=\frac{n}{n-d}\sum_{k\in S}\left(\frac{y_k}{\pi_k}-\frac{\mathbf{x}_k^T}{\pi_k}\hat{\beta}\right)^2,$$

où $\hat{\beta}=\left(\sum_{k\in S}(1-\pi_k)\frac{\mathbf{x}_k}{\pi_k}\frac{\mathbf{x}_k^T}{\pi_k}\right)^{-1}\left(\sum_{k\in S}(1-\pi_k)\frac{\mathbf{x}_k}{\pi_k}\frac{y_k}{\pi_k}\right)$.

- $\widehat{\mathbb{V}_u^{\text{SRS-strat}}}$ désigne l'estimateur de la variance sous le plan de tirage des logements au sein de l'unité primaire u, conditionnellement aux unités primaires tirées. Cet estimateur appliqué à l'estimation de $\sum_{l\in u}\frac{\hat{y}_l^{\log}}{\pi_l^{\log|\mathrm{UP}}}$ est donné par

$$\widehat{\mathbb{V}_u^{\text{SRS-strat}}}\left(\sum_{l\in u}\frac{\hat{y}_l^{\log}}{\pi_l^{\log|\mathrm{UP}}}\right)=\sum_{h=1}^{H}N_{uh}^2\frac{(1-f_{uh})}{n_{uh}}s^2_{\hat{y}_l^{\log},uh},$$

où H désigne le nombre de strates dans l'unité primaire u[2], n_{uh} (resp. N_{uh}) désigne la taille de l'échantillon (resp. de la population) de logements tirés dans l'UP u et dans la strate h, $f_{uh}=\frac{n_{uh}}{N_{uh}}=\pi_l^{\log|\mathrm{UP}}$, et $s^2_{\hat{y}_l^{\log},uh}$ désigne la dispersion de la variable $\hat{y}_l^{\log}$ dans cet échantillon.
- $q_u^{(\mathrm{UP})}=\frac{n_{\mathrm{UP}}}{n_{\mathrm{UP}}-d}\frac{1-(\pi_u^{\mathrm{UP}})^2}{\pi_u^{\mathrm{UP}}}\left(1-\frac{1-(\pi_u^{\mathrm{UP}})^2}{\pi_u^{\mathrm{UP}}}(\mathbf{x}_u^{\mathrm{eq}})^TB_{\mathbf{x}}\mathbf{x}_u^{\mathrm{eq}}\right)$, où $\mathbf{x}_u^{\mathrm{eq}}\in\mathbb{R}^d$ désigne le vecteur des variables d'équilibrage pour l'unité primaire u,

$$B_{\mathbf{x}}=\left(\sum_{v\in S_{\mathrm{UP}}}\frac{1-(\pi_v^{\mathrm{UP}})^2}{\pi_v^{\mathrm{UP}}}\mathbf{x}_v^{\mathrm{eq}}(\mathbf{x}_v^{\mathrm{eq}})^\top\right)^{-1}$$

et $n_{\mathrm{UP}}=\mathrm{Card}(S_{\mathrm{UP}})$.
- $q_l^{(\mathrm{UP,log})}=\frac{q_{u(l)}^{(\mathrm{UP})}}{(\pi_l^{\log|\mathrm{UP}})^2}+\left(\frac{1}{(\pi_{u(l)}^{\mathrm{UP}})^2}-q_{u(l)}^{\mathrm{UP}}\right)\frac{1-f_{uh}}{f_{uh}^2}$.
- $q_k^{(\text{UP, log, log-NR, ind})}=\frac{q_{l(k)}^{(\mathrm{UP,log})}}{(\hat{p}_{l(k)}^{\log})^2}+\left(\frac{1}{\pi_{l(k)}^2}-q_{l(k)}^{(\mathrm{UP,log})}\right)\frac{1-\hat{p}_{l(k)}}{\hat{p}_{l(k)}^2}$, où $l(k)$ désigne le logement associé à l'individu k.

3.1.1 Prise en compte du calage sur marges $\widehat{\mathbb{V}^{\text{ind-cal}}}$

Le calage sur marges permet d'utiliser de l'information auxiliaire afin d'améliorer la précision des estimateurs. Cette information auxiliaire est représentée par des variables disponibles pour chaque individu de l'échantillon et dont les totaux sur l'ensemble de la population sont connus.

[2]il est identique pour chaque unité primaire.

Le calage sur marges consiste à modifier le moins possible les poids de telle manière à ce que les estimateurs de totaux de variables auxiliaires soient égaux aux totaux sur la population. La pondération dépendant de l'échantillon, il n'est plus possible d'utiliser les estimateurs de la variance *à la Horvitz-Thompson*. Néanmoins, un estimateur de la variance d'un estimateur calé peut être calculé en utilisant les résultats asymptotiques proposés par Deville et Särndal (1992) : la variance de l'estimateur $\sum_{k \in S^t_{\text{ind,rep}}} w^{\text{cal}}_k y_k$ dont la pondération est calée sur la variable $\{\mathbf{x}^{\text{cal}}_k\}_{k \in S^t_{\text{ind,rep}}}$ est approximativement celle de l'estimateur d'Horvitz-Thompson des résidus $\{\hat{\varepsilon}_k\}_{k \in S^t_{\text{ind, rep}}}$ de la régression pondérée (par les poids avant calage) de $\{y_k\}_{k \in S^t_{\text{ind,rep}}}$ sur $\{\mathbf{x}_k\}_{k \in S^t_{\text{ind, rep}}}$ estimée sur $S^t_{\text{ind,rep}}$.

Il en vient que :

$$\widehat{\mathbb{V}^{\text{ind-cal}}}\left(\sum_{k \in S^t_{\text{ind,rep}}} w^{\text{cal}}_k y_k\right) \approx \widehat{\mathbb{V}^{\text{ind}}}\left(\sum_{k \in S^t_{\text{ind,rep}}} w^{\text{ind}}_k \hat{\varepsilon}_k\right) \tag{3.2}$$

où $\widehat{\mathbb{V}^{\text{ind}}}$ est un estimateur de la variance d'un estimateur linéaire d'une variable d'intérêt définie sur $S^t_{\text{ind,rep}}$.

3.1.2 Prise en compte du partage des poids $\widehat{\mathbb{V}^{\text{analyt}}}$

La prise en compte du partage des poids dans l'estimation de variance de $\hat{t}_{y,\text{men}}$ découle de deux remarques :

- $\hat{t}_{y,\text{men}} = \sum_{m \in S^t_{\text{men}}} w^{\text{men}}_m y_m$ peut être réécrit comme un total sur l'échantillon $S^t_{\text{ind,rep}}$: $\hat{t}_{y,\text{men}} = \sum_{m \in S^t_{\text{men}}} w^{\text{men}}_m y_m = \sum_{k \in S^t_{\text{ind,rep}}} w^{\text{cal}}_k z_k$ où $z_k = \sum_{m \in S^t_{\text{men}}} \frac{L_{km}}{L_{\bullet m}} y_m$
- le tirage de S^t_{men} conditionnellement à celui de $S^t_{\text{ind,rep}}$ n'est pas aléatoire : en effet, un ménage m appartient à S^t_{men} s'il contient au moins un individu-panel de $S^t_{\text{ind,rep}}$.

Il en vient que :

$$\widehat{\mathbb{V}^{\text{analyt}}}\left(\hat{t}_{y,\text{men}}\right) = \widehat{\mathbb{V}^{\text{ind-cal}}}\left(\sum_{k \in S^t_{\text{ind,rep}}} w^{\text{cal}}_k z_k\right) \tag{3.3}$$

où $\widehat{\mathbb{V}^{\text{ind-cal}}}$ désigne un estimateur de la variance d'un estimateur linéaire calé d'une variable d'intérêt définie sur $S^t_{\text{ind,rep}}$.

En combinant les estimateurs calculés dans l'équation 3.1, l'équation 3.2 et l'équation 3.3, un estimateur de la variance de $\hat{t}_{y,\text{men}}$ est :

$$\widehat{\mathbb{V}^{\text{analyt}}}(\hat{t}_{y,\text{men}}) = \hat{\mathbb{V}}\left(\sum_{m\in S^t_{\text{men}}} w^{\text{men}}_m y_m\right) \approx \widehat{\mathbb{V}^{\text{ind}}}\left(\sum_{k\in S^t_{\text{ind}}} w^{\text{ind,rep}}_k \tilde{\varepsilon}_k\right) \tag{3.4}$$

où $\tilde{\varepsilon}_k$ est le résidu de l'individu k de la régression de $\{z_k\}_{k\in S^t_{\text{ind,rep}}}$ sur les variables de calage $\{\mathbf{x}^{\text{cal}}_k\}_{k\in S^t_{\text{ind}}}$ pondérée par les poids des individus avant calage $\{w^{\text{ind}}_k\}_{k\in S^t_{\text{ind,rep}}}$.

3.2 Estimation d'autres paramètres sous le plan de l'enquête HVP

3.2.1 Linéarisation dans le cas différentiable

L'estimateur proposé dans l'équation 3.4 permet d'estimer la variance d'un estimateur calé du total d'une variable d'intérêt. Pour l'estimation de la variance d'estimateurs de la forme $\hat{\theta} = f(\hat{t}_{y_1,\text{men}}, \dots, \hat{t}_{y_d,\text{men}})$ où $f : \mathbb{R}^d \to \mathbb{R}$ est une fonction différentiable, il n'est pas possible d'utiliser les résultats sur le total.

Pour autant, le principe de linéarisation basé sur les développements de Taylor (Woodruff 1971) permet d'obtenir une approximation de la variance de $\hat{\theta}$:

$$\hat{\mathbb{V}}(\hat{\theta}) \approx \widehat{\mathbb{V}^{\text{analyt}}}(\hat{t}_{z,\text{men}})$$

où $\{z_k\}$ est la variable linéarisée estimée définie pour tout $k \in S^t_{\text{men}}$ par $z_k = \nabla f(\hat{t}_{y_1,\text{HT}}, ..., \hat{t}_{y_d,\text{HT}})(y_{1,k}, ..., y_{d,k})^T$

3.2.1.1 Exemple

Soient un échantillon S et deux variables d'intérêt $\{a_k\}$ et $\{b_k\}$. Il est possible d'obtenir un estimateur de la variance de l'estimateur du ratio $\hat{r} = \dfrac{\hat{t}_{a,\text{HT}}}{\hat{t}_{b,\text{HT}}} = \dfrac{\sum_{k\in S}\frac{a_k}{\pi_k}}{\sum_{k\in S}\frac{b_k}{\pi_k}}$ en utilisant le principe de linéarisation.

En effet, $\hat{r} = f(\hat{t}_{a,\text{HT}}, \hat{t}_{b,\text{HT}})$ avec $f : (a,b) \mapsto \frac{a}{b}$. La variable linéarisée estimée vaut $z_k = \nabla f(\hat{t}_{a,\text{HT}}, \hat{t}_{b,\text{HT}})(a_k, b_k)^T = \dfrac{1}{\hat{t}_{b,\text{HT}}}(a_k - \hat{r}_k b_k)$ où $\nabla f(a,b) = \left(\frac{\partial f}{\partial a}, \frac{\partial f}{\partial b}\right) = \left(\frac{1}{b}, -\frac{a}{b^2}\right)$.

L'utilisation du package *gustave* (Chevalier 2018) en R permet d'éviter d'implémenter les opérations de linéarisation les plus fréquemment utilisées.

3.2.2 Linéarisation d'un estimateur d'un quantile

L'approche par linéarisation basée sur les développements de Taylor admet des limites : certaines fonctions d'intérêt ne peuvent s'exprimer comme une fonction de totaux comme par exemple, les quantiles d'une distribution. Néanmoins, Deville (1999) propose une approche basée sur la fonction d'influence de l'estimateur considéré. Ces travaux ont été repris et appliqués dans le cadre de l'enquête européenne SILC (Osier 2009; Graf et Tillé 2014),

3.2.2.1 Notations et définitions

La linéarisation de paramètres complexes repose sur l'utilisation de fonctions d'influence. La fonction d'influence peut être appréhendée comme une généralisation de la dérivation à des fonctionnelles.

Une fonctionnelle T d'une mesure M est une fonction de M.

Par exemple, le total d'une variable d'intérêt sur l'ensemble de la population est une fonctionnelle pour la mesure $M_U = \sum_{k \in U} \delta_k$. En effet, $t_y = \sum_{k \in U} y_k = \int_{\mathbb{R}} y(k) \, \mathrm{d}M_U$. De même, l'estimateur linéaire $\hat{t}_y = \sum_{k \in S} w_k y_k$ est une fonctionnelle pour la mesure $M_S = \sum_{k \in S} w_k \delta_k$.

3.2.2.2 Fonction d'influence

La fonction d'influence d'une fonctionnelle T notée $\mathrm{IF}(T)$ est l'application $x \to \mathrm{IF}(T)(x) = \lim_{t \to 0} \frac{T(M + t\delta_x) - T(M)}{t}$. La fonction d'influence peut être vue comme une dérivée de la fonctionnelle par rapport à la mesure : cette fonction quantifie le comportement de la fonctionnelle lorsque les variables d'intérêt sont *contaminées* de manière infinitésimale.

Deville (1999) et Demnati et Rao (2004) indiquent que la fonction d'influence peut être utilisée afin de linéariser une fonctionnelle : la variance d'un estimateur $\hat{\theta} = T(M_S)$ est approximativement égale à la variance de l'estimateur du total de Horvitz-Thompson associé pour la variable linéarisée $\{z_k\}$ définie pour tout $k \in S$ par $z_k = \mathrm{IF}(T(M_S))(e_k)$ (z_k peut être une fonction de M) où $e_k \in \mathbb{R}^n$ est le vecteur constitué de zéros sur toutes ses composantes sauf la k-ième valant 1.

Comme dans le cas de la linéarisation utilisant le développement de Taylor, la variable linéarisée est rarement calculable directement : il est nécessaire de passer par un estimateur.

3.2.2.3 Fonctions de répartition et quantile

La fonction de répartition d'une variable $\{y_k\}$ associée à une mesure de probabilité M est la fonction $F_{1,y}(M)$ telle que pour tout $x \in \mathbb{R}$, $F_{1,y}(M) : x \to \int_{\mathbb{R}} \mathbb{1}_{y(k)\leq x}\, \mathrm{d}M$. Dans le cas où $M = M_U = \frac{1}{N}\sum_{k\in U}\delta_k$ (resp. $M = M_S = \frac{1}{\sum_{k\in S} w_k}\sum_{k\in S} w_k\delta_k$), $F_{1,y}(M) : x \to \frac{1}{N}\sum_{k\in U}\mathbb{1}_{y_k\leq x}$ (resp $F_{1,y}(M) : x \to \frac{1}{\sum_{k\in S} w_k}\sum_{k\in S} w_k\mathbb{1}_{y_k\leq x}$).

La fonction $F_{1,y}(M)$ est une fonction en escalier : cette dernière ne permet pas de donner une définition simple de la fonction de quantile. Une autre formulation consiste à interpoler linéairement la fonction de répartition : $F_{2,y}(M) : x \to \frac{1}{N}\sum_{k\in U\setminus\{1\}} H\left(\frac{x - y_{(k-1)}}{y_{(k)} - y_{(k-1)}}\right)$ avec $H : x \to x\mathbb{1}_{x\in[0,1]} + \mathbb{1}_{x\in]1,+\infty]}$.

Cette autre formulation permet de définir de manière unique la fonction quantile[3].

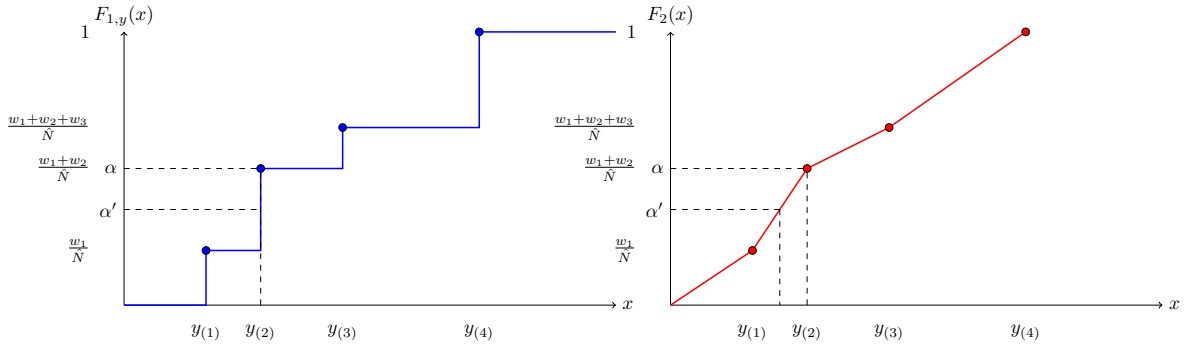


Figure 3.2: Comparaison de deux définitions des quantiles empiriques

À partir de la définition de la fonction de répartition, il est possible de proposer plusieurs définitions des quantiles (Hyndman et Fan 1996). La définition retenue ici est celle préconisée par Eurostat : $Q_\alpha = y_{(k-1)} + (y_{(k)} - y_{(k-1)})(\alpha N - (k-1))$[4] où k est tel que $\alpha N < k \leq \alpha N + 1$ et $\{y_{(k)}\}$ correspond à la variable $\{y_k\}$ ordonnée. Un estimateur $\hat{Q}_\alpha$ basé sur l'échantillon S^t_{men} est $\hat{Q}_\alpha = y_{(k-1)} + (y_{(k)} - y_{(k-1)})\frac{\alpha\hat{N} - \hat{N}_{k-1}}{w_k}$ où $\hat{N}_k = \sum_{j\in S^t_{\text{men}}} w_j^{\text{men}}\mathbb{1}_{y_j\leq y_k}$.

L'estimateur $\hat{Q}_\alpha$ peut être linéarisé en utilisant la linéarisation par fonction d'influence. Il en

[3] En supposant que chaque individu prend une valeur différente de la variable $\{y_k\}$.
[4] Il s'agit de l'unique solution à l'équation $F_2(M)(u) = \alpha$.

vient que pour $k \in S$,

$$\boxed{z_k = \text{IF}(\hat{Q}_\alpha)(e_k) = -\frac{1}{f_y(\hat{Q}_\alpha)} \frac{1}{\hat{N}} \left[H\left(\frac{\hat{Q}_\alpha - y_{(k-1)}}{y_{(k)} - y_{(k-1)}} \right) - \alpha \right]}$$

où f_y désigne la densité de la variable $\{y_k\}$.

La variable linéarisée estimée de $\hat{Q}_\alpha$ nécessite l'évaluation de f_y en $\hat{Q}_\alpha$. Or le calcul de f_y à partir de $F_y(M)$ conduit à une fonction nulle sauf aux points de discontinuité de $F_y(M)$. Osier (2009) propose d'estimer f_y en utilisant des méthodes d'estimation de la densité par noyau gaussien. Cette méthode est dépendante d'un paramètre de **bandwidth** h permettant de quantifier quel voisinage d'un point est utilisé pour estimer la densité. Le choix de ce paramètre peut être réalisé en utilisant des méthodes de validation croisée, des règles empiriques, … Dans ce travail, nous utilisons deux règles empiriques :

- règle de Silverman (inspirée[5] de Charpentier (2015)) : $\hat{h} = 0.9 \min\left(\hat{\sigma}, \frac{\hat{Q}_{0.75} - \hat{Q}_{0.25}}{1.349} \right) \hat{N}^{-\frac{1}{5}}$
- règle provenant de Osier (2009) : $\hat{h} = \frac{\hat{\sigma}}{\hat{N}^{-\frac{1}{5}}}$

Charpentier (2015) et Graf et Tillé (2014) proposent d'estimer la densité f_y des variables $\{y_k\}$ ayant une distribution à queue épaisse en estimant la fonction de densité $f_{\log(y)+a}$ de la variable $\{\log(y_k) + a\}$ puis d'utiliser le fait que pour toute variable aléatoire X positive admettant une densité f_X et pour tout $x > 0$, $f_X(t) = \frac{f_{\log(X+a)}(\log(t+a))}{t+a}$ où a désigne une valeur positive.

Étant donné que les performances de ces estimations de densité sont plus sensibles au choix de la fenêtre qu'au choix du noyau, nous décidons dans ce document de travail (voir Section 5.2) de comparer quatre versions de la linéarisation des quantiles en utilisant soit la règle de Silverman soit la règle provenant de Osier (2009) et en utilisant l'estimation de $\{y_k\}$ ou de $\{\log(y_k)\}$.

[5] Les régles décrites par Charpentier (2015) le sont dans un contexte hors sondage. Nous estimons N par $\hat{N}$ comme dans Osier (2009).

4 Estimation de la variance par bootstrap

L'approche par bootstrap consiste à produire plusieurs jeux de poids, appelés poids répliqués ou poids bootstrap, permettant de construire des réplications de l'estimateur $\widehat{\theta}$. L'estimateur de la variance de $\widehat{\theta}$ s'obtient alors en calculant la dispersion des estimateurs répliqués.

Cette approche basée sur des poids répliqués permet d'estimer la précision simplement pour une panoplie de statistiques sans avoir recours à des procédés de linéarisation. De plus, ces poids peuvent être diffusés plus simplement qu'une fonction permettant de calculer l'estimation analytique de la précision d'un estimateur : en effet, une telle fonction doit contenir *a minima* les informations auxiliaires comme les variables de calage potentiellement confidentielles. L'approche par bootstrap présente ainsi l'avantage de séparer les responsabilités d'estimations entre producteur et utilisateur :

- Le producteur est en mesure de fournir des poids répliqués incorporant implicitement tous les mécanismes mis en œuvre dans la construction de l'échantillon, sans avoir à connaitre les indicateurs qui pourraient être construits par un utilisateur.
- L'utilisateur a la possibilité d'estimer la précision de son indicateur sans avoir à connaître précisément les détails associés au plan de sondage.

Chauvet et al. (2022) a développé une méthode permettant de construire des poids répliqués basée sur l'approche de J. N. K. Rao et Wu (1988). Cette méthode permet de fournir des poids répliqués dans un contexte de plan à plusieurs degrés : seul un rééchantillonnage des unités primaires est nécessaire ainsi qu'une prise en compte du rééchantillonnage dans les redressements (traitement de la non-réponse, calage). Cette méthode se base en premier lieu sur l'hypothèse que le tirage des unités primaires est réalisé avec une faible fraction de sondage, permettant ainsi de le voir comme un tirage avec remise (ou tirage multinomial). Sous l'hypothèse d'un plan plus efficace[1] qu'un plan multinomial, cette méthode permet d'obtenir une estimation conservatrice de la variance d'estimateurs linéaires.

4.1 Estimation de variance bootstrap et intervalles de confiance

Cette section présente comment un utilisateur des données peut produire ses propres estimations de variance et intervalles de confiance à partir des poids répliqués fournis.

[1] Au sens de la variance.

En supposant que l'estimateur soit de la forme $\hat{\theta} = f(\{y_m\}_{m \in S^t_{\text{men}}}, \{w^{\text{men}}_m\}_{m \in S^t_{\text{men}}})$, la procédure à suivre est résumée dans l'algorithme 1. Il suffit donc pour un utilisateur de remplacer les poids de sondage classiques $\{w^{\text{men}}_m\}_{m \in S^t_{\text{men}}}$ par des poids de sondage répliqués $\{w^{\text{men},*b}_m\}_{m \in S^t_{\text{men}}}$ pour construire un estimateur répliqué $\hat{\theta}^{*b} = f(\{y_m\}_{m \in S^t_{\text{men}}}, \{w^{\text{men},*b}_m\}_{m \in S^t_{\text{men}}})$.

En particulier, lorsque $\hat{\theta}$ s'écrit comme une fonction de totaux, $\hat{\theta} = f(\hat{t}_{y_1,\text{men}}, \dots, \hat{t}_{y_d,\text{men}})$ avec $f : \mathbb{R}^d \to \mathbb{R}$, la construction de $\hat{\theta}^{*b}$ passe par la construction d'estimateurs répliqués des totaux, $\hat{\theta}^{*b} = f(\hat{t}^{*b}_{y_1,\text{men}}, \dots, \hat{t}^{*b}_{y_d,\text{men}})$ où pour chaque variable d'intérêt y,

$$\hat{t}^{*b}_{y,\text{men}} = \sum_{m \in S^t_{\text{men}}} w^{\text{men},*b}_m y_m$$

Les propriétés théoriques de convergence asymptotique du bootstrap présupposent en général que la fonction f soit lisse. Néanmoins la méthode en elle-même peut tout de même être appliquée en dehors de ce cadre théorique. La partie 5 de ce document étudie plus en détails comment les estimations par bootstrap se comportent en pratique.

Par ailleurs, l'approche par bootstrap reproduisant plutôt un tirage avec remise, menant à des estimateurs avec une variance plus élevée que le plan de sondage de HVP, il est ainsi attendu que l'approche par bootstrap surestime la variance des estimateurs. De même, il est attendu que l'intervalle de confiance construit à partir de cette estimation de variance soit plutôt conservatif.

Algorithme 1 Estimation de variance bootstrap et construction d'intervalles de confiance

On suppose que les poids répliqués $\{w_m^{\text{men},*1}\}_{m\in S_{\text{men}}^t}, \dots, \{w_m^{\text{men},*B}\}_{m\in S_{\text{men}}^t}$ sont fournis. On suppose également que l'estimateur $\hat{\theta}$ s'écrit sous la forme $\hat{\theta} = f(\{y_m\}_{m\in S_{\text{men}}^t}, \{w_m^{\text{men}}\}_{m\in S_{\text{men}}^t})$.

1. Pour chaque $b \in \{1, \dots, B\}$, on construit un estimateur répliqué $\hat{\theta}^{*b}$ à l'aide du jeu de poids répliqués $\{w_m^{\text{men},*b}\}_{m\in S_{\text{men}}^t}$:

$$\hat{\theta}^{*b} = f(\{y_m\}_{m\in S_{\text{men}}^t}, \{w_m^{\text{men},*b}\}_{m\in S_{\text{men}}^t})$$

2. L'estimateur bootstrap de variance de $\hat{\theta}$ est

$$\hat{\mathbb{V}}^{\text{boot}}(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b}\right)^2$$

3. L'intervalle de confiance de niveau $1-2\alpha$ associé, sous hypothèse de normalité asymptotique, se calcule par

$$IC_{1-2\alpha}^{\text{norm}}(\hat{\theta}) = \left[\hat{\theta} - u_{1-\alpha}\sqrt{\hat{\mathbb{V}}^{\text{boot}}(\hat{\theta})}, \hat{\theta} + u_{1-\alpha}\sqrt{\hat{\mathbb{V}}^{\text{boot}}(\hat{\theta})}\right]$$

avec $u_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la distribution normale centrée réduite.

4.2 Construction des poids répliqués

Cette section a maintenant pour but de présenter la méthodologie associée à la construction des poids répliqués dans les estimations bootstrap. L'algorithme 2 donne une vue d'ensemble de la méthode pour générer un jeu de poids répliqués.

Algorithme 2 Calcul des poids répliqués des ménages en t (Vue d'ensemble)

1. On calcule les poids répliqués $\{w_{u*}^{\text{UP}}\}_{u\in S_{\text{UP}}}$ des unités primaires selon l'algorithme 3.
2. On calcule ensuite les poids répliqués $\{w_{k*}^{\text{ind}}\}_{k\in S_{\text{ind,rep}}^{t_0\to t}}$ associés au tirage et suivi des individus-panels en appliquant l'algorithme 4 aux temps $t_0 = t-9, t-6, t-3, t$. En regroupant ces poids, on obtient les poids répliqués $\{w_{k*}^{\text{ind}}\}_{k\in S_{\text{ind,rep}}^t}$ pour l'ensemble des individus-panels présents en t.
3. On applique enfin un calage et un partage des poids selon l'algorithme 5 pour obtenir les poids répliqués $\{w_{m*}^{\text{men}}\}_{m\in S_{\text{men}}^t}$ des ménages en t.

Le rééchantillonnage est ensuite répété de manière indépendante pour obtenir B jeux de poids $\{w_m^{\text{men},*1}\}_{m\in S_{\text{men}}^t}, \dots, \{w_m^{\text{men},*B}\}_{m\in S_{\text{men}}^t}$.

4.2.1 Calcul des poids répliqués des unités primaires

Comme annoncé en début de section, le coeur de la méthode par bootstrap repose sur l'idée que le tirage des unités primaires puisse se voir comme un plan multinomial. Cette approximation semble raisonnable lorsque les unités primaires ont une faible probabilité d'inclusion.

Dans ce cas, la méthode bootstrap sélectionne $n_{\text{UP},h} - 1$ unités primaires selon un tirage simple avec remise et à probabilités égales dans l'échantillon $S_{\text{UP},h}$ issu de la strate $U_{\text{UP},h}$. Nous introduisons alors pour tout $u \in S_{\text{UP},h}$ un facteur d'ajustement de pondération donné par

$$G_u^{\text{UP}} = \frac{n_{\text{UP},h}}{n_{\text{UP},h} - 1} \times m_u^{\text{UP}}$$

où m_u^{UP} correspond à la multiplicité de l'unité primaire u, c'est-à-dire le nombre de fois que l'unité primaire u a été sélectionnée dans le rééchantillon de $S_{\text{UP},h}$. Ces facteurs d'ajustement de pondération introduisent ainsi une variabilité dans la pondération des unités primaires, qui vont ensuite se répercuter en une variabilité des autres poids de sondages. Pour la suite, la dépendance en la stratification h ne sera plus répétée.

Plusieurs défis apparaissent dans cette première étape :

- Toutes les unités primaires n'ont en réalité pas une faible probabilité d'inclusion. Certaines unités primaires sont mêmes sélectionnées avec exhaustivité. Appliquer directement la méthode précédente mènerait à une forte surestimation de la variance puisqu'il n'y a aucune variabilité d'échantillonnage dans la sélection de ces unités. Pour la suite on distinguera trois types d'unités primaires :
  - Les unités primaires dites "exhaustives" qui ont une probabilité d'inclusion égale à 1. Elles forment l'échantillon $S_{\text{UP}}^{\text{exh}}$.
  - Les unités primaires dites "quasi-exhaustives" qui ont une probabilité d'inclusion supérieure à un seuil δ. Elles forment l'échantillon $S_{\text{UP}}^{\text{quasi-exh}}$.
  - Les unités primaires dites "normales", correspondant au reste des unités primaires, et qui ont donc une probabilité d'inclusion inférieure à δ. Elles forment l'échantillon $S_{\text{UP}}^{\text{norm}}$.

  L'approche proposée ici est de considérer les unités primaires exhaustives et quasi-exhaustives comme tirées avec certitude. Elles ne font donc pas partie de la phase de rééchantillonnage des unités primaires. À la place, le rééchantillonnage est appliqué à l'intérieur de celles-ci, au niveau des logements. Toutefois, il n'existe *a priori* pas de méthodes permettant de définir quantitativement le seuil δ : nous proposons de tester différentes valeurs de celui-ci.

- Comme indiqué dans la description du plan de sondage, le tirage des unités primaires résulte d'un tirage équilibré régionalement. Or, la méthode initiale de J. N. K. Rao et Wu (1988) ne permet pas de prendre en compte les gains apportés par l'équilibrage. Pour pouvoir tenir compte de cette phase d'équilibrage, Chauvet et al. (2022) propose d'intégrer à la place une étape de calage dans chaque région après le rééchantillonnage des unités primaires. Ici, seules les unités primaires exhaustives sont retirées du calage.

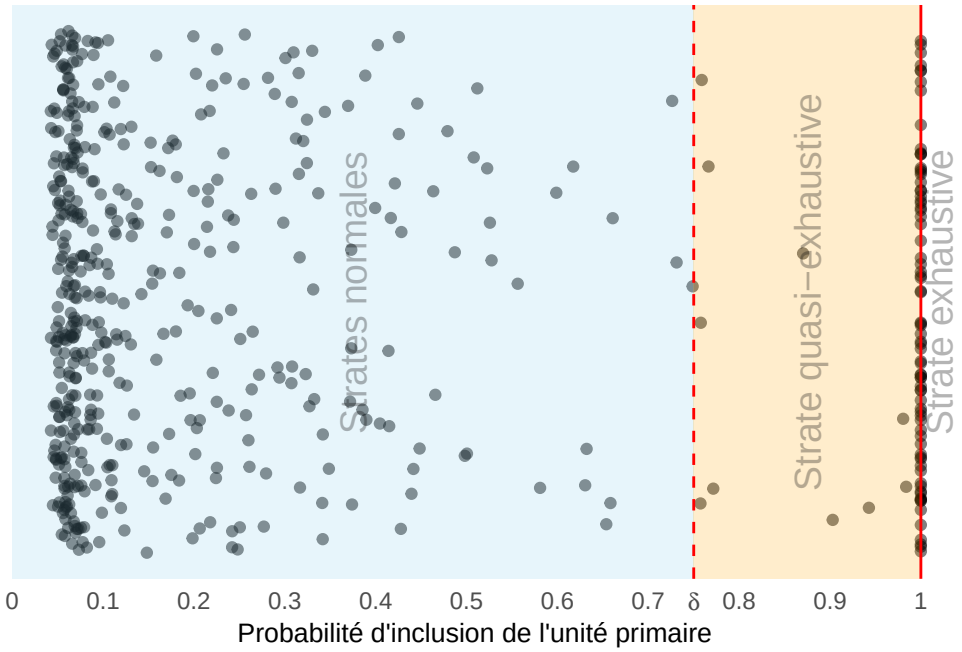


Figure 4.1: Partition des unités primaires de l'échantillon-maître Nautile avec un seuil d'exhaustivité de 0,75

Algorithme 3 Calcul des poids répliqués des unités primaires pour une réplication b donnée

Par souci de clarté, l'indice b est omis dans la suite.

1. On considère les $n_{\text{UP}}^{\text{norm}}$ unités primaires de $S_{\text{UP}}^{\text{norm}}$, et on tire $n_{\text{UP}}^{\text{norm}} - 1$ unités avec remise parmi celles-ci. On note alors m_u la multiplicité du tirage de l'unité u.

2. Les poids de sondage répliqués, hors équilibrage, des unités primaires sont donnés par

$$d_{u*}^{\text{UP}} = G_u^{\text{UP}} \frac{1}{\pi_u^{\text{UP}}} \text{ pour tout } u \in S_{\text{UP}}, \tag{4.1}$$

$$\text{avec } G_u^{\text{UP}} = \begin{cases} \frac{n_{\text{UP}}^{\text{norm}}}{n_{\text{UP}}^{\text{norm}}-1} \times m_u^{\text{UP}} & \text{pour } u \in S_{\text{UP}}^{\text{norm}}, \\ 1 & \text{sinon.} \end{cases} \tag{4.2}$$

3. L'équilibrage réalisé sur les variables auxiliaires $\{\mathbf{x}_u^{\text{eq}}\}_{u \in S_{\text{UP}}}$ dans le premier degré est pris en compte en réalisant à la place un calage des poids répliqués. Si l'on note $S'_{\text{UP}} = S_{\text{UP}}^{\text{norm}} \cup S_{\text{UP}}^{\text{quasi-exh}}$, alors les poids de sondages répliqués des unités primaires $(w_{u*}^{\text{UP}})_{u \in S'_{\text{UP}}}$ sont donnés par

$$w_{u*}^{\text{UP}} = d_{u*}^{\text{UP}} \left(1 + \mathbf{x}_u^{\text{eq}\top} \lambda_*^{\text{UP}}\right), \tag{4.3}$$

$$\text{avec } \lambda_*^{\text{UP}} = \left(\sum_{u \in S'_{\text{UP}}} d_{u*}^{\text{UP}} \mathbf{x}_u^{\text{eq}} \mathbf{x}_u^{\text{eq}\top} \right)^{-1} \left(t_{\mathbf{x}_u^{\text{eq}}} - \hat{t}_{\mathbf{x}_{u*}^{\text{eq}}}\right) \text{ et } \hat{t}_{\mathbf{x}_{u*}^{\text{eq}}} = \sum_{u \in S'_{\text{UP}}} d_{u*}^{\text{UP}} \mathbf{x}_u^{\text{eq}}.$$

   Enfin les poids des unités exhaustives sont inchangés. On pose $w_{u*}^{\text{UP}} = d_{u*}^{\text{UP}}$ pour $u \in S_{\text{UP}}^{\text{exh}}$.

4.2.2 Calcul des poids répliqués des individus-panels et des ménages

La phase de calcul des poids répliqués des individus-panels correspond à une reproduction des étapes de redressements dans HVP. Ainsi, toutes les phases de correction de la non-réponse ou du statut de réponse inconnue se basent sur la même méthode que HVP, en mobilisant entre autres les mêmes groupes homogènes de réponse. De même, la procédure de suivi des individus au cours du temps, le calage des poids ainsi que le partage des poids pour la construction des poids du ménage sont reproduits.

Il faut toutefois noter que la phase de tirage initial des logements incorpore un rééchantillonnage lorsque celui-ci n'a pas été réalisé au moment de la sélection des unités primaires. La procédure est résumée dans les algorithmes 4 et 5.

Algorithme 4 Calcul des poids répliqués en t des individus-panels tirés en t_0 ($t_0 \in \{t-9, t-6, t-3, t\}$)

Pour les tirages associés au temps t_0 :

1. Pour les ménages $m \in S_{\text{log}}^{t_0}$ appartenant à une unité primaire exhaustive ou quasi-exhaustive, on introduit un facteur d'ajustement de pondération G_{m*}^{log} de la même forme que dans l'équation 4.2, les multiplicités du ménages étant obtenues après un tirage avec remise à l'intérieur de l'unité primaire. Pour les ménages m appartenant à une unité primaire normale, on pose $G_{m*}^{\text{log}} = 1$.

2. On calcule les probabilités de réponses estimées $\hat{p}_{m*}$ des ménages en utilisant les mêmes groupes homogènes de réponses que dans l'échantillonnage d'origine. Ceci donne des poids répliqués au temps t_0 pour les ménages corrigés de la non-réponse

$$w_{m*}^{\text{men},t_0 \to t_0} = \frac{G_{m*}^{\text{log}} w_{u*}^{\text{UP}}}{\pi_m^{\text{log}|\text{UP}} \hat{p}_{m*}} \text{ pour tout } m \in S_{\text{log}}^{t_0}. \tag{4.4}$$

3. On en déduit les poids répliqués individuels au temps t_0 avant redressement

$$w_{k*}^{\text{ind},t_0} = w_{m(k)*}^{\text{men},t_0 \to t_0} \text{ pour tout } k \in S_{\text{ind,rep}}^{t_0 \to t_0} \text{ avec } m(k) \text{ le ménage de l'individu } k \text{ au temps } t_0. \tag{4.5}$$

4. On estime la probabilité de réponse et de statut de réponse connu $\hat{\rho}_{k*}$ associée au suivi de l'individu k entre t_0 et t, en utilisant les mêmes groupes homogènes de réponses que dans l'échantillonnage d'origine. Ceci donne des poids répliqués au temps t corrigés de la non-réponse de réinterrogation :

$$w_{k*}^{\text{ind}} = \frac{w_{k*}^{\text{ind},t_0}}{\hat{\rho}_{k*}} \text{ pour } k \in S_{\text{ind,rep}}^{t_0 \to t} \tag{4.6}$$

Algorithme 5 Calcul des poids répliqués des ménages en t

On dispose des poids de sondage $\{w_k^{\text{ind}}\}_{k \in S^t_{\text{ind,rep}}}$.

1. Les poids répliqués calés correspondent à

$$w_{k*}^{\text{ind,cal}} = w_{k*}^{\text{ind}} \left(1 + \{\mathbf{x}_k^{\text{cal}}\}^\top \lambda_*\right), \tag{4.7}$$
$$\text{with } \lambda_* = \left(\sum_{k \in S^t_{\text{ind,rep}}} w_{k*}^{\text{ind}} \{\mathbf{x}_k^{\text{cal}}\} \{\mathbf{x}_k^{\text{cal}}\}^\top \right)^{-1} \left(t_{\mathbf{x}}^{\text{cal}} - \widehat{t}_{\mathbf{x}^{\text{cal}}*}\right)$$
$$\text{et } \widehat{t}_{\mathbf{x}^{\text{cal}}*} = \sum_{k \in S^t_{\text{ind,rep}}} w_{k*}^{\text{ind}} \mathbf{x}_k^{\text{cal}}.$$

2. Les poids répliqués des ménages au temps t est obtenu après partage des poids

$$w_{m*}^{\text{men}} = \frac{\sum_{k \in S^t_{\text{ind,rep}}} w_{k*}^{\text{ind,cal}} L_{km}}{L_{\bullet m}}. \tag{4.8}$$

4.3 Expérimentation

Cette méthode a été testée en utilisant deux jeux de variables d'équilibrage différents :

- **Jeu 1** : le poids des unités primaires et la variable constamment égale à 1. Ces deux variables assurent que les poids rééchantillonnés calés respectent les conditions d'un plan de taille fixe et que la taille de la population des unités primaires est correctement estimée.
- **Jeu 2** : les variables d'équilibrage utilisées lors du tirage des unités primaires, correspondant majoritairement à des variables sociodémographiques.

5 Comparaison des approches

Les estimations de variance sont réalisées sur des variables issues de l'enquête HVP 2020. Les variables d'intérêt sont de nature différente : quantitative (patrimoine par exemple) ou qualitative (détention d'un livret A, …). Certaines variables présentent des coefficients d'asymétrie importants (comme les variables de patrimoine).

Afin de ne pas être sensible aux échelles, nous comparons des estimations des coefficients de variation des variables $\widehat{\mathrm{CV}^{\mathrm{met}}}(\hat{\theta}) = \dfrac{\hat{\theta}}{\widehat{\mathbb{V}^{\mathrm{met}}}(\theta)}$ où $\widehat{\mathbb{V}^{\mathrm{met}}}(\theta)$ décrit l'estimateur de variance obtenu selon la méthode met (soit met vaut analyt ou boot).

La méthode de bootstrap a été appliquée en faisant varier deux facteurs :

- le seuil à partir duquel le rééchantillonnage est effectué au niveau logement plutôt qu'au niveau des unités primaires. Ce seuil peut prendre les valeurs 0, 0,1, 0,5 ou 1.
- les variables de calage utilisées pour prendre en compte l'équilibrage. Ces variables sont réparties en deux jeux de variables : variable égale à 1 et poids de tirage des unités primaires (jeu 1) et variables utilisées dans les échantillons-maîtres (jeu 2).

La table 5.1 décrit les huit scénarios obtenus en croisant les deux facteurs.

Les estimations de variance bootstrap se basent sur 1000 réplications.

Les estimations de variance sont réalisées sur 17 variables (voir la table 5.2) ayant des natures différentes : variables à distribution fortement asymétriques (patrimoine), variable binaire, …

Table 5.1: Liste des scénarios pour l'estimation par approche bootstrap.

Identifiant du scénario	Seuil d'exhaustivité	Jeu de variables
1	0	Jeu 1
2	0,1	Jeu 1
3	0,5	Jeu 1
4	1	Jeu 1
5	0	Jeu 2
6	0,1	Jeu 2
7	0,5	Jeu 2
8	1	Jeu 2

Table 5.2: Liste des variables d'intérêt

Nom de Variable	Description
PATFISOM	Montant du patrimoine financier du ménage (en clair).
PATMM	Montant du patrimoine immobilier du ménage (en clair).
PATRI_BRUT	Patrimoine brut du ménage (en clair).
PATRI_BRUT_HORSRESTE	Patrimoine brut du ménage hors patrimoine immobilier privé.
PATRI_NET	Patrimoine net du ménage.
R_CEL_1	Détention d'un Compte Épargne Logement (CEL).
R_LDD_1	Détention d'un Livret de Développement Durable (LDD).
R_LIVABL_1	Détention d'un Livret A.
R_LIVJEUN_1	Détention d'un Livret Jeune.
R_PATIMMO_1	Détention d'un patrimoine immobilier privé.
R_PATPROF_1	Détention d'un patrimoine professionnel.
R_PEA_1	Détention d'un Plan d'Épargne en Actions (PEA).
R_RESPRIN_1	Détention d'une résidence principale.
R_VMOB_1	Détention de valeurs mobilières.
ZNITVIE	Niveau de vie (source fiscale).
ZREVDEC	Revenu déclaré (source fiscale).
ZREVDISP	Revenu disponible (source fiscale).

5.1 Approche analytique contre approche bootstrap

Les figures 5.1, 5.2 et 5.3 illustrent les résultats obtenus pour les moyennes, médianes et premiers déciles de différentes variables, en utilisant pour le bootstrap un calage national[1] des unités primaires, et en utilisant pour l'approche analytique une estimation de la densité basée sur la méthode décrite par Osier (2009). Des résultats complémentaires associés aux autres quantiles sont disponibles en annexe A.

Conformément aux attentes, les estimations de variance données par l'approche bootstrap sont en général plus élevées qu'avec l'approche analytique. Cette surestimation demeure entre 5 % et 40 % pour la moyenne et la médiane, mais elle peut être très importante dans le cas du premier quartile, en donnant des estimations deux fois plus grandes pour certaines variables. Il ne semble toutefois pas y avoir de lien immédiat entre l'écart des estimations par approche analytique et par bootstrap et la variable utilisée.

Pour le cas du premier décile, les estimations de variance par bootstrap peuvent en revanche être inférieures à celles obtenues par approche analytique, l'estimation de variance pouvant être jusqu'à 6 fois moins importante. Ce résultat a été également relevé par Lamarche et Salembier

[1]Initialement, la méthode proposée par Chauvet et al. (2022) était basée sur un calage par région. Néanmoins, le nombre d'unités primaires par région avait conduit à des poids très dissipés. Nous avons décidé de réaliser le calage au niveau national.

(2015). Il est possible que la linéarisation de la méthode analytique soit trop sensible aux paramètres utilisés pour l'estimation de densité, ce qui sera testé dans la suite.

L'estimation de variance semble peu sensible :

- au seuil d'exhaustivité : l'estimation de variance augmente légèrement lorsque le seuil augmente. Pour autant, il semble difficile ici d'affirmer que cette augmentation est significative (cette augmentation pourrait être liée à l'aléa propre aux réplications bootstrap).
- au choix des variables de calage pour la prise en compte de l'équilibrage : les résultats semblent sensiblement les mêmes avec ou sans l'utilisation du calage au niveau national.

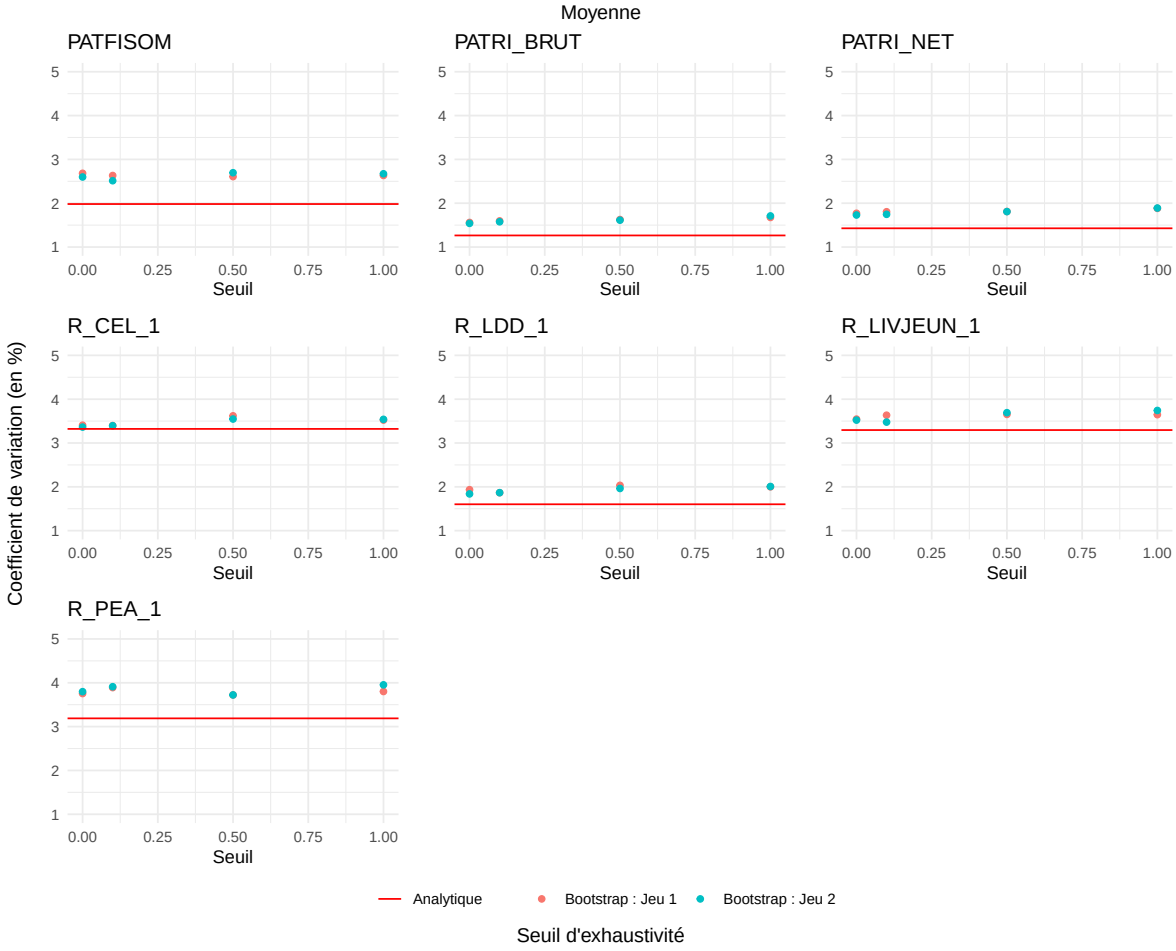


Figure 5.1: Résultats des estimations des coefficients de variation des estimations de moyenne pour quelques variables.

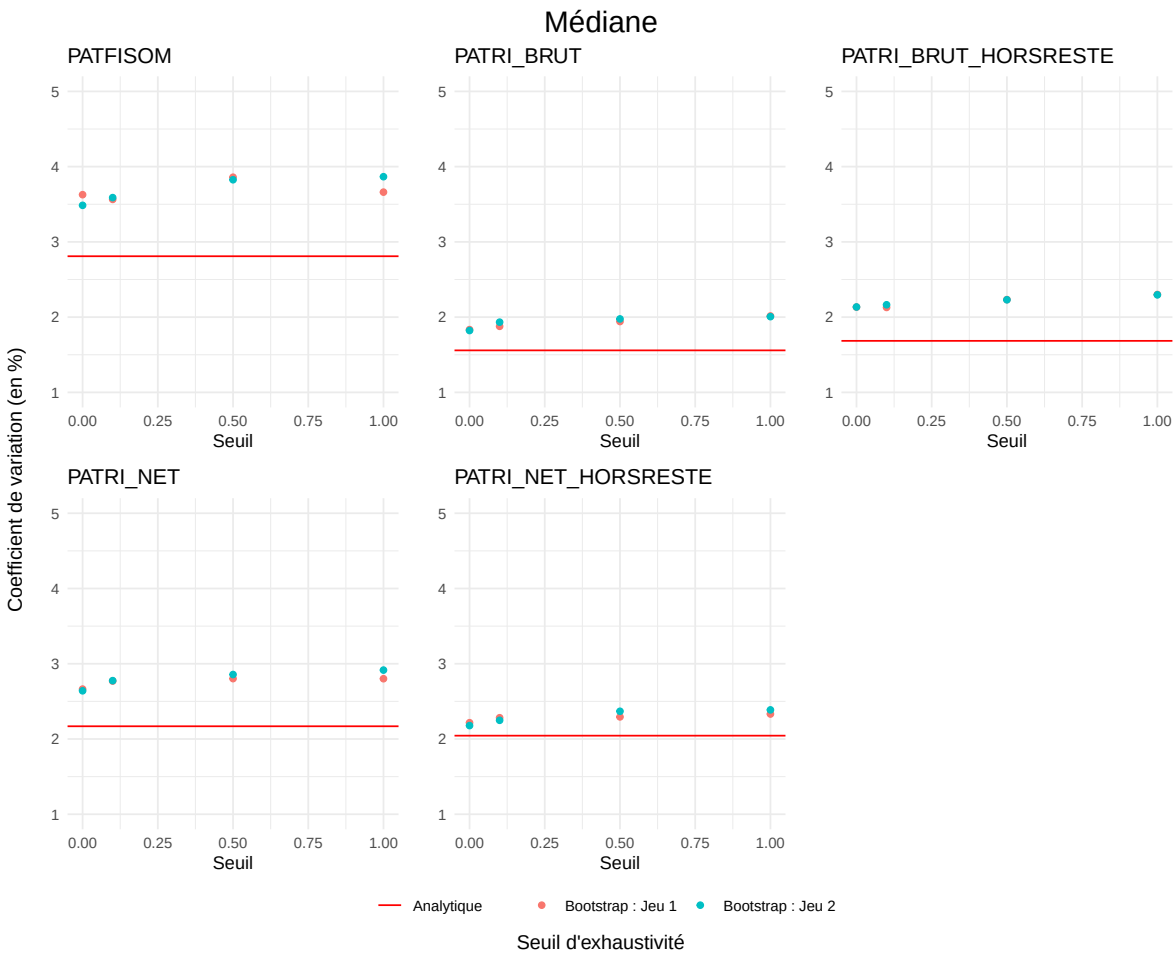


Figure 5.2: Résultats des estimations des coefficients de variation des estimations de médiane pour quelques variables.

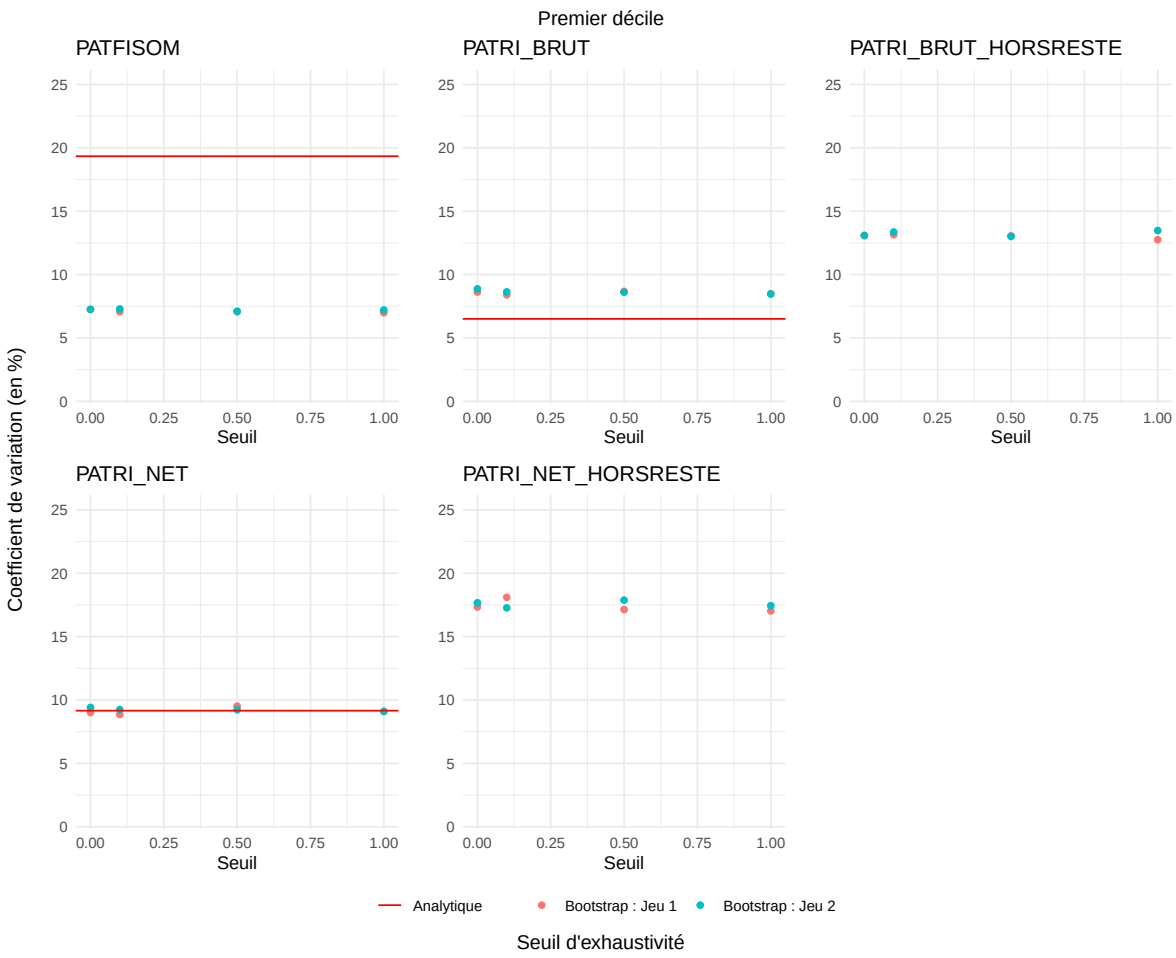


Figure 5.3: Résultats des estimations des coefficients de variation pour des estimations du premier décile. Par souci de clarté, l'estimation analytique n'est pas affichée pour les variables PATRI_BRUT_HORSRESTE et PATRI_NET_HORSRESTE.

5.2 Influence de la méthode d'estimation de la bandwidth dans l'approche analytique

La variable linéarisée estimée des estimateurs de quantile $\hat{Q}_\alpha$ est donnée pour tout individu k par $z_k = \text{IF}(\hat{Q}_\alpha)(e_k) = -\dfrac{1}{f_y(\hat{Q}_\alpha)}\dfrac{1}{\hat{N}}\left[H\left(\dfrac{\hat{Q}_\alpha - y_{(k-1)}}{y_{(k)} - y_{(k-1)}}\right) - \alpha\right]$. La densité f_y est inconnue : elle est donc estimée en mobilisant des méthodes d'estimation de densité par noyau. Dans cette partie, nous nous intéressons à l'influence du choix du paramètre de **bandwidth** sur l'estimation de la variance des estimateurs des quantiles relatifs à des variables de patrimoine.

Nous comparons quatre méthodes d'estimation : l'approche 1 correspond à l'approche décrite dans Graf et Tillé (2014) et l'approche 2 correspond à celle de Osier (2009).

Il semblerait que :

- la variance estimée par approche analytique est plus faible que celle estimée par bootstrap pour le premier quartile et la médiane. Cela est également observé pour la moyenne et les quantiles d'ordre supérieur à 0.25.
- l'approche combinant une estimation basée sur la méthode proposée par Osier (2009) et une modélisation directe conduit à une estimation de la variance plus faible qu'avec les trois autres approches dans le cas de l'estimation de variance du premier décile des variables patrimoniales. Ces estimations s'approchent de celles obtenues par estimation bootstrap.
- les fluctuations dans les estimations de variance en fonction de l'approche utilisée sont moins importantes lorsqu'on considère un quantile d'ordre supérieur.
- pour certaines variables, l'approche basée sur une modélisation du logarithme de la variable d'intérêt conduit à une estimation de la variance plus forte que celle par modélisation directe (par exemple, dans le cas de la variable PATFISOM).

Ces travaux montrent que les résultats obtenus par l'approche bootstrap sont comparables à ceux de la méthode analytique. Comme attendu, la méthode par bootstrap conduit à des estimations supérieures sauf pour les estimations de variance associées aux quantiles d'ordre inférieur à 0.1.

Table 5.3: Comparaison des coefficients de variation estimés par plusieurs approches dans le cadre analytique. Ici, l'approche par bootstrap correspond au scénario 8.

	Approche 1 - modélisation directe	Approche 2 - modélisation directe	Approche 1 - modélisation logarithme	Approche 2 - modélisation logarithme	Bootstrap
Premier décile					
PATFISOM	19.34	5.40	22.69	7.94	7.20
PATRI_BRUT	6.50	4.63	3.65	2.91	8.47
PATRI_BRUT_HORSRESTE	30.70	12.81	8.52	8.58	13.47
PATRI_NET	9.16	5.66	6.22	5.63	9.10
PATRI_NET_HORSRESTE	113.82	49.66	68.66	51.68	17.44
Premier quartile					
PATFISOM	5.18	4.40	5.99	4.40	6.28
PATRI_BRUT	3.19	3.48	3.32	3.39	5.57
PATRI_BRUT_HORSRESTE	3.94	5.66	7.78	8.11	10.40
PATRI_NET	3.46	3.90	3.80	3.90	6.37
PATRI_NET_HORSRESTE	5.01	6.65	5.58	6.58	10.24
Médiane					
PATFISOM	2.81	2.84	2.78	2.71	3.87
PATRI_BRUT	1.56	1.52	1.59	1.46	2.01
PATRI_BRUT_HORSRESTE	1.68	2.01	1.91	2.08	2.30
PATRI_NET	2.17	2.89	2.44	2.71	2.92
PATRI_NET_HORSRESTE	2.04	1.93	1.97	1.94	2.39

6 Conclusion

La comparaison des deux méthodes d'estimation de la variance appliquées à l'enquête Histoire de vie et Patrimoine 2020 apporte plusieurs enseignements essentiels. Dans l'ensemble, les résultats issus de l'approche analytique se révèlent d'un ordre de grandeur comparable à ceux obtenus par la méthode du bootstrap lorsqu'on considère des moyennes ou des quantiles dont l'ordre est plus grand que 0,25. Toutefois, certaines limites demeurent pour chacune d'entre elles.

L'approche bootstrap produit des jeux de poids relativement simples d'utilisation au prix dans la majorité des cas d'une surestimation de la variance, dont l'ampleur est plus ou moins grande selon la statistique étudiée. De plus, l'étude met en lumière la difficulté de prendre en compte l'équilibrage des unités primaires dans cette approche, puisque cela peut conduire à des poids disproportionnés. En outre, l'augmentation du seuil d'exhaustivité entraîne une légère hausse de la variance avec la méthode bootstrap, tandis que le choix du jeu de variables pour l'équilibrage n'a qu'un effet mineur sur les estimations.

L'approche analytique s'appuie sur des formules exactes ou des formules d'approximation précises des différents mécanismes, ainsi que sur des approximations par linéarisation raisonnables pour des fonctions d'intérêt régulières. Elle fait donc office de valeur de référence par rapport à l'approche bootstrap. Cependant, dans l'estimation des quantiles d'ordre faible, l'approche analytique conduit à des résultats nettement plus élevés pour certaines variables patrimoniales. Ces disparités étaient observées par Lamarche et Salembier (2015) et semblent provenir de la sensibilité de l'approche analytique à la manière d'estimer la fonction de densité pour la variable linéarisée. Ainsi, il est difficile dans le cas des quantiles d'ordre faible de déterminer dans quelle mesure l'approche analytique et l'approche bootstrap sont valides : la forte sous-estimation de la variance dans l'approche bootstrap relativement à l'approche analytique peut tout aussi bien venir d'une sous-estimation effective de la variance du côté de l'approche bootstrap que d'une sur-estimation de la variance avec l'approche analytique. D'autre part, l'approche analytique repose sur des formules nettement plus complexes d'utilisation pour un non-producteur que l'approche bootstrap. Le package `gustave` a justement pour but de réduire ce coût d'utilisation en mettant à dispostion des fonctions pour réaliser relativement simplement ces estimations (Chevalier 2023; Larbi 2024).

Il faut également rappeler que les estimations de variance réalisées ici n'intègrent pas toutes les sources d'erreurs. La non-réponse partielle et les techniques d'imputation associées rajoutent également une difficulté dans les calculs de variance et sont souvent ignorées dans les enquêtes de la statistique publique. Les analyses ont été réalisées sur des données déjà imputées et ne

prennent donc pas en compte l'aléa inhérent à cette phase. Il serait possible d'intégrer ces aléas dans de futurs travaux, toutefois l'étude de Lamarche et Salembier (2015) suggère que l'incertitude inhérente à l'imputation est faible relativement aux autres sources d'erreurs.

En perspective de l'enquête HVP 2023, ces limites et sources de variabilité identifiées pourraient être étudiées plus en détail pour améliorer les estimations de variance. Par exemple, pour le cas du bootstrap, des travaux sur la prise en compte de l'équilibrage dans le rééchantillonnage pourraient être envisagés en s'appuyant non pas sur un calage, mais sur des méthodes de rééchantillonnages équilibrés (Rubin 2024).

7 Annexe

A Résultats détaillés

Les tableaux suivants présentent les comparaisons d'estimations de la variance pour la moyenne et certains quantiles de plusieurs variables d'intérêt. Ces estimations ont été réalisées dans plusieurs scénarios décrits dans la table 5.1. Elles sont ensuite rapportées à celles de l'approche analytique : un rapport inférieur à 1 indique que la méthode conduit à des estimations de la variance plus faible que dans l'approche analytique.

Figure A.1: Estimation de la variance pour la moyenne

Méthode	Scénario	PATFISOM	PATIMM	PATRI-BRUT	PATRI-BRUT-HORSRESTE
Analytique	NA	1	1	1	1
Bootstrap	1	1.35	1.13	1.23	1.25
Bootstrap	2	1.33	1.18	1.26	1.28
Bootstrap	3	1.32	1.22	1.29	1.3
Bootstrap	4	1.33	1.28	1.32	1.33
Bootstrap	5	1.31	1.12	1.22	1.23
Bootstrap	6	1.27	1.17	1.25	1.25
Bootstrap	7	1.36	1.18	1.28	1.29
Bootstrap	8	1.35	1.26	1.35	1.37

Méthode	Scénario	PATRI-NET	PATRI-NET-HORSRESTE	ZREVDEC	ZREVDISP
Analytique	NA	1	1	1	1
Bootstrap	1	1.24	1.26	1.41	1.28
Bootstrap	2	1.26	1.28	1.42	1.33
Bootstrap	3	1.27	1.28	1.45	1.3
Bootstrap	4	1.32	1.33	1.47	1.34
Bootstrap	5	1.21	1.23	1.37	1.25
Bootstrap	6	1.22	1.23	1.44	1.33
Bootstrap	7	1.27	1.28	1.49	1.36
Bootstrap	8	1.32	1.34	1.43	1.28

Méthode	Scénario	ZNIVVIE	R-CEL-1	R-LDD-1	R-LIVABL-1
Analytique	NA	1	1	1	1
Bootstrap	1	1.49	1.03	1.21	1.19
Bootstrap	2	1.52	1.02	1.16	1.13
Bootstrap	3	1.48	1.09	1.27	1.16
Bootstrap	4	1.52	1.06	1.25	1.14
Bootstrap	5	1.48	1.01	1.15	1.17
Bootstrap	6	1.54	1.02	1.17	1.18
Bootstrap	7	1.59	1.07	1.22	1.18
Bootstrap	8	1.48	1.07	1.25	1.15

Méthode	Scénario	R-LIVJEUN-1	R-PATIMMO-1	R-PATPROF-1	R-PEA-1
Analytique	NA	1	1	1	1
Bootstrap	1	1.08	1.3	1.19	1.18
Bootstrap	2	1.1	1.33	1.25	1.22
Bootstrap	3	1.11	1.36	1.28	1.17
Bootstrap	4	1.11	1.39	1.26	1.19
Bootstrap	5	1.07	1.34	1.22	1.19
Bootstrap	6	1.06	1.37	1.27	1.23
Bootstrap	7	1.12	1.4	1.29	1.17
Bootstrap	8	1.14	1.41	1.29	1.24

Méthode	Scénario	R-RESPRIN-1	R-VMOB-1
Analytique	NA	1	1
Bootstrap	1	1.07	1.18
Bootstrap	2	1.06	1.18
Bootstrap	3	1.07	1.18

Figure A.2: Estimation de la variance pour la médiane

Méthode	Scénario	PATFISOM	PATIMM	PATRI-BRUT	PATRI-BRUT-HORSRESTE
Analytique	NA	1	1	1	1
Bootstrap	1	1.35	1.13	1.23	1.25
Bootstrap	2	1.33	1.18	1.26	1.28
Bootstrap	3	1.32	1.22	1.29	1.3
Bootstrap	4	1.33	1.28	1.32	1.33
Bootstrap	5	1.31	1.12	1.22	1.23
Bootstrap	6	1.27	1.17	1.25	1.25
Bootstrap	7	1.36	1.18	1.28	1.29
Bootstrap	8	1.35	1.26	1.35	1.37

Méthode	Scénario	PATRI-NET	PATRI-NET-HORSRESTE	ZREVDEC	ZREVDISP
Analytique	NA	1	1	1	1
Bootstrap	1	1.24	1.26	1.41	1.28
Bootstrap	2	1.26	1.28	1.42	1.33
Bootstrap	3	1.27	1.28	1.45	1.3
Bootstrap	4	1.32	1.33	1.47	1.34
Bootstrap	5	1.21	1.23	1.37	1.25
Bootstrap	6	1.22	1.23	1.44	1.33
Bootstrap	7	1.27	1.28	1.49	1.36
Bootstrap	8	1.32	1.34	1.43	1.28

Méthode	Scénario	ZNIVVIE
Analytique	NA	1
Bootstrap	1	1.49
Bootstrap	2	1.52
Bootstrap	3	1.48
Bootstrap	4	1.52
Bootstrap	5	1.48
Bootstrap	6	1.54
Bootstrap	7	1.59
Bootstrap	8	1.48

Figure A.3: Estimation de la variance pour le premier décile

Méthode	Scénario	PATFISOM	PATRI-BRUT	PATRI-BRUT-HORSRESTE	PATRI-NET
Analytique	NA	1	1	1	1
Bootstrap	1	0.37	1.33	0.43	0.98
Bootstrap	2	0.37	1.29	0.43	0.97
Bootstrap	3	0.36	1.33	0.43	1.04
Bootstrap	4	0.36	1.3	0.42	0.99
Bootstrap	5	0.38	1.36	0.43	1.03
Bootstrap	6	0.38	1.33	0.44	1.01
Bootstrap	7	0.37	1.32	0.42	1.01
Bootstrap	8	0.37	1.3	0.44	0.99

Méthode	Scénario	PATRI-NET-HORSRESTE	ZREVDEC	ZREVDISP	ZNIVVIE
Analytique	NA	1	1	1	1
Bootstrap	1	0.15	1.88	1.27	1.3
Bootstrap	2	0.16	1.87	1.33	1.35
Bootstrap	3	0.15	1.79	1.23	1.25
Bootstrap	4	0.15	1.8	1.26	1.28
Bootstrap	5	0.16	1.8	1.23	1.26
Bootstrap	6	0.15	1.81	1.26	1.28
Bootstrap	7	0.16	1.89	1.37	1.34
Bootstrap	8	0.15	1.84	1.28	1.31

Figure A.4: Estimation de la variance pour le premier quartile

Méthode	Scénario	PATFISOM	PATRI-BRUT	PATRI-BRUT-HORSRESTE	PATRI-NET
Analytique	NA	1	1	1	1
Bootstrap	1	1.12	1.62	2.45	1.75
Bootstrap	2	1.09	1.68	2.42	1.77
Bootstrap	3	1.2	1.7	2.52	1.82
Bootstrap	4	1.12	1.69	2.49	1.79
Bootstrap	5	1.09	1.68	2.51	1.76
Bootstrap	6	1.11	1.63	2.44	1.73
Bootstrap	7	1.22	1.65	2.51	1.76
Bootstrap	8	1.21	1.75	2.64	1.84

Méthode	Scénario	PATRI-NET-HORSRESTE	ZREVDEC	ZREVDISP	ZNIVVIE
Analytique	NA	1	1	1	1
Bootstrap	1	1.92	1.39	1.21	1.27
Bootstrap	2	1.84	1.37	1.17	1.26
Bootstrap	3	2.02	1.3	1.16	1.21
Bootstrap	4	1.94	1.44	1.22	1.29
Bootstrap	5	1.91	1.37	1.21	1.2
Bootstrap	6	1.87	1.42	1.25	1.28
Bootstrap	7	2.01	1.48	1.25	1.29
Bootstrap	8	2.04	1.38	1.24	1.27

Figure A.5: Estimation de la variance pour le troisième quartile

Méthode	Scénario	PATFISOM	PATIMM	PATRI-BRUT	PATRI-BRUT-HORSRESTE
Analytique	NA	1	1	1	1
Bootstrap	1	1.21	1.03	1.14	1.08
Bootstrap	2	1.25	1.06	1.18	1.1
Bootstrap	3	1.23	1.13	1.23	1.17
Bootstrap	4	1.24	1.14	1.18	1.13
Bootstrap	5	1.19	1.04	1.11	1.03
Bootstrap	6	1.25	1.05	1.17	1.08
Bootstrap	7	1.27	1.13	1.19	1.12
Bootstrap	8	1.24	1.11	1.19	1.14

Méthode	Scénario	PATRI-NET	PATRI-NET-HORSRESTE	ZREVDEC	ZREVDISP
Analytique	NA	1	1	1	1
Bootstrap	1	1.07	1.06	0.9	0.83
Bootstrap	2	1.09	1.09	0.91	0.88
Bootstrap	3	1.14	1.08	0.94	0.89
Bootstrap	4	1.15	1.08	0.94	0.92
Bootstrap	5	1.06	1.03	0.89	0.88
Bootstrap	6	1.08	1	0.95	0.96
Bootstrap	7	1.12	1.07	0.91	0.88
Bootstrap	8	1.12	1.05	0.98	0.9

Méthode	Scénario	ZNIVVIE
Analytique	NA	1
Bootstrap	1	1.13
Bootstrap	2	1.11
Bootstrap	3	1.15
Bootstrap	4	1.14
Bootstrap	5	1.12
Bootstrap	6	1.13
Bootstrap	7	1.11
Bootstrap	8	1.12

Figure A.6: Estimation de la variance pour le neuvième décile

Méthode	Scénario	PATFISOM	PATIMM	PATRI-BRUT	PATRI-BRUT-HORSRESTE
Analytique	NA	1	1	1	1
Bootstrap	1	1.17	1.03	1.11	1.4
Bootstrap	2	1.12	1.05	1.14	1.42
Bootstrap	3	1.1	1.08	1.21	1.51
Bootstrap	4	1.06	1.08	1.2	1.5
Bootstrap	5	1.14	0.99	1.08	1.39
Bootstrap	6	1.13	1.07	1.13	1.41
Bootstrap	7	1.13	1.04	1.15	1.47
Bootstrap	8	1.14	1.1	1.23	1.56

Méthode	Scénario	PATRI-NET	PATRI-NET-HORSRESTE	ZREVDEC	ZREVDISP
Analytique	NA	1	1	1	1
Bootstrap	1	1.28	1.16	1.17	1.13
Bootstrap	2	1.3	1.17	1.13	1.11
Bootstrap	3	1.33	1.15	1.21	1.15
Bootstrap	4	1.32	1.17	1.17	1.1
Bootstrap	5	1.23	1.1	1.14	1.08
Bootstrap	6	1.27	1.14	1.21	1.13
Bootstrap	7	1.28	1.13	1.23	1.22
Bootstrap	8	1.36	1.21	1.23	1.13

Méthode	Scénario	ZNIVVIE
Analytique	NA	1
Bootstrap	1	1.04
Bootstrap	2	1.03
Bootstrap	3	1.07
Bootstrap	4	1.07
Bootstrap	5	1
Bootstrap	6	1.05
Bootstrap	7	1.09
Bootstrap	8	1.07

Figure A.7: Estimation de la variance pour le quatre-vingt-neuvième centile

Méthode	Scénario	PATFISOM	PATIMM	PATRI-BRUT	PATRI-BRUT-HORSRESTE
Analytique	NA	1	1	1	1
Bootstrap	1	1.24	0.69	2.14	2.7
Bootstrap	2	1.2	0.76	2.12	2.68
Bootstrap	3	1.23	0.78	2.15	2.67
Bootstrap	4	1.27	0.84	2.19	2.75
Bootstrap	5	1.25	0.77	2.14	2.7
Bootstrap	6	1.29	0.76	2.17	2.73
Bootstrap	7	1.27	0.77	2.19	2.74
Bootstrap	8	1.27	0.83	2.24	2.81

Méthode	Scénario	PATRI-NET	PATRI-NET-HORSRESTE	ZREVDEC	ZREVDISP
Analytique	NA	1	1	1	1
Bootstrap	1	2.08	1.64	1.5	5.49
Bootstrap	2	2.01	1.61	1.49	5.59
Bootstrap	3	2.02	1.6	1.57	5.26
Bootstrap	4	2.06	1.62	1.63	5.6
Bootstrap	5	2.09	1.62	1.62	5.5
Bootstrap	6	2.14	1.66	1.61	5.57
Bootstrap	7	2.05	1.63	1.6	5.86
Bootstrap	8	2.19	1.71	1.61	5.46

Méthode	Scénario	ZNIVVIE
Analytique	NA	1
Bootstrap	1	0.49
Bootstrap	2	0.46
Bootstrap	3	0.47
Bootstrap	4	0.48
Bootstrap	5	0.48
Bootstrap	6	0.51
Bootstrap	7	0.49
Bootstrap	8	0.49

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD,
X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS
K. MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS
K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.

M
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOUR
R. LE SAOUT
B. ROUPPERT

M2016/05 : Les modèles multiniveaux
P. GIVORD
M. GUILLERM

M2016/06 : Econométrie spatiale : une introduction pratique
P. GIVORD
R. LE SAOUT

M2016/07 : La gestion de la confidentialité pour les données individuelles
M. BERGEAT

M2016/08 : Exploitation de l'enquête expérimentale Logement internet-papier
T. RAZAFINDROVONA

M2017/01: Exploitation de l'enquête expérimentale Qualité de vie au travail
T. RAZAFINDROVONA

M2018/01: Estimation avec le score de propension sous R
S. QUANTIN

M2018/02: Modèles semi-paramétriques de survie en temps continu sous R
S. QUANTIN

M2019/01 : Les méthodes de décomposition appliquées à l'analyse des inégalités
B. BOUTCHENIK
E. COUDIN
S. MAILLARD

M2020/01 : L'économétrie en grande dimension
J. L'HOUR

M2021/01 : R Tools for JDemetra+ - Seasonal adjustment made easier
A. SMYK
A. TCHANG

M2021/02 : Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman
L. CASTELL
P. SILLARD

M2021/03 : Conception de questionnaires auto-administrés
H. KOUMARIANOS
A. SCHREIBER

M2022/01 : Introduction à la géomatique pour le statisticien : quelques concepts et outils innovants de gestion, traitement et diffusion de l'information spatiale
F. SEMECURBE
E. COUDIN

M2022/02 : Le zonage en unites urbaines 2020
V. COSTEMALLE
S. OUJIA
C. GUILLO
A. CHAUVET

M2023/01 : Les réseaux de neurones appliqués à la statistique publique : méthodes et cas d'usages
D. BABET
Q. DELTOUR
T. FARIA
S. HIMPENS

M2023/02 : Redressements de la première vague de l'enquête epicov : un exemple de correction des effets de sélection dans les enquêtes multimodes
L.CASTELL
C. FAVRE-MARTINOZ
N. PALIOD
P. SILLARD

M2023/03 : Appariements de données individuelles : concepts, méthodes, conseils
L.MALHERBE

M2023/04 : Victimations déclarées et effets de mode : enseignements de l'expérimentation panel multimode de l'enquête cadre de vie et sécurité
L. CASTELL
M. CLERC
D. CROZE
S. LEGLEYE
A. NOUGARET

M2024/01 : Estimation en temps réel de la tendance-cycle : apport de l'utilisation des filtres asymétriques dans la détection des points de retournement
A.QUARTIER-LA-TENTE

M2024/02 : La disponibilité des coordonnées de contact dans fidéli-nautile - quels enseignements pour les protocoles de collecte ?
G. CHARRANCE (INED)

M2024/03 : Discuter l'existence d'un effet de sélection dans un cadre multimode grâce à une analyse de sensibilité - Application aux enquêtes annuelles de recensement
L. COURT
S. QUANTIN

M2024/04 : Vers une désaisonnalisation des séries temporelles infra-mensuelles avec JDemetra+
A. SMYK
K. WEBEL

M2025/01 : Les estimations par capture-recapture ou par système multiple : quelques éléments théoriques
P. ARDILLY
H. KOUMARIANOS

M2025/02 : Tests cognitifs pour les enquêtes auto-administrées : quelques éléments de méthode
D. GUILLEMOT
J. DIRAND
C. FLUXA

M2025/03 : Statistiques fondées sur des données administratives - esquisse d'un cadre général
H. KOUMARIANOS
P. RIVIÈRE

M2025/04 : Peut-on estimer un effet de mesure sur une enquête à partir d'un essai croisé ab/ba : la question de la non-réponse non ignorable dans l'enquête test emploi du temps
Loreline COURT
Simon QUANTIN

M2025/05 : L'apport des technologies cloude pour industrialiser le processus d'innovation statistique
R. AVOUAC
T. FARIA
F. COMTE

M2025/06 : Le data editing : Définition et principes généraux
N.CARON

M2025/07 : Estimation de la variance pour l'enquête longitudinale Histoire de vie et Patrimoine : une comparaison des approches analytique et par bootstrap
K. LARBI
J. RUBIN