Enquête santé européenne (EHIS) 2019 : calculs de précision

France métropolitaine

Thomas Deroyon

Introduction

Les estimateurs qui peuvent être calculés à partir de l'échantillon d'une enquête sont affectés d'une série d'erreurs qui les éloignent des vrais paramètres de la population qu'ils sont supposés refléter. Le processus d'enquête, de la constitution du questionnaire et du choix de l'échantillon à la diffusion des résultats, est conçu pour minimiser et contrôler ces erreurs, mais elles sont pour une large part inévitables.

Ces erreurs sont le plus souvent décrites et catégorisées via l'approche de l'erreur d'enquête totale[1] (*Total Survey Error*). Cette approche distingue deux types d'erreurs principales :

- les erreurs d'échantillonnage, liées au fait que les réponses à l'enquête ne sont obtenues que sur une fraction, sélectionnée aléatoirement, de la population ;
- les erreurs non liées à l'échantillonnage.

Les erreurs d'échantillonnage ont la particularité de ne générer que de la variance et pas de biais : en moyenne sur l'ensemble des échantillons et pourvu que le plan de sondage avec lequel a été sélectionné l'échantillon soit conçu conformément aux standards de la statistique publique[2], les estimateurs calculés sur l'échantillon sont égaux au vrai total de la population. Autrement dit, l'estimateur calculé sur chaque échantillon présente un écart avec le paramètre de la population qu'il cherche à approcher, mais la moyenne de ces écarts est nulle.

Les erreurs non liées à l'échantillonnage correspondent :

- aux erreurs de spécification (*specification error*) : ces erreurs surviennent quand il existe un écart entre le concept qui devrait être mesuré dans l'enquête et celui qui est effectivement mesuré. Le questionnaire de l'enquête santé européenne est fortement encadré par les exigences et recommandations méthodologiques européennes, desquelles le plus souvent le questionnaire français ne s'écarte que pour obtenir des réponses plus détaillées. Aussi les erreurs de spécification peuvent être considérées comme négligeables dans l'EHIS ;
- aux erreurs liées à la base de sondage (*frame error*) : ces erreurs correspondent aux écarts entre la base de sondage dans laquelle est tiré l'échantillon et la population cible de l'enquête. Elles peuvent être de deux natures :
  - erreurs de défaut de couverture : ces erreurs correspondent aux individus de la population cible de l'enquête qui n'appartiennent pas à la base de sondage et ce faisant ne peuvent être

[1] Voir P. Biemer, *Total Survey Error: design, implementation and evaluation*, Public Opinion Quarterly, 2010, R. Groves, L. Lyberg, *Total Survey Error: past, present and future*, Public Opinion Quarterly, 2010 et T. Razafindranovona, *La collecte multimode et le paradigme de l'erreur d'enquête totale*, Document de Travail de l'Insee n°M2015/01, 2015.

[2] i.e. que toutes les unités de la population aient une probabilité non nulle d'appartenir à l'échantillon, ce qui est le cas pour l'EHIS 2019.

atteints par l'échantillon. Ce point a été discuté dans la section décrivant le calcul des pondérations et notamment le calage sur marges. Rappelons rapidement les enjeux de cette discussion. Le défaut de couverture de l'échantillon est vraisemblablement très faible, et correspond aux personnes de 15 ans ou plus qui vivent en France dans un logement ordinaire au moment de la collecte de l'enquête, mais qui n'y vivaient pas au moment de la constitution de la base de sondage, soit des expatriés revenus s'installer en France, ou des personnes qui vivaient en institution et ont emménagé en résidence principale entre temps[3]. Le calage sur marges permet de réduire voire annuler ces biais de couverture, car il se fait sur des marges calculées dans l'Enquête Emploi, qui ne présente pas ce problème de défaut de couverture ;

  - erreurs de surcouverture : ces erreurs sont liées au fait que la base de sondage et partant l'échantillon contiennent des individus qui n'appartiennent pas au champ de l'enquête. De fait, l'échantillon contient des individus hors champ, dont la plupart sont identifiés lors de la collecte. Il est possible que quelques individus hors champ n'aient pas été identifiés, principalement parmi les personnes non contactées. Mais l'emploi pour la collecte de modes hétéroadministrés (collectes réalisées par des enquêteurs, au téléphone ou en face à face) rend ces situations vraisemblablement rares ;

- aux erreurs liées à la non réponse (*nonresponse error*) : comme on l'a expliqué dans la section sur le calcul des pondérations, la non réponse introduit une source d'erreur supplémentaire par rapport aux erreurs d'échantillonnage, puisque de son fait les données ne sont pas disponibles sur l'intégralité de l'échantillon sélectionné. La non réponse augmente la variance d'échantillonnage en diminuant la taille de l'échantillon disponible pour les exploitations ; elle génère un biais qui tient au fait qu'une partie de la population n'est pas représentée par l'échantillon. Ce biais peut être important si les non répondants diffèrent des répondants, notamment sur leurs réponses à l'enquête. Les corrections de la non réponse totale à l'enquête par repondération et à la non réponse partielle par imputation visent à réduire voire supprimer le biais introduit par la non réponse ;

- aux erreurs de mesure (*measurement error*) : ces erreurs correspondent aux cas où la donnée collectée ne correspond pas à la réalité de la situation de la personne interrogée ; elles peuvent être causées par un défaut de compréhension des questions, ou des effets plus complexes, liés à la désirabilité sociale ou au *satisfacing* [4], en fonction du mode de collecte ;

- aux erreurs de traitement des données (*data processing error*) : ces erreurs interviennent une fois les réponses à l'enquête collectées et désignent toutes les altérations des données collectées qui se produisent lors du traitement de celles-ci : erreurs de saisie ou de codage des questionnaires, altération des données informatiques contenant les réponses à l'enquête.. Les erreurs de traitement des données intègrent également l'effet des redressements des données permettant d'améliorer la précision de celles-ci (i.e. principalement le calage sur marges).

Idéalement, le calcul de précision devrait tenir compte de l'intégralité de ces sources d'erreur. Cela est néanmoins impossible, certaines erreurs étant difficiles à détecter et à quantifier. La pratique est plutôt de produire des estimations de précision rendant compte des erreurs pour lesquelles des méthodes d'estimation standard sont proposées par la littérature méthodologique, et de décrire de manière détaillée dans le rapport qualité de l'enquête[5] le processus de réalisation de celle-ci, notamment les mesures protectrices mises en oeuvre

[3] par exemple des jeunes qui vivaient en résidence étudiante et se sont installés dans un logement ordinaire, prisonniers libérés, personnes âgées ayant quitté un ehpad pour reemménager chez elles, personnes hospitalisées en soin de suite et de réadaptation rentrées chez elle…

[4] La désirabilité sociale désigne la tendance chez l'enquêté à vouloir se présenter sous un jour favorable à l'enquêteur, dans les modes hétéroadministrés. Ainsi l'enquêté choisit les réponses non en fonction de la réalité de sa situation, mais pour qu'elles renvoient une image de lui positive à l'enquêteur. Le satisficing renvoie aux situations où l'enquêté ne fournit pas l'effort nécessaire pour fournir une réponse de qualité et répond très rapidement à la question. Alors que la désirabilité sociale est plus fréquente dans les modes de collecte administrés par un enquêteur, le *satisficing* est plus fréquent dans les modes autoadministrés, où aucun enquêteur ne peut soutenir l'effort de l'enquêté.

[5] Disponible sur la page de l'enquête santé européenne sur le site de la Drees.

dans le processus d'enquête pour prévenir les autres. Ceci permet aux utilisateurs de l'enquête d'évaluer, en fonction de leurs sujets d'intérêt, les erreurs pouvant affecter les données qu'ils mobilisent.

En pratique, le calcul de précision de l'enquête prend en compte les sources d'erreurs suivantes :

- l'erreur d'échantillonnage ;
- l'erreur liée à la non réponse totale, qui, comme on le verra, est intégrée à la variance d'échantillonnage ;
- les gains de précision permis par le calage sur marges de l'échantillon de l'enquête.

Le choix des sources d'erreur prises en compte dans les calculs de variance est classique dans les enquêtes auprès des ménages de la statistique publique. En particulier, l'impact de la non réponse partielle et de sa correction sont rarement pris en compte, car ils sont complexes à intégrer aux calculs et dépendent étroitement de la technique d'imputation mise en oeuvre. Pour l'EHIS 2019, la non réponse partielle est par ailleurs limitée et son ampleur est beaucoup plus faible que celle de la non réponse totale, aussi son impact sur la variance d'estimation, restreint, ne paraissait pas justifier la complexité et le temps nécessaire à sa prise en compte.

Organisation du calcul de précision

On suppose que les trois sources d'erreur prises en compte dans le calcul de précision génèrent de la variance d'estimation, mais pas de biais, ou alors un biais négligeable. Pour le plan de sondage, il ne s'agit pas d'une hypothèse mais d'un fait, le plan de sondage étant construit de manière à n'empêcher la sélection d'aucun individu présent dans la base de sondage.

Pour la non réponse totale, cette hypothèse revient à considérer que la correction de la non réponse totale décrite plus haut suffit à rendre négligeable le biais de non réponse résiduel. Pour le calage sur marges, cela revient à négliger le biais à distance finie de l'estimateur calé et à se placer dans le cadre asymptotique sous lequel l'estimateur après calage est sans biais[6].

Aussi le calcul de précision de l'EHIS 2019 se réduit-il à un problème d'estimation de variance.

Pour réaliser cette estimation de variance, on utilise une approche classique en statistique d'enquête : on développe une formule analytique permettant l'estimation de la variance de l'estimateur du total d'une variable d'intérêt avec les pondérations de l'enquête. On utilise cette formule pour estimer la variance d'autres indicateurs plus complexes, en mobilisant la technique de linéarisation.

Linéarisation d'un paramètre complexe

La linéarisation d'un paramètre[7], en théorie des sondages, est une méthode qui permet d'approximer la variance de l'estimateur d'un paramètre complexe comme la variance avec laquelle est estimé le total d'une variable dépendant de la formule du paramètre et appelée linéarisée. La linéarisation permet donc de ramener la grande majorité des problèmes d'estimation de variance sur données d'enquête à l'estimation de la variance de l'estimateur du total d'une variable.

Ainsi, si on note S l'échantillon des répondants à l'EHIS 2019 et w_k^C les poids finaux calés de ces répondants, l'estimation de variance mise en oeuvre repose sur la construction d'une formule de variance exprimant à partir des paramètres décrivant le plan de sondage, la correction de la non réponse et le calage sur marges et des réponses à l'enquête, $\mathbb{V}(\sum_{k\in S}\ w_k^C\ y_k)$, la variance de l'estimateur du total de la variable d'intérêt y sous le plan de sondage, la non réponse totale et le calage sur marges.

On s'intéresse à présent à l'estimation de variance d'un paramètre plus complexe θ, qui peut s'exprimer comme une fonction régulière de totaux sur la population : $\theta\ = f\left(\sum_{k\in U}\ y_{1,k},\ldots,\sum_{k\in U}\ y_{p,k}\right)$, où U désigne le champ de l'EHIS 2019, et θ est estimé par :

---

[6] Ce qui, compte tenu des tailles élevées de la population et de l'échantillon, est une hypothèse crédible.

[7] voir J.C. Deville, *Estimations de variance pour des statistiques et des estimateurs complexes : linéarisation et technique des résidus* , Techniques d'enquête, 1999 et A. Demnati, J.N.K. Rao, *Estimation de variance par linéarisation pour des paramètres d'un modèle à partir de données d'enquête*, Recueil du Symposium 2005 de Statistiques Canada, 2005.

$$\hat{\theta} = f\left(\sum_{k\in S} w_k^C\, y_{1,k}, \ldots, \sum_{k\in S} w_k^C\, y_{p,k}\right)$$

f est une fonction à p paramètres, notés u_1 à u_p.

θ peut par exemple être le ratio de deux variables d'intérêt, dans ce cas $R(X,Y) = \left(\sum_{k\in U}\, y_k\right) / \left(\sum_{k\in U}\, x_k\right)$. La fonction f est alors la fonction qui a u et v associe le ratio v/u et l'estimateur avec l'enquête de $R(X,Y)$ est égal à :

$$\hat{R}(X,Y) = \frac{\sum_{k\in S}\, w_k^C\, y_k}{\sum_{k\in S}\, w_k^C\, x_k}$$

Considérons à présent la variable dite linéarisée de θ[8], égale à :

$$\hat{L}(\theta)_k = \sum_{j=1}^{p} \frac{\partial f}{\partial u_j}\left(\sum_{k\in S} w_k^C\, y_{1,k}, \ldots, \sum_{k\in S} w_k^C\, y_{p,k}\right) y_{p,k}$$

Alors, l'estimateur de la variance de l'estimateur du total de $\hat{L}(\theta)_k$ converge asymptotiquement vers la variance de $\hat{\theta}$:

$$\mathbb{V}\left(\sum_{k\in S} w_k^C\, \hat{L}(\theta)_k\right) \underset{|U|,|S|\to+\infty}{\to} \mathbb{V}\left(\hat{\theta}\right)$$

Dès lors qu'on sait estimer la variance de l'estimateur du total d'une variable avec notre enquête, on peut donc estimer la variance de n'importe quel paramètre θ s'exprimant comme une fonction régulière de totaux dans la population[9].

Pour reprendre l'exemple du ratio des totaux de deux variables d'intérêt, l'application de la formule générale de la linéarisée permet d'exprimer la formule de la linéarisée du ratio :

$$L\left(R(X,Y)\right)_k = \frac{y_k - \hat{R}(X,Y)\, x_k}{\sum_{k\in S}\, w_k^C\, x_k}$$

Pour mettre en oeuvre en pratique les calculs de variance à partir de la formule analytique, on utilise le package R gustave développé par Martin Chevalier au département des méthodes statistiques de l'Insee. Ce package fonctionne de la manière suivante :

- l'utilisateur développe la formule de variance analytique pour l'estimation du total d'une variable d'intérêt ;
- il programme cette formule dans une fonction sous R, appelée fonction de variance de base. Cette fonction prend comme arguments une variable d'intérêt y de l'enquête ainsi que l'ensemble des paramètres techniques décrivant le plan de sondage et les redressements nécessaires à l'estimation de variance ;
- on applique enfin la fonction gustave::define_variance_wrapper du package gustave à la fonction de variance de base et aux paramètres techniques rendant compte du plan de sondage. gustave::define_variance_wrapper produit une fonction de variance qui enveloppe la fonction de variance de base dans le sens où :

8

Plus précisément, $\hat{L}(\theta)_k$ est l'estimateur avec l'échantillon de l'enquête de la variable linéarisée $L(\theta)_k$ égale à :

$$L(\theta)_k = \sum_{j=1}^{p} \frac{\partial f}{\partial u_j}\left(\sum_{k\in U} y_{1,k}, \ldots, \sum_{k\in U} y_{p,k}\right) y_{p,k}$$

9 Cette méthode peut également être appliquée à des paramètres plus complexes, comme des quantiles ou des indices de Gini, voir F. Dell, X. D'Haultfoeuille, P. Février, E. Massé, *Mise en oeuvre de calcul de variance par linéarisation*, Actes des Journées de Méthodologie Statistique, 2002

– elle prend en charge la linéarisation pour l'estimation de variance de certains paramètres, comme les moyennes, les ratios, les ratios de ratios ;

– elle inclut dans son environnement les valeurs des paramètres techniques, que l'utilisateur n'a ainsi plus à renseigner pour chaque calcul de variance et auquel il n'a pas besoin d'avoir accès ; il est par contre nécessaire de disposer d'un fichier contenant les variables d'intérêt de l'enquête collectées auprès des répondants ;

– elle prend également en charge d'autres fonctionnalités, comme les calculs de variance sur des domaines (des sous-populations spécifiques).

- la fonction de variance ainsi développée peut être mise à disposition des utilisateurs de l'enquête avec un fichier contenant les réponses à l'enquête, la complexité des aspects technique nécessaires à la mise en oeuvre du calcul de variance étant pris en charge pour l'utilisateur dans la fonction.

On va à présent détailler comment est construite la formule de variance analytique pour l'estimation du total d'une variable y.

Prise en compte du calage sur marges

L'impact du calage sur marges sur la variance de l'estimateur du total d'une variable y peut être pris en compte également par une technique de linéarisation[10].

Si on note w_k^C le poids calé de l'individu k et w_k^{CNR} le poids de ce même individu en entrée du calage sur marges (i.e. son poids corrigé de la non réponse), si on note enfin $\hat{\epsilon}_k$ le résidu de la régression linéaire de y_k sur les variables de calage x_k estimée sur l'échantillon et pondérée par les poids w_k^C [11], alors

$$\mathbb{V}\left(\sum_{k\in S} w_k^{CNR}\,\hat{\epsilon}_k\right) \underset{|U|,|S|\to+\infty}{\to} \mathbb{V}\left(\sum_{k\in S} w_k^C\,y_k\right)$$

Tout se passe ainsi comme si le calage sur marges purgeait la variable d'intérêt de la partie de sa dispersion due aux variables de calage. Si variables d'intérêt et variables de calage sont fortement corrélées, le calage sur marges peut ainsi se traduire par des gains de précision substantiels.

Ainsi, une fois pris en compte l'effet du calage sur marges, on se ramène à l'estimation de la variance de l'estimateur du total d'une variable avec l'échantillon de l'enquête et les poids corrigés de la non réponse. On va détailler dans la section suivante comment estimer cette variance avec une formule analytique.

Prise en compte du plan de sondage

Représentation du plan de sondage pour le calcul de variance

Le plan de sondage de l'EHIS 2019 comprend plusieurs étapes de sélection, déjà présentées dans la section du DREES Méthodes consacré à la méthodologie de l'EHIS 2019 relative au plan de sondage. On les rappelle rapidement pour décrire la manière dont elles sont prises en compte dans le calcul de variance :

- d'abord, un échantillon de zones géographiques, appelées unités primaires, est tiré ;
- dans chaque résidence principale de ces unités primaires, un individu est sélectionné parmi l'ensemble des personnes dans le champ de l'enquête (i.e. âgées de 15 ans ou plus au moment de l'enquête) suivant un sondage aléatoire simple ;
- ensuite, un échantillon des individus sélectionnés à l'étape précédente est tiré suivant un tirage systématique à probabilités inégales sur fichier trié, les probabilités de tirage dépendant du nombre d'habitants du logement dans le champ de l'enquête.

---

[10] voir J.C. Deville, C.E. Särndal, *Calibration estimators in survey sampling*, Journal of the American Statistical Association, 1993

[11] Il est également possible de pondérer la régression par les poids w_k^{CNR}.

L'échantillon de l'enquête mis en collecte est formé de l'ensemble des personnes sélectionnées au terme de ces trois étapes.

L'échantillon final à partir duquel sont calculés les estimateurs de l'enquête est cependant obtenu par des phases de sélection additionnelles, correspondant à la non réponse totale. Comme on l'a décrit dans la section consacrée au calcul des pondérations de l'enquête, la non réponse totale est assimilée à un plan de sondage aléatoire dont les probabilités d'inclusion sont inconnues et sont estimées via les modèles de correction de la non réponse. Ce plan de sondage est de plus supposé poissonien, i.e. les comportements de réponse des différents membres de l'échantillon sont supposés indépendants les uns des autres.

L'échantillon des répondants est ainsi sélectionné dans l'échantillon initialement sélectionné via les étapes suivantes :

- pour les personnes orientées vers la collecte en face à face, par une étape supplémentaire de sélection des répondants dans l'échantillon mis en collecte selon un plan de sondage poissonien avec des probabilités d'inclusion estimées par les probabilités de répondre en face à face $\hat{\rho}_k^{(1)}$;
- pour les personnes orientées vers la collecte téléphonique :
  - par une première étape de sélection correspondant au tirage aléatoire des personnes répondantes ou impossibles à joindre dans l'ensemble de l'échantillon mis en collecte suivant ce mode, avec des probabilités d'inclusion estimées $\hat{\rho}_k^2$;
  - pour les personnes impossibles à joindre, par une nouvelle étape de sélection des individus basculés vers la collecte en face à face selon un tirage poissonien avec une probabilité de 0,5 ;
  - enfin, pour les personnes basculées en face à face, par une étape de sélection décrivant le tirage des individus répondants, suivant un plan de sondage poissonien avec des probabilités d'inclusion estimées $\hat{\rho}_k^{(3)}$.

Avant de décrire comment ces différentes étapes de sélection sont prises en compte dans le calcul de variance, rappelons rapidement ce que sont les sondages à plusieurs phases et à plusieurs degrés.

Dans un sondage à deux phases, un premier échantillon est tiré dans une population, puis un deuxième échantillon dans l'échantillon initial. Les sondages à deux degrés sont un type particulier de sondage à deux phases. Dans un sondage à deux degrés, la population est découpée en groupes, appelés unités primaires, d'unités de base de la population. Un échantillon d'unités primaires est tiré, puis dans chaque unité primaire, indépendamment d'une unité primaire à l'autre, un échantillon d'unités de base de la population (appelées alors unités secondaires). Grâce à l'indépendance du tirage des unités secondaires d'une unité primaire à l'autre, le calcul de variance des plans de sondage à plusieurs degrés est plus simple que dans le cas des sondages à deux phases plus généraux.

La première étape du plan de sondage, la sélection des UP, se fait suivant un plan de sondage doublement équilibré.

La deuxième étape du tirage prise en compte dans le calcul de variance correspond à la sélection des individus mis en collecte dans les unités primaires. Elle regroupe donc deux étapes du tirage décrit plus haut, i.e. le tirage dans chaque logement des UP de l'individu du champ interrogé, puis la sélection parmi ces individus de ceux effectivement conservés dans l'échantillon final.

La sélection de l'individu du champ interrogé dans chaque logement se fait par un sondage aléatoire simple stratifié par l'identifiant du logement de 1 individu par logement ; le tirage des individus finalement retenus et mis en collecte est ensuite réalisé via un tirage stratifié suivant le critère de résidence dans un quartier prioritaire de la politique de la ville (QPV) et dans chaque strate par un tirage systématique sur fichier trié par l'identifiant de l'unité primaire et d'autres caractéristiques des individus. Le tirage systématique sur fichier trié établit une stratification implicite sur le critère de tri, aussi pour le calcul de variance on va faire comme si le tirage était en fait stratifié par unité primaire et résidence dans un QPV.

Aussi, avec ces simplifications, la deuxième étape de tirage est un degré de tirage supplémentaire, puisque le tirage des individus mis en collecte est stratifié par unité primaire ou par logement dans les unités primaires, donc indépendant d'une unité primaire à l'autre.

La dernière étape de sélection de l'échantillon correspond à la non réponse. Toutes les différentes étapes de sélection des répondants dans l'échantillon initialement mis en collecte sont décrites comme des phases de

tirage poissoniennes. Plusieurs plans de sondage poissoniens successifs sont en fait identiques à un seul tirage poissonien : si on tire un premier échantillon S_1 par un plan de sondage poissonien avec des probabilités $\pi_{i,1}$, puis un deuxième échantillon S_2 dans S_1 suivant un nouveau plan de sondage poissonien avec des probabilités de tirage $\pi_{i,2}$, alors tout se passe comme si S_2 était directement sélectionné dans la base de sondage par un plan de sondage poissonien de probabilités d'inclusion $\pi_{i,1}\,\pi_{i,2}$.

La sélection des répondants dans l'échantillon mis en collecte est donc assimilable à une phase de tirage poissonienne avec les probabilités d'inclusion $\hat{\rho}_k$ égales à :

- $\hat{\rho}_k^{(1)}$ pour les individus répondants à la collecte directe en face à face ;
- $\hat{\rho}_k^{(2)}$ pour les individus répondants au téléphone ;
- $2\,\hat{\rho}_k^{(2)}\,\hat{\rho}_k^{(3)}$ pour les individus répondants en face à face après bascule depuis la collecte au téléphone.

Le plan de sondage est poissonien, aussi les comportements de réponse sont indépendants d'un individu à l'autre. Ceci implique que cette dernière étape de sélection de l'échantillon peut être décrite comme un nouveau degré de sondage, les "unités primaires" étant cette fois les individus mis en collecte, le tirage des unités secondaires dans les unités primaires revenant à déterminer pour chaque individu s'il est répondant ou pas.

Au final, pour le calcul de variance, la sélection de l'échantillon des individus répondants à l'enquête est donc assimilée à un plan de sondage à trois degrés.

On va à présent rappeler rapidement certains résultats théoriques sur l'estimation de variance dans les plans de sondage à plusieurs degrés qui vont permettre de développer la formule de variance analytique de l'EHIS 2019.

Formule de Rao[12]

La formule de Rao permet d'obtenir des expressions analytiques des variances générées par les plans de sondage à plusieurs degrés. Elle concerne les plans de sondage à deux degrés, mais peut s'appliquer, de manière itérative, aux plans de sondage à plus de deux degrés.

On se place dans le cas général d'un sondage à deux degrés.

On note S_{UP} l'échantillon d'unités primaires sélectionné dans la population, avec des poids de sondage notés w_i^{UP}, suivant un plan de sondage Π_{UP}. Dans chaque unité primaire, on sélectionne un échantillon d'unités secondaires S_i, indépendamment d'une unité primaire à l'autre, suivant un plan de sondage Π_{US}.

Avec cet échantillon, on peut dans chaque unité primaire i estimer le total $Y_i = \sum_{j\in i}\ y_j$ d'une variable d'intérêt y par un estimateur sans biais $\hat{Y}_i$. Soit V_i la variance avec laquelle $\hat{Y}_i$ estime Y_i, cette variance dépend du plan de sondage avec lequel l'échantillon d'unités secondaire est sélectionné dans chaque unité primaire : $V_i = V_{\Pi_{US}}\left(\hat{Y}_i/S_{UP}\right)$.

Si les données de toutes les unités secondaires étaient disponibles dans chaque unité primaire, nous pourrions estimer le total d'une variable y sur l'ensemble de la population par $\sum_{i\in S_{UP}}\ w_i^{UP}\,Y_i$ avec $Y_i = \sum_{j\in i}\ y_j$ le total de la variable d'intérêt dans l'unité primaire i. En pratique, cependant, seul l'estimateur $\hat{Y} = \sum_{i\in S_{UP}}\ w_i^{UP}\,\hat{Y}_i$ est accessible.

On suppose qu'on dispose de V, un estimateur sans biais de la variance de $\sum_{i\in S_{UP}}\ w_i^{UP}\,Y_i$ liée au tirage des unités primaires.

V est, dans la majeure partie des cas, une forme quadratique, i.e. a la forme $V = \sum_{i\in S_{UP}}\ Q_i\,Y_i^2 + \sum_{i\in S_{UP}}\ \sum_{j\in S_{UP},j\neq i}\ Q_{ij}\,Y_i\,Y_j$. avec Q_i et Q_{ij} des termes dépendant du plan de sondage par lequel sont sélectionnées les unités primaires.

En utilisant la formule des variances conditionnelles, il est possible de décomposer la variance de l'estimateur du total de y sous le plan de sondage à deux degrés en isolant l'effet de chaque degré du plan de sondage :

---

[12] voir N. Caron, J.C. Deville, O.Sautory, *Estimation de variance de données issues d'enquêtes*, Document de Travail de l'Unité de Méthodologie Statistique de l'Insee n°9806, 1996

$$V_{\Pi_{UP}\,\Pi_{US}}(\hat{Y}) = V_{\Pi_{UP}\,\Pi_{US}}\left(\sum_{i\in S_{UP}} w_i^{UP}\,\hat{Y}_i\right)$$

$$= V_{\Pi_{UP}}\left(E_{\Pi_{US}}\left(\sum_{i\in S_{UP}} w_i^{UP}\,\hat{Y}_i\,/\,S_{UP}\right)\right) + E_{\Pi_{UP}}\left(V_{\Pi_{US}}\left(\sum_{i\in S_{UP}} w_i^{UP}\,\hat{Y}_i\,/\,S_{UP}\right)\right)$$

Or

$$E_{\Pi_{US}}\left(\sum_{i\in S_{UP}} w_i^{UP}\,\hat{Y}_i\,/\,S_{UP}\right) = \sum_{i\in S_{UP}} w_i^{UP}\,Y_i$$

et, du fait que les tirages de second degré sont indépendants d'une unité primaire à l'autre :

$$V_{\Pi_{US}}\left(\sum_{i\in S_{UP}} w_i^{UP}\,\hat{Y}_i\,/\,S_{UP}\right) = \sum_{i\in S_{UP}} (w_i^{UP})^2\,V_{\Pi_{US}}(\hat{Y}_i/S_{UP}) = \sum_{i\in S_{UP}} (w_i^{UP})^2\,V_i$$

Aussi, la variance de l'estimateur du total de y sous les deux degrés du plan de sondage s'exprime comme la somme de deux termes :

$$V_{\Pi_{UP}\,\Pi_{US}}(\hat{Y}) = V_{UP}\left(\sum_{i\in S_{UP}} w_i^{UP}\,Y_i\right) + E_{UP}\left(\sum_{i\in S_{UP}} (w_i^{UP})^2\,V_i\right)$$

Le premier terme représente la variance résultant du tirage des unités primaires, le second terme la variance additionnelle due au tirage des unités secondaires dans les unités primaires.

Ce deuxième terme peut être estimé simplement et sans biais sous le plan de sondage par $\sum_{i\in S_{UP}} (w_i^{UP})^2\,V_i$.

Le premier terme peut être estimé en remplaçant, dans V, Y_i par son estimateur avec l'échantillon d'unités secondaires S_i tiré dans l'unité primaire i, $\hat{Y}_i$. L'estimateur ainsi obtenu, $\hat{V} = \sum_{i\in S_{UP}} Q_i\,\hat{Y}_i^2 + \sum_{i\in S_{UP}} \sum_{j\in S_{UP}, i\neq j} Q_{ij}\,\hat{Y}_i\,\hat{Y}_j$ n'est cependant pas sans biais sous le plan de sondage.

En effet,

$$E_{\Pi_{US}}(\hat{V}/S_{UP}) = \sum_{i\in S_{UP}} Q_i\,E_{\Pi_{US}}(\hat{Y}_i^2/S_{UP}) + \sum_{i\in S_{UP}} \sum_{j\in S_{UP}, i\neq j} Q_{ij}\,E_{\Pi_{US}}(\hat{Y}_i\hat{Y}_j/S_{UP})$$

Or, si $i \neq j$, les tirages dans les unités primaires i et j sont indépendants, donc conditionnellement à l'échantillon d'unités primaires, $\hat{Y}_i$ et $\hat{Y}_j$ sont indépendants : $E_{\Pi_{US}}(\hat{Y}_i\hat{Y}_j/S_{UP}) = E_{\Pi_{US}}(\hat{Y}_i/S_{UP})\,E_{\Pi_{US}}(\hat{Y}_j/S_{UP}) = Y_i\,Y_j$

Par contre, $E_{\Pi_{US}}(\hat{Y}_i/S_{UP})^2 = V_{\Pi_{US}}(\hat{Y}_i/S_{UP}) + \left(E_{\Pi_{US}}(\hat{Y}_i/S_{UP})\right)^2 = V_i + Y_i^2$.

Au final, $E_{\Pi_{US}}(\hat{V}/S_{UP}) = V + \sum_{i\in S_{UP}} Q_i\,V_i$.

$\hat{V} - \sum_{i\in S_{UP}} Q_i\,V_i$ est donc un estimateur sans biais de $V_{UP}\left(\sum_{i\in S_{UP}} w_i^{UP}\,Y_i\right)$ sous le plan de sondage.

On dispose ainsi d'un estimateur de la variance de $\sum_{i\in S_{UP}} w_i^{UP}\,\hat{Y}_i$ sans biais sous le plan de sondage :

$$\hat{V}_{\Pi_{UP},\Pi_{US}}\left(\sum_{i\in S_{UP}} w_i^{UP}\,\hat{Y}_i\right) = \hat{V} + \sum_{i\in S_{UP}} ((w_i^{UP})^2 - Q_i)\,V_i$$

Cette formule est appelée formule de Rao. Q_i est appelé terme diagonal de la forme quadratique à laquelle correspond V.

Application de la formule de Rao au plan de sondage de l'EHIS 2019

Pour appliquer la formule de Rao au plan de sondage de l'EHIS 2019, on reprend les notations utilisées dans la section précédente : S_{UP} désigne l'échantillon d'unités primaires utilisé pour l'EHIS 2019, indicées par i, w_i^{UP} la probabilité de tirage de l'unité primaire i. Pour une variable d'intérêt y, Y_i désigne le total de y dans l'unité

primaire i. V_1 désigne l'estimateur de la variance résultant du tirage des unités primaires, et $Q_i^{(1)}$ désigne son terme diagonal.

Commençons par ne pas prendre en compte la non réponse : les réponses ont alors été obtenues pour l'ensemble de l'échantillon mis en collecte. Soit $w_{j/i}$ la probabilité de tirage de l'individu j de l'unité primaire i. Y_i est inconnu mais peut être estimé par $\hat{Y}_i = \sum_{j \in S_i} w_{j/i}\, y_{ij}$, où S_i est l'échantillon d'individus tirés dans l'UP i. Si V_i désigne l'estimateur de la variance avec laquelle $\hat{Y}_i$ estime Y_i, et si $\hat{V}_1$ désigne la valeur obtenue en remplaçant dans V_1 chaque Y_i par son estimateur $\hat{Y}_i$, alors la variance de l'estimateur du total de la variable y à partir de l'échantillon complet de l'EHIS 2019 peut être estimée par :

$$V_{12} = \mathbb{V}\left(\sum_{i \in S_{UP}} \sum_{j \in S_i} w_i^{UP}\, w_{j/i}\, y_{ij}\right) = \hat{V}_1 + \sum_{i \in S_{UP}} \left((w_i^{UP})^2 - Q_i^{(1)}\right) V_i$$

Il reste cependant à prendre en compte le dernier degré de tirage, i.e. la non réponse. En pratique, seul l'estimateur $\sum_{i \in S_{UP}} \sum_{j \in R_i} \frac{w_i^{UP}\, w_{j/i}}{\hat{\rho}_{ij}}\, y_{ij}$ est disponible, où R_i désigne l'échantillon des individus répondants de l'unité primaire i et $\hat{\rho}_{ij}$ les estimateurs de leurs probabilités de réponse.

La non réponse est prise en compte à l'aide de la formule de Rao : tout se passe comme si, pour chaque individu, y_{ij} était estimé par $\hat{y}_{ij} = \frac{r_{ij}}{\hat{\rho}_{ij}}\, y_{ij}$, où r_{ij} est la variable égale à 1 si l'individu j a répondu et 0 sinon. Ainsi, comme la détermination du statut de réponse est effectuée indépendamment d'un individu à l'autre, il est possible de prendre en compte la non réponse dans la variance de l'EHIS 2019 en remplaçant dans V_{12} y_{ij} par $\hat{y}_{ij}$. Si $Q_{ij}^{(12)}$ désigne le terme diagonal de V_{12}, i.e. le facteur de y_{ij}^2 dans la formule de V_{12}, alors la variance de l'estimateur du total de y dans l'EHIS 2019 peut être estimée par :

$$\hat{\mathbb{V}}\left(\sum_{i \in S_{UP}} \sum_{j \in R_i} \frac{w_i^{UP}\, w_{j/i}}{\hat{\rho}_{ij}}\, y_{ij}\right) = \hat{V}_{12} + \sum_{i \in S_{UP}} \sum_{j \in S_i} \left(\left(w_i^{UP}\, w_{j/i}\right)^2 - Q_{ji}^{(12)}\right) V_{ij}$$

On peut remarquer que dans $\hat{V}_1$ le terme associé à y_{ij} est $Q_i^{(1)}\, w_{j/i}^2$. Dans $\sum_{i \in S_{UP}} \left((w_i^{UP})^2 - Q_i^{(1)}\right) V_i$, ce terme est égal à $\left((w_i^{UP})^2 - Q_i^{(1)}\right) Q_{ij}^{(2)}$, où $Q_{ij}^{(2)}$ est le terme diagonal associé à l'individu j dans l'estimateur de la variance de deuxième degré V_i.

Ainsi, en tenant compte de ces différents éléments, on peut exprimer $Q_{ij}^{(12)}$ et développer la formule de variance précédente :

$$\hat{\mathbb{V}}\left(\sum_{i \in S_{UP}} \sum_{j \in R_i} \frac{w_i^{UP}\, w_{j/i}}{\hat{\rho}_{ij}}\, y_{ij}\right)$$
$$= \tilde{V}_1 + \sum_{i \in S_{UP}} \left((w_i^{UP})^2 - Q_i^{(1)}\right) \hat{V}_i$$
$$+ \sum_{i \in S_{UP}} \sum_{j \in S_i} \left[\left(w_i^{UP}\, w_{ij}\right)^2 - (w_i^{UP})^2\, Q_i^{(1)} - \left((w^{UP})^2 - Q_i^{(1)}\right) Q_{ij}^{(2)}\right] V_{ij}$$

$\tilde{V}_1$ désigne l'estimation de V_1 obtenue en remplaçant pour chaque UP Y_i par l'estimateur de Y_i obtenu avec les seuls individus répondants de cette unité primaire, $\hat{V}_i$ désigne l'estimateur de V_i obtenu en remplaçant pour chaque individu y_{ij} par $\frac{r_{ij}}{\hat{\rho}_{ij}}\, y_{ij}$.

Il reste à présent à décrire les formules des estimateurs de variance de premier degré V_1, de deuxième degré V_i puis liée à la non réponse V_{ij}, ainsi que leurs termes diagonaux $Q_i^{(1)}$ et $Q_{ij}^{(2)}$.

Prise en compte de la non réponse totale

Pour estimer la variance générée par un plan de sondage poissonien, on utilise la formule de variance de Horvitz-Thompson : si on dispose d'un échantillon S sélectionné dans une population U suivant un plan de sondage

générant des probabilités d'inclusion simples π_i et des probabilités d'inclusion doubles π_{ij}[13], le total d'une variable y étant estimé à l'aide de l'estimateur d'Horvitz-Thompson : $\hat{Y} = \sum_{i \in S} y_i / \pi_i$, alors la variance de $\hat{Y}$ est estimée sans biais sous le plan de sondage par :

$$\hat{V}\left(\sum_{i \in S} \frac{y_i}{\pi_i}\right) = \sum_{i \in S} \sum_{j \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} = \sum_{i \in S} \frac{1 - \pi_i}{\pi_i^2} y_i^2 + \sum_{i \in S} \sum_{j \neq i \in S} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j}$$

Dans le cas d'un sondage poissonien, les tirages sont indépendants entre deux individus, aussi $\pi_{ij} = \pi_i \pi_j$, si bien que l'estimateur d'Horvitz-Thompson se simplifie en :

$$\hat{V}\left(\sum_{i \in S} \frac{y_i}{\pi_i}\right) = \sum_{i \in S} \frac{1 - \pi_i}{\pi_i^2} y_i^2$$

Si on applique cette formule à la prise en compte de la non réponse, alors la population U désigne chaque individu échantillonné et l'échantillon S est égal à cet individu s'il est répondant, et à un échantillon vide s'il ne répond pas. Aussi, si r_{ij} désigne la variable égale à 1 si l'individu est répondant et 0 sinon

$$V_{ij} = \frac{1 - \hat{\rho}_{ij}}{\hat{\rho}_{ij}^2} y_{ij}^2 r_{ij}$$

Prise en compte du tirage des individus dans les unités primaires

On cherche ici à donner une formule du terme V_i représentant la variance avec laquelle l'échantillon d'individus S_i tiré dans l'unité primaire i permet d'estimer le total d'une variable d'intérêt dans cette unité primaire.

L'estimation de la variance générée par la sélection des individus interrogés dans les unités primaires est complexe.

Cela tient en particulier à l'étape de tirage d'un individu dans chaque logement des unités primaires de premier degré. Il est en effet impossible d'estimer sans biais la variance générée par ce tirage, suivant un plan de sondage aléatoire simple stratifié de 1 individu dans chaque logement. Pour estimer la variance d'un sondage aléatoire simple stratifié, il faut en effet pouvoir estimer avec l'échantillon la dispersion de la variable d'intérêt dans chaque strate, ce qui suppose que l'échantillon contienne au moins 2 observations dans chacune d'elles.

Pour estimer cette variance, on simplifie le plan de sondage par lequel sont sélectionnés les individus dans chaque unité primaire : on suppose que ces individus sont sélectionnés directement en une étape, par un sondage à probabilités inégales stratifié par le croisement de l'unité primaire et du fait de résider dans un QPV. Ce faisant, on néglige le fait que l'échantillon est en réalité sélectionné en deux phases. Le poids de tirage w_{ij} des individus de l'échantillon mis en collecte dans les unités primaires de l'échantillon de premier degré est égal à l'inverse du produit d'une part des probabilités d'inclusion des individus dans chaque logement et d'autre part des individus finalement retenus pour l'échantillon mis en collecte dans l'ensemble des individus tirés à l'étape précédente.

Les strates de tirage sont formées par l'intersection des unités primaires et du statut de résidence en QPV, ce qui implique que chaque unité primaire est divisée en deux strates : la première contient l'ensemble des logements QPV de l'unité primaire situés dans un QPV, et la seconde strate les autres logements. Pour un individu j donné de l'unité primaire i, si on note h_j la strate de tirage à laquelle il appartient dans l'UP, L_{h_j} le nombre de résidences principales dans la strate h_j et enfin α_{h_j} le nombre d'individus à tirer dans cette strate suivant la procédure décrite dans la section sur le plan de sondage de l'enquête en métropole, alors :

$$w_{ij} = \frac{L_{h_j}}{\alpha_{h_j}}$$

[13] les probabilités d'inclusion simples décrivent les probabilités qu'à chaque individu de la population d'appartenir à l'échantillon compte tenu du plan de sondage, les probabilités d'inclusion doubles les probabilités qu'ont deux individus donnés d'appartenir simultanément à l'échantillon.

Pour estimer la variance associée à ce plan de sondage, on utilise la formule de Deville[14] à l'intérieur de chacune des strates de tirage. Si on indice par h les strates de tirage dans l'unité primaire i et si on note S_h l'échantillon d'individus sélectionnés dans la strate h, V_i s'exprime comme :

$$V_i = \sum_{h=1}^{2} \sum_{j \in S_h} \frac{\alpha_h}{1-\alpha_h} \left(1 - \frac{1}{w_{ij}}\right) \left(w_{ij}\, y_{ij} - \frac{\sum_{k \in S_h} \left(1 - \frac{1}{w_{ik}}\right) w_{ik}\, y_{ik}}{\sum_{k \in S_h} \left(1 - \frac{1}{w_{ik}}\right)} \right)^2$$

Le terme diagonal associé de cette formule de variance est égal à :

$$Q_{ij}^{(2)} = \frac{\alpha_h}{1-\alpha_h} w_{ij}^2 \left(1 - \frac{1}{w_{ij}}\right) \left[1 - 2 \frac{1 - \frac{1}{w_{ij}}}{\sum_{k \in S_h} \left(1 - \frac{1}{w_{ik}}\right)} + \frac{\left(1 - \frac{1}{w_{ij}}\right)^2}{\left(\sum_{k \in S_h} \left(1 - \frac{1}{w_{ik}}\right)\right)^2} \right]$$

Prise en compte du tirage des unités primaires

Le plan de sondage par lequel sont sélectionnées les unités primaires est particulièrement complexe : il s'agit d'un plan de sondage doublement équilibré inspiré des travaux de Y.Tillé et A.Grafström. Pour estimer la variance générée par ce plan de sondage, on utilise comme dans la section précédente un estimateur de variance de Deville. Ceci conduit à ne pas tenir compte de l'apport de l'équilibrage dans la variance de l'estimateur et conduit ce faisant à surestimer la variance d'échantillonnage.

Cette formule facilite cependant fortement l'implémentation du calcul de variance d'une part ; d'autre part, les variables d'équilibrage sont pour beaucoup également des variables utilisées pour le calage sur marges, aussi le gain de variance qu'elles apportent, et qui est dans les formules des estimateurs proposés par J.C. Deville et Y. Tillé[15] pris en compte en calculant la régression du total de la variable d'intérêt dans les UP sur les variables d'équilibrage, comme pour le calage sur marges, est-il sans doute faible une fois pris en compte le calage sur marges.

La formule qu'on utilise pour V_1 est donc :

$$V_1 = \sum_{l=1}^{L} \frac{n_l}{1-n_l} \sum_{i \in S_l^{UP}} (1-\pi_i) \left(\frac{Y_i}{\pi_i} - \frac{\sum_{t \in S_l^{UP}} (1-\pi_t)\, Y_t}{\sum_{t \in S_l^{UP}} (1-\pi_t)} \right)^2$$

où les l désignent les strates de tirage des unités primaires (i.e les régions), n_l désigne le nombre d'UP à tirer dans la strate l, S_l^{UP} l'échantillon d'unités primaire de la strate l, π_i la probabilité d'inclusion de l'unité primaire i et Y_i le total de la variable d'intérêt y dans l'UP i.

Le terme diagonal associé à la formule de V_1 est égal à :

$$Q_i^{(1)} = \frac{n_l}{1-n_l} \frac{1-\pi_i}{\pi_i^2} \left[1 - 2 \frac{1-\pi_i}{\sum_{t \in S_l^{UP}} (1-\pi_t)} + \frac{(1-\pi_i)^2}{\left(\sum_{t \in S_l^{UP}} (1-\pi_t)\right)^2} \right]$$

Résumé de la procédure d'estimation de variance

Les différentes étapes pour l'estimation de variance, sur la base des éléments détaillés dans les sections précédentes, sont donc :

- si l'on cherche à estimer la variance d'un paramètre autre qu'un total, le calcul de la linéarisée du paramètre ;

---

[14] voir N. Caron, J.C. Deville, O.Sautory, *Estimation de variance de données issues d'enquêtes*, Document de Travail de l'Unité de Méthodologie Statistique de l'Insee n°9806, 1996

[15] voir J.C. Deville, Y. Tillé, *Variance approximation under balanced sampling*, Journal of Statistical Planning and Inference, 2003

- le calcul des résidus de la régression de la variable d'intérêt si l'on s'intéresse à l'estimation de son total, ou de la linéarisée calculée à l'étape précédente, sur les variables de calage ;
- puis la prise en compte de la variance d'échantillonnage et de la variance causée par la non réponse totale en appliquant la formule détaillée dans la section précédente aux résidus de la régression.