Révision du plan de sondage pour l'enquête suisse sur la structure des salaires

Lionel Qualité [1]

[1] *Office Fédéral de la statistique, Espace de l'Europe 10, CH-2010 Neuchâtel, Suisse. lionel.qualite@bfs.admin.ch.*

Résumé. L'enquête suisse sur la structure des salaires permet d'obtenir des informations sur les salaires versés par les entreprises, sur les emplois concernés et sur les caractéristiques des employés qui perçoivent ces salaires. Pour l'enquête de 2018, l'Office fédéral de la statistique (OFS) a entrepris de réviser le plan d'échantillonnage. Le nouveau plan doit être défini en fonction d'objectifs de précision donnés, et permettre d'utiliser les données auxiliaires sur les revenus qui proviennent de la Centrale de Compensation (CdC, partie du système de sécurité sociale suisse). C'est également l'occasion d'adapter la procédure au tirage de Poisson avec probabilités de sélection uniformes dans des domaines qui remplace depuis 2012 le tirage stratifié. L'utilisation d'un plan de Poisson a permis de simplifier le calcul d'une allocation optimale. En effet, contrairement au plan stratifié, ce calcul ne nécessite pas de traiter des problèmes d'arrondis ou d'imposer des tailles minimales d'échantillons dans les domaines. Certains des objectifs de précision concernaient l'estimation de médianes. Pour les traiter, des calculs de variables linéarisées ont été effectués puis comparés à l'aide de simulations. Enfin, la précision attendue avec la nouvelle allocation peut être évaluée en utilisant les données de l'enquête 2016.

Mots-clés. Plan de Poisson, Allocation optimale, Médiane.

Abstract. The Swiss Earnings Structure Survey provides information on the wages paid by companies, the jobs concerned and the characteristics of the employees who receive these wages. For the 2018 survey, the Federal Statistical Office (FSO) has undertaken to revise the sampling design. The new design should be defined so as to respect given precision objectives, and allow the use of auxiliary data regarding incomes that were collected from the Central Compensation Office (part of the swiss social security system). It is also an opportunity to adapt the procedure to the Poisson sampling with uniform selection probabilities within domains that has replaced the stratified sampling design since 2012. The use of a Poisson selection method has simplified the calculation of an optimal allocation. Indeed, unlike for the stratified design, this calculation does not require dealing with rounding problems or forcing minimum sample sizes within domains. Some of the precision objectives concerned the estimation of medians. To process them, several linearized variables were calculated and compared using simulations. Finally, the expected precision with the new allocation can be assessed using the 2016 survey data.

Keywords. Poisson Sampling, Optimal allocation, Median.

1 Enquête sur la structure des salaires

L'enquête suisse sur la structure des salaires (Lohnstrukturerhebung - LSE) permet d'obtenir des informations sur les salaires versés par les entreprises et sur les caractéristiques des employés qui perçoivent ces salaires. Elle permet ainsi par exemple d'estimer le coût total de la main d'œuvre dans différentes branches d'activité, les salaires médians dans des domaines choisis, et de mettre en regard ces salaires avec les caractéristiques sociodémographiques (âge, sexe, niveau de formation,...) des personnes qui les reçoivent. Les résultats de l'enquête sont aussi utilisés pour l'élaboration du calculateur de salaires "Salarium". L'enquête est réalisée tous les deux ans auprès d'un échantillon d'entreprises sélectionnées aléatoirement dans un extrait du registre des entreprises et établissements.

Selon leur taille, ces entreprises sont invitées à renseigner l'enquête pour une partie de leurs employés : tous les employés pour les *petites* entreprises (1-19 employés), au moins la moitié des employés pour les entreprises *de taille moyenne* (20-49 employés), et au moins le tiers des employés pour les *grandes* entreprises (50 employés ou plus). Les recommandations qui accompagnent l'enquête visent à obtenir une sélection aléatoire de ces employés mais les entreprises sont, en fin de compte, libres de s'organiser comme elles l'entendent.

Avant 2012, le plan de sondage utilisé pour la sélection de l'échantillon d'entreprises était un plan stratifié. Les strates étaient un croisement des classes de taille, de catégories d'activité selon la nomenclature "NOGA" (transposition des nomenclatures européennes "NACE"), et des zones géographiques (grandes régions - "NUTS2" ou cantons d'extension de l'échantillon). Depuis 2012, les échantillons de la LSE sont sélectionnés à l'aide d'un programme de tirages coordonnés qui impose d'utiliser un plan de Poisson (voir Qualité 2009).

Pour la LSE 2018, l'Office fédéral de la statistique a entrepris de réviser le calcul des probabilités de sélection des unités, c'est-à-dire "l'allocation" de l'échantillon. Cette révision doit permettre d'obtenir un plan adapté à des objectifs de précision donnés pour différentes statistiques d'intérêt : le coût total de la main d'œuvre et les salaires médians dans des domaines choisis. Il doit également permettre de séparer explicitement l'allocation réalisée pour les besoins de la statistique fédérale et les extensions réalisées sur demande des cantons pour obtenir leurs propres objectifs de précision. Enfin, cette révision permet de considérer l'utilisation des données sur les revenus qui proviennent de la Centrale de Compensation.

Ce travail a été l'occasion de visiter ou de revisiter certains problèmes d'échantillonnage qui sont évoqués dans les sections suivantes :

1. *L'allocation d'échantillons pour un plan stratifié et pour un plan de Poisson.* Si l'on veut utiliser un plan simple stratifié, on est classiquement confronté au problème d'optimisation de l'allocation sous des contraintes de tailles minimales et de tailles maximales d'échantillons dans des strates. Il faut de plus résoudre un problème

d'arrondis. Une grande partie de la difficulté est levée lorsque l'on considère des plans de Poisson avec des probabilités de sélection uniformes dans des domaines. Il est alors en contrepartie nécessaire d'anticiper au moins grossièrement le calage qui sera utilisé lors de la pondération de l'échantillon.

2. *L'allocation d'échantillons pour estimer des médianes (ou des quantiles).* On souhaite obtenir une allocation optimale en vue d'estimer des salaires médians dans certains domaines. On pense alors naturellement à utiliser une technique de linéarisation pour pouvoir appliquer les algorithmes d'allocation usuels pour des totaux. Mais différentes options sont possibles pour calculer une variable linéarisée. Toutes demandent en outre de choisir une fonction "noyau" et une fenêtre adaptée. Des simulations montrent que, pour de petits échantillons en tout cas, le choix de la variable linéarisée n'est pas indifférent.
3. *L'estimation de la variance attendue avec le nouveau plan en utilisant les données d'une enquête passée.* La LSE peut être appréhendée comme une enquête à deux degrés : sélection aléatoire et contrôlée (à la non-réponse près) des entreprises puis sélection des salaires observés au sein de chaque entreprise. L'estimation de variance pour un plan à deux degrés demande habituellement de calculer deux termes. Si le plan de deuxième degré de la nouvelle enquête n'est pas le même que celui de l'enquête utilisée pour l'estimation de variance, il faut introduire un troisième terme dans le calcul.

2 Allocation optimale : du plan stratifié au plan de Poisson

L'allocation optimale d'un échantillon entre les strates d'un plan stratifié a fait et continue de faire l'objet de nombreuses études. On citera évidemment Neyman (1934), jusqu'à récemment Aeberhardt et Marcus (2006), Koubi et Mathern (2009), Gabler et al. (2012), Wright (2014), Friedrich et al. (2015). Les difficultés rencontrées sont les suivantes :

— La somme des probabilités de sélection dans chaque strate doit être un nombre entier pour pouvoir appliquer un plan stratifié, avec des tailles d'échantillons fixes dans les strates,
— La taille d'échantillon ne doit pas dépasser la taille de la strate, ou même une fraction de cette taille pour prendre en compte la non-réponse attendue,
— Chaque strate doit contenir une ou deux unités échantillonnées au minimum si l'on veut avoir des estimateurs sans biais des totaux et de la variance d'estimation, et ce même si la variable d'intérêt est uniforme dans la strate. Par ailleurs, on veut ici aussi éventuellement anticiper la non-réponse,
— Si l'on veut calculer un budget et une allocation en fonction d'une précision voulue, il faut de plus disposer d'algorithmes d'allocation suffisamment rapides car on est

amené à les répéter plusieurs fois pour trouver la taille d'échantillon ou le budget minimal qui permet de respecter la précision visée.

Certaines de ces contraintes sont simplement incompatibles : tailles entières d'échantillons avant et après non-réponse et taux de réponse imposé, petites strates dans la population avec au moins deux unités observées et non-réponse, arrondis et respect des taux de réponse anticipés.

Même en négligeant ces incompatibilités, le problème reste ardu. Le résultat de Gabler et al. (2012), par exemple, ne peut être juste : les auteurs proposent classiquement de chercher la solution en partant de la solution sans contraintes sur les tailles. Puis ils commencent systématiquement par traiter le cas d'une strate où la taille maximale d'échantillon est dépassée. Or, on peut facilement produire des exemples de situations où la solution optimale n'est obtenue qu'en commençant par traiter une strate où c'est la contrainte de taille minimale qui est activée. Aeberhardt et Marcus (2006) et Koubi et Mathern (2009) remarquent, à l'instar de Gabler et al. (2012) que l'on peut donner un ordre d'activation des contraintes de tailles d'échantillons dans les strates. Cependant, ils ne prouvent pas que les algorithmes proposés conduisent à la solution optimale. Plus récemment, Wright (2014) propose une solution exacte en nombres entiers mais qui peut demander des calculs conséquents, puis Friedrich et al. (2015) avancent disposer d'une solution à la fois exacte, spectaculairement rapide, et en nombres entiers. Les arguments utilisés par ces derniers sont toutefois hors de ma portée.

Ces difficultés techniques sont potentiellement sources d'erreurs qui peuvent être conséquentes, par exemple lorsqu'il y a des petites strates de grandes entreprises dans lesquelles les arrondis influent fortement sur le résultat car ils changent les taux de réponse effectivement utilisés. Après que l'enquête a eu lieu, il faut en outre souvent procéder à des regroupements de strates lorsque la non-réponse n'a pas été exactement ce qui était attendu. Ces regroupements ont pour but de continuer à pouvoir faire comme si le mécanisme de génération de l'échantillon observé était un plan simple stratifié.

Ces difficultés n'ont plus lieu d'être lorsque l'on utilise un plan de Poisson avec des probabilités de sélection uniformes dans des domaines à la place d'un plan stratifié : il n'y a plus besoin d'avoir des "tailles" entières allouées dans chaque domaine, il n'y a plus besoin non plus d'imposer des tailles minimales par domaine pour obtenir des estimateurs sans biais, il n'y a plus de problèmes de "petites strates".

Si la population U est partitionnée en domaines $h = 1, \ldots, H$ et si la probabilité de tirage des unités est uniforme égale à π_h dans chaque domaine, la variance de l'estimateur de Horvitz-Thompson du total d'une variable y_k, $k \in U$ vaut

$$\mathrm{var}(\widehat{Y}) = \sum_{h=1}^{H} \frac{1-\pi_h}{\pi_h} \sum_{k \in h} y_k^2 = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{R_h^2}{n_h},$$

où $n_h = N_h \pi_h$ n'est pas nécessairement entier et $R_h^2 = \sum_{k \in h} y_k^2 / N_h$. Cependant, un plan de Poisson ne s'utilise en général pas sans calage car il faut pouvoir éliminer la variance qui

provient de la taille aléatoire de l'échantillon. Une poststratification sur tous les domaines h conduit à remplacer R_h^2 par $\sum_{k\in h}(y_k - \bar{y}_h)^2/N_h$. Plus généralement, un calage peut être pris en compte en remplaçant R_h^2 par

$$\tilde{S}_h^2 = \frac{1}{N_h}\sum_{k\in h} e_k^2,$$

où les e_k sont les erreurs de régression de y sur les variables de calage. Dans le cas de petites "strates", une postratification complète n'est pas un scénario crédible. Il faut donc anticiper au moins grossièrement le calage qui sera effectué après enquête. En particulier, on peut réfléchir à quels domaines seront regroupés lors du calage, et ce n'est pas sans rappeler le problème de regroupement de strates post-enquêtes rencontré avec les plans stratifiés.

Les n_h optimaux sont maintenant beaucoup plus faciles à trouver : ils ne doivent plus être entiers, et sont seulement contraints par leur valeur maximale. Dans ce cas, une recherche itérative de la meilleure solution ne pose pas de problème, mais on peut aussi trouver directement la solution.

3 Allocation d'échantillons pour estimer des médianes

Une grande part des résultats de l'enquête LSE consiste en des estimations de salaires médians dans des domaines : par secteur d'activité, catégorie socio-professionnelle, genre, etc. Les programmes utilisés pour les précédentes LSE cherchaient une allocation optimale uniquement pour l'estimation de totaux. Pour la LSE 2018, nous avons intégré des objectifs de précision pour les salaires médians dans le calcul d'allocation.

Le principe de linéarisation permet d'approcher la variance d'un estimateur qui n'est pas une fonction linéaire des variables observées par la variance d'un estimateur du total d'une variable calculée, appelée "linéarisée" (voir Deville 1999). Différentes manières de calculer une telle variable ont été proposées (voir par exemple Demnati et Rao 2004). Pour le cas d'une médiane, on peut citer Osier (2009), et plus récemment Graf et Tillé (2014), Tillé et Vallée (2017). Les variables linéarisées proposées pour des quantiles font intervenir un estimateur "à noyau" de la distribution de la variable d'intérêt. Ils supposent donc de choisir une fonction noyau et un paramètre de lissage. La forme proposée par Osier (2009) est la suivante :

$$z_k = -\frac{1}{Nf(q_\alpha)}\left[I(y_k \leq q_\alpha) - \alpha\right],$$

où q_α est le quantile d'ordre α dans la population, N la taille de la population, I la fonction indicatrice, et f est une "estimation de densité" de y par la méthode des noyaux,

c'est à dire

$$f(x) = \frac{1}{\eta N} \sum_{k \in U} \phi \left(\frac{x - x_k}{\eta} \right),$$

où ϕ est une fonction noyau, et η un nombre positif qui sert de paramètre de lissage. Osier (2009) propose d'utiliser le noyau gaussien avec le choix classique $\eta = 1.06 \cdot N^{-1/5} \cdot \sigma$, ou σ est l'écart-type de la variable d'intérêt. Tillé et Vallée (2017) proposent la forme :

$$\tilde{z}_k = -\frac{1}{N f(q_\alpha)} \left[\Phi \left(\frac{q_\alpha - x_k}{\eta} \right) - F(q_\alpha) \right],$$

où Φ et F sont des primitives de ϕ et f. Tous les termes utilisés pour calculer ces linéarisées doivent être estimés lorsque l'on travaille sur un échantillon.

Notre objectif est de calculer une variable linéarisée dans la base de sondage en utilisant les revenus de la Centrale de Compensation comme proxy des salaires à enquêter, dans le but d'utiliser cette variable pour définir une allocation d'échantillon. Pour valider cette démarche, nous avons réalisé des simulations avec différents noyaux, linéarisations et choix de paramètres η. Il ne nous était malheureusement pas possible de faire des simulations avec un plan semblable à celui de la LSE (30'000 entreprises, plus d'un million et demi de salaires). Cette expérience a permis de tirer les enseignements suivants :

- Comme l'ont remarqué Graf et Tillé (2014) parmi d'autres, le choix de la fenêtre a une influence sur les résultats. Le choix $\eta = 1.06 \cdot N^{-1/5} \cdot \sigma$ est sensible aux valeurs extrêmes. Nous avons appliqué une recette classique en remplaçant σ par le minimum entre σ et l'écart inter-quartiles divisé par 1.34.
- Une fois ce problème de robustesse traité, la linéarisée calculée sur la base de sondage permet d'évaluer correctement la variance des estimateurs, et les calculs de variance par linéarisation sur chaque échantillon simulé fonctionnent correctement en termes de variance espérée et de "taux de couverture" observés. Mais ceci uniquement à la condition que l'on utilise l'estimateur de médiane qui correspond à la méthode de linéarisation ! C'est à dire pour $\tilde{z}_k$ que la médiane estimée $\hat{q}_{.5}$ doit être telle que $\widehat{F}(\hat{q}_{.5}) = 0.5$, si

  $$\widehat{F}(x) = \frac{1}{\eta \widehat{N}} \sum_{k \in s} w_k \Phi \left(\frac{x - x_k}{\eta} \right),$$

  où s est l'échantillon, w_k les poids d'extrapolation et $\widehat{N} = \sum_{k \in s} w_k$. Pour la linéarisée z_k, la médiane calculée par défaut par SAS pour un échantillon pondéré semble convenir.
- Les estimateurs de médianes calculés pour les linéarisées $\tilde{z}_k$ semblent plus précis que ceux calculés par défaut par SAS, en particulier quand les simulations comportent une étape de calage.
- Enfin, le calcul de la variable linéarisée devrait en théorie comporter un terme supplémentaire qui rend compte de la dépendance de η aux données mais dans la pratique on ne voit pas de différence sur les résultats.

Les tableaux 1 et 2 illustrent les trois derniers points. Ils sont le résultat de 10'000 répétitions de la sélection d'un échantillon de 5'000 "revenus CdC" par sondage aléatoire simple. Dans le tableau 1, les poids d'extrapolation utilisés sont les inverses des probabilités de tirage. Dans le tableau 2, les échantillons sélectionnés sont calés au niveau entreprise sur les nombres d'entreprises et le nombre d'employés par classe de taille et au niveau employé sur des classes de revenus. L'effet du calage est appréhendé par le calcul de résidus de régression. Les méthodes de linéarisation comparées pour la médiane sont :

- m0 la méthode Osier (2009) avec le noyau gaussien,
- m1 la méthode Tillé et Vallée (2017) avec le noyau gaussien,
- m2 la méthode Tillé et Vallée (2017) avec le noyau dont la fonction de répartition est $x \mapsto 1/[1 + \exp(-x)]$,
- m3 la méthode Tillé et Vallée (2017) avec un terme correctif pour la dérivée du paramètre de lissage et le noyau gaussien,
- m4 la méthode Tillé et Vallée (2017) avec un terme correctif pour la dérivée du paramètre de lissage et le noyau dont la fonction de répartition est $x \mapsto 1/[1 + \exp(-x)]$.

Les colonnes des tableaux 1 et 2 sont respectivement la méthode utilisée, la médiane calculée dans la base de sondage selon la méthode choisie, la moyenne des médianes estimées sur les 10'000 échantillons, l'écart-type de l'estimateur de la médiane approché par linéarisation sur toute la base de sondage, la moyenne des écarts-types estimés sur les échantillons, l'écart-type empirique sur les 10'000 échantillons et le "taux de couverture" : proportion des intervalles de confiance naifs ($\widehat{q}_{.5} \pm 1.96 \cdot \widehat{S}$) qui contiennent la médiane.

Méthode	Mediane	E($\widehat{q}_{.5}$)	S calculé	E($\widehat{S}$)	E.T. simul.	Tx. Couv.
m0	55'948	55'982	754.54	754.50	757.11	0.9533
m1	55'958	55'960	736.03	737.46	732.59	0.9491
m2	55'956	55'958	737.35	738.40	728.00	0.9540
m3	55'958	55'955	736.07	737.05	728.93	0.9523
m4	55'956	55'938	737.40	739.05	744.55	0.9473

TABLE 1 – Résultats de 10'000 simulations, échantillon simple de taille 5'000, poids Horvitz-Thompson.

Méthode	Mediane	E($\widehat{q}_{.5}$)	S calculé	E($\widehat{S}$)	E.T. simul.	Tx. Couv.
m0	55'948	55'974	350.92	351.08	347.80	0.9473
m1	55'958	55'963	314.88	315.60	317.72	0.9470
m2	55'956	55'959	315.86	316.50	315.40	0.9483
m3	55'958	55'960	314.88	315.41	314.81	0.9485
m4	55'956	55'951	315.85	316.78	317.77	0.9495

TABLE 2 – Résultats de 10'000 simulations, échantillon simple de taille 5'000, poids calés.

Une dernière constatation, d'intérêt pour notre problème d'allocation, est que les estimateurs de médianes présentent, avec nos données, des coefficients de variation bien plus petits que les estimateurs de totaux. Si nos objectifs de précision ne changent pas, l'allocation finale sera vraisemblablement déterminée entièrement par les objectifs pour les estimations de totaux.

4 Estimation de la variance anticipée à partir de données d'enquêtes

Les paramètres de dispersion S_h^2 nécessaires pour calculer l'allocation optimale de l'échantillon étaient jusqu'à la LSE 2016 estimés par la dispersion des totaux estimés à l'enquête précédente pour les entreprises de chaque strate h :

$$\widehat{S}_h^2 = \frac{1}{n_h - 1} \sum_{k \in s_h} \left(\widehat{T}_k - \widehat{\overline{T}}_h \right)^2,$$

où s_h est l'échantillon de la précédente LSE dans la strate h (dont la définition n'a pas changé), n_h la taille de s_h, $\widehat{T}_k$ est l'estimateur du total de la variable d'intérêt pour l'entreprise k et $\widehat{\overline{T}}_h$ est la moyenne de ces totaux dans la strate h. Pour la LSE 2018, nous allons utiliser les revenus fournis par la Centrale de Compensation (CdC) comme variables proxy et éviter ainsi l'aléa d'échantillonnage qui allait avec ces paramètres $\widehat{S}_h^2$. Par précaution, nous voudrions tout de même pouvoir estimer la précsion de l'allocation calculée à l'aide des données de la LSE 2016, pour vérifier que les modifications du plan ne sont pas trop radicales et que l'utilisation des revenus de la CdC peut être acceptée.

Avec les revenus de la CdC, la prise en compte de la seconde étape de sondage, des salaires au sein des entreprises peut se faire simplement en rappelant que la variance de l'estimateur de Horvitz-Thompson d'un total dans un plan à deux degrés s'écrit

$$\text{var}(\widehat{T}) = \sum_{i \in UP} \sum_{j \in UP} \frac{T_i}{\pi_i} \frac{T_j}{\pi_j} \Delta_{ij} + \sum_{i \in UP} \frac{\text{var}_2(\widehat{T}_i)}{\pi_i}, \tag{1}$$

où $\widehat{T}_i$ est l'estimateur du total de la variable d'intérêt pour l'unité primaire $i \in UP$ et $\text{var}_2(\widehat{T}_i)$ est la variance de second degré de cet estimateur,

$$\widehat{T} = \sum_{i \in s} \frac{\widehat{T}_i}{\pi_i},$$

i et j sont des unités primaires dans l'échantillon s, π_i, π_j sont les probabilités de sélection de ces unités primaires et Δ_{ij} est la covariance des indicatrices de sélection de i et j. Si les deux degrés de sondage sont des plans simples de taille fixe, on a

$$\text{var}(\widehat{T}) = \frac{N^2 S^2}{n} - N S_{ext}^2,$$

où

$$S^2_{ext} = \frac{1}{N-1}\sum_i (T_i - \overline{T})^2$$

et

$$S^2 = S^2_{ext} + \frac{1}{N}\sum_i m_i^2 \left(1 - \frac{n_i}{m_i}\right)\frac{S_i^2}{n_i},$$

m_i est la taille de l'unité primaire i, n_i la taille du tirage de second degré dans i, S_i^2 la variance corrigée de la variable d'intérêt dans i et $\overline{T}$ est la moyenne des T_i. Une modification simple permet en général de réutiliser les programmes d'allocation écrits pour un plan stratifié à un degré.

Lors des enquêtes précédentes, on ne cherchait pas activement à tenir compte du deuxième degré de tirage mais il est bien connu que le terme $\widehat{S}_h^2$ capture une bonne partie de l'aléa de second degré. Si les probabilités d'inclusion jointes π_{ij} sont positives et si la variance de second degré peut être estimée sans biais, alors

$$\widehat{\mathrm{var}}(\widehat{T}) = \sum_{i\in s}\sum_{j\in s}\frac{\widehat{T}_i}{\pi_i}\frac{\widehat{T}_j}{\pi_j}\frac{\Delta_{ij}}{\pi_{ij}} + \sum_{i\in s}\frac{\widehat{\mathrm{var}}_2(\widehat{T}_i)}{\pi_i},$$

est un estimateur sans biais de la variance (1).

Dans le cas où on veut estimer la variance pour une nouvelle enquête à deux degrés à partir des données d'une enquête à deux degrés précédente, une difficulté supplémentaire apparait. Il faut distinguer les probabilités d'inclusion et les variances de second degré de la nouvelle enquête et de l'ancienne. Soit π_i^n, Δ_{ij}^n les probabilités d'inclusion et covariances des indicatrices de sélection dans la nouvelle enquête, et $\mathrm{var}_2^n(\widehat{T}_i)$ la variance de second degré de la nouvelle enquête. L'estimateur

$$\widehat{V}_1 + \widehat{V}_2 = \sum_{i\in s}\sum_{j\in s}\frac{\widehat{T}_i}{\pi_i^n}\frac{\widehat{T}_j}{\pi_j^n}\frac{\Delta_{ij}^n}{\pi_{ij}} + \sum_{i\in s}\frac{\widehat{\mathrm{var}}_2(\widehat{T}_i)}{\pi_i}$$

n'estime pas sans biais la variance associée à la nouvelle enquête sauf si $\widehat{\mathrm{var}}_2(\widehat{T}_i)$ est un estimateur sans biais de $\mathrm{var}_2^n(\widehat{T}_i)$, c'est-à-dire habituellement si le plan de deuxième degré est le même dans les deux enquêtes. Dans le cas général, il faut considérer un troisième terme pour définir l'estimateur :

$$\widehat{\mathrm{var}}^n(\widehat{T}) = \sum_{i\in s}\sum_{j\in s}\frac{\widehat{T}_i}{\pi_i^n}\frac{\widehat{T}_j}{\pi_j^n}\frac{\Delta_{ij}^n}{\pi_{ij}} + \sum_{i\in s}\frac{\widehat{\mathrm{var}}_2(\widehat{T}_i)}{\pi_i} + \sum_{i\in s}\frac{1}{\pi_i\pi_i^n}\left[\widehat{\mathrm{var}}_2^n(\widehat{T}_i) - \widehat{\mathrm{var}}_2(\widehat{T}_i)\right], \quad (2)$$

où $\widehat{\mathrm{var}}_2^n(\widehat{T}_i)$ est un estimateur sans biais de $\mathrm{var}_2^n(\widehat{T}_i)$ calculé à l'aide des données de l'enquête précédente. L'estimateur (2) est alors sans biais, pour autant que la nouvelle enquête est sélectionnée dans la même population que celle qui a subi l'enquête précédente. Il se simplifie notablement lorsque le plan de la nouvelle enquête est un plan de Poisson, ou bien lorsque l'ancienne et la nouvelle enquête ont des plans simples stratifiés avec la même stratification.

Bibliographie

Aeberhardt, R. et Marcus, V. (2006). *Mesure et Contrôle de la Précision dans un Plan de Sondage Complexe. Cas de l'Enquête sur la Structure des Salaires de 2006.* Présentation à un atelier méthode de la DSE, INSEE.

Demnati, A. et Rao, J. N. K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, 30, pp. 17-34.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators : linearization and residual techniques. *Survey Methodology*, 25, pp. 193-204.

Friedrich, U., Münnich, R., de Vries, S. et Wagner, M. (2015). Fast integer-valued algorithms for optimal allocations under constraints in stratified sampling. *Computational Statistics and Data Analysis*, 92, pp. 1-12.

Gabler, S., Ganninger, M. et Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika*, 75, pp. 151-161.

Graf, E. et Tillé, Y. (2014). Variance Estimation Using Linearization for Poverty and Social Exclusion Indicators. *Survey Methodology*, 40, pp. 61-79.

Koubi, M. et Mathern, S. (2009). *La nouvelle méthode d'échantillonnage de l'enquête trimestrielle ACEMO depuis 2006. Amélioration de l'allocation de Neyman.* Document d'études de la Direction de l'animation de la recherche, des études et des statistiques, 146.

Neyman, J. (1934). On the two different aspects of the representative method : The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, pp. 558-606.

Osier, G. (2009). Variance estimation for complex indicators of poverty and inequality using linearization techniques. *Survey Research Methods*, 3, pp. 167-195.

Qualité, L. (2009). *Unequal probability sampling and repeated surveys*, Université de Neuchâtel, Neuchâtel.

Tillé, Y. et Vallée, A.-A. (2017). Variance Estimation by Linearization via the Sampling Indicators With Application to Nonresponse. *Document de travail.*

Wright, T. (2014). *A Simple Method of Exact Optimal Sample Allocation under Stratification with Any Mixed Constraint Patterns.* Research Report Series, Statistics # 2014-07, U.S. Census Bureau, Washington D.C.