Calcul des termes diagonaux de plusieurs estimateurs de variances usuels

Martin Chevalier (Insee, DMS) – 12 juillet 2016

Pour un échantillon s et une variable Y quelconque, on note $\hat{T}(Y)$ l'estimateur de son total dans la population et $\hat{V}(\hat{T}(Y))$ l'estimateur de la variance de $\hat{T}(Y)$. $\hat{V}(\hat{T}(Y))$ est une forme quadratique, c'est-à-dire qu'il existe $(q_k)_{k\in s}$ et $(q_{k,\ell})_{k\in s,\ell\in s}$ tels que :

$$\hat{V}(\hat{T}(Y)) = \sum_{k\in s} q_k y_k^2 + \sum_{k\in s}\sum_{\substack{\ell\in s\\ \ell\neq k}} q_{k,\ell} y_k y_\ell$$

Le terme q_k est appelé terme diagonal de la forme quadratique $\hat{V}(\hat{T}(Y))$. Celui-ci est particulièrement important dans la mesure où **il intervient systématiquement dans l'application de la formule de Rao utilisée pour calculer la variance d'un sondage à plusieurs degrés (Caron, Deville, Sautory, 1998)**.

En effet, dans un plan de sondage à deux degrés comptant m unités primaires, si on dispose d'un estimateur sans biais de la variance au premier degré $\hat{V}_{UP}(T_1(Y), ..., T_m(Y))$ à partir des vrais totaux des unités primaires $T_1(Y), ..., T_m(Y)$, alors on peut estimer sans biais la variance associée aux deux degrés de sondage par :

$$\hat{V}(\hat{T}(Y)) = \hat{V}_{UP}(\hat{T}_1(Y), ..., \hat{T}_m(Y)) + \sum_{k=1}^{m}\left[\frac{1}{\pi_k^2} - q_k\right]\hat{V}_{US,k}(y_1, ..., y_{n_k})$$

où π_k est la probabilité de sélection de l'unité primaire k, $\hat{V}_{US,k}$ un estimateur sans biais de la variance associée au tirage des unités secondaires au sein de l'unité primaire k et q_k le terme diagonal de $\hat{V}_{UP}$ associé à l'unité primaire k.

En pratique, cette formule est appliquée dans l'ensemble des enquêtes réalisées par l'Insee auprès des ménages (Gros, Mousallam, 2015) : d'où l'importance de connaître l'expression de ces termes diagonaux dans les cas les plus fréquents. **L'objet de cette note est de calculer les termes diagonaux de plusieurs estimateurs de variance usuels**.

Dans le contexte d'un plan de sondage à deux degrés, la forme quadratique à laquelle on s'intéresse est $\hat{V}_{UP}$, dans la mesure où c'est celle-ci qui détermine l'expression du terme diagonal q_k. Les probabilités d'inclusion simple π_k la probabilité et doubles $\pi_{k,\ell}$ utilisées dans la suite renvoient donc dans ce contexte [1] au plan de sondage des unités primaires. De même, la taille (fixe) de l'échantillon d'unités tirées est notée n.

1. La formule de Rao peut être utilisée itérativement dans le cas de plan de sondage à plus de deux degrés. On se place ici dans le cas à exactement deux degrés pour faciliter l'exposé.

Terme diagonal de l'estimateur de Horvitz-Thompson

$$\hat{V}_{HT}(\hat{T}(Y)) = \sum_{k\in s}\sum_{\ell\in s}\frac{\pi_{k,\ell}-\pi_k\pi_\ell}{\pi_{k,\ell}}\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell}$$

Le terme diagonal est le coefficient devant y_k^2, autrement dit celui qui apparaît dans la double somme pour $\ell = k$:

$$\begin{aligned} q_{HT,k} &= \frac{\pi_{k,k}-\pi_k\pi_k}{\pi_{k,k}}\times\frac{1}{\pi_k^2} \\ &= \frac{\pi_k-\pi_k^2}{\pi_k}\times\frac{1}{\pi_k^2} \\ &= \frac{1}{\pi_k^2}-\frac{1}{\pi_k} \end{aligned}$$

Remarque La forme du terme diagonal de l'estimateur de Horvitz-Thompson conduit à une simplification de la formule de Rao :

$$\begin{aligned} \hat{V}(\hat{T}(Y)) &= \hat{V}_{UP,HT}(\hat{T}_1(Y),\ldots,\hat{T}_m(Y)) + \sum_{k=1}^{m}\left[\frac{1}{\pi_k^2}-\left(\frac{1}{\pi_k^2}-\frac{1}{\pi_k}\right)\right]\hat{V}_{US,k}(y_1,...,y_{n_k}) \\ &= \hat{V}_{UP,HT}(\hat{T}_1(Y),\ldots,\hat{T}_m(Y)) + \sum_{k=1}^{m}\frac{1}{\pi_k}\times\hat{V}_{US,k}(y_1,...,y_{n_k}) \end{aligned}$$

Dans ce contexte, la formule de Rao coïncide avec la formule de Des Raj (Caron, Deville, Sautory, 1998).

Terme diagonal de l'estimateur de Sen-Yates-Grundy

$$\hat{V}_{SYG}(\hat{T}(Y)) = -\frac{1}{2}\sum_{k\in s}\sum_{\substack{\ell\in s\\ \ell\neq k}}\frac{\pi_{k,\ell}-\pi_k\pi_\ell}{\pi_{k,\ell}}\left(\frac{y_k}{\pi_k}-\frac{y_\ell}{\pi_\ell}\right)^2$$

En développant le carré cette expression se réécrit :

$$\begin{aligned} \hat{V}_{SYG}(\hat{T}(Y)) = -\frac{1}{2}\Bigg[&\sum_{k\in s}\sum_{\substack{\ell\in s\\ \ell\neq k}}\frac{\pi_{k,\ell}-\pi_k\pi_\ell}{\pi_{k,\ell}}\left(\frac{y_k}{\pi_k}\right)^2+\sum_{k\in s}\sum_{\substack{\ell\in s\\ \ell\neq k}}\frac{\pi_{k,\ell}-\pi_k\pi_\ell}{\pi_{k,\ell}}\left(\frac{y_\ell}{\pi_\ell}\right)^2 \\ &-2\sum_{k\in s}\sum_{\substack{\ell\in s\\ \ell\neq k}}\frac{\pi_{k,\ell}-\pi_k\pi_\ell}{\pi_{k,\ell}}\left(\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell}\right)\Bigg] \end{aligned}$$

En inversant par symétrie k et ℓ dans le deuxième terme, cette expression se simplifie :

$$\hat{V}_{SYG}(\hat{T}(Y)) = -\sum_{k\in s}\sum_{\substack{\ell\in s\\ \ell\neq k}} \frac{\pi_{k,\ell}-\pi_k\pi_\ell}{\pi_{k,\ell}}\left(\frac{y_k}{\pi_k}\right)^2 + \sum_{k\in s}\sum_{\substack{\ell\in s\\ \ell\neq k}} \frac{\pi_{k,\ell}-\pi_k\pi_\ell}{\pi_{k,\ell}}\left(\frac{y_k}{\pi_k}\frac{y_\ell}{\pi_\ell}\right)$$

Il ne reste alors plus qu'à chercher la valeur du terme diagonal $q_{SYG,k}$, autrement dit la valeur du coefficient devant y_k^2 quand toutes les expressions sont développées au maximum :

$$q_{SYG,k} = -\frac{1}{\pi_k^2}\sum_{\substack{\ell\in s\\ \ell\neq k}} \frac{\pi_{k,\ell}-\pi_k\pi_\ell}{\pi_{k,\ell}}$$

Terme diagonal de l'estimateur de Deville

Pour une variable Y quelconque, un estimateur de la variance de l'estimateur du total de Y peut également être obtenu en utilisant l'approximation de Deville :

$$\hat{V}_D(\hat{T}(Y)) = \frac{n}{n-1}\sum_{k\in s}(1-\pi_k)\left[\frac{y_k}{\pi_k} - \frac{\sum_{\ell\in s}(1-\pi_\ell)\frac{y_\ell}{\pi_\ell}}{\sum_{\ell\in s}(1-\pi_\ell)}\right]^2$$

À noter que plusieurs autres formes proches sont qualifiées d'approximations ou formules de Deville (Caron, Deville, Sautory, 1998, p. 8).

En développant le carré cette expression se réécrit :

$$\begin{aligned}\hat{V}_D(\hat{T}(Y)) = \frac{n}{n-1}\Bigg[&\sum_{k\in s}(1-\pi_k)\left(\frac{y_k}{\pi_k}\right)^2 + \sum_{k\in s}(1-\pi_k)\left(\frac{\sum_{\ell\in s}(1-\pi_\ell)\frac{y_\ell}{\pi_\ell}}{\sum_{\ell\in s}(1-\pi_\ell)}\right)^2\\ &-2\sum_{k\in s}(1-\pi_k)\frac{y_k}{\pi_k}\times\frac{\sum_{\ell\in s}(1-\pi_\ell)\frac{y_\ell}{\pi_\ell}}{\sum_{\ell\in s}(1-\pi_\ell)}\Bigg]\end{aligned}$$

En simplifiant les deuxième et troisième termes, on obtient alors :

$$\hat{V}_D(\hat{T}(Y)) = \frac{n}{n-1}\left[\sum_{k\in s}(1-\pi_k)\left(\frac{y_k}{\pi_k}\right)^2 - \frac{\left(\sum_{k\in s}(1-\pi_k)\frac{y_k}{\pi_k}\right)^2}{\sum_{k\in s}(1-\pi_k)}\right]$$

Il ne reste alors plus qu'à chercher la valeur du terme diagonal $q_{D,k}$, autrement dit la valeur du coefficient devant y_k^2 quand toutes les expressions sont développées au maximum :

$$\begin{aligned}q_{D,k} &= \frac{n}{n-1}\left[(1-\pi_k)\frac{1}{\pi_k^2} - \frac{(1-\pi_k)^2\frac{1}{\pi_k^2}}{\sum_{\ell\in s}(1-\pi_\ell)}\right]\\ &= \frac{n}{n-1}\times(1-\pi_k)\times\frac{1}{\pi_k^2}\times\left[1-\frac{(1-\pi_k)}{\sum_{\ell\in s}(1-\pi_\ell)}\right]\end{aligned}$$

Terme diagonal de l'estimateur de Wolter en cas de *collapse* de strates

Dans le cadre d'un **sondage aléatoire simple stratifié**, le nombre d'unités tirées par strate peut être faible. Dans ce contexte, l'estimateur de variance de Deville peut être peu performant (Caron, Deville, Sautory, 1998,p. 8). Plus encore, si moins de 2 unités sont échantillonnées dans certaines strates alors aucun estimateur direct de la variance n'est calculable.

Dans cette configuration, il est fréquent de fusionner les strates problématiques par groupes de 2 ou 3 selon un certain critère[2] : on parle de *collapse* de strates. On part de H strates indicées par h pour aboutir à P pseudo-strates indicées par p. N_h,n_h,N_p,n_p désignent respectivement le nombre d'unités appartenant à la strate h dans la population et dans l'échantillon, le nombre d'unités appartenant à la pseudo-strate p dans la population et dans l'échantillon. On note de plus H_p le nombre de strates dans la pseudo-strate p.

On note enfin $\hat{T}_h(Y)$ l'estimateur du total de Y dans la strate h et $\hat{T}_p(Y)$ l'estimateur du total de y dans la pseudo-strate p. Comme on a affaire à des sondages aléatoires simples au sein de chaque strate, l'estimateur d'Horvitz-Thompson vaut :

$$\hat{T}_h(Y) = \frac{N_h}{n_h} \sum_{k \in s_h} y_k \quad \text{et} \quad \hat{T}_p(Y) = \sum_{h=1}^{H_p} \frac{N_h}{n_h} \sum_{k \in s_h} y_k$$

Sous l'hypothèse que chaque pseudo-strate est constituée d'au moins 2 strates ($\forall k \quad H_k > 1$), l'estimateur de variance suivant (Wolter, 2008, p. 53) a la propriété de majorer la variance de l'estimateur sous le vrai plan de sondage (tirage de 1 unité dans certaines strates) :

$$V_C(\hat{T}(Y)) = \sum_{p=1}^{P} \frac{H_p}{H_p - 1} \sum_{h=1}^{H_p} \left[\hat{T}_h(Y) - \frac{\hat{T}_p(Y)}{H_p} \right]^2$$

En développant le carré cette expression se réécrit :

$$\begin{aligned}
V_C(\hat{T}(Y)) &= \sum_{p=1}^{P} \frac{H_p}{H_p - 1} \left[\sum_{h=1}^{H_p} \hat{T}_h(Y)^2 + \sum_{h=1}^{H_p} \left(\frac{\hat{T}_p(Y)}{H_p} \right)^2 - 2 \left(\sum_{h=1}^{H_p} \hat{T}_h(Y) \right) \frac{\hat{T}_p(Y)}{H_p} \right] \\
&= \sum_{p=1}^{P} \frac{H_p}{H_p - 1} \left[\sum_{h=1}^{H_p} \hat{T}_h(Y)^2 + H_p \times \left(\frac{\hat{T}_p(Y)}{H_p} \right)^2 - 2 \frac{\hat{T}_p(Y)^2}{H_p} \right] \\
&= \sum_{p=1}^{P} \frac{H_p}{H_p - 1} \left[\sum_{h=1}^{H_p} \hat{T}_h(Y)^2 - \frac{\hat{T}_p(Y)^2}{H_p} \right] \\
&= \sum_{p=1}^{P} \frac{H_p}{H_p - 1} \left[\sum_{h=1}^{H_p} \left(\frac{N_h}{n_h} \sum_{k \in s_h} y_k \right)^2 - \frac{\left(\sum_{h=1}^{H_p} \frac{N_h}{n_h} \sum_{k \in s_h} y_k \right)^2}{H_p} \right]
\end{aligned}$$

2. Le choix du critère n'est pas examiné ici. Pour un début de discussion, voir (Wolter, 2008, p. 50 *sqq*).

Il ne reste alors plus qu'à chercher la valeur du terme diagonal $q_{C,k}$, autrement dit la valeur du coefficient devant y_k^2 quand toutes les expressions sont développées au maximum :

$$q_{C,k} = \frac{H_p}{H_p - 1}\left[\left(\frac{N_h}{n_h}\right)^2 - \left(\frac{N_h}{n_h}\right)^2 \times \frac{1}{H_p}\right] = \left(\frac{N_h}{n_h}\right)^2 \frac{H_p}{H_p - 1} \times \left(1 - \frac{1}{H_p}\right) = \left(\frac{N_h}{n_h}\right)^2$$

avec h et p respectivement la strate et la pseudo-strate auxquelles appartient l'unité k.

Références

Caron N., Deville J.-C., Sautory O. (1998) *Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE*, Document de travail de l'Insee n 9806, 44 p.

Gros E., Moussallam K. (2015) *Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse*, Document de travail de l'Insee M 2015/03, 96 p.

Wolter K. (2008) *Introduction to Variance Estimation*, Springer, 450 p.