

ASSIGNATION ALÉATOIRE DE QUESTIONNAIRES PILOTES EN PRÉSENCE DE GRAPPES

Michel Hidiroglou et Pierre Lavallée¹

RÉSUMÉ

Depuis 1970, l'Enquête annuelle sur les heures et les gains (*Annual Survey of Hours and Earnings*, ASHE) est la principale source de données pour la distribution du revenu au Royaume-Uni. Elle mesure le revenu d'emploi des travailleurs du Royaume-Uni pour l'ensemble des secteurs économiques. La taille d'échantillon actuelle est d'environ 240 000 employés. Jusqu'à l'année de référence 2004, les données ont été collectées en utilisant un questionnaire de deux pages. Plusieurs questions sont cependant sur le point d'être ajoutées, ce qui devrait mener à l'utilisation d'un nouveau questionnaire de six pages. De manière à avoir une idée de la différence des réponses entre les deux questionnaires, un sous-échantillon a été tiré et 5 000 nouveaux questionnaires ont été envoyés aux employés pour l'année de référence 2004. L'envoi des questionnaires a été effectué en respectant la contrainte voulant que les employés sélectionnés d'une même entreprise (ou grappe) reçoivent tous le même questionnaire. Cet article donne les calculs visant à mesurer les différences entre les deux questionnaires pilotes, compte tenu du fait que les questionnaires sont assignés aléatoirement tout en respectant les grappes.

1. INTRODUCTION

L'Enquête annuelle sur les heures et les gains (*Annual Survey of Hours and Earnings*, ASHE) est la principale source de données pour la distribution du revenu au Royaume-Uni. L'ASHE a été effectuée annuellement depuis 1970 sans changement notable. Elle mesure le revenu d'emploi des travailleurs du Royaume-Uni pour l'ensemble des secteurs économiques. L'ASHE est menée par l'*Office for National Statistics* (ONS) de Grande-Bretagne et par le Département de l'entreprise, du commerce et de l'investissement de l'Irlande du Nord.

La majeure partie de l'échantillon de 1% d'employés est tiré par le *Inland Revenue*, le département du Royaume-Uni responsable de la collecte de l'impôt sur le revenu, à partir du système PAYE (*Pay As You Earn*) par le biais des deux derniers chiffres du numéro d'assurance nationale des employés. Une petite proportion de l'échantillon est identifiée à partir du même critère de sélection par des employeurs qui retournent l'information électroniquement à l'ONS. Ce plan de sondage crée en fait un panel d'employés. La taille d'échantillon actuelle est d'environ 240 000 employés. Pour plus de détail concernant l'ASHE, on peut consulter Pont (2003).

Les numéros d'assurance nationale sélectionnés sont appariés au Registre des entreprises interdépartemental (*Interdepartmental Business Register*, IDBR) au niveau entreprise. Les entreprises identifiées doivent alors remplir un questionnaire pour fournir des données pour chacun de leurs employés qui font partie de l'échantillon de l'ASHE. Ces données détaillées portent sur leurs revenus, leurs heures travaillées et une description de leur emploi, pour ne nommer que celles-ci. Pour simplifier, on suppose ici qu'un employé n'appartient qu'à une seule entreprise.

Jusqu'à l'année de référence 2004, les données ont été collectées en utilisant un questionnaire de deux pages. Plusieurs questions sont cependant sur le point d'être ajoutées, ce qui devrait mener à l'utilisation d'un nouveau questionnaire de six pages. De manière à avoir une idée de la différence des réponses entre les deux questionnaires, un sous-échantillon a été tiré et 5 000 nouveaux questionnaires ont été envoyés aux employés pour l'année de référence 2004. L'envoi des questionnaires a été effectué en respectant la contrainte voulant que les employés sélectionnés d'une même entreprise (ou grappe) reçoivent tous le même questionnaire. Cette contrainte complique cependant le calcul des estimations, ainsi que de leur variance.

¹ Michel Hidiroglou, *Office for National Statistics*, Royaume-Uni (mike.hidiroglou@ons.gov.uk), Pierre Lavallée, Division des méthodes d'enquêtes sociales, Statistique Canada, K1A 0T6 (pierre.lavallee@statcan.ca)

2. ASSIGNATION ALÉATOIRE DES DEUX VERSIONS DU QUESTIONNAIRE

Soit U^A , la base de sondage contenant N^A numéros d'assurance nationale pour la sélection des employés. Soit U^B , l'univers des N^B entreprises du IDBR associées aux employés. Chaque entreprise i de la base U^B peut être vue comme une grappe d'employés liés à la base U^A .

Un premier échantillon s^A de m employés est tiré par sondage aléatoire simple sans remise de U^A . On identifie les employés par l'indice k . Chaque échantillon s^A est alors divisé en deux sous-échantillons s_{2h}^A et s_{6h}^A destinés à recevoir les questionnaires de deux et de six pages, respectivement. Les deux sous-échantillons contiennent m_{2h} et m_{6h} employés chacun.

La répartition des employés de s^A aux deux sous-échantillons est faite de sorte que les employés d'une même entreprise i (ou grappe) reçoivent nécessairement le même questionnaire de deux ou de six pages. Plus précisément, on procède comme suit : l'échantillon s^A est premièrement amené au niveau des grappes, ce qui définit l'échantillon s^B de n grappes où $s^B = \{i \mid k \in i, i \in U^B, k \in s^A\}$. L'échantillon s^B est alors stratifié en H strates où chaque strate est identifiée par l'indice h , c'est-à-dire $s^B = \bigcup_{h=1}^H s_h^B$. Ensuite, dans chaque s_h^B , on tire un échantillon aléatoire simple sans remise s_{6h}^B de n_{6h} grappes qui recevront un questionnaire de six pages, et le reste de l'échantillon s_h^B des grappes recevra un questionnaire de deux pages. On dénote ce sous-échantillon de n_{2h}^B grappes par s_{2h}^B . On ramène finalement les sous-échantillons s_{2h}^B et s_{6h}^B au niveau des employés, ce qui nous donne les sous-échantillons s_{2h}^A et s_{6h}^A de m_{2h} et m_{6h} employés, où $s_{2h}^A = \{k \mid k \in i, i \in s_{2h}^B, k \in s^A\}$ et $s_{6h}^A = \{k \mid k \in i, i \in s_{6h}^B, k \in s^A\}$, $h=1, \dots, H$. On note que $s_h^A = s_{2h}^A \cup s_{6h}^A$, $s_h^B = s_{2h}^B \cup s_{6h}^B$, $m_h = m_{2h} + m_{6h}$ et $n_h = n_{2h} + n_{6h}$, $h=1, \dots, H$.

À partir des sous-échantillons s_{2h}^B et s_{6h}^B , on enquête auprès des employés de s_{2h}^A et s_{6h}^A en utilisant la version du questionnaire rattachée à chaque sous-échantillon. Une fois les données recueillies, on cherche alors à savoir s'il existe une différence significative entre les réponses des deux versions du questionnaire. On s'intéresse ainsi aux totaux $Y_2 = \sum_{k=1}^M y_{2k}$ et $Y_6 = \sum_{k=1}^M y_{6k}$, ainsi qu'à la différence $D = Y_2 - Y_6$ pour la même variable d'intérêt y mesurée par les questionnaires de deux et de six pages.

3. ESTIMATION DES TOTAUX Y_2 ET Y_6

Dans cette section, nous nous limiterons à l'estimation du total Y_2 puisque que l'estimation du total Y_6 est parfaitement similaire.

L'obtention de l'échantillon s_{2h}^A peut être vu comme un sondage à deux phases. À la première phase, on tire l'échantillon s^A qui est alors ramené au niveau des grappes. On travaille alors avec l'échantillon s^B qui est divisé en H strates et où chaque échantillon s_h^B est divisé de nouveau en s_{2h}^B et s_{6h}^B . L'échantillon s_{2h}^A correspond alors simplement aux employés choisis des grappes de s_{2h}^B . Il est important de noter que l'échantillon s^A détermine complètement l'échantillon s^B .

Pour estimer Y_2 , on peut utiliser l'estimateur suivant :

$$\hat{Y}_2 = \frac{M}{m} \sum_{h=1}^H \frac{n_h}{n_{2h}} \sum_{i=1}^{n_{2h}} y_{2hi} \quad (1)$$

où $y_{2hi} = \sum_{k=1}^{m_{2hi}} y_{2hik}$. La quantité y_{2hi} correspond au total des m_{2hi} employés de la strate h de la grappe i qui recevront le questionnaire de deux pages. Notons que lors de son application à l'ASHE, l'estimateur utilisé sera relativement plus complexe que l'estimateur (2) parce qu'il y aura un ajustement pour la non-réponse, ainsi que du calage sur marges. L'estimateur (2) sert ici à montrer une partie de la problématique reliée au choix des deux versions du questionnaire. On peut démontrer que l'estimateur (1) est sans biais. Soit $t_k = 1$ si $k \in s^A$, et 0 sinon, et soit $u_{2hi} = 1$ si $i \in s_{2h}^B$, et 0 si $i \in s_{6h}^B$. On peut alors écrire l'estimateur (1) de la forme

$$\hat{Y}_2 = \frac{M}{m} \sum_{h=1}^H \frac{n_h}{n_{2h}} \sum_{i=1}^{n_h} u_{2hi} y_{2hi} \quad (2)$$

En conditionnant sur s^A (et donc sur s^B), on peut démontrer que

$$E(\hat{Y}_2 | s^A) = \frac{M}{m} \sum_{k=1}^M t_k y_{2k} \quad (3)$$

Puisque $E(t_k) = m/M$, on obtient finalement que $E(\hat{Y}_2) = Y_2$.

Pour la variance de l'estimateur (1), on utilise de nouveau l'approche conditionnelle en partant de l'expression suivante :

$$Var(\hat{Y}_2) = E(Var(\hat{Y}_2 | s^A)) + Var(E(\hat{Y}_2 | s^A)) \quad (4)$$

À partir de (3), on obtient directement

$$Var(E(\hat{Y}_2 | s^A)) = M^2 \left(1 - \frac{m}{M}\right) \frac{S_{2y}^2}{m} \quad (5)$$

où $S_{2y}^2 = \sum_{k=1}^M (y_{2k} - \bar{Y}_2)^2 / (M - 1)$ et $\bar{Y}_2 = \sum_{k=1}^M y_{2k} / M$. Pour obtenir $Var(\hat{Y}_2 | s^A)$, on utilise l'expression (2) et on obtient

$$\begin{aligned} Var(\hat{Y}_2 | s^A) &= \frac{M^2}{m^2} \sum_{h=1}^H \frac{n_h^2}{n_{2h}} \left(1 - \frac{n_{2h}}{n_h}\right) \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(y_{2hi} - \frac{\sum_{i=1}^{n_h} y_{2hi}}{n_h}\right)^2 \\ &= \frac{M^2}{m^2} \sum_{h=1}^H \frac{n_h(n_h - n_{2h})}{n_{2h}} \tilde{S}_{2h}^2 \end{aligned} \quad (6)$$

où $\tilde{S}_{2h}^2 = \sum_{i=1}^{n_h} (y_{2hi} - \bar{y}_{2h})^2 / (n_h - 1)$ et $\bar{y}_{2h} = \sum_{i=1}^{n_h} y_{2hi} / n_h$. Il ne reste alors qu'à prendre l'espérance de (6) pour obtenir $E(Var(\hat{Y}_2 | s^A))$. On obtient ainsi la variance de l'expression (4) donnée par

$$Var(\hat{Y}_2) = M^2 \left(1 - \frac{m}{M}\right) \frac{S_{2y}^2}{m} + E \left[\frac{M^2}{m^2} \sum_{h=1}^H \frac{n_h(n_h - n_{2h})}{n_{2h}} \tilde{S}_{2h}^2 \right] \quad (7)$$

Pour estimer $Var(\hat{Y}_2)$, on peut utiliser

$$\begin{aligned} var(\hat{Y}_2) &= \frac{M(M-m)}{m(m-1)} \left\{ \sum_{h=1}^H \frac{n_h}{n_{2h}} \sum_{i=1}^{n_{2h}} \sum_{k=1}^{m_{2hi}} y_{2hik}^2 - \frac{1}{m} \left(\sum_{h=1}^H \frac{n_h}{n_{2h}} \sum_{i=1}^{n_{2h}} y_{2hi} \right)^2 \right\} \\ &+ \frac{M(M-1)}{m(m-1)} \sum_{h=1}^H \frac{n_h^2}{n_{2h}} \left(1 - \frac{n_{2h}}{n_h}\right) \frac{1}{n_{2h} - 1} \sum_{i=1}^{n_{2h}} \left(y_{2hi} - \frac{\sum_{i=1}^{n_{2h}} y_{2hi}}{n_{2h}}\right)^2 \end{aligned} \quad (8)$$

Le terme entre accolades est obtenu à partir du résultat 9.3.1 de Särndal, Swennson et Wretman (19920)

4. ESTIMATION DE LA DIFFÉRENCE D

Pour estimer la différence $D = Y_2 - Y_6$, on utilise l'estimateur (1) pour les variables d'intérêt y_2 et y_6 . On obtient alors

$$\begin{aligned}\hat{D} = \hat{Y}_2 - \hat{Y}_6 &= \frac{M}{m} \sum_{h=1}^H \frac{n_h}{n_{2h}} \sum_{i=1}^{n_{2h}} y_{2hi} - \frac{M}{m} \sum_{h=1}^H \frac{n_h}{n_{6h}} \sum_{i=1}^{n_{6h}} y_{6hi} \\ &= \frac{M}{m} \sum_{h=1}^H \left(\frac{n_h}{n_{2h}} \sum_{i=1}^{n_{2h}} y_{2hi} - \frac{n_h}{n_{6h}} \sum_{i=1}^{n_{6h}} y_{6hi} \right)\end{aligned}\quad (9)$$

Puisque $E(\hat{Y}_2) = Y_2$ et $E(\hat{Y}_6) = Y_6$, on a $E(\hat{D}) = D$. Pour obtenir la variance de l'estimateur (9), on procède de nouveau en conditionnant sur s^A . En s'inspirant de (3), on obtient

$$\begin{aligned}E(\hat{D} | s^A) &= \frac{M}{m} \sum_{h=1}^H \left(\sum_{i=1}^{n_{2h}} y_{2hi} - \sum_{i=1}^{n_{6h}} y_{6hi} \right) \\ &= \frac{M}{m} \sum_{h=1}^H \sum_{i=1}^{n_h} d_{hi} = \frac{M}{m} \sum_{k=1}^M t_k d_k\end{aligned}\quad (10)$$

où $d_{hi} = y_{2hi} - y_{6hi}$ et $d_k = y_{2k} - y_{6k}$. À partir de (10), on obtient directement

$$\text{Var}(E(\hat{D} | s^A)) = M^2 \left(1 - \frac{m}{M} \right) \frac{S_d^2}{m}\quad (11)$$

où $S_d^2 = \sum_{k=1}^M (d_k - \bar{D})^2 / (M - 1)$ et $\bar{D} = \sum_{k=1}^M d_k / M$. De plus, en partant de (9), on a

$$\begin{aligned}\text{Var}(\hat{D} | s^A) &= \frac{M^2}{m^2} \sum_{h=1}^H \left[\text{Var} \left(\frac{n_h}{n_{2h}} \sum_{i=1}^{n_{2h}} y_{2hi} \mid s^A \right) + \text{Var} \left(\frac{n_h}{n_{6h}} \sum_{i=1}^{n_{6h}} y_{6hi} \mid s^A \right) \right. \\ &\quad \left. - 2 \text{Cov} \left(\frac{n_h}{n_{2h}} \sum_{i=1}^{n_{2h}} y_{2hi}, \frac{n_h}{n_{6h}} \sum_{i=1}^{n_{6h}} y_{6hi} \mid s^A \right) \right]\end{aligned}\quad (12)$$

Les deux premiers termes de (12) s'obtiennent directement de (6). Pour le troisième terme, à partir des résultats de Tam (1984), on obtient

$$\begin{aligned}\text{Cov} \left(\frac{n_h}{n_{2h}} \sum_{i=1}^{n_{2h}} y_{2hi}, \frac{n_h}{n_{6h}} \sum_{i=1}^{n_{6h}} y_{6hi} \mid s^A \right) &= -\frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(y_{2hi} - \frac{\sum_{i=1}^{n_h} y_{2hi}}{n_h} \right) \left(y_{6hi} - \frac{\sum_{i=1}^{n_h} y_{6hi}}{n_h} \right) \\ &= -n_h \tilde{S}_{26h}\end{aligned}\quad (13)$$

En mettant (11), (12) et (13) ensembles, on obtient finalement

$$\text{Var}(\hat{D}) = M^2 \left(1 - \frac{m}{M} \right) \frac{S_d^2}{m} + E \left[\frac{M^2}{m^2} \sum_{h=1}^H n_h \left(\frac{n_{6h}}{n_{2h}} \tilde{S}_{2h}^2 + \frac{n_{2h}}{n_{6h}} \tilde{S}_{6h}^2 - \tilde{S}_{26h} \right) \right]\quad (14)$$

Malheureusement, il n'existe pas d'estimateur sans biais de la variance de \hat{D} . En effet, l'estimation sans biais de (14) requiert l'estimation de la covariance \tilde{S}_{26h} qui demande, à son tour, d'avoir les valeurs de y_{2hi} et y_{6hi} (ou y_{2hik} et y_{6hik}) mesurées au sein de la même grappe i . Par construction, le plan de sondage force l'utilisation d'une seule version du questionnaire par grappe. Ainsi, il n'y a pas de grappe où on retrouve les mesures de y_{2hi} et y_{6hi} simultanément.

Pour estimer la covariance (12), il existe plusieurs approches possibles, mais toutes ces approches demandent l'utilisation de certaines hypothèses. Une approche simpliste est de supposer que la covariance (12) est nulle. Malheureusement, comme cette covariance est en général négative, il s'avère que la variance (13) sera alors sous-estimée, ce qui pose un problème dans les intervalles de confiance.

Dumais et Lavallée (1990) se sont servis de l'imputation afin d'obtenir les données nécessaires pour l'estimation de \tilde{S}_{26h} . Il ont ainsi imputé les valeurs de y_{2hi} pour les n_{6h} grappes où le questionnaire de deux pages n'a pas été donné et, à l'inverse, ils ont aussi imputé les valeurs de y_{6hi} pour les n_{2h} grappes où le questionnaire de six pages n'a pas été donné. Ils ont donc obtenu des valeurs — réelles ou imputées — pour les n_h grappes entrant dans le calcul de \tilde{S}_{26h} . En suivant l'approche de Dumais et Lavallée (1990), on obtient

$$\hat{S}_{26h} = \frac{1}{n_h - 1} \left[\frac{\tilde{y}_{6h}}{\tilde{y}_{2h}} \frac{1}{(n_{2h} - 1)} \sum_{i=1}^{n_{2h}} \left(y_{2hi} - \frac{\sum_{i=1}^{n_{2h}} y_{2hi}}{n_{2h}} \right)^2 + \frac{\tilde{y}_{2h}}{\tilde{y}_{6h}} \frac{1}{(n_{6h} - 1)} \sum_{i=1}^{n_{6h}} \left(y_{6hi} - \frac{\sum_{i=1}^{n_{6h}} y_{6hi}}{n_{6h}} \right)^2 \right] \quad (15)$$

où $\tilde{y}_{2h} = \sum_{i=1}^{n_{2h}} y_{2hi} / n_{2h}$ et $\tilde{y}_{6h} = \sum_{i=1}^{n_{6h}} y_{6hi} / n_{6h}$.

Plus récemment, Ardilly (2004) a aussi proposé une approche pour l'estimation de la variance (11) sans toutefois passer par l'estimation explicite de la covariance (12). Il propose de jumeler les échantillons $s_2^A = \bigcup_{h=1}^H s_{2h}^A$ et $s_6^A = \bigcup_{h=1}^H s_{6h}^A$ en un seul échantillon $s^A = s_2^A \cup s_6^A$ et de calculer la variance de $\hat{Y} = (M/m) \sum_{k=1}^m y_k$ où $y_k = y_{2k}$ si $k \in s_2^A$, $y_k = y_{6k}$ si $k \in s_6^A$. Notons que le calcul de cette variance ignore le caractère aléatoire de l'assignation des deux versions du questionnaire.

5. SOMMAIRE ET CONCLUSION

L'estimation des totaux Y_2 et Y_6 , ainsi que la différence D permettra de comparer les résultats des deux versions du questionnaire de l'ASHE, soient les versions de deux pages et de six pages. L'utilisation des variances estimées de \hat{Y}_2 , \hat{Y}_6 et \hat{D} seront alors utiles afin de savoir si les différences observées sont significatives ou non. Si les différences s'avèrent non significatives, l'ONS pourra alors utiliser sans trop de problème le nouveau questionnaire de six pages et ainsi obtenir plus d'information que le précédent questionnaire de deux pages.

RÉFÉRENCES

- Ardilly, P. (2004), "Calcul de précision des variables annuelles structurelles dans l'enquête « emploi » : une piste", Note de service de l'Unité « Méthodes statistiques » de l'INSEE, Paris, 30 mars 2004.
- Dumais, J., Lavallée, P. (1990), "Une approche pour l'échantillonnage et l'estimation d'enquêtes trimestrielles: l'Enquête sur le camionnage pour compte d'autrui", *Actes du colloque sur les méthodes et domaines d'application de la statistique*, Bureau de la statistique du Québec, Québec, 1990, p. 51-58.
- Pont, M. (2003), "Redesigning the UK's Annual Structural Earning Survey", article présenté au Federal Committee on Statistical Methods, Washington, D.C., novembre 2003.
- Särndal, C.-E., Swensson, B., Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.
- Tam, S.M. (1984), "On Covariances From Overlapping Samples", *The American Statistician*, Vol. 38, No. 4, novembre 1984.