Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

ESTIMATION DE LA VARIANCE DES ERREURS DE MESURE SIMPLES ET CORRÉLÉES LORSQUE LA PREMIÈRE ET LA SECONDE INTERVIEW UTILISENT DES MODES DE COLLECTE DIFFÉRENTS

Piero Demetrio Falorsi, Marco Fortini, Alessandro Pallara[1]

Aldo Russo[2]

RÉSUMÉ

La présente communication propose une *méthode des moments* pour estimer les *composantes simple et corrélée de la variance des erreurs de mesure* lorsqu'on procède à une réinterview auprès d'un sous-échantillon de répondants, mais qu'on ne peut pas considérer les deux mesures comme étant effectuées dans les mêmes conditions; elles sont donc sujettes à une *variance différente des erreurs de mesure.* Cette hypothèse semble plus réaliste lorsqu'il s'avère impossible d'assurer les mêmes conditions de mesure dans les deux interviews, par exemple lorsque, à cause de contraintes opérationnelles et budgétaires, on doit adopter un mode de collecte différent pour la réinterview.

MOTS-CLÉS : réinterview; erreurs de mesure; variance des erreurs de mesure simples et corrélées; méthode des moments.

1. INTRODUCTION

Les erreurs de mesure au moment de la collecte attribuables aux répondants, aux intervieweurs et à la relation entre les premiers et les seconds peuvent avoir une incidence importante sur l'exactitude des estimations d'enquête.

Pour estimer l'incidence des erreurs de mesure, on peut procéder à une *réinterview* auprès d'un sous-échantillon de répondants afin d'obtenir des mesures répétées et d'appliquer des modèles pertinents.

Une autre méthode, plutôt axée sur l'estimation de l'effet de l'intervieweur, fait appel à un plan d'échantillonnage spécial fondé sur un réseau de sous-échantillons (Bailar et Dalenius, 1969; Lessler, 1984).

Lorsqu'on utilise la méthode fondée sur la *réinterview* et qu'on obtient *deux* mesures auprès des mêmes répondants, on envisage habituellement les hypothèses suivantes (Forsman, 1989; Hansen et coll., 1961) :

- les deux mesures sont modélisées comme des variables aléatoires;
- les mesures répétées à l'égard de la même unité sont indépendantes;
- les conditions de mesure sont identiques ou aussi identiques que possible dans les deux cas (Särndal *et coll.*, 1992, p. 614 à 616), ce qui suppose que les variables aléatoires liées aux deux mesures sont sujettes à la *même variance des erreurs de mesure*.

Dans nos travaux, en supposant toujours que les deux mesures sont indépendantes, nous essayons d'estimer les *composantes simple et corrélée de la variance* lorsqu'on procède à une réinterview auprès d'un sous-

[1] Institut national italien de la statistique, Via Depretis 74/b, 00184, Rome, Italie.
[2] Département des institutions politiques et des sciences sociales, Université Roma Tre, Rome, Italie.

échantillon de répondants, mais qu'on ne peut pas considérer les deux mesures comme ayant été effectuées dans les mêmes conditions; elles sont donc sujettes à une *variance différente des erreurs de mesure*.

Cette hypothèse semble plus réaliste lorsqu'il s'avère impossible d'assurer les mêmes conditions de mesure dans les deux interviews; en effet, à cause de contraintes opérationnelles et budgétaires, on doit souvent adopter un mode de collecte différent pour la réinterview. Lors du Recensement italien de l'agriculture, par exemple, on a mené l'interview selon la méthode de l'*interview directe*. Or, en raison de contraintes budgétaires, le programme de réinterview pour l'évaluation de la qualité des données du recensement était fondé sur une *enquête de validation* qui a été menée selon la méthode de l'interview téléphonique assistée par ordinateur (ITAO).

2. LE MODÈLE STATISTIQUE

Prenons les conditions suivantes :

(a) une population finie *P* est constituée de *N* unités;

(b) dans cette population *P*, on prélève un échantillon aléatoire initial *s* (enquête I_1) de taille n selon le plan d'échantillonnage $p(\cdot)$, π_k étant la probabilité que l'élément *k* soit compris dans l'échantillon et π_{kl}, la probabilité que les deux éléments *k et l* soient compris dans l'échantillon;

(c) pour chaque élément $k \in s$, $y_{k,1}$ représente la valeur observée da la variable *y* qui nous intéresse;

(d) on prélève un second échantillon r ($r \subseteq s$) (enquête I_2) de taille n_r selon le plan d'échantillonnage $p(\cdot|s)$, de telle sorte que $p(r|s)$ est la probabilité conditionnelle de choisir *r*. Selon ce plan d'échantillonnage, les probabilités d'inclusion sont appelées $\pi_{k|s}$ et $\pi_{kl|s}$ pour les éléments *k et* $l \in s$;

(e) pour chaque élément $k \in r$, $y_{k,2}$ représente la valeur observée de la variable *y* qui nous intéresse.

Le modèle de mesure *m* est défini comme suit :

$$y_{k,t} = \mu_k + \varepsilon_{k,t} \qquad (t = 1,2) \tag{1}$$

où μ_k est la *valeur réelle et* $\varepsilon_{k,t}$ est la composante *aberrante*.

Étant donné les conditions générales d'enquête de I_1 et de I_2, les valeurs espérées $E_m(\cdot)$ selon le modèle (1) sont les suivantes :

$$E_m(y_{k,t}) = \mu_k \qquad \text{pour } k \in s \text{ si } t=1 \text{ et } k \in r \subseteq s \text{ si } t=2$$

$$E_m(y_{k,t} - \mu_k)^2 = \sigma^2_{k,t} \qquad \text{pour } k \in s \text{ si } t=1 \text{ et } k \in r \subseteq s \text{ si } t=2$$

$$E_m[(y_{k,t} - \mu_k)(y_{l,t} - \mu_l)] = \sigma_{kl,t} \qquad \text{pour } k \in s \text{ si } t=1 \text{ et } k \in r \subseteq s \text{ si } t=2$$

$$E_m[(y_{k,1} - \mu_k)(y_{l,2} - \mu_l)] = 0\,. \qquad \text{pour } k,l \in r \subseteq s\,.$$

Il importe de noter que dans le modèle présenté ci-dessus, chacune des valeurs μ_k, $\sigma^2_{k,t}$, est liée à l'unité spécifique $k \in P$ et que $\sigma^2_{kl,t}$ est une valeur liée au couple spécifique $(k,l) \in P$, indépendamment de l'échantillon spécifique choisi dans l'enquête I_t (t =1 ou 2).

L'objectif de l'enquête I_1 consiste à estimer la population totale des *valeurs réelles* suivantes :

$$t_y = \sum\nolimits_P \mu_k$$

où nous représentons par $\sum_P$ la somme des *N* unités de la population *P*.

On peut obtenir une estimation de t_y à l'aide de l'estimateur de Horvitz-Thompson :

$$\tilde{t}_y = \sum_s \frac{y_{k,1}}{\pi_k} . \tag{2}$$

3. DÉCOMPOSITION DE L'ERREUR QUADRATIQUE MOYENNE

Comme mesure de l'exactitude de l'estimateur (2), nous utilisons l'erreur quadratique moyenne de $\tilde{t}_y$ définie en fonction de $p(\cdot)$ et de *m*. Soit

$$B_{pm}(\tilde{t}_y) = E_{pm}(\tilde{t}_y) - t_y = 0$$

l'erreur quadratique moyenne de $\tilde{t}_y$ est représentée par l'équation suivante :

$$MSE_{pm}(\tilde{t}_y) = V_{pm}(\tilde{t}_y) = E_{pm}\left[(\tilde{t}_y - t_y)^2\right] \tag{3}$$

où les indices *p* et *m* définissent les opérateurs *E* et *V* à l'égard du plan d'échantillonnage et du modèle.

Le terme de variance $V_{pm}(\tilde{t}_y)$ peut être décomposé comme suit :

$$V_{pm}(\tilde{t}_y) = E_p\left[V_m(\tilde{t}_y|s)\right] + V_p\left[E_m(\tilde{t}_y|s)\right] =$$

$$= V_A + V_B .$$

La composante V_A, appelée *variance de mesure*, peut être décomposée comme suit :

$$V_A = E_p\left[V_m(\tilde{t}_y|s)\right] = \sum_k \sum_l {}_P \frac{\pi_{kl}}{\pi_k \pi_l} \sigma_{kl,1} . \tag{4}$$

La composante V_B est exprimée comme suit :

$$V_B = V_p\left[E_m(\tilde{t}_y|s)\right] = \sum_k \sum_l {}_P \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \mu_k \mu_l .$$

À l'instar de Särndal *et coll.* (1992, p. 610), on peut aussi exprimer la variance de mesure (4) comme suit :

$$V_A = V_{A1} + V_{A2}$$

où

$$V_{A1} = \sum_P \sigma^2_{k,1} + \sum_k \sum_{l \neq k} {}_P \ \sigma_{kl,1} \tag{5}$$

$$V_{A2} = \sum_P \frac{(1-\pi_k)}{\pi_k} \sigma^2_{k,1} + \sum_k \sum_{l \neq k} {}_P \ \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_k \pi_l} \sigma_{kl,1} . \tag{6}$$

4. ESTIMATION DE L'ERREUR QUADRATIQUE MOYENNE

En l'absence d'erreurs de mesure, l'estimateur de la variance totale $V_{pm}(\tilde{t}_y)$ est représenté par l'équation suivante :

$$\tilde{V}(\tilde{t}_y) = \sum_k \sum_l {}_s \; \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{y_{k,1}}{\pi_k} \frac{y_{l,1}}{\pi_l} . \tag{7}$$

Comme le montrent Koch *et coll*. (1975), en présence d'erreurs de mesure, l'estimateur (7) est biaisé, car le résultat suivant est valide :

$$E_{pm}\left(\tilde{V}(\tilde{t}_y)\right) = V_{pm}(\tilde{t}_y) - V_{A1} .$$

Pour obtenir une estimation sans biais de $V_{pm}(\tilde{t}_y)$, il suffit donc de calculer une estimation sans biais de la composante V_{A1}.

Afin de résoudre ce problème, selon le modèle statistique énoncé dans la section 2, prenons le calcul suivant :

$$E_m[V(y_t)] = E_m\left[\frac{1}{N}\sum\nolimits_P \left(y_{k,t} - \frac{1}{N}\sum\nolimits_P y_{k,t}\right)^2\right] =$$

(t =1,2)

$$= E_m\left[\frac{1}{N}\sum\nolimits_P \left[\left(\mu_k - \frac{1}{N}\sum\nolimits_P \mu_k\right) + \left(\varepsilon_{k,t} - \frac{1}{N}\sum\nolimits_P \varepsilon_{k,t}\right)\right]^2\right].$$

Après quelques calculs simples, il est possible d'obtenir

$$E_m[V(y_1)] = V(\mu) + \frac{N-1}{N^2}\sum\nolimits_P \sigma^2_{k,1} - \frac{1}{N^2}\sum_k \sum_{l \neq k} {}_P \; \sigma_{kl,1} \tag{8}$$

où

$$V(\mu) = \frac{1}{N}\sum\nolimits_P \left(\mu_k - \frac{1}{N}\sum\nolimits_P \mu_k\right)^2 .$$

Un résultat semblable vaut pour $E_m[V(y_2)]$.

En outre, considérons les résultats suivants :

$$E_m\left[\sum\nolimits_P (y_{k,1} - y_{k,2})^2\right] = E_m\left[\sum\nolimits_P \left((\mu_k + \varepsilon_{k,1}) - (\mu_k + \varepsilon_{k,2})\right)^2\right] =$$

$$= \sum\nolimits_P \sigma^2_{k,1} + \sum\nolimits_P \sigma^2_{k,2} \tag{9}$$

$$E_m\left[\left(\sum\nolimits_P y_{k,1} - \sum\nolimits_P y_{k,2}\right)^2\right] = E_m\left[\left(\sum\nolimits_P (\mu_k + \varepsilon_{k,1}) - \sum\nolimits_P (\mu_k + \varepsilon_{k,1})\right)^2\right] =$$

$$= \sum\nolimits_P \sigma^2_{k,1} + \sum_k \sum_{l \neq k} {}_P \; \sigma_{kl,1} + \sum\nolimits_P \sigma^2_{k,2} + \sum_k \sum_{l \neq k} {}_P \; \sigma_{kl,2} \tag{10}$$

$$E_m\left[\sum\nolimits_P\left(y_{k,1}-\frac{1}{N}\sum\nolimits_P y_{k,1}\right)\left(y_{k,2}-\frac{1}{N}\sum\nolimits_P y_{k,2}\right)\right]=$$

$$=E_m\left[\sum\nolimits_P\left(\mu_k+\varepsilon_{k,1}-\frac{1}{N}\sum\nolimits_P \mu_l+\varepsilon_{l,1}\right)\left(\mu_k+\varepsilon_{k,2}-\frac{1}{N}\sum\nolimits_P \mu_l+\varepsilon_{l,2}\right)\right]=V(\mu). \quad (11)$$

À l'aide des résultats ci-dessus, nous présentons le système d'équations suivant :

$$\begin{cases} V(\mu)+\frac{N-1}{N^2}\sum\nolimits_P \sigma_{k,1}^2-\frac{1}{N^2}\sum_k\sum_{l\neq k}\nolimits_P \sigma_{kl,1}=E_m[V(y_1)] \\ V(\mu)+\frac{N-1}{N^2}\sum\nolimits_P \sigma_{k,2}^2-\frac{1}{N^2}\sum_k\sum_{l\neq k}\nolimits_P \sigma_{kl,2}=E_m[V(y_2)] \\ \sum\nolimits_P \sigma_{k,1}^2+\sum\nolimits_P \sigma_{k,2}^2=E_m\left[\sum\nolimits_P (y_{k,1}-y_{k,2})^2\right] \\ \sum\nolimits_P \sigma_{k,1}^2+\sum_k\sum_{l\neq k}\nolimits_P \sigma_{kl,1}+\sum\nolimits_P \sigma_{k,2}^2+\sum_k\sum_{l\neq k}\nolimits_P \sigma_{kl,2}=E_m\left[\left(\sum\nolimits_P y_{k,1}-\sum\nolimits_P y_{k,2}\right)^2\right] \\ V(\mu)=E_m\left[\sum\nolimits_P\left(y_{k,1}-\frac{1}{N}\sum\nolimits_P y_{k,1}\right)\left(y_{k,2}-\frac{1}{N}\sum\nolimits_P y_{k,2}\right)\right] \end{cases} \quad (12)$$

En adoptant une notation matricielle, on peut exprimer le système d'équations ci-dessus comme suit :

$$\mathbf{AX}=\mathbf{b} \quad (13)$$

où **A** est la matrice des coefficients, soit :

$$\mathbf{A}=\begin{bmatrix} 1 & (N-1)/N^2 & 0 & -1/N^2 & 0 \\ 1 & 0 & (N-1)/N^2 & 0 & -1/N^2 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad (14)$$

X=$\{X_i\}$ (i =1,…,5), est le vecteur de *termes inconnus*, soit :

$$X_1=V(\mu) \quad , \quad X_2=\sum\nolimits_P \sigma_{k,1}^2 \quad , \; X_3=\sum\nolimits_P \sigma_{k,2}^2$$

$$X_4=\sum_k\sum_{l\neq k}\nolimits_P \sigma_{kl,1} \quad , \quad X_5=\sum_k\sum_{l\neq k}\nolimits_P \sigma_{kl,2}, \quad (15)$$

et **b**=$\{b_i\}$ (i =1,…,5), est le vecteur de *termes connus*, soit :

$$b_1=V(y_1) \quad , \quad b_2=V(y_2) \quad , \quad b_3=\sum\nolimits_P (y_{k,1}-y_{k,2})^2$$

$$b_4=\left(\sum\nolimits_P y_{k,1}-\sum\nolimits_P y_{k,2}\right)^2, \; b_5=\sum\nolimits_P\left(y_{k,1}-\frac{1}{N}\sum\nolimits_P y_{k,1}\right)\left(y_{k,2}-\frac{1}{N}\sum\nolimits_P y_{k,2}\right).$$

Afin d'obtenir la solution du système (13), il est nécessaire de remplacer les *termes connus* b_i (i=1,…,5) par les estimations d'échantillonnage correspondantes. À l'aide des résultats présentés dans Särndal et coll. (1992, ch. 9), il est possible de montrer que des estimations sans biais des termes connus sont représentées par les équations suivantes :

$$\tilde{b}_1 = \frac{1}{N^2}\left[\sum_s \frac{y_{k,1}^2}{\pi_k}(N-1) - \sum_k \sum_{l\neq k}{}_s \frac{y_{k,1}\, y_{l,1}}{\pi_{kl}}\right]$$

$$\tilde{b}_2 = \frac{1}{N^2}\left[\sum_r \frac{y_{k,2}^2}{\pi_{k,2}}(N-1) - \sum_k \sum_{l\neq k}{}_r \frac{y_{k,2}\, y_{l,2}}{\pi_{kl,2}}\right]$$

$$\tilde{b}_3 = \sum_r \frac{1}{\pi_{k,2}}\left(y_{k,1} - y_{k,2}\right)^2$$

$$\tilde{b}_4 = \sum_s \frac{y_{k,1}^2}{\pi_k} + \sum_k \sum_{l\neq k}{}_s \frac{y_{k,1}\, y_{l,1}}{\pi_{kl}} + \sum_r \frac{y_{k,2}^2}{\pi_{k,2}} + \sum_k \sum_{l\neq k}{}_r \frac{y_{k,2}\, y_{l,2}}{\pi_{kl,2}} + -2\sum_k \sum_l{}_r \frac{y_{k,1}\, y_{l,2}}{\pi_{kl,2}}$$

$$\tilde{b}_5 = \sum_r \frac{y_{k,1}\, y_{k,2}}{\pi_{k,2}}\left(1-\frac{1}{N}\right) - \frac{1}{N}\sum_k \sum_{l\neq k}{}_r \frac{y_{k,1}\, y_{l,2}}{\pi_{kl,2}}$$

où

$$\pi_{k,2} = \pi_k\, \pi_{k|s} \quad , \quad \pi_{kl,2} = \pi_{kl}\, \pi_{kl|s}\,.$$

Soit $\tilde{\mathbf{b}} = \{\tilde{b}_i\}'$ (i =1,…,5) le vecteur des estimations de termes connus, on peut obtenir le vecteur $\hat{\mathbf{X}}$, qui renferme les estimations $\hat{X}_i$ (i =1,…,5) des termes inconnus, en posant l'équation suivante :

$$\hat{\mathbf{X}} = \mathbf{A}^{-1}\tilde{\mathbf{b}}$$

Sur la base des estimations $\hat{X}_i$ (i =1,…,5), il est possible d'obtenir l'estimation suivante de la variance V_{A1} :

$$\hat{V}_{A1} = \hat{X}_2 + \hat{X}_4\,. \tag{16}$$

Enfin, une estimation de $V_{pm}\left(\tilde{t}_y\right)$ peut être représentée par l'équation suivante :

$$\tilde{V}_{pm}\left(\tilde{t}_y\right) = \tilde{V}\left(\tilde{t}_y\right) + \hat{V}_{A1} \quad . \tag{17}$$

BIBLIOGRAPHIE

Bailar B.A. et Dalenius T. (1969), "Estimating the Response Variance Components of the U.S. Bureau of the Census' Survey Model", *Sankhya* 20, 287-294.

Forsman G. (1989), " Early Survey Models and Their Use in Survey Quality Work", *Journal of Official Statistics*, 5, pp. 41-55.

Hansen M., Hurwitz W., Bershad M.A. (1961), “Measurements Errors in Census and Surveys”, *Bulletin of the International Statistical Institute*, 38 :2, pp. 359-374.

Koch G. G., Freeman D.H. et Freeman J.L. (1975), “Strategies in the Multivariate Analysis of Data from Complex Surveys”, *International Statistical Review*, 43, pp. 59-78.

Lessler J.T. (1984), "Measurement Errors in Surveys", dans C.F. Turner et E. Martin (éds), *Surveying Subjective Phenomena*, Vol. 2, New York, Russell Sage Foundation, pp. 405-440.

Särndal C.E., Swensson B. et Wretman J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York.