ESTIMATION ET CALCUL DE PRÉCISION POUR DES ÉCHANTILLONS ROTATIFS NON CHEVAUCHANTS

Pierre LAVALLÉE

Statistique Canada

Journées de méthodologie Statistique

INSEE, mars 2004

Remerciements

Chantal Grondin

Carl Särndal

Jean Dumais

Contenu

Introduction

Sondage rotatif :

- Remplacer une partie de l'échantillon à chaque vague de consultation

Rotation de 100% :

- Renouvellement complet de l'échantillon à chaque vague de consultation
- Échantillons rotatifs non chevauchants (RNC)
  - Échantillons distincts ou disjoints
  - Échantillons avec renouvellement complet
  - *Rolling Samples*

Kish (1981) :

- Échantillons RNC pour les recensements

- Recensement par étape qui consiste en T échantillons RNC conçus de sorte qu'en faisant la somme des données des T échantillons, on obtient un dénombrement complet de la population visée

Ex : Recensement français

Échantillons RNC aussi utilisés pour des estimations obtenues par cumul

Ex : Estimations annuelles obtenues en cumulant des échantillons RNC trimestriels

Problème abordé :

- **Estimation des covariances** provenant de l'utilisation d'échantillons RNC

- Covariances difficiles à estimer en pratique parce qu'une estimation sans biais requiert l'utilisation d'unités communes sur au moins deux vagues

Plan de sondage et estimation

U : Population de N unités destinée à être enquêtée à plus d'une occasion

Constante dans le temps

s : Échantillon de taille n tiré de U à partir d'un plan de sondage $p(s) > 0$ quelconque

$\pi_k > 0$: Probabilité de sélection de l'unité k

Échantillon s divisé en T sous-échantillons s_t de tailles n_t

où $s = \bigcup_{t=1}^{T} s_t$ et $s_t \cap s_{t'} = \varnothing$ pour $t \neq t'$

On suppose que les T sous-échantillons s_t sont enquêtés à différents temps t

Si les T sous-échantillons s_t sont de tailles égales,

$\pi_{tk} = \pi_k / T$

y_{tk} : Variable d'intérêt mesurée à différents temps t

Quantités d'intérêt :

- $Y_t = \sum_{k=1}^{N} y_{tk}$: total au temps t
- $\tau = \sum_{t=1}^{T} Y_t$: total cumulé pour l'ensemble de la période T
- $\bar{\tau} = \sum_{t=1}^{T} Y_t / T$: moyenne sur la période T
- $\Delta_{t,t'} = Y_t - Y_{t'}$: différence entre les totaux des temps t et t'

Précisions à apporter :

τ : Total cumulé de la variable d'intérêt y au cours du temps

Ex : Revenu annuel par cumul des revenus trimestriels

Pour le recensement de la population, τ n'a pas d'intérêt proprement dit

$\bar{\tau}$: Moyenne sur la période T

Ex : Pour le recensement, population moyenne de la France pour cinq années consécutives

$w_{tk} = 1/\pi_{tk}$: Poids de sondage de l'unité k de s_t

$$\sum_{k=1}^{n_t} w_{tk} \approx N$$

Chaque s_t nous ramène à la population U

Estimation de Y_t :

$$\hat{Y}_t = \sum_{k=1}^{n_t} w_{tk} y_{tk}$$

$$\hat{\tau} = \sum_{t=1}^{T} \hat{Y}_t = \sum_{t=1}^{T} \sum_{k=1}^{n_t} w_{tk} y_{tk}$$

$$Var(\hat{\tau}) = \sum_{t=1}^{T} Var(\hat{Y}_t) + \sum_{t=1}^{T} \sum_{t' \neq t} Cov(\hat{Y}_t, \hat{Y}_{t'})$$

Estimation de $Var(\hat{Y}_t)$: Estimateurs habituels…

Estimation de $Cov(\hat{Y}_t, \hat{Y}_{t'})$: Problématique parce pas d'unités communes entre s_t et $s_{t'}$

Dans la littérature, estimation des covariances $Cov(\hat{Y}_t, \hat{Y}_{t'})$ **pratiquement ignorée**

Raison : Parce que les échantillons sont disjoints, on considère souvent ces covariances comme étant nulles (Kish, 1965, et Kish, 1999)

Disjoints ≠ indépendants !

$Cov(\hat{Y}_t, \hat{Y}_{t'}) \neq 0$

$Cov(\hat{Y}_t, \hat{Y}_{t'})$ généralement négatives

Si on ignore les covariances, on obtient une surestimation de la variance d'une somme

Conservateur dans les tests d'hypothèses

Dans certaines enquêtes, les fractions des sondages sont suffisamment élevées pour que la surestimation de la variance pose problème

Ex : Pour le recensement français, chaque s_t correspond à 1/5 de U (petites communes)

Certains auteurs proposent d'ignorer le fait que les y_{tk} sont mesurées à différents temps *t*

Ne tient plus compte que s ait été divisé aléatoirement en s_t, t=1,…,T

Estimateur de $\bar{\tau}$:

$$\hat{\bar{\tau}}_{comb} = \sum_{k=1}^{n} w_k y_k$$

où $w_k = 1/\pi_k$ et $y_k = y_{tk}$ pour $k \in s_t$

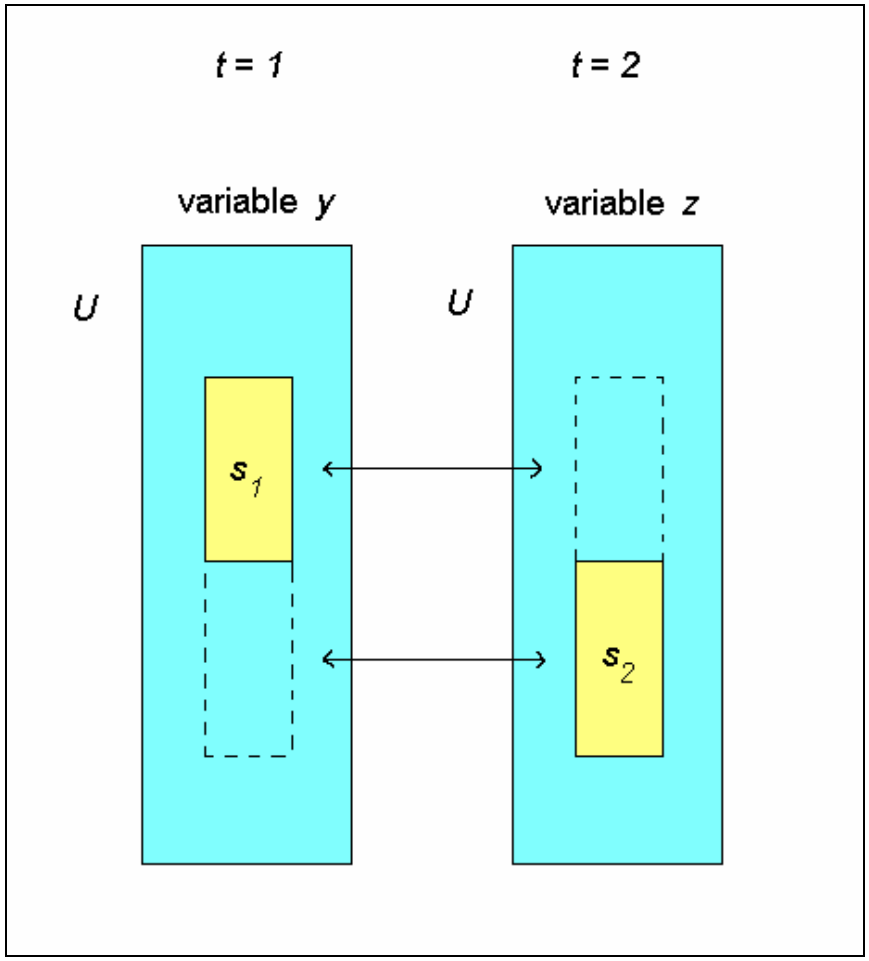
$Var(\hat{\bar{\tau}}_{comb})$ reliée simplement à la sélection de s

$\hat{Var}(\hat{\bar{\tau}}_{comb})$ sous-estime $Var(\hat{\bar{\tau}})$ puisqu'on ne tient pas compte du fait que y a été mesurée différemment selon s_t

Une solution basée sur l'utilisation de l'imputation

Simplifications :

- Sélection de s faite par SASSR
- T=2 : s divisé en s_1 et s_2 de taille n_1 et n_2
- y_k : variable d'intérêt (y_{1k}) au temps *t*=1
- z_k : variable d'intérêt (y_{2k}) au temps *t*=2



Avec T=2 et SASSR,

$$Var(\hat{\tau}) = Var(\hat{Y}) + Var(\hat{Z}) + 2Cov(\hat{Y},\hat{Z})$$
$$= N^2\left(1-\frac{n_1}{N}\right)\frac{S_y^2}{n_1} + N^2\left(1-\frac{n_2}{N}\right)\frac{S_z^2}{n_2} - 2N\,S_{yz}$$

où $S_y^2 = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{Y})^2$, $S_z^2 = \frac{1}{N-1}\sum_{k=1}^{N}(z_k - \bar{Z})^2$, et

$$S_{yz} = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \bar{Y})(z_k - \bar{Z})$$

$$Cov(\hat{Y},\hat{Z}) = -N\,S_{yz}$$

Pour estimer S_{yz} sans biais, on doit disposer d'unités communes entre s_1 et s_2

Pas de valeurs de y_k pour $k \in s_2$, ni de valeurs de z_k pour $k \in s_1$

<u>Solution proposée :</u>

Imputer les données « manquantes » à s_1 et s_2

$\hat{y}_k$: Valeur imputée de y_k pour $k \in s_2$
$\hat{z}_k$: Valeur imputée de z_k pour $k \in s_1$

$$y_k^* = \begin{cases} y_k & \text{pour } k \in s_1 \\ \hat{y}_k & \text{pour } k \in s_2 \end{cases} \quad \text{et} \quad z_k^* = \begin{cases} \hat{z}_k & \text{pour } k \in s_1 \\ z_k & \text{pour } k \in s_2 \end{cases}$$

Estimation de S_{yz} :

$$\hat{S}_{yz} = \frac{1}{n-1}\sum_{k=1}^{n}(y_k^* - \bar{y}^*)(z_k^* - \bar{z}^*)$$

$$= \frac{1}{n-1}\left[\sum_{k=1}^{n_1}(y_k - \bar{y}^*)(\hat{z}_k - \bar{z}^*) + \sum_{k=1}^{n_2}(\hat{y}_k - \bar{y}^*)(z_k - \bar{z}^*)\right]$$

où $\bar{y}^* = \sum_{k=1}^{n} y_k^* / n$ et $\bar{z}^* = \sum_{k=1}^{n} z_k^* / n$

Donc,

$$\hat{Var}(\hat{\tau}) = \hat{Var}(\hat{Y}) + \hat{Var}(\hat{Z}) + 2\hat{Cov}(\hat{Y}, \hat{Z})$$

$$= N^2\left(1 - \frac{n_1}{N}\right)\frac{s_y^2}{n_1} + N^2\left(1 - \frac{n_2}{N}\right)\frac{s_z^2}{n_2} - 2N\hat{S}_{yz}$$

où $s_y^2 = \frac{1}{n_1 - 1}\sum_{k=1}^{n_1}(y_k - \bar{y})^2$, $\bar{y} = \sum_{k=1}^{n_1} y_k / n_1$,

$$s_z^2 = \frac{1}{n_2 - 1}\sum_{k=1}^{n_2}(z_k - \bar{z})^2 \text{ et } \bar{z} = \sum_{k=1}^{n_2} z_k / n_2$$

$\hat{C}ov(\hat{Y},\hat{Z}) = -N\hat{S}_{yz}$ biaisé

En général, tendance à surestimer $Cov(\hat{Y},\hat{Z})$, et donc $Var(\hat{\tau})$, mais pas autant que si on suppose $Cov(\hat{Y},\hat{Z}) = 0$

Important :

- Les valeurs imputées $\hat{y}_k$ et $\hat{z}_k$ ne servent ici qu'à l'estimation de S_{yz}, et donc de $Cov(\hat{Y},\hat{Z})$
- Ces valeurs imputées ne rentrent pas dans l'estimation $\hat{Y}_t$, $\hat{\tau}$ ou $\hat{\bar{\tau}}$

Différentes méthodes d'imputation

Imputation par la moyenne :

$\hat{y}_k = \bar{y}$ pour $k \in s_2$

$\hat{z}_k = \bar{z}$ pour $k \in s_1$

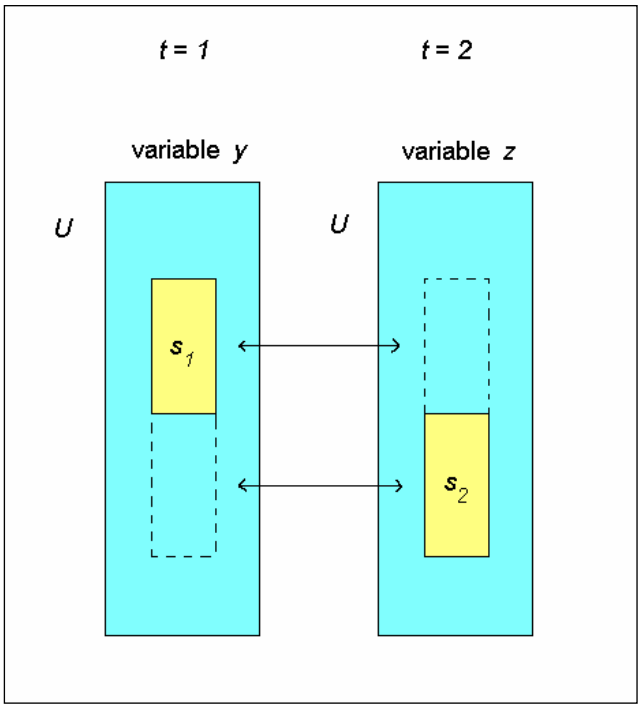
$\hat{S}_{yz} = 0$ et donc, $\hat{C}ov(\hat{Y}, \hat{Z}) = 0$

N.B. L'estimateur reste $\hat{\tau} = \sum_{t=1}^{T} \hat{Y}_t = \sum_{t=1}^{T} \sum_{k=1}^{n_t} w_{tk} y_{tk}$

Imputation historique :

$\hat{y}_k = z_k$ pour $k \in s_2$

$\hat{z}_k = y_k$ pour $k \in s_1$



N.B. L'estimateur reste

$$\hat{\tau} = \sum_{t=1}^{T} \hat{Y}_t = \sum_{t=1}^{T} \sum_{k=1}^{n_t} w_{tk} y_{tk}$$

$$\bar{y}^* = \bar{z}^* = \mu = (n_1 \bar{y} + n_2 \bar{z}) / n$$

$$\hat{S}_{yz} = \frac{1}{(n-1)} \left[(n_1 - 1) s_y^2 + (n_2 - 1) s_z^2 + \frac{n_1 n_2}{n} (\bar{y} - \bar{z})^2 \right]$$

$$\hat{Cov}(\hat{Y}, \hat{Z}) = -\frac{N}{(n-1)} \left[(n_1 - 1) s_y^2 + (n_2 - 1) s_z^2 + \frac{n_1 n_2}{n} (\bar{y} - \bar{z})^2 \right]$$

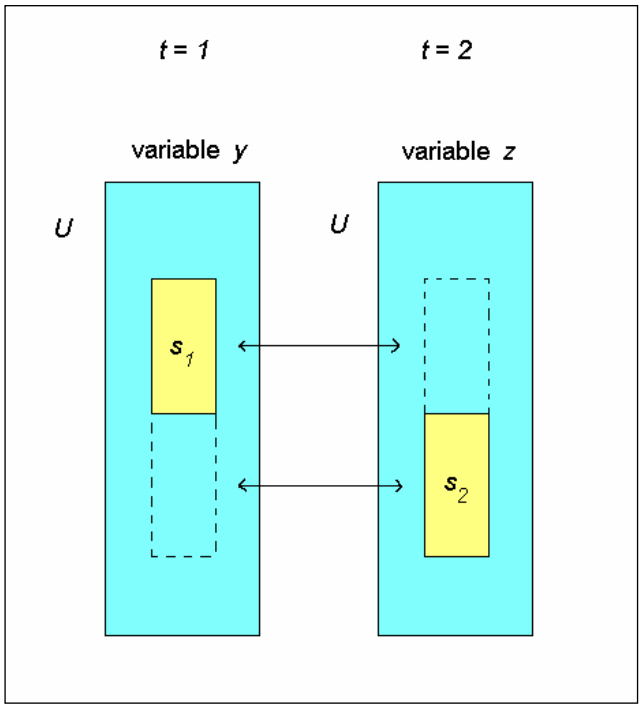
$\hat{Cov}(\hat{Y}, \hat{Z})$ estimée à partir d'une moyenne pondérée des s_y^2 et s_z^2, en plus d'un terme qui devient nul si $\bar{y} = \bar{z}$

Pratiquement dans la même situation que si on ignore que les y_{tk} sont mesurées à différents temps *t*

Imputation historique avec tendance :

$$\hat{y}_k = \frac{\bar{y}}{\bar{z}} z_k \text{ pour } k \in s_2$$

$$\hat{z}_k = \frac{\bar{z}}{\bar{y}} y_k \text{ pour } k \in s_1$$



N.B. L'estimateur reste

$$\hat{\tau} = \sum_{t=1}^{T} \hat{Y}_t = \sum_{t=1}^{T} \sum_{k=1}^{n_t} w_{tk} y_{tk}$$

$$\bar{y}^* = \bar{y} \text{ et } \bar{z}^* = \bar{z}$$

$$\hat{Cov}(\hat{Y}, \hat{Z}) = -\frac{N}{(n-1)} \left[(n_1 - 1) \frac{\bar{z}}{\bar{y}} s_y^2 + (n_2 - 1) \frac{\bar{y}}{\bar{z}} s_z^2 \right]$$

$\hat{Cov}(\hat{Y}, \hat{Z})$ estimée à partir d'une moyenne pondérée des quantités s_y^2 et s_z^2

Le recensement de France

Pistes qui pourraient s'avérer utiles pour le calcul de la précision des estimations issues du recensement

Population U divisée en deux groupes:

- Communes de moins de 10 000 habitants :

  Sélection de régions géographiques appelées groupes de rotation

- Communes d'au moins 10 000 habitants :

  Sélection d'adresses tirées du « répertoire d'immeubles localisés »

Utilisation de l'indice k pour l'identification des deux types d'unités

U divisée en T=5 échantillons RNC s_t

où $U = \bigcup_{t=1}^{5} s_t$ et $s_t \cap s_{t'} = \varnothing$ pour $t \neq t'$

Grosso modo, couverture en cinq ans de l'ensemble du territoire français (petites communes)

s_t obtenus à partir d'un plan de sondage équilibré pour obtenir une certaine représentativité de la population

- Données pour l'équilibrage obtenues du recensement de 1999

π_{tk} de chaque unité k approximativement de 1/5

y_{tk} : Variable d'intérêt

Deux statistiques d'intérêt :

$Y_t = \sum_{k=1}^{N} y_{tk}$: Total de la population au temps t

$\bar{\tau} = \sum_{t=1}^{5} Y_t / 5$: Moyenne de la population sur la période de cinq ans

- Mesure de la population française à l'année médiane T=3

$w_{tk} = 1/\pi_{tk}$: Poids de sondage

Rappel : $\sum_{k=1}^{n_t} w_{tk} \approx N$

$\hat{Y}_t = \sum_{k=1}^{n_t} w_{tk} y_{tk}$

Parce qu'on ne dispose que de s_t de 1/5 de la population, $\hat{Y}_t$ n'est destiné qu'à des estimations globales (nationales et régionales)

$\hat{\bar{\tau}} = \sum_{t=1}^{5} \hat{Y}_t / 5$

Parce qu'on a « balayé » la population française, $\hat{\bar{\tau}}$ peut servir à des statistiques détaillées (niveaux communal)

En pratique, au temps t, imputation des variables y_{tk} des échantillons $s_{t'}$ pour $t' \neq t$

$$Var(\hat{\bar{\tau}}) = \frac{1}{25}\left[\sum_{t=1}^{5} Var(\hat{Y}_t) + \sum_{t=1}^{5}\sum_{t'\neq t} Cov(\hat{Y}_t, \hat{Y}_{t'})\right]$$

Pour estimer $Cov(\hat{Y}_t, \hat{Y}_{t'})$, imputation des données « manquantes » aux différents s_t

Pour un recensement de la population, imputation historique avec tendance semble appropriée

Tendances calculées à partir de :

- Données déjà collectées
- Projections démographiques provenant d'une source externe (ex : recensements précédents, données fiscales)

Utilisation de valeurs imputées utile afin de calculer la précision du recensement de France

Si on utilise l'approche visant à ignorer que les données ont été recueillies à des temps différents, on a $\hat{V}ar(\hat{\bar{\tau}}) = 0$

$\therefore$ Surévaluation de la précision du recensement