Inference for superpopulation parameters using sample surveys by Barry I. Graubard and Edward L. Korn

Guillaume Chauvet et Cyril Favre-Martinoz

École Nationale de la Statistique et de l'Analyse de l'Information

Groupe de travail Sondages et Econométrie
Insee
09/02/2015

En bref

Les auteurs s'intéressent à l'estimation de paramètres de superpopulation pour des données issues d'une enquête par sondage (plan stratifié, plan à plusieurs degrés). Plus précisément, ils s'intéressent à l'estimation de variance pour des estimateurs de ces paramètres.

Une façon habituelle d'estimer la variance consiste à "oublier" que l'échantillon est issu d'un plan de sondage, comme si les données étaient directement générées selon le modèle de superpopulation. Les auteurs montrent que cette approche peut conduire à sous-estimer la variance globale.

Les auteurs proposent également une revue de la littérature, en comparant leur approche avec des approches existantes.

Contexte

Les données dans la population U sont générées selon un modèle de super-population

$$F : Y_i \sim_{iid} \mathcal{L}(\mu, \sigma^2) \quad \text{pour} \quad i = 1, \ldots, K$$

avec K la taille de la population finie U.

Un échantillon est ensuite tiré dans U selon un plan de sondage $p(\cdot)$:

- un sondage aléatoire simple stratifié (STSRS) est considéré dans la Section 2 : "Model without clusters";
- un sondage à plusieurs degrés, avec tirage à probabilités inégales (pps) au 1er degré, est considéré dans la Section 3 : "Model with clusters".

Section 2
Model without clusters

Le modèle

Les données dans la population U sont générées selon un modèle à deux niveaux :

$$F : (Y_i, \eta_i) \sim_{iid} F \quad \text{pour} \quad i = 1, \ldots, K,$$

avec

- η_i une indicatrice de strate, générée selon une loi discrète de support $\{1, \ldots, L\}$,
- Y_i générée conditionnellement à $\eta_i = h$ selon une loi $\mathcal{L}(\mu_h, \sigma_h^2)$.

On note K_h le nombre total d'observations obtenues dans la strate h (telles que $\eta_i = h$). On note également

$$\bar{Y} = \sum_{h=1}^{L} \frac{K_h}{K} \bar{Y}_h \quad \text{avec} \quad \bar{Y}_h = \frac{1}{K_h} \sum_{i=1}^{K_h} y_{hi}.$$

Un échantillon est ensuite sélectionné selon un STSRS de taille k_h dans la strate h (l'allocation peut dépendre des tailles de strates).

Estimation du paramètre $\mu = E_F[Y_i]$

On considère l'estimateur (sans biais)

$$\bar{y} = \sum_{h=1}^{L} \frac{K_h}{K} \bar{y}_h \quad \text{avec} \quad \bar{y}_h = \frac{1}{k_h} \sum_{i=1}^{k_h} y_{hi}.$$

La variance de cet estimateur est donnée par

$$\begin{aligned} Var(\bar{y}) &= E_F Var_{RS}(\bar{y}) + Var_F E_{RS}(\bar{y}) \\ &= E_F Var_{RS}(\bar{y}) + Var_F(\bar{Y}). \end{aligned} \tag{1}$$

L'estimateur de variance habituel pour un STSRS

$$\widehat{var}_{wo}(\bar{y}) = \sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{K_h - k_h}{K_h} \frac{s_h^2}{k_h}$$

estime sans biais le 1er terme de (1). Son biais vaut donc $-Var_F(\bar{Y})$.

Estimation du paramètre $\mu = E_F[Y_i]$ (2)

On considère l'estimateur (sans biais)

$$\bar{y} = \sum_{h=1}^{L} \frac{K_h}{K} \bar{y}_h \quad \text{avec} \quad \bar{y}_h = \frac{1}{k_h} \sum_{i=1}^{k_h} y_{hi}.$$

La variance de cet estimateur est donnée par

$$\begin{aligned} Var(\bar{y}) &= E_F Var_{RS}(\bar{y}) + Var_F(\bar{Y}) \\ &= E_F Var_{RS}(\bar{y}) + E_F Var_{F|\eta}(\bar{Y}) + Var_F E_{F|\eta}(\bar{Y}). \end{aligned} \quad (2)$$

La variance sous le modèle est due :

- à la variabilité des y_{hi} : second terme de (2);
- à la variabilité des tailles de strates K_h : troisième terme de (2).

Estimation du paramètre $\mu = E_F[Y_i]$ (3)

On considère l'estimateur (sans biais)

$$\bar{y} = \sum_{h=1}^{L} \frac{K_h}{K} \bar{y}_h \quad \text{avec} \quad \bar{y}_h = \frac{1}{k_h} \sum_{i=1}^{k_h} y_{hi}.$$

La variance de cet estimateur est donnée par

$$\begin{aligned} Var(\bar{y}) &= E_F Var_{RS}(\bar{y}) + Var_F(\bar{Y}) \\ &= E_F Var_{RS}(\bar{y}) + E_F Var_{F|\eta}(\bar{Y}) + Var_F E_{F|\eta}(\bar{Y}). \end{aligned} \quad (3)$$

L'estimateur de variance qui "oublie" le caractère sans remise

$$\widehat{var}_{wr}(\bar{y}) = \sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{s_h^2}{k_h}$$

estime sans biais les 2 1ers termes de (3). Son biais vaut $-Var_F E_{F|\eta}(\bar{Y})$.

Estimation du paramètre $\mu = E_F[Y_i]$ (4)

Graubard et Korn proposent d'ajouter un estimateur non biaisé de la partie manquante. Par exemple, on estime

$$\begin{aligned} Var_F E_{F|\eta}(\bar{Y}) &\equiv \frac{\Delta_{betw,y}}{K} \\ &= \frac{1}{K}\sum_{h=1}^{L} E\left(\frac{K_h}{K}\right)(\mu_h - \mu)^2. \end{aligned}$$

en utilisant

$$\widehat{\Delta}_{betw,y} = \frac{K}{K-1}\sum_{h=1}^{L}\frac{K_h}{K}(\bar{y}_h - \bar{y})^2 - \sum_{h=1}^{L}\frac{K_h(K-K_h)}{K(K-1)}\frac{s_h^2}{k_h}$$

(formule 2.9). On obtient l'estimateur de variance SB

$$\widehat{var}_{SP}(\bar{y}) = \widehat{var}_{wr}(\bar{y}) + \frac{\widehat{\Delta}_{betw,y}}{K}$$

donné en (2.11). Notons que $\Delta_{betw,y} = 0$ dans le cas d'une seule strate.

Remarques

S 2.2 Le terme manquant $\frac{\Delta_{betw,y}}{K}$ correspond à une variabilité inter-strates. On pourrait penser à utiliser l'estimateur de variance

$$\widetilde{var}_{wr}(\bar{y}) \quad = \quad \frac{s^2}{k}$$

qui "oublie" la stratification. Graubard et Korn montrent dans le cas d'une allocation proportionnelle que cette approche conduit à sur-estimer la variance.

S 2.8 Graubard et Korn établissent un parallèle avec le contexte d'une enquête en deux phases pour une stratification. Ne pas tenir compte du second terme de

$$Var(\bar{y}) \quad = \quad E_F Var_{RS}(\bar{y}) + Var_F(\bar{Y})$$

conduit à ignorer la variance des tailles de strate dues à la 1ère phase de tirage.

Estimation d'un ratio

L'estimation de paramètres complexes peut se faire en utilisant la technique de linéarisation. Dans le cas de données

$$(U_i, X_i, \eta_i) \sim_{iid} F \quad \text{avec} \quad \rho = \frac{E_F(U)}{E_F(X)},$$

l'estimateur du ratio ρ vaut $\bar{r} = \dfrac{\bar{u}}{\bar{x}}$.

On peut utiliser l'estimateur de variance approximativement sans biais

$$\widehat{var}_{SP}(\bar{r}) \quad = \quad \widehat{var}_{wr}(\bar{z}) + \frac{\widehat{\Delta}_{betw,z}}{K}$$

avec $z_{hi} = \dfrac{1}{\bar{x}}(u_{hi} - \bar{r}\ x_{hi})$ (Section 2.3).

L'estimation de variance pour les paramètres d'une régression linéaire ou d'une régression logistique s'effectue de la même manière (Section 2.4).

Etude par simulations

Afin de quantifier le biais de l'estimateur de variance $\widetilde{var}_{wr}(\bar{y})$, Korn et Graubard (1998) donnent les résultats d'une petite étude par simulations. On note $\pi_h = E(K_h/K)$, et :

- variance inter-strates : $\sum_{h=1}^{L} \pi_h(\mu_h - \mu)^2$,
- variance intra-strates : $\sum_{h=1}^{L} \pi_h \sigma_h^2$.

	Ratio variance inter/variance intra		
Fraction de sondage	0.1	1	2
1%	$< 1\%$	1%	2%
10%	1%	9%	17%
25%	2%	20%	33%

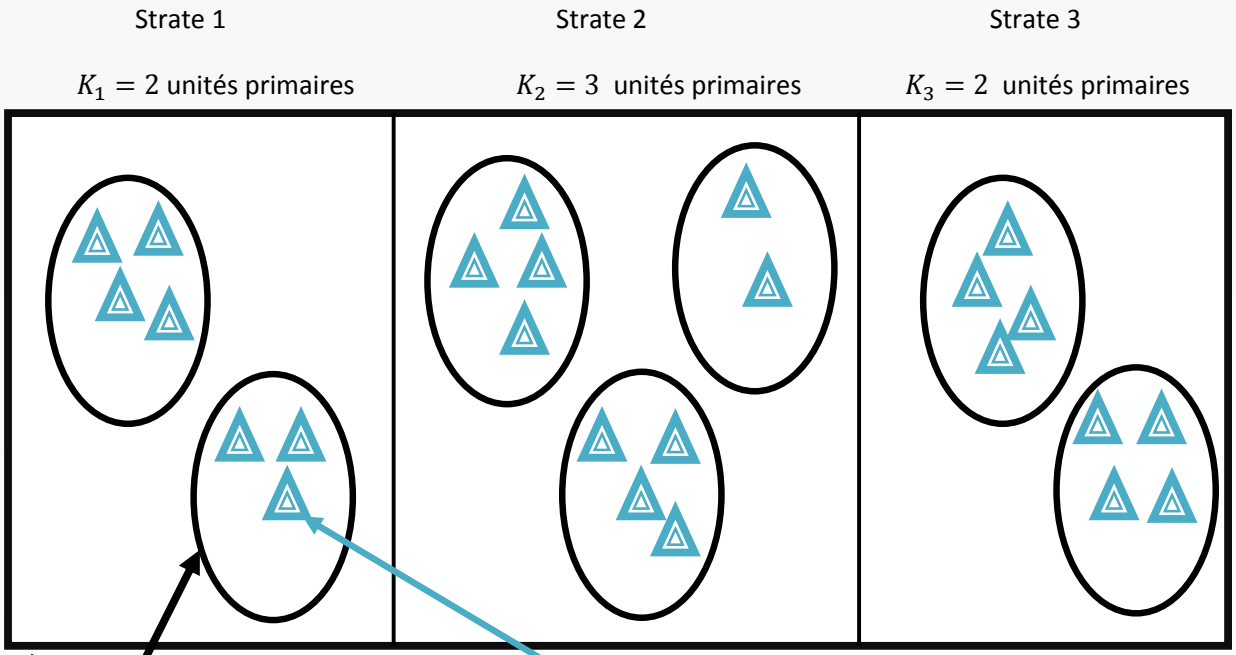
Table: Biais relatif de l'estimateur de variance $\hat{var}_{wr}(\bar{y})$ pour un STSRS avec un taux de sondage identique dans les strates

Section 3
Model with clusters

Le modèle

Population (U) : K unités primaires

Elle est subdivisée en L strates contenant K_l unités primaires



Unité primaire n° i

De taille $N_i = 3$ unités secondaires

Caractérisé par une variable de taille Z_i

Et une variable d'appartenance à la strate $\eta_i \in \{1..L\}$

Unité secondaire n° j de taille M_{ij} et de total T_{ij}

Le modèle

Les données dans la population U sont générées selon le modèle suivant :

$$F : \begin{pmatrix} M_{ij} \\ T_{ij} \end{pmatrix} \quad \sim_{iid} \quad \mathcal{L}\left(\begin{pmatrix} \alpha_i \\ \tau_i \end{pmatrix}, \begin{pmatrix} \sigma_{11i} & \sigma_{12i} \\ & \sigma_{22i} \end{pmatrix} \right)$$

$$(\alpha_i, \tau_i, \sigma_{11i}, \sigma_{22i}, \sigma_{12i}, N_i, Z_i, \eta_i) \quad \sim_{iid} F \quad \text{pour } i = 1, \ldots, K,$$

avec

- M_{ij} le nombre d'UT dans l'US j de l'UP i,
- T_{ij} le total de la variable d'intérêt dans l'US j de l'UP i,
- η_i une indicatrice de strate, générée selon une loi discrète de support $\{1, \ldots, L\}$,
- Z_i une variable aléatoire représentant une variable de taille pour l'UP i.

On note K_h le nombre total d'unités primaires obtenues dans la strate h (telles que $\eta_i = h$).

Le plan de sondage

- Au premier degré, un échantillon de k_h UP est sélectionné dans la strate h, avec des probabilités proportionnelles à la variable de taille Z_i.
- Au second degré, un échantillon de n_{hi} US est sélectionné dans l'UP i appartenant à la strate h, selon un sondage aléatoire simple **avec** remise.
- Les degrés suivants d'échantillonnage n'ont pas besoin d'être spécifiés. L'article requiert seulement la connaissance des poids de sondages w_{hijl} des UT appartenant à l'US j.

Le paramètre de superpopulation

- Le paramètre de superpopulation à estimer est :

$$\mu = \frac{E_F\left(N_i \tau_i\right)}{E_F(N_i \alpha_i)}$$

- estimé asymptotiquement sans biais par :

$$\bar{y} = \frac{t}{d} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{k_h} t_{hi}}{\sum_{h=1}^{L} \sum_{i=1}^{k_h} d_{hi}}$$

- où

$$t_{hi} = \sum_{j=1}^{n_{hi}} t_{hij}, \qquad d_{hi} = \sum_{j=1}^{n_{hi}} d_{hij},$$

$$t_{hij} = \sum_{l=1}^{m_{hij}} w_{hijl} y_{hijl}, \qquad d_{hij} = \sum_{l=1}^{m_{hij}} w_{hijl}.$$

Le paramètre de superpopulation

- La variance de cet estimateur est donnée par

$$Var(\bar{y}) \quad \approx \quad E_F Var_{RS}(\bar{y}) + Var_F(\bar{Y}) \tag{4}$$

car $E_{RS}(\bar{y}) \approx \bar{Y}$ et $\bar{Y} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{K_h} T_{hi}}{\sum_{h=1}^{L} \sum_{i=1}^{K_h} N_{hi}}$

- Le premier terme de (4) peut être estimé asymptotiquement sans biais par :

$$\widehat{var}_{wo}(\bar{y}) = \frac{1}{d^2} \left\{ \sum_{h=1}^{L} \sum_{i=1}^{k_h} \sum_{j<i}^{k_h} \left[\frac{\lambda_{hi}\lambda_{hj}}{\lambda_{hij}} - 1 \right] \left[(t_{hi} - \bar{y}d_{hi}) - (t_{hj} - \bar{y}d_{hj}) \right]^2 + K s_w^2 \right\}$$

- avec

$$s_w^2 = \frac{1}{K} \sum_{h=1}^{L} \sum_{i=1}^{k_h} \lambda_{hi} n_{hi} s_{hi}^2,$$

$$s_{hi}^2 = \frac{1}{n_{hi} - 1} \sum_{j=1}^{n_{hi}} \left[(t_{hij} - \bar{y}d_{hij}) - \frac{(t_{hi} - \bar{y}d_{hi})}{n_{hi}} \right]^2 .$$

Le paramètre de superpopulation

- La variance de cet estimateur est donnée par

$$Var(\bar{y}) \approx E_F Var_{RS}(\bar{y}) + Var_F(\bar{Y}) \quad (4)$$

car $E_{RS}(\bar{y}) \approx \bar{Y}$ et $\bar{Y} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{K_h} T_{hi}}{\sum_{h=1}^{L} \sum_{i=1}^{K_h} N_{hi}}$

- Un estimateur sans biais du deuxième terme de variance est donné par :

$$\widehat{var}\left[\bar{Y}\right] = \frac{1}{d^2}\left[\frac{K}{K-1}\sum_{h=1}^{L}\sum_{i=1}^{k_h}\lambda_{hi}\left(t_{hi} - \bar{y}d_{hi}\right)^2 - Ks_w^2\right]$$

- L'estimateur final de la variance est :

$$\widehat{var}_{SP}(\bar{y}) = \widehat{var}_{wo}\left(\bar{y}\right) + \widehat{var}\left[\bar{Y}\right]$$

- Cet estimateur nécessite la connaissance des probabilités d'inclusion d'ordre 2 des UP, notées λ_{hij}.

Le paramètre de superpopulation

- Afin de se libérer de cette contrainte, Graubard et Korn proposent de construire un estimateur de variance à partir de l'estimateur de variance avec remise.
- Pour cela, ils utilisent la décomposition suivante :

$$Var(\bar{y}) \quad = \quad E_F Var(\bar{y}|strateobs) + Var_F(E(\bar{y}|strateobs))$$

- Le premier terme peut être estimé asymptotiquement sans biais par :

$$\widehat{var}_{wr}(\bar{y}) = \frac{1}{d^2} \sum_{h=1}^{L} \frac{k_h}{k_h - 1} \sum_{i=1}^{k_h} \left[(t_{hi} - \bar{y} d_{hi}) - \frac{1}{k_h} \sum_{j=1}^{k_h} (t_{hj} - \bar{y} d_{hj}) \right]^2 .$$

Le paramètre de superpopulation

- Un estimateur sans biais du second terme de variance est donné par :

$$\hat{\Delta}_{st-mpps} = \widehat{var}_b - \widehat{var}_w$$

- où

$$\widehat{var}_b = \frac{1}{d^2} \sum_{h=1}^{L} \frac{1}{K_h} \left[\sum_{i=1}^{k_h} (t_{hi} - \bar{y} d_{hi}) \right]^2,$$

$$\widehat{var}_w = \frac{1}{d^2} \sum_{h=1}^{L} \frac{k_h}{K_h (k_h - 1)} \sum_{i=1}^{k_h} \left[(t_{hi} - \bar{y} d_{hi}) - \frac{1}{k_h} \sum_{j=1}^{k_h} (t_{hj} - \bar{y} d_{hj}) \right]^2.$$

- L'estimateur final de la variance est :

$$\widehat{var}_{SP}(\bar{y}) = \widehat{var}_{wr}(\bar{y}) + \widehat{var}_b - \widehat{var}_w.$$

- Une heuristique est donnée dans Korn et Graubard (1998).

Les conditions asymptotiques

- On considère que l'on tire un nombre d'UP croissant dans un nombre fixe de strates.
- On fait croître le nombre de strates en gardant dans chaque strate un taux de sondage faible pour les UP.
- Les conditions plus techniques sont données par Korn et Graubard (1998).

Remarques

- Le fait d'avoir une fraction de sondage faible au 1er degré, ne permet pas d'obtenir un estimateur de variance "avec remise" approximativement sans biais.
- En effet, même avec une fraction de sondage globale faible au 1er degré, ils montrent que le biais de l'estimateur de variance "avec remise" $\hat{var}_{wr}$ peut être élevé si une proportion (même petite) de grosses UP est échantillonnée, dans une strate avec une fraction de sondage importante.

Sections 4 et 5

- Section 4 : Revue de la littérature d'articles portant sur des sujets connexes,
- Section 5 : Application des estimateurs de variance proposés sur trois enquêtes (NHIS, NHANES III, NHDS).

Section 6
Discussion

Discussion

- **Choix du paramètre d'intérêt**
  Paramètres d'un modèle de superpopulation relativement simple. D'autres approches (complexes) permettant d'incorporer dans le modèle toutes les caractéristiques de la population et du plan de sondage sont possibles.
- **Choix de l'estimateur**
  L'utilisation d'un estimateur basé sur le plan de sondage permet de limiter les hypothèses du modèle.
- **Prise en compte de l'aléa de sondage**
  Possible de définir des estimateurs de variance basés sur le modèle, mais au prix d'hypothèses supplémentaires.
- **Réaliser une analyse conditionnelle?**