Estimation simplifiée de la variance dans le cas de l'échantillonnage à deux phases

Audrey Béliveau

Simon Fraser University

Travail en collaboration avec

David Haziza et Jean-François Beaumont

Université de Montréal et Statistique Canada

Colloque francophone sur les sondages 2012

Rennes, France

6 novembre 2012

Plan de la présentation

Échantillonnage à deux phases

- U : population finie de taille N
- s_1 : échantillon de $1^{\text{ère}}$ phase, de taille n_1
- s_2 : échantillon de 2^{e} phase, de taille n_2, sélectionné à partir de s_1

Échantillonnage à deux phases

- U : population finie de taille N
- s_1 : échantillon de $1^{\text{ère}}$ phase, de taille n_1
- s_2 : échantillon de 2^{e} phase, de taille n_2, sélectionné à partir de s_1
- I_{1i} : variable indicatrice de sélection de l'unité i dans s_1
- I_{2i} : variable indicatrice de sélection de l'unité i dans s_2
- Vecteurs d'indicatrices : $\mathbf{I}_1 = (I_{11}, \cdots, I_{1N})'$ and $\mathbf{I}_2 = (I_{21}, \cdots, I_{2N})'$

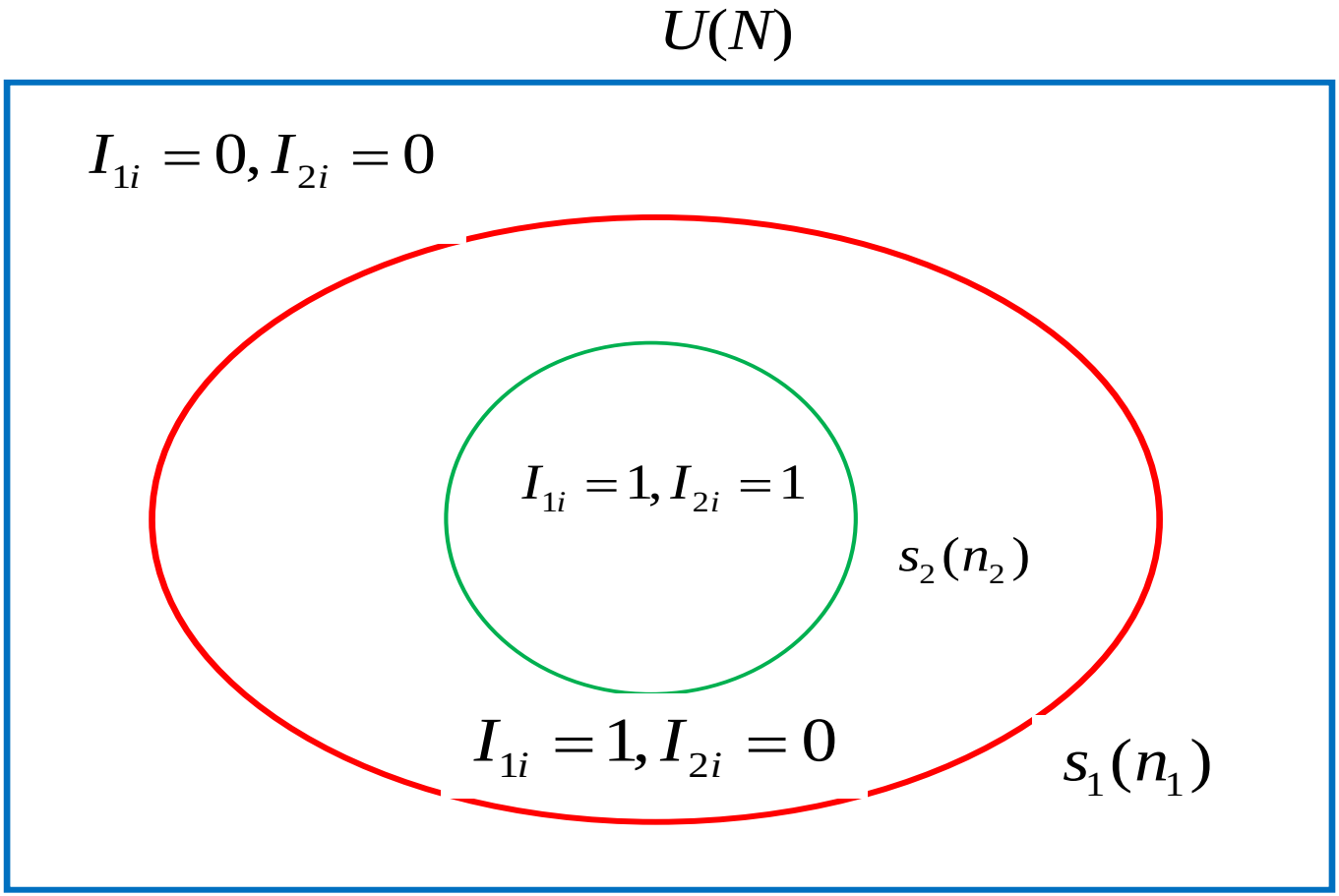
Échantillonnage à deux phases

- U : population finie de taille N
- s_1 : échantillon de $1^{\text{ère}}$ phase, de taille n_1
- s_2 : échantillon de 2^{e} phase, de taille n_2, sélectionné à partir de s_1
- I_{1i} : variable indicatrice de sélection de l'unité i dans s_1
- I_{2i} : variable indicatrice de sélection de l'unité i dans s_2
- Vecteurs d'indicatrices : $\mathbf{I}_1 = (I_{11}, \cdots, I_{1N})'$ and $\mathbf{I}_2 = (I_{21}, \cdots, I_{2N})'$
- Probabilité d'inclusion de l'unité i dans s_1 : $\pi_{1i} = P(I_{1i} = 1)$
- Probabilité d'inclusion conjointe des unités i et j dans s_1 : $\pi_{1ij} = P(I_{1i} = 1, I_{1j} = 1)$

Échantillonnage à deux phases

- U : population finie de taille N
- s_1 : échantillon de $1^{\text{ère}}$ phase, de taille n_1
- s_2 : échantillon de 2^{e} phase, de taille n_2, sélectionné à partir de s_1
- I_{1i} : variable indicatrice de sélection de l'unité i dans s_1
- I_{2i} : variable indicatrice de sélection de l'unité i dans s_2
- Vecteurs d'indicatrices : $\mathbf{I}_1 = (I_{11}, \cdots, I_{1N})'$ and $\mathbf{I}_2 = (I_{21}, \cdots, I_{2N})'$
- Probabilité d'inclusion de l'unité i dans s_1 : $\pi_{1i} = P(I_{1i} = 1)$
- Probabilité d'inclusion conjointe des unités i et j dans s_1 : $\pi_{1ij} = P(I_{1i} = 1, I_{1j} = 1)$
- Probabilité d'inclusion de l'unité i dans s_2 : $\pi_{2i}(\mathbf{I}_1) = P(I_{2i} = 1 | \mathbf{I}_1; I_{1i} = 1)$
- Probabilité d'inclusion conjointe des unités i et j dans s_2 : $\pi_{2ij}(\mathbf{I}_1) = P(I_{2i} = 1, I_{2j} = 1 | \mathbf{I}_1; I_{1i} = 1, I_{1j} = 1)$

Échantillonnage à deux phases



Invariance

- Un plan à deux phases possède la propriété d'invariance si

$$P(\mathbf{I}_2|\mathbf{I}_1) = P(\mathbf{I}_2)$$

- Invariance $\Rightarrow \pi_{2i}(\mathbf{I}_1) = \pi_{2i}$ et $\pi_{2ij}(\mathbf{I}_1) = \pi_{2ij}$
- Exemple de non-invariance :
  - Échantillon aléatoire simple sans remise (EASSR) à la première phase
  - Échantillon proportionnel à la taille à la deuxième phase, c'est-à-dire,

$$\pi_{2i}(\mathbf{I}_1) = n_2 \frac{x_i}{\sum_{i \in s_1} x_i},$$

où x_i représente la taille de l'unité i et est disponible pour tout $i \in s_1$.

Estimation pontuelle

- But : estimer le total de la variable d'intérêt y au niveau de la population :

$$Y = \sum_{i \in U} y_i$$

- Disponibilité de y : seulement pour $i \in s_2$
- Estimateur par double dilatation :

$$\hat{Y}_{DE} = \sum_{i \in s_2} \frac{y_i}{\pi_{1i}\pi_{2i}(\mathbf{I}_1)} = \sum_{i \in s_2} \frac{y_i}{\pi_i^*}$$

- $\hat{Y}_{DE}$ est sans biais sous le plan pour Y ; i.e.,

$$E_1 E_2(\hat{Y}_{DE}|\mathbf{I}_1) = Y$$

Variance totale

- Erreur totale de $\hat{Y}_{DE}$:

$$\hat{Y}_{DE} - Y = \underbrace{\left(\hat{Y}_E - Y\right)}_{\substack{\text{Erreur due à la} \\ 1^{\text{ère}} \text{ phase}}} + \underbrace{\left(\hat{Y}_{DE} - \hat{Y}_E\right)}_{\substack{\text{Erreur due à la} \\ 2^{\text{e}} \text{ phase}}} \quad (1)$$

où $\hat{Y}_E = \sum_{i \in s_1} \pi_{1i}^{-1} y_i$ est l'estimateur qui aurait été utilisé dans le cas d'un plan de sondage à une phase

- Variance totale de $\hat{Y}_{DE}$:

$$\begin{aligned} V\left(\hat{Y}_{DE}\right) &= V_1 E_2\left(\hat{Y}_{DE}|\mathbf{I_1}\right) + E_1 V_2\left(\hat{Y}_{DE}|\mathbf{I_1}\right) \\ &= \sum_{i \in U}\sum_{j \in U}(\pi_{1ij} - \pi_{1i}\pi_{1j})\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1j}} \\ &+ E_1\left\{\sum_{i \in s_1}\sum_{j \in s_1}\left(\pi_{2ij}(\mathbf{I}_1) - \pi_{2i}(\mathbf{I}_1)\pi_{2j}(\mathbf{I}_1)\right)\frac{y_i}{\pi_i^*}\frac{y_j}{\pi_j^*}\right\} \end{aligned}$$

Estimation de la variance totale

- Un estimateur sans biais de $V\left(\hat{Y}_{DE}\right)$ est obtenu en estimant séparément les deux termes contribuant à la variance :

$$\begin{aligned}\hat{V}\left(\hat{Y}_{DE}\right) &= \sum\sum_{i\in s_2\ j\in s_2} \frac{\Delta_{1ij}}{\pi_{2ij}(\mathbf{I}_1)} y_i y_j + \sum\sum_{i\in s_2\ j\in s_2} \Delta_{2ij}(\mathbf{I}_1)\frac{y_i}{\pi_{1i}}\frac{y_j}{\pi_{1j}} \\ &\equiv \hat{V}_1 + \hat{V}_2,\end{aligned}$$

où

$$\Delta_{1ij} = \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1ij}\pi_{1i}\pi_{1j}}$$

et

$$\Delta_{2ij}(\mathbf{I}_1) = \frac{\pi_{2ij}(\mathbf{I}_1) - \pi_{2i}(\mathbf{I}_1)\pi_{2j}(\mathbf{I}_1)}{\pi_{2ij}(\mathbf{I}_1)\pi_{2i}(\mathbf{I}_1)\pi_{2j}(\mathbf{I}_1)}.$$

Estimation de la variance totale

- $\hat{V}_1$ et $\hat{V}_2$ dépendent des probabilités d'inclusion jointes à la 2^{e} phase, $\pi_{2ij}(\mathbf{I}_1)$, qui peuvent être difficiles (voire impossibles) à obtenir.
- Le calcul de $\hat{V}\left(\hat{Y}_{DE}\right) = \hat{V}_1 + \hat{V}_2$ requiert un logiciel spécialisé conçu spécialement pour l'estimation de la variance dans le cas des plans à deux phases.
- But : proposer un estimateur de la variance simplifié qui
  - ne dépend pas des $\pi_{2ij}(\mathbf{I}_1)$;
  - peut être calculé à partir des logiciels conçus pour l'estimation de la variance dans le cas des plans à une phase.

Un estimateur de la variance simplifié

- En notant que

$$\frac{1}{\pi_{1ij}} = \frac{1}{\pi_{1i}\pi_{1j}} - \Delta_{1ij}$$

et

$$\frac{1}{\pi_{2ij}(\mathbf{I}_1)} = \frac{1}{\pi_{2i}(\mathbf{I}_1)\pi_{2j}(\mathbf{I}_1)} - \Delta_{2ij}(\mathbf{I}_1)$$

on peut réarranger les termes de sorte que l'estimateur de la variance totale, $\hat{V}\left(\hat{Y}_{DE}\right)$, s'exprime comme

$$\begin{aligned}\hat{V}\left(\hat{Y}_{DE}\right) &= \sum_{i\in s_2}\sum_{j\in s_2}\Delta_{1ij}\frac{y_i}{\pi_{2i}(\mathbf{I}_1)}\frac{y_j}{\pi_{2j}(\mathbf{I}_1)} + \sum_{i\in s_2}\sum_{j\in s_2}\frac{\Delta_{2ij}(\mathbf{I}_1)}{\pi_{1ij}}y_i y_j \\ &\equiv \hat{V}_1^R + \hat{V}_2^R.\end{aligned}$$

Un estimateur de la variance simplifié

- Estimateur de la variance simplifié : $\hat{V}_1^R$
  - ne dépend pas des $\pi_{2ij}(\mathbf{I}_1)$;
  - peut être calculé à partir des logiciels conçus pour l'estimation de la variance dans le cas des plans à une phase, car peut être exprimé comme

$$\hat{V}_1^R = \sum_{i \in s_1} \sum_{j \in s_1} \Delta_{1ij} z_i z_j,$$

où

$$z_i = \frac{y_i}{\pi_{2i}(\mathbf{I}_1)} I_{2i};$$

  - Cette expression correspond à l'estimateur de la variance dans le cas des plans à une phase appliqué à la variable z.
  - Ne requiert pas la propriété d'invariance.
- Question : Quand $\hat{V}_2^R$ est-il négligeable ?

Évaluer la contribution de $\hat{V}_2^R$

- Contribution de $\hat{V}_2^R$ à la variance totale :

$$\widehat{C}_2^R \equiv \frac{\hat{V}_2^R}{\hat{V}\left(\hat{Y}_{DE}\right)}$$

- Nous considérons également

$$\widetilde{C}_2^R \equiv \frac{\hat{V}_2^R}{\hat{V}_2} \geq \widehat{C}_2^R$$

- Si la propriété d'invariance est satisfaite, on peut aussi utiliser

$$C_2^R \equiv \frac{E_1 E_2(\hat{V}_2^R|\mathbf{I}_1)}{E_1 E_2\left(\hat{V}(\hat{Y}_{DE})|\mathbf{I}_1\right)}$$

$\hat{V}_2^R$ est-il négligeable ? Le cas d'un plan de Poisson à la 2e phase

- Utile dans le contexte de la non-réponse.
- Plan de Poisson à la 2e phase : consiste à procéder à des épreuves de Bernoulli independantes à la 2e phase avec probabilité $\pi_{2i}(\mathbf{I}_1)$.

$$\Rightarrow I_{2i} \text{ and } I_{2j} \text{ sont indépendants si } i \neq j$$

- Contribution de $\hat{V}_2^R$: $\widetilde{C}_2^R \equiv \frac{\hat{V}_2^R}{\hat{V}_2}$

$$|\widetilde{C}_2^R| \leq max(\pi_{1i})$$

- Condition usuelle : $\max(\pi_{1i}) = O(n_1/N) \Rightarrow$ l'estimateur de la variance simplifié $\hat{V}_1^R$ est approprié si n_1/N est négligeable.

$\hat{V}_2^R$ est-il négligeable ? Le cas d'un plan à deux degrés

- La population d'éléments est partitionnée en N grappes.
- 1^{er} degré : un échantillon de grappes de taille n est sélectionné ;
- 2^{e} degré : un échantillon est sélectionnée dans chaque grappe sélectionnée au 1^{er} degré ;
- Échantillonnage à deux degrés : cas particulier de l'échantillonnage à deux phases ;
- 2^{e} phase : indépendance de la sélection entre les grappes.

$\hat{V}_2^R$ est-il négligeable ? Le cas d'un plan à deux degrés

- Contribution de $\hat{V}_2^R$: $\widetilde{C}_2^R \equiv \frac{\hat{V}_2^R}{\hat{V}_2}$

$$|\widetilde{C}_2^R| \leq max(\pi_{1i})$$

- Condition usuelle : $\max(\pi_{1i}) = O(n/N) \Rightarrow$ l'estimateur de la variance simplifié $\hat{V}_1^R$ est approprié si n/N est négligeable.
- Estimateur de la variance simplifié : identique à l'estimateur simplifié dans Särndal, Swensson et Wretman (1992, Chapitre 4)

$\hat{V}_2^R$ est-il négligeable ? Le cas de l'échantillonnage aléatoire simple sans remise à la 2e phase

- Contribution de $\hat{V}_2^R$: $C_2^R \equiv \frac{E_1 E_2(\hat{V}_2^R|\mathbf{I}_1)}{E_1 E_2\left(\hat{V}(\hat{Y}_{DE})|\mathbf{I}_1\right)}$

$$C_2^R = \frac{-\left(\sum_{i\in U} y_i\right)^2 + n_1 \sum_{i\in U} y_i^2}{-\sum_{i\in U}\sum_{j\in U} \Delta_{1ij}\pi_{1ij} y_i y_j - \left(\sum_{i\in U} y_i\right)^2 + n_1 \sum_{i\in U} \frac{y_i^2}{\pi_{1i}}}.$$

- Sous de faibles conditions de régularité, le numérateur et le dénominateur sont $O(N^2)$.
- $\hat{V}_2^R$ n'est généralement pas négligeable dans le cas de l'EASSR à la 2e phase.
- On ne peut pas utiliser l'estimateur de la variance simplifié $\hat{V}_1^R$, même si n_1/N est négligeable.

$\hat{V}_2^R$ est-il négligeable ? Le cas de l'échantillonnage aléatoire simple sans remise à la 2^e phase

- Supposons que $\sum_{i \in U} y_i = 0$. Alors, C_2^R est $O(n_1/N)$
  $\Rightarrow \hat{V}_2^R$ est négligeable lorsque la fraction de sondage à la 1$^{\text{ère}}$ phase est négligeable.
- Lien avec le calage : lorsque $\sum_{i \in U} E_i = 0$ on peut utiliser un estimateur de la variance simplifié si n_1/N est négligeable
- Cas particulier : l'estimateur de Hájek

$$\hat{Y}_C = \frac{\sum_{i \in s_1} \frac{1}{\pi_{1i}}}{\sum_{i \in s_2} \frac{1}{\pi_i^*}} \hat{Y}_{DE}$$

- EASSR à la 2^e phase : étudié par Kott et Stukel (1997, Survey Methodology) et Kim, Navarro et Fuller (2006, JASA)

Quelques liens intéressants

- Variance totale de $\hat{Y}_{DE}$ en utilisant l'approche renversée (en supposant l'invariance) :

$$V\left(\hat{Y}_{DE}\right) = E_2 V_1\left(\hat{Y}_{DE}|\mathbf{I_2}\right) + V_2 E_1\left(\hat{Y}_{DE}|\mathbf{I_2}\right)$$

- La contribution de $V_2 E_1\left(\hat{Y}_{DE}|\mathbf{I_2}\right)$ à la variance totale est

$$\frac{V_2 E_1\left(\hat{Y}_{DE}|\mathbf{I_2}\right)}{V\left(\hat{Y}_{DE}\right)} = O\left(\frac{n_1}{N}\right)$$

- $V_2 E_1\left(\hat{Y}_{DE}|\mathbf{I_2}\right)$ est négligeable lorsque la fraction de sondage à la $1^{\text{ère}}$ phase est négligeable.
- Estimation de $E_2 V_1\left(\hat{Y}_{DE}|\mathbf{I_2}\right)$: il suffit de trouver un estimateur sans biais de $V_1\left(\hat{Y}_{DE}|\mathbf{I_2}\right)$
- Ex : linéarisation de Taylor ou méthode de rééchantillonnage

Quelques liens intéressants

- Estimateur de la variance simplifié : estimateur de $E_2 V_1 \left(\hat{Y}_{DE} | \mathbf{I_2} \right)$ obtenu en traitant les π_{2i} comme fixes.
- Problème : étant donné $\mathbf{I_2}$, les π_{2i} sont des variables aléatoires
  $\Rightarrow$ l'estimateur de la variance simplifié n'est généralement pas valide (ex : estimateur par double dilatation + EASSR à la 2^e phase)
- Kott et Stukel (1997, SM) : estimation de la variance par le jackknife en traitant les π_{2i} comme fixes.

Quelques liens intéressants

- Estimateur par double dilatation + EASSR à la 2^{e} phase :

$$\begin{aligned}
\hat{Y}_{DE} &= \sum_{i \in s_2} \frac{1}{\pi_{1i}} \frac{1}{\pi_{2i}} y_i \\
&= \frac{n_1}{n_2} \sum_{i \in s_2} w_{1i} y_i \\
&= \frac{\sum_{i \in s_1} w_{1i} \pi_{1i}}{\sum_{i \in s_1} w_{1i} (\pi_{1i} I_{2i})} \sum_{i \in s_1} w_{1i} (I_{2i} y_i)
\end{aligned}$$

- Estimation de $V_1\left(\hat{Y}_{DE} | \mathbf{I_2}\right)$: on peut utiliser une linéarisation de Taylor de premier ordre.

Quelques liens intéressants

- Kim, Navarro et Fuller (2006, JASA) : utilisent une méthode de rééchantillonnage (bootstrap)

$$\hat{Y}_{DE}^{(j)} = \frac{\sum_{i \in s_1} w_{1i}^{(j)} \pi_{1i}}{\sum_{i \in s_1} w_{1i}^{(j)} (\pi_{1i} I_{2i})} \sum_{i \in s_1} w_{1i}^{(j)} (I_{2i} y_i)$$

- Estimateur de la variance par répliques :

$$\hat{V}_{KNF} = \sum_{j=1}^{L} c_j \left(\hat{Y}_{DE}^{(j)} - \hat{Y}_{DE} \right)^2$$

- L'idée peut être appliquée au cas des estimateurs de calage (ex : l'estimateur de Hájek).

Résumé

- Nous avons proposé un estimateur de la variance simplifié :
  - valide pour certains plans de sondages et/ou estimateurs ponctuels
  - peut être utilisé pour n'importe quel estimateur de calage (pas uniquement l'estimateur de Hájek)
  - justifié par l'approche renversée
- Nous avons établi des liens entre l'estimateur de la variance simplifié et les résultats de Kott et Stukel (1997) et Kim, Navarro et Fuller (2006).