Estimation de la variance par linéarisation via l'indicatrice d'échantillonnage avec application à la non-réponse

Audrey-Anne Vallée
et
Yves Tillé

Université de Neuchâtel

Colloque Francophone sur les sondages
13 octobre 2016
Gatineau

Estimation d'un paramêtre

- U: population de N unités;
- But: estimer un paramètre d'intérêt θ, avec variable d'intérêt y;
- Échantillon: $\mathbf{a} = (a_1 \dots a_k \dots a_N)^\top$ où a_k vaut 1 si l'unité k est sélectionnée, 0 sinon;
- $\boldsymbol{\pi} = (\pi_1 \dots \pi_k \dots \pi_N)^\top$: où π_k est la probabilité d'inclusion d'ordre 1 de k;
- $\widehat{\theta} = \widehat{\theta}(\mathbf{y}, \mathbf{a})$: estimateur de θ.

Estimation de la variance de $\widehat{\theta}$

- Estimateur obtenu directement dans les cas simples;
- Méthodes de ré-échantillonnage:
  - Bootstrap (Efron, 1979; Shao et Steel, 1999),
  - Jackknife (Quenouille, 1949; Rao et Shao, 1992);
- Linéarisation de $\widehat{\theta}$ en fonction de:
  - totaux estimés (Binder, 1983, 1996; Woodruff, 1971),
  - poids de sondages (Demnati et Rao, 2004, 2010).

Estimation de la variance de $\widehat{\theta}$

- totaux estimés (Binder, 1983, 1996; Woodruff, 1971),
- poids de sondages (Demnati et Rao, 2004, 2010).

 Linéarisation pas en fonction des éléments aléatoires directement.

$\rightarrow$ Rien n'assure que $\widehat{\theta}$ est linéaire en a_k.

$\rightarrow$ Rien n'assure la possibilité de calculer une variance.

$\rightarrow$ Difficilement applicable en présence de non-réponse.

Approche proposée pour linéariser

Motivation: On désire calculer $\mathrm{V}_p(\widehat{\theta})$, où l'élément aléatoire de $\widehat{\theta} = \widehat{\theta}(\mathbf{y}, \mathbf{a})$ est $\mathbf{a}$. Rendre $\widehat{\theta}$ linéaire en termes des a_k, $k = 1 \ldots N$.

Linéarisation proposée, (Graf, 2011)

$$\widehat{\theta}(\mathbf{y}, \mathbf{a}) = \widehat{\theta}(\mathbf{y}, \boldsymbol{\pi}) + \sum_{\ell \in U} z_\ell (a_\ell - \pi_\ell) + R$$

où $z_\ell = \frac{\partial \widehat{\theta}}{\partial a_\ell}\Big|_{\mathbf{a}=\boldsymbol{\pi}}$ et R dépend des dérivées secondes.

Variance de l'estimateur

$$\mathrm{V}_p(\widehat{\theta}) \approx \mathrm{V}_p\left(\sum_{\ell \in U} z_\ell a_\ell\right) = \sum_{k \in U} \sum_{\ell \in U} (\pi_{k\ell} - \pi_k \pi_\ell) z_k z_\ell$$

Approche proposée pour linéariser

Variance de l'estimateur

$$\mathrm{V}_p(\widehat{\theta}) \approx \mathrm{V}_p\left(\sum_{\ell \in U} z_\ell a_\ell\right) = \sum_{k \in U}\sum_{\ell \in U}(\pi_{k\ell} - \pi_k\pi_\ell)z_k z_\ell$$

Estimation de la variance

On a que $z_\ell = \frac{\partial \widehat{\theta}}{\partial a_\ell}\Big|_{\mathbf{a}=\boldsymbol{\pi}}$ n'est pas disponible.

On propose deux estimations de z_ℓ: $\widehat{z}_\ell$ et $\widetilde{z}_\ell = \partial\widehat{\theta}/\partial a_\ell$.

On obtient l'estimateur

$$\widehat{\mathrm{V}}_p(\widehat{\theta}) = \sum_{k \in s}\sum_{\ell \in s}\frac{\pi_{k\ell} - \pi_k\pi_\ell}{\pi_{k\ell}}\widetilde{z}_k\widetilde{z}_\ell.$$

Estimateur calé

- $\widehat{\theta} = H(\mathbf{a}, \mathbf{w})$ l'estimateur calé de θ.
- $\mathbf{w} = (w_1 \dots w_k \dots w_N)^\top$, $w_k = F_k(\mathbf{x}_k^\top \boldsymbol{\lambda})/\pi_k$ est le poids de calage de l'unité k tel que $\sum_U a_k w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$.
- La variable linéarisée est $z_\ell = \frac{\partial \widehat{\theta}}{\partial a_\ell} = w_\ell e_\ell$.
  - e_ℓ: résidus de régression linéaire de $h_\ell(\mathbf{a}, \mathbf{w})$ sur les variables auxiliares $\mathbf{x}$, pondérés par $F'_\ell(\mathbf{x}_\ell^\top \boldsymbol{\lambda})$.
  - $h_\ell(\mathbf{a}, \mathbf{w})$: linéarisation comme si les poids de sondage étaient fixes.
- Estimation d'un total: Résultat cohérent avec Demnati et Rao (2004).

Traitements de la non-réponse

Deux types de non-réponse:

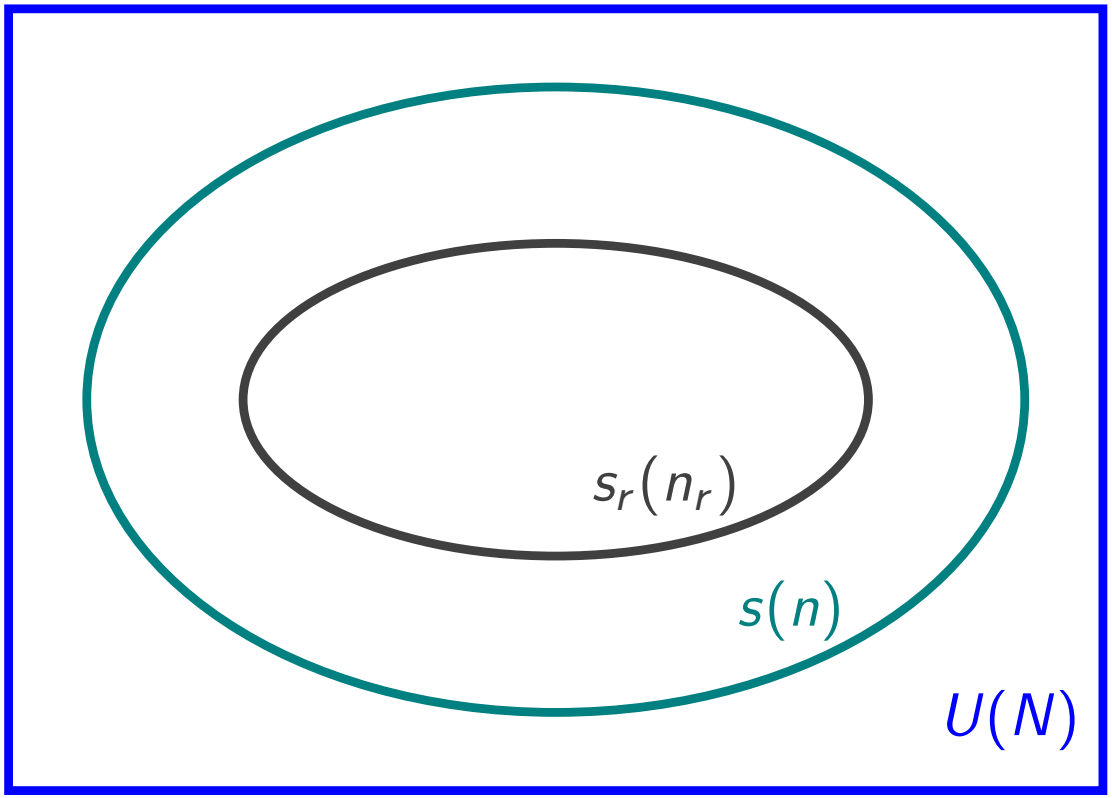
- Non-réponse partielle,
- Non-réponse totale.

Deux traitements de la non-réponse:

- Imputation des valeurs manquantes,
- Repondération des unités répondantes.

Soit $\widehat{\theta}_I$ l'estimateur de θ considérant le traitement de la non-réponse.

Sources d'aléa en présence de non-réponse traitée



3 sources d'aléa de $\widehat{\theta}_I$:

- Échantillon $\mathbf{a}$,
- Réponse $\mathbf{R} = (R_1 \dots R_k \dots R_N)^\top$,
- Variable d'intérêt $\mathbf{y}$.

On écrit $\widehat{\theta}_I = \widehat{\theta}_I(\mathbf{y}, \mathbf{a}, \mathbf{R})$

Cadres de travail pour l'inférence

- Basé sur le plan de sondage: Échantillon **a** et réponse **R** aléatoires. Variable d'intérêt **y** vue comme fixe.
- Basé sur un modèle: Échantillon **a** et modèle de **y** aléatoires. Hypothèses sur la réponse **R** qui est indépendante de **y** conditionnellement.

Décomposition de la variance

Deux approches:

- Deux-phases: $U \to s \to s_r$.
- Renversée: $U \to U_r \to S_r$.

Décompositions de la variance de $\widehat{\theta}_I$:

	Inférence basée sur	
Approche	Plan de sonsage	Modèle
Deux-phases	$\mathrm{E}_p\mathrm{E}_q(\widehat{\theta}_I - \theta)^2$	$\mathrm{E}_m\mathrm{E}_p\mathrm{E}_q(\widehat{\theta}_I - \theta)^2$
Renversée	$\mathrm{E}_q\mathrm{E}_p(\widehat{\theta}_I - \theta)^2$	$\mathrm{E}_q\mathrm{E}_m\mathrm{E}_p(\widehat{\theta}_I - \theta)^2$

 Comment calculer ces espérances quand l'estimateur et le traitement de la non-réponse sont complexes?

Méthodologie pour estimer la variance par linéarisation

Proposition: Pour chaque espérance, linéariser le paramètre en fonction de l'élément aléatoire directement.

Variance renversée basée sur le plan de sondage

Ici, le plan de sondage et la réponse sont aléatoires. Donc linéariser en fonction de a_ℓ, puis en fonction de R_ℓ.

1. Décomposion de la variance: $\mathrm{V}(\widehat{\theta}_I) = \mathrm{E}_q \mathrm{V}_p(\widehat{\theta}_I|\mathbf{R}) + \mathrm{V}_q \mathrm{E}_p(\widehat{\theta}_I|\mathbf{R})$.

Méthodologie pour estimer la variance par linéarisation

Variance renversée basée sur le plan de sondage

1. Décomposion de la variance: $\mathrm{V}(\widehat{\theta}_I) = \mathrm{E}_q \mathrm{V}_p(\widehat{\theta}_I|\mathbf{R}) + \mathrm{V}_q \mathrm{E}_p(\widehat{\theta}_I|\mathbf{R})$.
2. Linéarisation via a_ℓ:
$$\widehat{\theta}_I \approx \widehat{\theta}_I(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) + \sum_U z_\ell^a (a_\ell - \pi_\ell), \quad z_\ell^a = \frac{\partial \widehat{\theta}_I}{\partial a_\ell}.$$
3. Approximation des termes de la variance:
$$\mathrm{V}_p(\widehat{\theta}_I|\mathbf{R}) \approx \sum\sum_U (\pi_{k\ell} - \pi_k \pi_\ell) z_k^a z_\ell^a, \; \mathrm{E}_p(\widehat{\theta}_I|\mathbf{R}) \approx \widehat{\theta}_I(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}).$$
4. Linéarisation via R_ℓ:
$$\widehat{\theta}_I(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) \approx \widehat{\theta}_I(\mathbf{y}, \boldsymbol{\pi}, \mathbf{p}) + \sum_U z_\ell^{aR} (R_\ell - p_\ell), \quad z_\ell^{aR} = \partial \widehat{\theta}_I(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) / \partial R_\ell.$$
5. Approximation de la variance:
$$\mathrm{V}_q \mathrm{E}_p(\widehat{\theta}_I|\mathbf{R}) \approx \mathrm{V}_q \widehat{\theta}_I(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R}) \approx \sum\sum_U (p_{k\ell} - p_k p_\ell) z_k^{aR} z_\ell^{aR}.$$

Repondération par le calage pour estimer un total

- $\widehat{\theta}_I = \sum_U a_k R_k w_k y_k$ estimateur repondéré de $\theta = \sum_U y_k$.
- $w_k = F_k(\mathbf{x}_k^\top \boldsymbol{\lambda})/\pi_k$ tel que $\sum_U a_k R_k w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$.
- La linéarisation via a_ℓ:

$$z_\ell^a = \frac{\partial \widehat{\theta}_I}{\partial a_\ell} = w_\ell e_\ell.$$

- $\mathrm{E}_q \mathrm{V}_p(\widehat{\theta}_I) \approx \mathrm{E}_q \sum\sum_U (\pi_{k\ell} - \pi_k \pi_\ell) w_k w_\ell e_k e_\ell$.
- Linéarisation via R_ℓ:

$$z_\ell^{aR} = \frac{\partial \widehat{\theta}_I(\mathbf{y}, \boldsymbol{\pi}, \mathbf{R})}{\partial R_\ell} = F_\ell(\mathbf{x}_\ell^\top \boldsymbol{\lambda}) e_\ell.$$

- $\mathrm{V}_q \mathrm{E}_p(\widehat{\theta}_I) = \sum\sum_U (p_{k\ell} - p_k p_\ell) F_\ell(\mathbf{x}_\ell^\top \boldsymbol{\lambda}) F_k(\mathbf{x}_k^\top \boldsymbol{\lambda}) e_k e_\ell$.
- Si p_k est estimé par $F_k(\mathbf{x}_k^\top \boldsymbol{\lambda})$, résultat obtenu par Kott (2006).

Conclusion

La méthodologie d'approximation

- assure la linéarisation de l'estimateur en les éléments aléatoires,
- assure la possibilité de calculer espérance/variance,
- assure l'obtention d'un estimateur de variance explicite,
- est efficace pour tous les cas: estimateurs complexes, estimateurs calés, estimateurs imputés, estimateurs repondérés,
- est simple puisqu'il suffit de suivre une méthodologie,
- est cohérente et intuitive.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex survey. *International Statistical Review*, **51**, 279–292.

Binder, D. A. (1996). Linearization methods for single phase and two-phase samples: a cookbook approach. *Survey Methodology*, **22**, 17–22.

Demnati, A. et Rao, J. N. K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, **30**, 17–34.

Demnati, A. et Rao, J. N. K. (2010). Linearization variance estimators for model parameters from complex survey data. *Survey Methodology*, **36**, 193–201.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1–26.

Graf, M. (2011). Use of survey weights for the analysis of compositional data. In *Compositional Data Analysis: Theory and Applications* (eds. V. Pawlowsky-Glahn et A. Buccianti), 114–127. Chichester: Wiley.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, **32**, 133–142.

Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, 355–375.

Rao, J. N. K. et Shao, J. (1992). Jackknife variance estimation with survey data under hot-deck imputation. *Biometrika*, **79**, 811–822.

Shao, J. et Steel, P. (1999). Variance estimation for survey data with composite imputation and nonneglible sampling fractions. *Journal of the American Statistical Association*, **94**, 254–265.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, **66**,