Design-based and Model-based Methods for Estimating Model Parameters

David A. Binder, Georgia R. Roberts

Groupe de lecture économétrie/sondages

19 janvier 2015

Estimation de paramètres à partir de données d'enquête

Questions

- Faut-il utiliser les poids de sondage et si oui, comment ?
- Tests d'hypothèses, intervalles de confiance : comment estimer la variance ?

Deux types d'approches

- Approches fondées sur le plan (*design-based*)
- Approches fondées sur les modèles (*model-based*)

Binder et Roberts (2003) : étudient les biais des estimateurs dans les différents contextes ainsi que les conséquences d'un "mauvais" choix de spécification.

⇒ **approches opposées ou réconciliables ?**

Quelques références pour une perspective historique des débats :
Kish "*The hundred years' wars of survey sampling*" (1995)
Särndal "*Models in Survey Sampling*" (2010)

Särndal(2010)

Revue des débats sur le rôle des modèles dans les sondages, dans un cadre théorique comme pratique.

Réconcilier les deux approches ?

- revenir à des relations amicales et à l'harmonie
- rendre compatible, faire converger

→ **Les deux approches ne peuvent vivre l'une sans l'autre !**

(expliciter modélisation - ratio, régression généralisée -, intégrer phase de sélection dans les modèles, petits domaines, non réponse, calage, etc.)

Plan de la présentation

- Plan vs modèles : définitions et résultats élémentaires
- Variance totale : mécanisme d'échantillonnage en deux phases
- Statistiques non linéaires
- Conséquences d'une mauvaise spécification de modèle
- Recommandations et conclusion

Approche plan

Paramètres d'intérêt : θ_N de la **population finie** de taille N
Aléa : sélection d'un échantillon (selon un plan de sondage) de n unités parmi N
Valeurs prises par les unités considérées comme fixes.

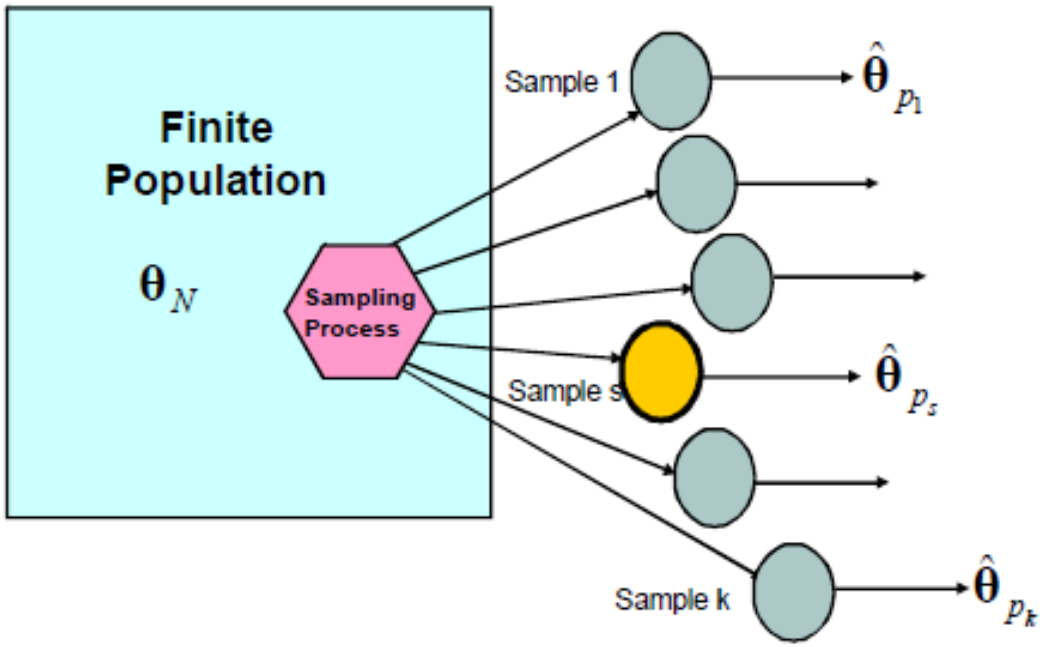

Fig. 1: Design-based Framework

Binder(2011)

Approche modèle

Paramètres d'intérêt : θ d'une **superpopulation infinie**
Aléa : observations sont n réalisations indépendantes issues de la superpopulation

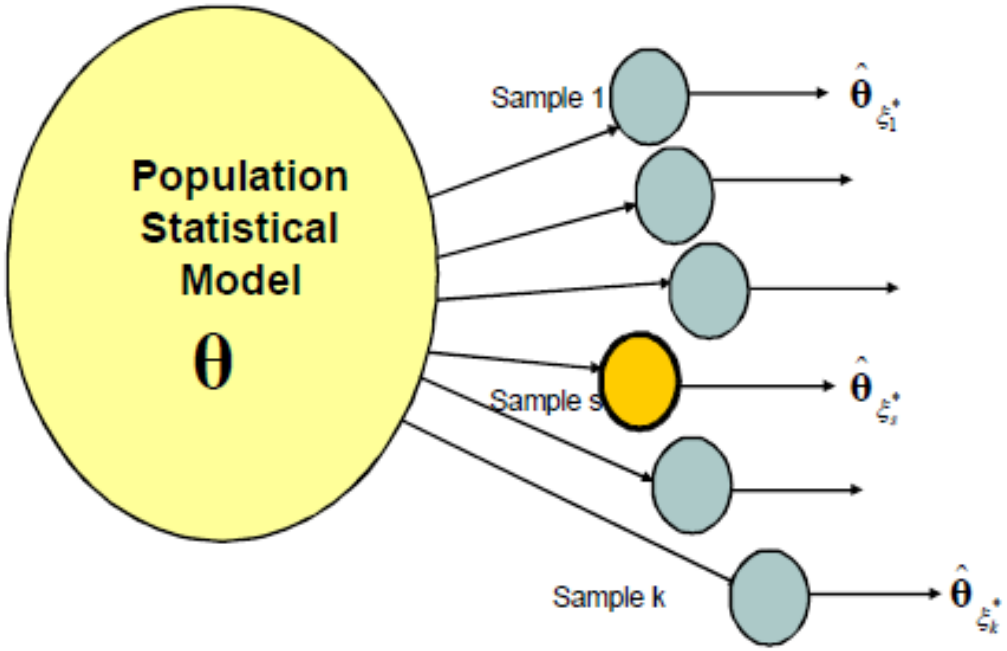

Fig. 2: Model-based Framework

Binder(2011)

Notations

$\bar{y}_s$ moyenne empirique de y sur l'échantillon s
$\bar{y_U}$ moyenne de y sur la population finie U
v_U variance de y sur la population U

I_t $(t = 1, \ldots, N)$ variables indicatrices de sélection dans l'échantillon s

$-\xi$ relatif au modèle
$-p$ relatif au plan
$-\xi p$ relatif aux deux mécanismes aléatoires

Estimation d'une moyenne

Approche modèle

Observations $y_1, \ldots, y_n$ réalisations indépendantes d'une superpopulation où les y_i sont de moyenne β et de variance σ^2.
Le paramètre d'intérêt est β et dans une approche modèle qui ignore le plan de sondage utilisé pour sélectionner les n unités, l'estimateur est la moyenne non pondérée sur l'échantillon s, $\hat{\beta} = \bar{y}_s$.

Approche plan

Les n observations proviennent des unités tirées (selon un plan de sondage) dans une population finie U de taille N.
Le paramètre d'intérêt b est la moyenne $\bar{y}_U$ sur la population U estimée dans l'approche plan par la moyenne sur s tenant compte des poids de sondage.

$\Rightarrow$ **approches *a priori* non réconciliables ?**

Si les données ont bien été générées selon le modèle spécifié, alors $\hat{\beta}$ est non biaisé selon l'approche plan pour b :

$$E_p(\hat{\beta}) - b = o(1)$$

Exemple d'un plan de sondage stratifié

On suppose que le modèle sous-jacent est que les y_t sont de moyenne μ, de variance σ^2 et non corrélées.
L'échantillonnage est réalisé selon un plan stratifié : n_1 unités sont sélectionnées dans une strate de taille N_1 et n_2 dans celle de taille N_2.

L'estimateur habituel pour l'approche plan de $b = \bar{y}_U$ est :

$$\hat{b} = \frac{N_1 \bar{y_1}_s + N_2 \bar{y_2}_s}{N} \quad \text{avec} \quad E_p[\hat{b}] = \frac{N_1 \bar{y_1}_U + N_2 \bar{y_2}_U}{N}$$

L'estimateur pour l'approche modèle est $\hat{\beta} = \bar{y}_s$ d'où

$$E_p[\hat{\beta}] = \frac{n_1 \bar{y_1}_U + n_2 \bar{y_2}_U}{n}$$

Dans le cas général, l'estimateur $\hat{\beta}$ ne converge pas (pour le plan) vers le paramètre d'intérêt de la population finie.
En revanche, si le modèle spécifié est "vrai", il y a convergence asymptotique pour $\hat{\beta}$ vers b.

Approche *model-design-based*

Alternative aux approches modèle et plan "pures" : mécanisme en deux phases avec un aléa de sélection aux deux phases, approche dite *model-design-based*.
Généralisation de l'approche plan pure qui conditionne sur la première phase (valeurs de la population finie considérées comme fixes) et de l'approche modèle pure qui conditionne sur la sélection de deuxième phase (variables de sélection I_t considérées comme fixes).

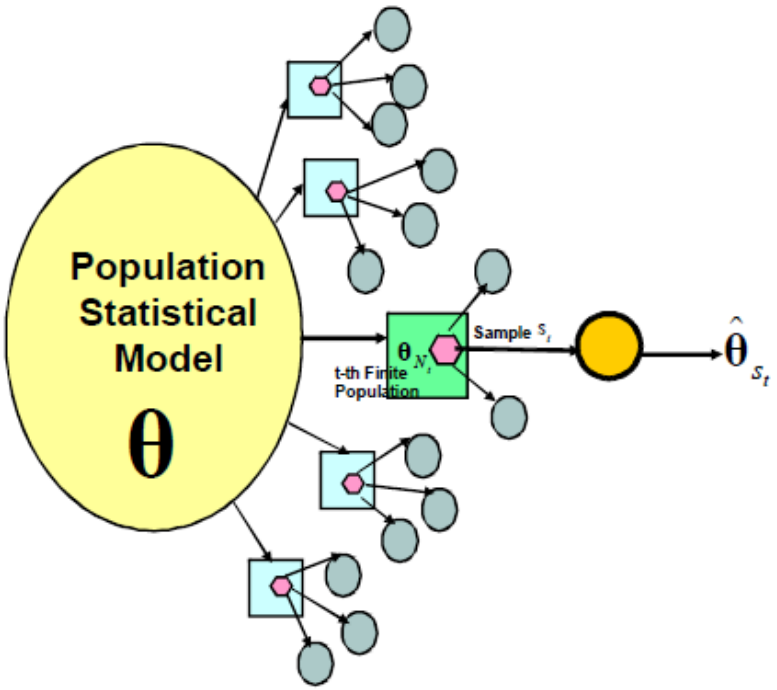

Fig. 3: Model-design-based Framework

Variance totale

Ainsi, si la variance de $\hat{b}$ (dans l'approche plan pure) est $var_p[\hat{b}]$, la variance totale, tenant compte des deux phases est :

$$var_{\xi p}[\hat{b}] = var_\xi[E_p[\hat{b}]] + E_\xi[var_p[\hat{b}]]$$

Les auteurs montrent alors (sous certaines hypothèses) qu'**un estimateur** $v_p(\hat{b})$ **de** $var_p[\hat{b}]$ **peut être utilisé pour estimer la variance totale** $var_{\xi p}[\hat{b}]$ si $v_p(\hat{b})$ est asymptotiquement non biaisé (pour le modèle) pour $E_\xi[var_p[\hat{b}]]$ ce qui est le cas quand $v_p(\hat{b})$ converge vers $var_p[\hat{b}]$.

Erreur quadratique moyenne

Lorsque **les données sont bien générées selon le modèle spécifié, l'estimateur "modèle" $\hat{\beta}$ peut être meilleur que l'estimateur "plan"** $\hat{b}$ du point de vue de l'erreur quadratique moyenne (espérance modèle de l'EQM plan).

Exemple : on reprend le plan stratifié et le modèle de l'exemple précédent.

$$\hat{\beta} = \bar{y}_s = \frac{n_1 \bar{y_1}_s + n_2 \bar{y_2}_s}{n}$$

$$E_p[\hat{\beta}] = \frac{n_1 \bar{y_1}_U + n_2 \bar{y_2}_U}{n} \quad \text{d'où} \quad E_p[\hat{\beta}] - b = \alpha(\bar{y_1}_U - \bar{y_2}_U) \quad \text{avec} \quad \alpha = \frac{n_1}{n} - \frac{N_1}{N}$$

Sachant que :

$$var_p[\hat{\beta}] = \frac{n_1 v_{1U} + n_2 v_{2U}}{n^2}$$

On montre que lorsque le modèle considéré est vrai :

$$E_\xi[mse_p(\hat{\beta})] \leq E_\xi[mse_p(\hat{b})]$$

Ratios et statistiques non linéaires

En utilisant des méthodes de linéarisation de Taylor, les auteurs montrent que pour :

- les ratios
- les statistiques non linéaires définies explicitement
- les statistiques non linéaires définies implicitements (MV)

lorsque le modèle est vrai, les ξp-espérances de $\hat{b}$ et $\hat{\beta}$ convergent vers le paramètre d'intérêt du modèle et la variance de $\hat{b}$ est asymptotiquement sans biais pour la variance totale.

Le plan de sondage n'est pas ignorable

Exemple : Pfeffermann (1996)

Population finie avec valeurs 0 ou 1 où le modèle serait $P(y_t = 1) = \mu$ avec $0 < \mu < 1$. Le paramètre d'intérêt, β est égal à $E_\xi[\bar{y_U}] = \mu$.

Considérons que pour chaque unité où $y_t = 0$, la probabilité de sélection est p_0 et qu'elle est p_1 pour les $y_t = 1$. Alors, d'après le théorème de Bayes :

$$\mu_{t1} = P(y_t = 1 | I_t = 1) = \frac{\mu p_1}{(1-\mu)p_0 + \mu p_1}$$

qui n'est pas nécessairement égal à μ.

En effet, avec ce plan non ignorable, le biais de $\bar{y}_s$ comme estimateur de μ est :

$$\mu_{t1} - \mu = \frac{\mu(1-\mu)(p_1 - p_0)}{(1-\mu)p_0 + \mu p_1}$$

et la moyenne de l'échantillon, $\bar{y}_s$, ne converge pas vers μ sous le modèle spécifié, dès lors que $p_0 \neq p_1$.

La structure du modèle pour les moyennes μ est incorrecte

Exemple

Deux observations y_1 et y_2 et le modélisateur suppose que y_1 et y_2 ont même moyenne et ont comme matrice de variance

$$\frac{\sigma^2}{n}\begin{pmatrix}1 & 0\\ 0 & 4\end{pmatrix}$$

Le meilleur estimateur linéaire sans biais de la moyenne est alors $0,8y_1 + 0,2y_2$.

Mais si les vraies moyennes $\mu_t = E_\xi(y_t)$ pour $t = 1,2$ sont différentes alors le biais de cet estimateur de $\beta = \frac{\mu_1+\mu_2}{2}$ est $0,3(\mu_1 - \mu_2)$ et donc $\hat{\beta}$ n'est pas convergent pour l'approche modèle.

La structure du modèle pour les variances σ est incorrecte

Exemple

On suppose que le modèle sous-jacent est que les observations $y_1, \ldots, y_n$ sont indépendantes et identiquement distribuées, de loi de Poisson de paramètre μ.

L'estimateur usuel pour μ est la moyenne de l'échantillon $\bar{y}_s$ dont la variance est $\frac{\mu}{n}$ sous le modèle de Poisson. Ainsi un estimateur de la variance de $\bar{y}_s$ (dans une approche modèle) communément utilisé serait $\frac{\bar{y}_s}{n}$.

Mais un tel estimateur de la variance peut être potentiellement biaisé si le modèle de Poisson n'est pas correct et que la vraie variance de y_t n'est pas μ.

Recommandations

Choix entre approche modèle et approche plan doit être fondé sur la pertinence de la modèle que l'on suppose.

→ si la spécification du modèle n'est pas correcte, l'estimation des paramètres et des variances peut être largement fausse en particulier si les estimations ne sont pas robustes à la mauvaise spécification.

→ ainsi **il est souvent plus pertinent de préférer l'approche plan (plutôt l'estimateur $\hat{b}$ que $\hat{\beta}$)** : $\hat{b}$ est asymptotiquement sans biais pour β et l'estimateur (plan) de la variance fournit un estimateur raisonnable de la variance totale de $\hat{b}$ quand l'échantillon est grand et que les taux de sondages sont faibles, même si le modèle n'est pas correctement spécifié. La perte d'efficacité peut exister si le modèle spécifié est en fait correct : mais pour de gros échantillons, cette perte d'efficacité n'est pas tant problématique que cela.

Conclusion

Quelques points de conclusion sur l'article :

- un bon panorama sur les approches modèle et plan (mais l'approche modèle est abordée dans un cadre assez restrictif), la réconciliation de ces différentes approches, l'importance de la modélisation et de la robustesse des estimations à cette phase préliminaire, etc.
- quelques points non abordés (cadre des fonctions estimantes pour les estimations de variance, propriétés conditionnelles).
- le cadre théorique reste très général : pas d'exemple pratique où les approches peuvent se compléter, le problème de la non réponse n'est pas évoqué, les poids de sondage sont à peine mentionnés et la régression n'est pas étudiée en tant que telle (alors que l'estimateur de l'approche modèle s'appelle $\hat{\beta}$) ... la présentation suivante sera sans doute plus éclairante sur ces différents points !

Bibliographie

BINDER (2011)
Estimating Model Parameters from a Complex Survey under a Model-Design Randomization Framework (*Pakistan Journal of Statistics*)

KISH (1995)
The hundred years' wars of survey sampling (*Statistics in Transition*)

SARNDAL (2010)
Models in Survey Sampling (*Official Statistics - Methodology and Applications in Honour of Daniel Thorburn*)