Variance estimation through influence function linearization technique : theory and applications

Camelia Goga

IMB, Université de Bourgogne
e-mail : camelia.goga@u-bourgogne.fr

Workshop on Survey Sampling Theory and Methodology
Kuressaare - August 25-29 2008

Parameter of interest : the one-sample case

Population $\mathbf{U}$, sample $s \subset \mathbf{U}$ according to $\mathbf{p(s)}$ with $\pi_k > 0$ and $\pi_{kl} > 0$, $k \neq l \in \mathbf{U}$

• Parameter of interest is a **nonlinear** function of population totals, $\phi = \phi(t_x, t_y, \ldots)$ such as :

- Ratio $R = t_y / t_x$
- Covariance $\text{Cov} = \sum_U x_k y_k / N - \sum_U x_k \sum_U y_k / N^2$
- functions of the empirical distribution function as the Gini index $G = \sum_U y_k \left(2F(y_k) - 1\right) / t_y$, the Theil index,...
- eigenfunctions of the functional principal components analysis : $\Gamma v_j(t) = \lambda_j v_j(t)$, $t \in [0,1]$ (Cardot *et al.*, 2007) with $\Gamma = \frac{1}{N} \sum_{k \in U} (Y_k - \mu) \otimes (Y_k - \mu)$.

Complex statistic : the one-sample case

• The substitution estimator $\hat{\phi}$ for ϕ : each total is substituted by the Horvitz-Thompson estimator :

$$\hat{\phi} = \phi \left(\sum_s \frac{x_k}{\pi_k}, \sum_s \frac{y_k}{\pi_k}, \dots \right)$$

- all indexes are nonlinear functions of population totals ;
- we have parameters depending on quantiles ;
- we have implicit parameters.

• Variance $\mathrm{Var}(\hat{\phi})$?

• Variance estimator $\widehat{\mathrm{Var}}(\hat{\phi})$?

Variance estimation : resampling methods

The most used :

1. **jackknife** : repeated computation leaving out one observation (Rao *et al.*, 1992, Berger & Skinner, 2005.)
2. **bootstrap** (Gross, 1980, Chauvet, 2007) : construct a pseudo-population U^* by individuals duplicating assumed to mimic U and draw independent samples from U^* according to the initial survey design
3. **balanced repeated replication**...

The jackknife and linearization methods are similar in the sense that the analytic derivative in linearization is replaced by a numerical approximation (Davison & Hinkley, 1997, page 50).

Variance estimation : linearization methods

1. **Estimation equations** : Kovačević & Binder, 1997 ;
2. **Influence function** : Deville, 1999 ;
3. **Taylor linearization** : Demnati & Rao, 2004.

Consist in finding a linearized variable u_k (unknown) and approximate

$$\text{Var}(\hat{\phi}) \simeq \text{Var}\left(\frac{\sum_s u_k}{\pi_k}\right) = \sum_U \sum_U \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}$$

$$\widehat{\text{Var}}(\hat{\phi}) = \widehat{\text{Var}}\left(\frac{\sum_s \hat{u}_k}{\pi_k}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}$$

Wolter (1985, p. 316) : "...it may be warranted that the Taylor series method is good, perhaps best in some circumstances, in terms of the MSE and bias criteria but the balanced half-samples and secondarily the jackknife methods are preferable from the point of view of confidence interval coverage probabilities"

Influence function linearization technique : functionals of measure M

A very general approach allowing to linearize statistics which are not Taylor linearisable (Gini index or eigenfunctions of the functional principal components analysis).

- Let be the population U and $\mathcal{X} \in \mathbf{R}^p$, x_k, $k \in U$ the variable of interest;
- Define on $\mathbf{R}^p$ a mesure M as follows

$$M = \sum_{k \in U} \delta_{x_k}, \quad \delta_{x_k}(x) = \begin{cases} 1 & \text{if} & x = x_k \\ 0 & \text{elsewhere} & \end{cases}$$

$(U, \mathcal{X})$ is identified by the measure M.

- Let be a homogeneous functional T of degree α and write

$$\phi = T(M).$$

1. Population total : $t_x = \sum_U x_k = \int \mathcal{X} dM$ of degree 1,
2. Ratio : $R = \frac{\sum_U x_{k,2}}{\sum_U x_{k,1}} = \frac{\int \mathcal{X}_2 dM}{\int \mathcal{X}_1 dM}$ of degree 0.

Substitution estimator $T(\hat{M})$

- Estimate M by $\hat{M}$ with weights $w_k = \frac{1}{\pi_k}$ for each individual $k \in s$ and zero elsewhere,

$$\hat{M} = \sum_U w_k \delta_{x_k} = \sum_s \frac{1}{\pi_k} \delta_{x_k}$$

- The substitution estimator is

$$\hat{\phi} = T(\hat{M})$$

- Examples :

1. $\hat{t}_x = \int \mathcal{X} d\hat{M} = \sum_s \frac{x_k}{\pi_k}$
2. $\hat{R} = \frac{\int \mathcal{X}_2 d\hat{M}}{\int \mathcal{X}_1 d\hat{M}} = \frac{\sum_s x_{k,2}/\pi_k}{\sum_s x_{k,1}/\pi_k}$

The linearized variables : the influence function

The linearized variable corresponds to the influence function of T at M and $x = \mathcal{X}(k) = x_k$, $k \in U$,

$$u_k = IT(M, x_k).$$

Influence function : Gâteaux derivative of $T(M)$ in the direction of the Dirac mass at x,

$$IT(M, x) = \lim_{h \to 0} \frac{T(M + h\delta_x) - T(M)}{h}$$

Examples : 1. For $T = t_x$, we have $u_k = x_k$ and
2. For $T = R = \dfrac{\sum_U x_{k,2}}{\sum_U x_{k,1}}$, we have $u_k = \frac{1}{t_{x_1}} \left(x_{k,2} - R \cdot x_{k,1} \right)$.

Asymptotic variance of $T(\hat{M})$: asymptotic framework

Let us suppose :

1. $\lim_{N\to\infty} N^{-1}\int \mathcal{X}dM$ exists ;
2. $\lim_{N\to\infty} N^{-1}\left(\int \mathcal{X}d\hat{M} - \int \mathcal{X}dM\right) = 0$ in probability ;
3. $\lim_{N\to\infty} n^{1/2}N^{-1}(\int \mathcal{X}d\hat{M} - \int \mathcal{X}dM) = N(0,\Sigma)$ in distribution.
4. T is Fréchet differentiable.

Result

Under the above assumptions, we have

$$\sqrt{n}N^{-\alpha}(\hat{\phi}-\phi) = \sqrt{n}N^{-\alpha}\sum_{U} u_k(w_k-1) + o(1).$$

The influence function approach is a theoretical justification of the Taylor linearization approach developed by Demnati & Rao (2004).

Remarks upon the remainder

- the requirement of T to be Fréchet differentiable is strong : it assures that the remainder is of order $o(d(\frac{\hat{M}}{N} - \frac{M}{N})) = o_p(n^{-1/2})$.
- we can relax this assumption by asking T to be only Hadamard differentiable : we obtain the "δ-method" (books of van der Vaart, 1998 and Luisa Fernholz, 1982) ;
- and if we have only the Gateaux differentiability, then one must make supplementary assumptions upon the sampling design (upon π_k and π_{kl}) (Cardot *et al*, 2007 and Chaouch & Goga, 2008).

Variance estimation

The result gives us that the asymptotic variance of $\hat{\phi} = T(\hat{M})$ is equal to the HT variance of $\sum_s \frac{u_k}{\pi_k}$, namely

$$\sum_U \sum_U \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l}, \quad \text{with} \quad \Delta_{kl} = \pi_{kl} - \pi_k \pi_l.$$

Drawbacks : we have sums on U and the linearized variables are unknown, so the variance estimator is :

$$\widehat{\mathrm{Var}}(\hat{\phi}) = \widehat{\mathrm{Var}}\left(\frac{\sum_s \hat{u}_k}{\pi_k}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{\hat{u}_k}{\pi_k} \frac{\hat{u}_l}{\pi_l}$$

where $\hat{u}_k = IT(\hat{M}, x_k)$.

Extensions or applications of Deville's approach

1. Applications :
   - **estimation of eigenelements of the functional principal components analysis (FPCA)** ; work in collaboration with Hervé Cardot, Mohamed Chaouch and Catherine Labruère from University of Burgundy, France and submitted to JSPI, 2007.
   - **estimation of the multidimensional quantile** ; work in collaboration with Mohamed Chaouch from University of Burgundy, preprint 2008.
2. Extension :
   - **partial influence functions approach for two-sample complex statistics** ; work in collaboration with Jean-Claude Deville from ENSAI/CREST Rennes and Anne Ruiz-Gazen from University Toulouse 1, France, in revision for Biometrika, 2008.

First application : Functional Data with Survey data

Functional Data Deville (1974), Dauxois *et al.* (1982), Besse & Ramsay (1986), Kirkpatrick & Heckman (1989), Ramsay & Silverman (2002, 2005), ...

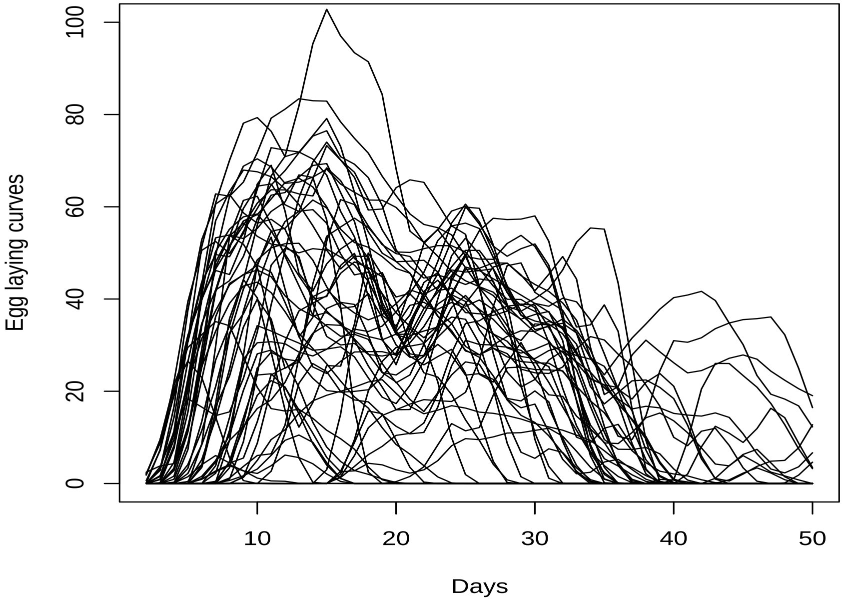
An observation is the realization of a random function $Y(t)$ (growth curve, temperature curve, ...) taking values in a function space $H = L^2([0, \mathcal{T}])$.

For instance n realizations $Y_1, Y_2, \ldots, Y_n$ of a continuous time process $Y = \{Y(t),\ t \in [0, \mathcal{T}]\}$

with discrete time measurements

$$\mathbf{Y}_i = \left(Y_i(t_1^i), Y_i(t_2^i), \ldots, Y_i(t_{p_i}^i)\right)'$$

An example : egg laying curves



Functional Data collected with survey sampling

The way data are collected is seldom taken into account

- Design of experiments in a functional setting : Cuevas *et al.* (2003).
- Multivariate PCA with survey data : Skinner, Holmes & Smith (1986), Deville (1999).

A study motivated by a project of the French electricity operator (EDF)
The aim is to have a precise idea of electricity consumption ("ideal" production, marketing, *etc*).
A population of more than 30 millions of electricity meters (for each firm or household) which will be able to deliver consumption curves for each household.

▷ Impossible to save and analyse online all this information.
A complex (balanced) survey approach to get a sample of electricity meters with measurements at a fine time scale (Dessertaine, 2006).

Survey sampling framework

We consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ with size N.

At each element k of the population U, we can associate a deterministic function

$$Y_k = (Y_k(t))_{t\in[0,\mathcal{T}]} \in L^2[0,\mathcal{T}].$$

Let us denote by $\mu \in L^2[0,1]$, the "mean" of the functions Y_k

$$\mu(t) \quad = \quad \frac{1}{N}\sum_{k\in U} Y_k(t), \quad t \in [0,\mathcal{T}]$$

and the "covariance function" by

$$\gamma(s,t) = \frac{1}{N}\sum_{k\in U} (Y_k(t) - \mu(t))\,(Y_k(s) - \mu(s))$$

The covariance operator Γ is defined, for all $f \in L^2[0,\mathcal{T}]$, by

$$\Gamma f(s) \quad = \quad \int \gamma(s,t) f(t)dt \ , \quad s \in [0,\mathcal{T}].$$

Functional PCA for finite populations

▷ The best linear approximation in a q dimensional functional space, according to a variance criterion, of the functions Y_k

$$Y_k(t) \;=\; \mu(t) + \sum_{j=1}^{q} \langle Y_k - \mu, v_j \rangle v_j(t) + R_{qk}(t), \;\; t \in [0, T].$$

The mean square of remainder terms R_{qk}

$$\frac{1}{N} \sum_{k \in U} \|R_{qk}\|^2$$

is minimum for

$$\Gamma v_j(t) \;=\; \lambda_j \; v_j(t), \quad t \in [0, \mathcal{T}].$$

The eigenfunctions v_j form an orthonormal system in $L^2[0, T]$, the eigenvalues satisfy $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N \geq 0$.

Estimation of eigenelements of FPCA

Generally, N is unknown, so we have nonlinear functional parameters :

$$\begin{aligned}\mu(t) &= \frac{1}{N}\sum_{k\in U} Y_k(t), \quad t\in[0,\mathcal{T}]\\ \gamma(s,t) &= \frac{1}{N}\sum_{k\in U}\left(Y_k(t)-\mu(t)\right)\left(Y_k(s)-\mu(s)\right)\end{aligned}$$

and (λ_j, v_j) are defined implicitly by

$$\Gamma v_j(t) = \lambda_j\ v_j(t), \quad t\in[0,\mathcal{T}].$$

We have only a sample s of "curves" selected from the whole population U according to a sampling design $p(\cdot)$.
We want to give estimations of the "mean curve" and of the main modes of variability of the data, $(\lambda_j, v_j)_{j=1}^q$ based on the sample s.

We apply the influence function approach.

Nonlinear functionals of a discrete measure

Let us introduce a discrete measure M defined on $L^2[0, \mathcal{T}]$ by

$$M = \sum_{k \in U} \delta_{Y_k}$$

where $\delta_{Y_k} = 1$ if $Y = Y_k$ and zero else.

Our quantities of interest are (nonlinear) functionals T of this measure

$$N(M) = \int dM, \quad \mu(M) = \frac{\int \mathcal{Y} dM}{\int dM}$$

$$\Gamma(M) = \frac{\int (\mathcal{Y} - \mu(M)) \otimes (\mathcal{Y} - \mu(M)) \, dM}{\int dM}$$

The eigenelements of Γ are also functionals of M defined in an implicit way.

Linearized variables for FPCA

Result

Let us suppose that $\sup_{k\in U} \|Y_k\| < \infty$. *The influence functions of* μ *and of* Γ *exist and they are given by*

$$
\begin{aligned}
I\mu(M, Y_k) &= \frac{1}{N}(Y_k - \mu) \\
I\Gamma(M, Y_k) &= \frac{1}{N}\left((Y_k - \mu) \otimes (Y_k - \mu) - \Gamma\right).
\end{aligned}
$$

If moreover the non null eigenvalues of Γ *are distinct*

$$
\begin{aligned}
I\lambda_j(M, Y_k) &= \frac{1}{N}\left(\langle Y_k - \mu, v_j\rangle^2 - \lambda_j\right) \\
Iv_j(M, Y_k) &= \frac{1}{N}\left(\sum_{\ell\neq j} \frac{\langle Y_k - \mu, v_j\rangle\langle Y_k - \mu, v_\ell\rangle}{\lambda_j - \lambda_\ell} v_\ell\right).
\end{aligned}
$$

This result is similar to the multivariate case of PCA (Croux & Ruiz-Gazen, 2005 with a robust statistics point of view).

Asymptotic variance

We estimate M by $\hat{M}$. Under broad assumptions upon the sampling design $p(\cdot)$, we have

$$\hat{\mu} - \mu = \sum_s \frac{I\mu(M, Y_k)}{\pi_k} + o_p(n^{-1/2}),$$

If moreover the non null eigenvalues of Γ are distinct

$$\hat{\lambda}_j - \lambda_j = \sum_s \frac{I\lambda_j(M, Y_k)}{\pi_k} + o_p(n^{-1/2}),$$

$$\hat{v}_j - v_j = \sum_s \frac{Iv_j(M, Y_k)}{\pi_k} + o_p(n^{-1/2}).$$

$\Longrightarrow$ one can obtain the asymptotic variance of $\hat{\mu}$, $\hat{\lambda}_j$ and $\hat{v}_j$.

Second Application : Multidimensional Quantile Estimation

The observations are vectors $Y_1,\ Y_2, \dots,\ Y_N$ from $\mathbf{R}^d$.

The multidimensional or geometric u-th quantile $Q(u)$ is a generalization of the uni-dimensional quantile :

$$Q(u) = \arg\min_{\theta \in \mathbf{R}^d} \sum_{k=1}^{N} \phi(u, Y_k - \theta) \quad \text{for} \quad u \in B^d = \{z \in \mathbf{R}^d : ||z|| < 1\}.$$

$\phi : \mathbf{R}^d \times B^d \to \mathbf{R}$ with

$$\phi(u, t) = ||t|| + \langle u, t \rangle$$

for $|| \cdot ||$ the usual Euclidean norm and $< \cdot, \cdot >$ the usual Euclidean inner product

Existence and uniqueness of $Q(u)$

The objective function

$$\sum_{k=1}^{N} \phi\left(u, Y_k - \theta\right) \quad \text{is}$$

- continuos and convex with respect of θ
- and it explodes to infinity when $\|\theta\| \to \infty$, then

the uth quantile $Q(u)$ is the unique solution of the following equation

$$\sum_{k=1}^{N} \frac{\partial \phi\left(u, Y_k - \theta\right)}{\partial \theta} = \sum_{k=1}^{N} \left[S\left(Y_k - \theta\right) + u\right] = 0$$

Estimation of $Q(u)$ with Survey Data

Let be u a direction and a sample s from U;
The sample uth quantile, $\hat{Q}(u)$, is the unique solution of the equation

$$\sum_s \frac{S\left(Y_k - \theta\right) + u}{\pi_k} = 0$$

How to linearize this complex statistic ?
We introduce the measure M on $\mathbf{R}^d$ and the functional T given by

$$T(M, \theta) = \sum_{k=1}^{N} \left[S\left(Y_k - \theta\right) + u\right] = \int \left[S\left(Y - \theta\right) + u\right] dM.$$

The population u-th quantile $Q(u)$ is the solution of

$$T(M, \theta) = 0.$$

Asymptotic variance

- The functional T is differentiable with respect to M and θ,
- the matrix $\partial T / \partial \theta$ is inversible
- then the implicit theorem assures the existence and uniqueness of a functional $\tilde{T}$ such that

$$\tilde{T}(M) = Q(u) \quad \text{the u-th quantile}$$

Moreover, $\tilde{T}$ is differentiable with respect to M and $\tilde{T}(\hat{M}) = \hat{Q}(u)$ the sample u-th quantile. Deville's result gives

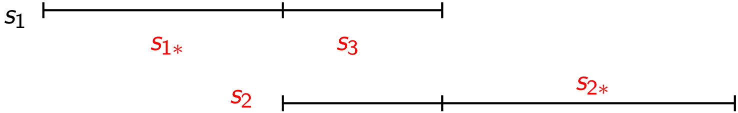
$$\begin{aligned} \tilde{T}(\hat{M}) - \tilde{T}(M) &= \int I\tilde{T}(M, Y) d(\hat{M} - M) + o_p(n^{-1/2}) \\ &= \sum_s \frac{I\tilde{T}(M, Y_k)}{\pi_k} + o_p(n^{-1/2}) \end{aligned}$$

with $I\tilde{T}(M, Y_k) = -(\partial T / \partial \theta)^{-1} \left[S\left(Y_k - \theta\right) + u \right].$

Complex statistic : the two-sample case

We consider a finite population U and

1. $s_1, s_2 \subset U$ selected according to $p_1(s_1), p_2(s_2)$ with π_k^1, π_k^2 ;
2. variables of interest $\mathcal{Z}_1 \in \mathbf{R}^{p_1}$ known on s_1 and $\mathcal{Z}_2 \in \mathbf{R}^{p_2}$ known on s_2



- Repeated sampling
- Nonresponse estimation

...

Complex statistic : the two-sample case

Estimate a nonlinear function of totals $t_{z_1} = \sum_{k \in U} z_{k_1}$, $t_{z_2} = \sum_{k \in U} z_{k_2}$,

$$\phi = \phi(t_{z_1}, t_{z_2})$$

taking into account the individuals from the common sample s_3.

On s_3, we know $\mathcal{Z}_3 = (\mathcal{Z}_1, \mathcal{Z}_2) \in \mathbf{R}^{p_3}$ with $p_3 = p_1 + p_2$.

1. **Gini index change** : $\Delta G = G_2 - G_1$.
2. Ratio estimation $R = t_y / t_x$ when nonresponse occurs differently for $\mathcal{X}$ and $\mathcal{Y}$.
3. **Covariance** : $\mathrm{Cov}(X, Y) = \sum_U x_k y_k / N - \sum_U x_k \sum_U y_k / N^2$ in a change estimation problem.

Two dimension linearization method through **partial** influence functions

$(\mathbf{U}, \mathcal{Z}_t)$ associated with a measure M_t for $t = 1, 2, 3$

$$M_t = \sum_{k \in U} \delta_{z_{k,t}}$$

and write $\Phi = T(M)$ with $M = (M_1, M_2, M_3)$

Examples : The ratio estimation $R = t_y / t_x$ with nonresponse, then

$(\mathbf{U}, \mathcal{X})$ associated with M_1 and $(\mathbf{U}, \mathcal{Y})$ associated with M_2,

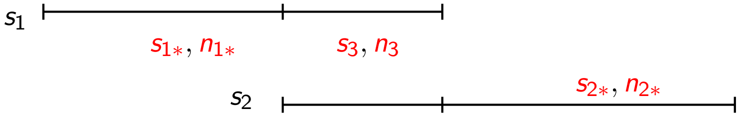
$$R = \frac{t_y}{t_x} = \frac{\int \mathcal{Y} dM_2}{\int \mathcal{X} dM_1}$$

Why different measures ? Because the variables of interest are measured on different samples.

Composite estimation

Two-dimension sampling design :
the probability $p(s = (s_1, s_2))$ of selecting a two-sample $s = (s_1, s_2) \in [\mathcal{P}(U)]^2$.

$$p(s) \geq 0 \quad \text{and} \quad \sum_s p(s) = 1.$$



Inclusion probabilities : for $d \in \{1*, 3, 2*\}$,

$$\pi_k^d = Pr(k \in s_d) = E(I_k^d) \quad \text{for} \quad I_k^d = \mathbf{1}_{\{k \in s_d\}}$$

$$\pi_{kl}^{d,d'} = Pr(k \in s_d \& l \in d') = E(I_k^d I_l^{d'})$$

Examples

1. Two-dimension Bernoulli (BE2) sampling :

$$\pi^{1*}, \quad \pi^3, \quad \pi^{2*}$$

$$p(s = (s_1, s_2)) = \pi_{1*}^{n_{1*}} \pi_3^{n_3} \pi_{2*}^{n_{2*}} (1 - \pi_{1*} - \pi_3 - \pi_{2*})^{N - n_1 - n_3 - n_{2*}}$$

2. Two-dimension simple random sampling without replacement (SRS2) :

$$p(s = (s_1, s_2)) = \frac{n_{1*}! n_3! n_{2*}! (N - n_1 - n_3 - n_{2*})!}{N!}$$

$$\pi_k^{1*} = \frac{n_{1*}}{N}, \quad \pi_k^3 = \frac{n_3}{N}, \quad \pi_k^{2*} = \frac{n_{2*}}{N}$$

- Composite estimator for M_1 and M_2 :

$$\hat{M}_t = \sum_U v_{k,t} \delta_{z_{k,t}}$$

$$v_{k,1} = \frac{a}{\pi_k^{1*}} I_k^{1*} + \frac{1-a}{\pi_k^3} I_k^3 \quad \text{and} \quad v_{k,2} = \frac{b}{\pi_k^{2*}} I_k^{2*} + \frac{1-b}{\pi_k^3} I_k^3$$

- HT estimator on the common sample s_3 for M_3

$$\hat{M}_3 = \sum_U \frac{\delta_{z_{k,t}}}{\pi_k^3} I_k^3$$

Particular cases :

1. $a = b = 0$, $\hat{M}_t = \sum_U \frac{\delta_{z_{k,t}}}{\pi_k^3} I_k^3$, $t = 1, 2$ are HT estimators on s_3
2. $a = \frac{\pi_k^{1*}}{\pi_k^1} = cst$ (for BE2 or SRS2), $b = \frac{\pi_k^{2*}}{\pi_k^2} = cst$ then

$M_t, t = 1, 2$ are HT estimators on s_1 and s_2 (resp.)

$$\hat{M}_1 = \sum_U \frac{\delta_{z_{k,t}}}{\pi_k^1} I_k^1, \quad \hat{M}_2 = \sum_U \frac{\delta_{z_{k,t}}}{\pi_k^2} I_k^2$$

Two-sample variance estimation

Substitution estimator of $T(M)$ is $T(\hat{M})$, $\hat{M} = (\hat{M}_1, \hat{M}_2, \hat{M}_3)$.

First partial influence function : $IT_1(M; z)$ of $T(M)$ is

$$IT_1(M; z) = \lim_{h \to 0} \frac{T(M_1 + h\delta_z, M_2, M_3) - T(M_1, M_2, M_3)}{h}$$

when this limits exists.

$IT_2(M; z)$ and $IT_3(M; z)$ defined similarly.

Linearized variables : $u_{k,t} = IT_t(M; z_{k,t})$ for $z_{k,t} \in \mathbf{R}^{p_t}$.

Example : For $R = \int \mathcal{Y} dM_2 / \int \mathcal{X} dM_1$ we have
$u_{k,1} = -x_k t_y / t_x^2$, $u_{k,2} = y_k / t_x$ and $u_{k,3} = 0$.

Asymptotic framework

For $t \in \{1, 2, 3\}$, we suppose that

1. $\lim\limits_{N\to\infty} n_2^{-1} n_1 = \lambda$ for $\lambda > 0$ and $\lim\limits_{N\to\infty} N^{-1} n_t = \pi_t \in (0, 1)$;
2. $\lim\limits_{N\to\infty} N^{-1} \int \mathcal{Z}_t dM_t$ exists ;
3. $\lim\limits_{N\to\infty} N^{-1} \left(\int \mathcal{Z}_t d\hat{M}_t - \int \mathcal{Z}_t dM_t \right) = 0$ in probability ;
4. $\lim\limits_{N\to\infty} \left({n_t}^{1/2} N^{-1} (\int \mathcal{Z}_t d\hat{M}_t - \int \mathcal{Z}_t dM_t) \right)_{t=1}^{3} = N(0, \Sigma)$ in distribution.
5. T is homogeneous, namely it exists a real number $\beta > 0$ dependent on T such that $T(rM) = r^{\beta} T(M)$ for any real $r > 0$;
6. $\lim_{N\to\infty} N^{-\beta} T(M) < \infty$.
7. T is Fréchet differentiable.

Main results : result 1

Variance approximation $T(\hat{M})$ is approximated by

$$\begin{aligned}\frac{\sqrt{n}}{N^{\beta}}(T(\hat{M})-T(M)) &= \frac{\sqrt{n}}{N^{\beta}}\sum_{t=1}^{3}\int I_t T(M;z)d(\hat{M}_t-M_t)(z)+o_p(1)\\ &= \frac{\sqrt{n}}{N^{\beta}}\sum_{t=1}^{3}\left(\sum_{k\in U}u_{k,t}(v_{k,t}-1)\right)+o_p(1).\end{aligned}$$

Then :

$$\mathrm{var}\left(\frac{\sqrt{n}}{N^{\beta}}(T(\hat{M})-T(M))\right)\simeq \mathrm{var}\left(\frac{\sqrt{n}}{N^{\beta}}\sum_{t=1}^{3}\left(\sum_{k\in U}u_{k,t}(v_{k,t}-1)\right)\right)$$

Main results : result 2

We consider the unbiased composite estimator $\hat{M}_t$ of M_t for $t = 1, 2$ and the HT estimator for M_3.
Let us denote by $\hat{t}^d_{u_t} = \sum_{k \in s_d} \frac{u_{k,t}}{\pi^d_k}$ for $t \in \mathcal{T}$ and $d = 1*, 2*, 3$.
Under the assumptions 1-7,

1. the substitution estimator $T(\hat{M})$ is approximated by the composite estimator

$$a\left(\hat{t}^{1*}_{u_1} - \hat{t}^{3}_{u_1}\right) + b\left(\hat{t}^{2*}_{u_2} - \hat{t}^{3}_{u_2}\right) + \sum_{t=1}^{3} \hat{t}^{3}_{u_t}.$$

2. $\mathrm{var}\left(\frac{\sqrt{n}}{N^\beta}(T(\hat{M}) - T(M)\right) \simeq \frac{n}{N^{2\beta}}\left(\theta'\Gamma\theta + 2\theta'\gamma + \mathrm{var}\sum_{t=1}^{3}\hat{t}^{3}_{u_t}\right)$ where $\theta = (a, b)'$ and Γ (resp. γ) is a matrix (resp. a vector) of variance terms. This variance is minimum for $\theta_{opt} = -\Gamma^{-1}\gamma$.

Two-dimension simple random sampling without replacement (SRS2) and $\Phi = T(M_1, M_2)$

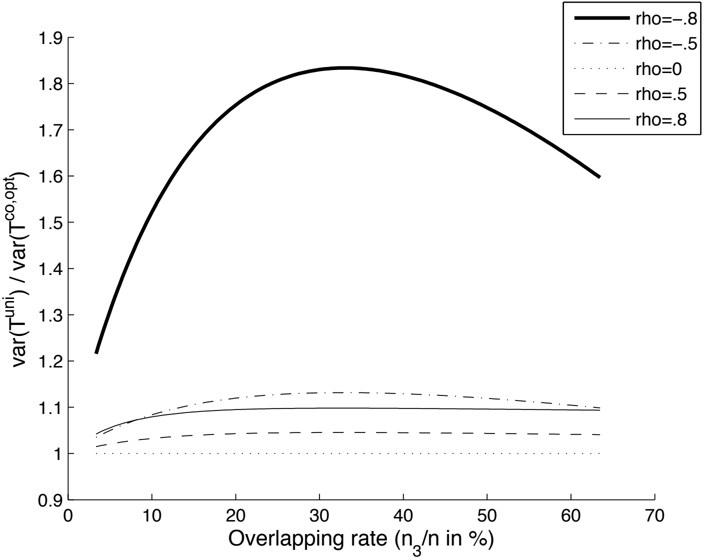
- Parameter $\phi = T(M_1, M_2)$, $u_{k,1}$ and $u_{k,2}$ the linearized variables ;
- Population $\mathbf{U}$ of size $N = 1000$;
- SRS2 sample $(s_1, s_2) \subset U \times U$ such that

$n = n_{1*} + n_3 + n_{2*} = 300$ and $n_{1*} = 100$;
- Comparison between

$$\begin{aligned}
T(\hat{M}_1^{opt}, \hat{M}_2^{opt}) &\simeq a_{opt}\left(\hat{t}_{u_1}^{1*} - \hat{t}_{u_1}^{3}\right) + b_{opt}\left(\hat{t}_{u_2}^{2*} - \hat{t}_{u_2}^{3}\right) + \sum_{t=1}^{2} \hat{t}_{u_t}^{3} \\
T(\hat{M}_1^{uni}, \hat{M}_2^{uni}) &\simeq \hat{t}_{u_1}^{1} + \hat{t}_{u_2}^{2} \quad a = \frac{n_{1*}}{n_1}, b = \frac{n_{2*}}{n_2} \\
T(\hat{M}_1^{int}, \hat{M}_2^{int}) &\simeq \hat{t}_{u_1}^{3} + \hat{t}_{u_2}^{3} \quad a = b = 0
\end{aligned}$$

for different values of ρ_{u_1,u_2} and of n_3/n;
- We suppose $S_{u_2}^2 / S_{u_1}^2 = 1$.

Plan SRS2 : the optimal estimator or the estimator on s_1 and s_2 ?



Plan SRS2 : the optimal estimator or the estimator on s_3 ?