Balanced sampling by two-stage cube method

Shoaib Ali, Li-Chun Zhang, Angela Luna

Department of Social Statistics & Demography
University of Southampton, UK

ITACOSM Florence

5 June, 2019

Notations

- Finite population: $U = \{1, ..., N\}$,
- Response variable: y with population values $y_1, ..., y_N$,
- Sampling distribution: $p(s)$ where s is random sample of fixed size n,
- Sample space: Ω such that $\sum_{s \in \Omega} p(s) = 1$,
- First-order inclusion probabilities: π_i, $i \in U$ defined as $\pi_i = \sum_{s \in \Omega} I_i p(s)$,
- Sample membership indicator variable: I_i, $i \in U$, where $I_i = 1$ if $i \in s$, $I_i = 0$ otherwise,
- Population total of response variable: $Y = \sum_{i \in U} y_i$,
- Horvitz-Thompson (HT) estimator for Y: $\hat{Y} = \sum_{i \in s} \frac{y_i}{\pi_i}$. [Horvitz and Thompson, 1952]

Balanced sampling design

- Let $x_1, ..., x_J$ are known auxiliary variables, related with y,
- A sampling design is balanced with respect to balancing variables $x_1, ..., x_J$ if it satisfy the balancing equations $\sum_{i \in s} \frac{\mathbf{x}_i}{\pi_i} - \sum_{i \in U} \mathbf{x}_i = \mathbf{0}$ for any sample s, where $\mathbf{x}_i = (x_{i1}, ..., x_{iJ})$,
- Relationship of y and $x_1, ..., x_J$ is defined by a population model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\epsilon_i \sim N(0, \sigma_i^2)$ and $X_{N \times J}$ is matrix of x_j's,
- The anticipated mean squared (AMSE) of HT-estimator $\hat{Y}$ under the linear model and sampling distribution $p(s)$ is given by

$$AMSE(\hat{Y}) = E_p \left[\left(\sum_{i \in s} \frac{\mathbf{x}_i}{\pi_i} - \sum_{i \in U} \mathbf{x}_i \right)^T \boldsymbol{\beta} \right]^2 + \sum_{i \in U} \left(\frac{1}{\pi_i} - 1 \right) \sigma_i^2$$

where E_p denotes expectation under the sampling distribution $p(s)$.

Cube method [Deville and Tillé, 2004]

- Cube method aims to select balanced sample with equal or unequal inclusion probabilities, it has two phases:
- Flight-phase transform the π_i's into sample membership indicator variable $I_i = \{0, 1\}$ through a random process, such that balancing equations are satisfied with fixed inclusion probabilities
- It does not always give a sample, that is, some π_i's are not integers $\{0, 1\}$ at the end of flight-phase,
- Landing-phase compromises the balancing equations in order to get the sample with fixed inclusion probabilities,
- Realized imbalance for a sample selected by cube: $(\hat{X}_j - X_j)^2$
- Expected imbalance: $V_p(\hat{X}_j) = E_p(\hat{X}_j - X_j)^2$ is not explicitly controlled by cube method.

Realized cube sample space

- Let K' samples of size n selected from U by cube method,
- Let Ω_K denote realized cube sample space of $K(\leq K')$ distinct samples,
- Let $\boldsymbol{\lambda}_{K\times 1}$ denote the empirical distribution of the K samples in Ω_K,
- The empirical estimate of the expected imbalance with respect to X_j based on $\boldsymbol{\lambda}$ can be calculated as $\Delta_j(\boldsymbol{\lambda}) = \sum_{k=1}^{K} \lambda_k(\hat{X}_{kj} - X_j)^2$

Reducing expected imbalance

- Do something better than Cube?
- There are good and bad samples in Ω_K in terms of balancing
- Choose the best sample from Ω_K? [π_i's are not achieved]
- Re-sample over Ω_K?
- Re-sample using $\boldsymbol{\lambda}$ [is equal to cube]
- Re-sample using a different sampling distribution over Ω_K which is expected to reduce the imbalance
- How to get this sampling distribution?

Adjusting empirical distribution under cube method

- Let $\boldsymbol{\lambda}^*(\neq \boldsymbol{\lambda})$ be another sampling distribution over Ω_K such that $\pi_i(\boldsymbol{\lambda}^*) = \pi_i(\boldsymbol{\lambda})$ for all $i \in U$, inclusion probabilities are not changed by re-sampling
- The estimated contribution of imbalance to AMSE under cube method is $E_{\boldsymbol{\lambda}}\left[\left(\sum_{i\in s}\frac{\mathbf{x}_i}{\pi_i} - \sum_{i\in U}\mathbf{x}_i\right)^T\boldsymbol{\beta}\right]^2$,
- Minimize: $\sum_{j=1}^{J} w_j \sum_{k=1}^{K}\lambda_k^*(\hat{X}_{jk} - X_j)^2 = \sum_{j=1}^{J} w_j\Delta_j(\boldsymbol{\lambda}^*)$ subjected to $\pi_i(\boldsymbol{\lambda}^*) = \pi_i(\boldsymbol{\lambda})$,
- where w_j are weights used in case of unknown β_j's, At the moment we are using $w_j = 1$,
- Using Simulated annealing, we obtain the adjusted empirical distribution $\boldsymbol{\lambda}^*$

Theoretical properties

- $\Delta_j(\boldsymbol{\lambda}^*) = \sum_{k=1}^{K} \lambda_k^*(\hat{X}_{jk} - X_j)^2 = MSE_{\boldsymbol{\lambda}^*}(\hat{X}_j|\Omega_K, \boldsymbol{\lambda})$
  $= V_{\boldsymbol{\lambda}^*}(\hat{X}_j|\Omega_K, \boldsymbol{\lambda}) + E^2_{\boldsymbol{\lambda}^*}(\hat{X}_j - X_j|\Omega_K, \boldsymbol{\lambda})$
- We propose to minimize the combined conditional imbalance
  $\sum_{j=1}^{J} w_j \Delta_j(\boldsymbol{\lambda}^*) \leq \sum_{j=1}^{J} w_j \Delta_j(\boldsymbol{\lambda}^*)$, β_j's are unknown,
- Therefore, 2s-cube improves the combined unconditional MSE of $\hat{X}_j$,
  i.e. $\sum_{j=1}^{J} w_j MSE_{2s-cube}(\hat{X}_j) \leq \sum_{j=1}^{J} w_j MSE_{cube}(\hat{X}_j)$
- Since $\pi_i(\boldsymbol{\lambda}^*) = \pi_i(\boldsymbol{\lambda})$ for all Ω_K and $\boldsymbol{\lambda}$,
  $E_{\boldsymbol{\lambda}^*}(\hat{X}_j|\Omega_K, \boldsymbol{\lambda}) = E_{\boldsymbol{\lambda}}(\hat{X}_j|\Omega_K, \boldsymbol{\lambda})$,
  $E^2_{\boldsymbol{\lambda}^*}(\hat{X} - X_j|\Omega_K, \boldsymbol{\lambda}) = E^2_{\boldsymbol{\lambda}}(\hat{X} - X_j|\Omega_K, \boldsymbol{\lambda})$
  $\sum_{j=1}^{J} w_j V_{\boldsymbol{\lambda}^*}(\hat{X}_j|\Omega_K, \boldsymbol{\lambda}) \leq \sum_{j=1}^{J} w_j V_{\boldsymbol{\lambda}}(\hat{X}_j|\Omega_K, \boldsymbol{\lambda})$
- Ω_K and $\boldsymbol{\lambda}$ based on K', $V[V_{\boldsymbol{\lambda}^*}(\hat{X}_j|\Omega_K, \boldsymbol{\lambda})]$ is smaller as $K' \to \infty$

Simulated data

$y = 0.1 + 2.44x_1 + 2.03x_2 + \epsilon$,

where $\sigma \propto x_1$, $x_1 \sim Gamma(4, 3)$, $x_2 \sim Normal(2, 1)$,

Generate population $N = 200$, Select $K' = 1000$ sample, $\pi_i \propto x_1$ by Cube, obtain $\boldsymbol{\lambda}^*$, calculate $\mathbb{M}_1$= 1st term and $\mathbb{M}_2$= 2nd term of AMSE and Δ_j, Repeat 50 times, calculate averages $\bar{\mathbb{M}}_1$, $\bar{\mathbb{M}}_2$ and $\bar{\Delta}_j$.

f	0.05		0.10		0.15		0.20	
	Cube	2s-Cube	Cube	2s-Cube	Cube	2s-Cube	Cube	2s-Cube
$\bar{\mathbb{M}}_1$	10668.86	3236.74	3204.05	978.55	1528.18	495.27	909.35	319.47
$\bar{\mathbb{M}}_2$	1313.65	1313.63	613.62	613.64	380.28	380.27	263.56	263.54
$\bar{\Delta}_0$	295.46	141.76	93.72	42.65	46.06	21.11	28.41	13.28
$\bar{\Delta}_1$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\bar{\Delta}_2$	2433.74	742.59	730.39	224.48	348.17	113.51	207.01	73.14

Real data [Särndal et al., 2003, p660-1]

- Swedish municipalities data `MU284`,
- Modified Clustered data used by [Deville and Tillé, 2004],
- Fit linear model and use regression estimates to calculate $\mathbb{M}_1$ and $\mathbb{M}_2$ where $\sigma \propto SIZE$,
- $N = 50$, Select $K' = 1000$, by probability proportional to size (PPS) sampling and Cube method, where $\pi_i \propto SIZE$,
- Obtain $\boldsymbol{\lambda}^*$
- Calculate Δ_j, $\mathbb{M}_1$ and $\mathbb{M}_2$ for Cube and 2s-Cube relative to PPS

f	0.05		0.1		0.2		0.4	
	Cube	2s-Cube	Cube	2s-Cube	Cube	2s-Cube	Cube	2s-Cube
$\mathbb{M}_1$	0.6459	0.2857	0.4266	0.2899	0.4062	0.2808	0.3211	0.2294
$\mathbb{M}_2$	0.9994	0.9994	0.9991	0.9991	1.0008	1.0008	0.9995	0.9995
Δ_0	0.5081	0.6745	0.1891	0.2054	0.1034	0.0961	0.0806	0.0815
$\Delta_{SIZE;X_1}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
$\Delta_{SS82;X_2}$	0.5819	0.5333	0.2611	0.2544	0.1642	0.1550	0.1147	0.1087
$\Delta_{CS82;X_3}$	0.6138	0.6513	0.2634	0.2690	0.1637	0.1602	0.1066	0.1018
Δ_{P75}	0.6362	0.2856	0.4099	0.2908	0.3886	0.2816	0.3264	0.2434
Δ_{RMT85}	0.6424	0.2857	0.4314	0.2924	0.4034	0.2762	0.3217	0.2299
Δ_{REV84}	0.6470	0.2397	0.4788	0.2834	0.4723	0.2785	0.4005	0.2627
Δ_{P85}	0.6282	0.2710	0.4156	0.2799	0.3888	0.2647	0.3154	0.2230
Δ_{ME82}	0.6555	0.2955	0.4390	0.2977	0.4211	0.2900	0.3273	0.2335
$\Delta_{S82;Y}$	0.6696	0.4452	0.3835	0.3500	0.3158	0.2865	0.2816	0.2622

Conclusion and future work

- Proposed method performs same as cube in terms of fixed inclusion probabilities,
- Proposed method reduces expected imbalance, given that the value of K' large enough,
- Simulated annealing may not be the best numerical solution for this problem, when population size is very large; For large population stratified or multistage sampling are used to divide population in small groups
- Extension for the population with correlated units?

References

Deville, J.-C. and Tillé, Y. (2004).

Efficient balanced sampling: the cube method.

Biometrika, 91(4):893–912.

Horvitz, D. G. and Thompson, D. J. (1952).

A generalization of sampling without replacement from a finite universe.

Journal of the American statistical Association, 47(260):663–685.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003).

Model assisted survey sampling.

Springer Science & Business Media.

Research Funding

This research is supported by ESRC through SC.DTP. in University of Southampton.

Questions?