Variance reduction using a non-informative sampling design

Thomas Zimmermann
Destatis – Division of Mathematical-Statistical Methods & Research Data Centre

IAOS Conference 2016, Abu Dhabi, December 6th – 8th

Status Quo in Official Statistics

- Conduct sample surveys by means of probability sampling
- Use design-based methods to produce estimates
- Procedure yields unbiased estimates by construction
- Very useful approach provided sample sizes are large
- Limitations: budget constraints, estimates for subgroups

Challenges for Official Statistics

- Data users struggle with survey weights
  - Many researchers simply run unweighted analysis
  - May lead to erroneous results due to informative sampling
- Official statistics started to move towards model-based methods, where incorporation of weights can be challenging
  - Small area estimation
  - Imputation

Aim of this contribution

Find a sampling method, which

1. enables precise design-based estimates for aggregates

2. is suitable for the application of model-based procedures

3. and is easy to communicate.

Notation

- Population U = {1, ..., k, ..., N}
- Sample $S \subset U$ of size n
- Variable of interest y with values $(y_1, \dots, y_k, \dots, y_N)$, known for sampled elements only
- Size variable z, with values $(z_1, \dots, z_k, \dots, z_N)$ that are known for all units in the population
- D mutually exclusive domains (or areas) $U_d \subset U, d = 1, \dots, D$ with domain sizes N_d
- Estimate population mean: $\mu = \frac{1}{N} \sum_{k \in S} y_k$ or area mean:

  $\mu_d = \frac{1}{N_d} \sum_{k \in S_d} y_k$

Antithetic Clustering (ATC)

1. Order the elements according to the size variable.
2. Assign largest and smallest unit to the first cluster.
3. Now assign second largest and second smallest unit to the next cluster.
4. Repeat procedure until all units are assigned to a cluster. This yields $L = [N/2]$clusters.
5. Draw $l > 1$ out of L clusters by means of a simple random sample.

Note: All units have the same inclusion probability, hence no bias due to informative sampling.

When does ATC work?

- Approach is based on cluster sampling
- Cluster sampling reduces the variance, if units from the same cluster are less similar than units from different clusters
- This is exactly what our method does for the size variable
- Further prerequisite: the size variable should be correlated with the variable of interest

Simulation study

- Comparison of ATC and SRS for mean estimates using design-based estimators (Direct and GREG estimator) and model-based small area estimator (BHF estimator)
- Population with $N = 12000$ units and $D = 30$ areas
- Sample size $n = 500$ in each of the 10000 replications
- Apply ATC and SRS on population level → random sample sizes in domains
- Population constructed as:

$$y_k = 6 + 3 \cdot z_k + v_d + \varepsilon_k, \quad k \in U_d$$

$$v_d \sim N(0,2), \varepsilon_k \sim N(0,4), \mathrm{z_k} \sim N(1,1)$$

Quality measures

- RB – relative bias
- AARB – average absolute relative bias
- RRMSE – relative root mean squared error
- ARRMSE – average relative root mean squared error
- ACR – average 95% confidence interval coverage rate

Results for domain estimates

		Direct		GREG		BHF	
	$E(n_d)$	AARB	ARRMSE	AARB	ARRMSE	AARB	ARRMSE
ATC	‹ 10	0.004	0.463	0.018	0.168	0.051	0.092
	10-30	0.002	0.259	0.000	0.059	0.013	0.050
	›30	0.001	0.161	0.001	0.035	0.004	0.031
SRS	‹ 10	0.005	0.465	0.019	0.169	0.051	0.092
	10-30	0.002	0.259	0.000	0.059	0.012	0.049
	›30	0.002	0.161	0.001	0.035	0.004	0.031

Results for national estimates

	Direct			GREG		
	RB	RRMSE	ACR	RB	RRMSE	ACR
ATC	-0.0001	0.0105	0.9493	-0.0001	0.0105	0.9490
SRS	-0.0002	0.0173	0.9529	-0.0001	0.0106	0.9473

Conclusion

- ATC yields variance reductions compared to SRS and does not interfere with models
- It permits an unbiased variance estimation
- It is easy to implement and does not require more information than other approaches to variance reduction
- Could be extended to account for multiple size variables

Thank you very much for your attention!