Balanced Sampling: Comparisons between INCA and Cube Method

Yang Cheng[1], Lu Chen[1,2], Luca Sartore[1,2], and Valbona Bejleri[1]

1. National Agricultural Statistics Service
2. National Institute of Statistical Sciences

Thursday, October 24, 2024, 8:30 AM - 10:00 AM

2024 FCSM Research and Policy Conference

Chesapeake A, University of Maryland, College Park, MD.

Outline

1. Background
2. Balanced sampling in general
3. Integer-calibration (INCA) method
4. Case study
5. Conclusion

Background

1. Many National Agricultural Statistics Service (NASS) surveys employ the Multivariate Probability Proportional to Size (MPPS) sample design for multipurpose surveys because MPPS allows for the use of multiple measures of size.
2. NASS faces challenges in the use of MPPS sampling due to lack of control over the sample size, which puts NASS at risk of workload and losing efficiency in estimators.
3. Current NASS MPPS sample design consists of two steps:
   - Construct MPPS inclusion probabilities that compromise among Measure of Size (MOS).
   - Apply Poisson sampling method by using those MPPS inclusion probabilities.

Notation

- Let U be the population with K $(K > 1)$ study variables.
- Define $y_i = (y_{1,i}, \cdots, y_{K,i})$, $i \in U$ as study survey variable where $y_{k,i}$ is the information on kth $(k = 1, \cdots, K)$ study variable of ith unit.
- Let U_k be the population with kth study variable, and $U_k \subset U$.
- N is the size of U and N_k is the size of U_k.
- $Y_k = \sum_{i \in U_k} y_{k,i}$ is the total of kth study variable. It is unknow parameter.
- S is a sample drawn from U and n is the size of S.
- $S_k = S \cap U_k$ and $S_k \subset S$.
- $n_k \leq n$ denotes the size of S_k.
- πps means probability proportional to size sampling without replacement.

MPPS inclusion probability

The details for deriving overall inclusion probability π_i, $i \in U$, are as follows:

1. Construct *K* frames U_k from *U*.
2. Identify MOS variable $x_{k,i}$ for each frame *k*, *k* = 1, …, *K*.
3. Determine the target sample size n_k^t to meet survey precision requirement. n_k^t is a function of survey precision, population size, and auxiliary data, where *t* stands for "target".
4. Calculate inclusion probability on U_k based on πps setting, $\pi_{k,i} = n_k^t \frac{x_{k,i}^p}{\sum_{i \in U_k} x_{k,i}^p}$, where $0 < p \leq 1$. If $\pi_{k,i} \geq 1$, then $\pi_{k,i} = 1$.
5. MPPS inclusion probability is

$$\pi_i = \max_{1 \leq k \leq K} \pi_{k,i}.$$

Poisson (PO) and MPPS sampling

PO sampling: In survey methodology, PO sampling includes each unit of the population based on the outcome of an independent Bernoulli trial:

1. Each unit in frame generates a random number $\varepsilon_i \sim U(0, 1)$ for every population unit $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$.
2. Unit i is sampled if $\varepsilon_i \leq \pi_i$, where $0 < \pi_i < 1$ is the desired inclusion probability for each unit. π_i is predetermined and may vary with index i.

MPPS sampling: apply MPPS inclusion probability in the PO sampling.

MPPS sampling

- Sample indicator function

$$I_i = \begin{cases} 1 & if\ \varepsilon_i \leq \pi_i \\ 0 & otherwise \end{cases}$$

for $i = 1, \ldots, N$, is an independent Bernoulli(π_i).

- The size of MPPS sampling is $n_s = \sum_{i \in U} I_{[\varepsilon_i \leq \pi_i]}$. n_s is random.

Motivation for this small talk is to address the **random sample size issue.**

Balanced sampling

- In general, match sample moments of auxiliaries to population moments.
- One of the most controlled sampling methods with respect to the set of inclusion probabilities.
- If choosing the first moment, balanced sampling is similar to a calibration in the sampling design.

Balanced sampling (cont.)

A random sample must satisfy the following balancing equations:

$$\sum_{i \in S} \frac{x_i}{\pi_i} = \sum_{i \in U} x_i$$

In other words, in a balanced sample, the total of the x-variables are estimated without error.

For MPPS setting, $x_i = \left(x_{1,i}, \cdots, x_{K,i}, \pi_{MPPS,i}\right)$ satisfies balanced equation:

$$\sum_{i \in S} \frac{x_i}{\pi_{MPPS,i}} = \sum_{i \in U} x_i.$$

Cube method (Deville & Tille, 2004)

- The cube method gives a sample that is nearly balanced but respects exactly the inclusion probabilities.
- **The flight phase**: A random walk begins at the vector of inclusion probabilities and remains in the intersection of the cube and the constraint subspace.

  This random walk stops at a vertex of the intersection of the cube and the constraint subspace.
- **The landing phase:** At the end of the flight phase, if a sample is not obtained, a sample is selected as close as possible to the constraint subspace.

Integer-calibration (INCA)

- NASS developed INCA method to produce integer calibrated weights (Sartore et al., 2019).
- Apply the discrete coordinate descent algorithm to optimize objective functions on a constrained lattice.
- Presented for the first time at the US Census Bureau during FedCASIC (https://www.census.gov/fedcasic/fc2016/ppt/1_5_Integer.pdf).

Consists of two phases like the cube method:

1. **Rounding** produces a vector of integer numbers.
2. **Calibration** adjusts integers to satisfies benchmarks.

Integer-calibration (INCA) (cont.)

- For sampling, the selection probabilities are used to initialize integer weights with constraints in $\{0, 1\}^N$.
- Benchmarks are provided for expected sample size, and balancing equations.
- Selection probabilities are transformed in binary values during the first INCA phase.
- The second phase adjusts the binary values to improve the approximation of balancing equations.

Case study

- State of Minnesota 2017 Census of Agriculture (COA) record-level data.
- Top 13 commodities by acreages.
- N = 23,528, all farms contain land in field crops.
- U_k is the frame on commodities k = 1, 2, 3, 4, which are soybean, corn, sunflower, and barley.
- Auxiliary variables ($x_{k,i}$) are the acreages; study variables ($y_{k,i}$) are the productions.
- Software: R packages:
  - Cube Method: library(sampling).
  - INCA Method: library(inca).

	Harvested Acres	Number of Farms
Soybean	5,567,246	17,924
Corn	5,049,186	17,698
Spring Wheat	948,038	2,753
Potato	380,907	356
Sugarbeet	361,916	856
Dry Bean	106,557	338
Barley	58,742	417
Oat	56,370	1,568
Sunflower	30,786	128
Winter Wheat	3,925	108
Durum Wheat	1,162	12
Sorghum	252	6
Sweet Potato	187	43

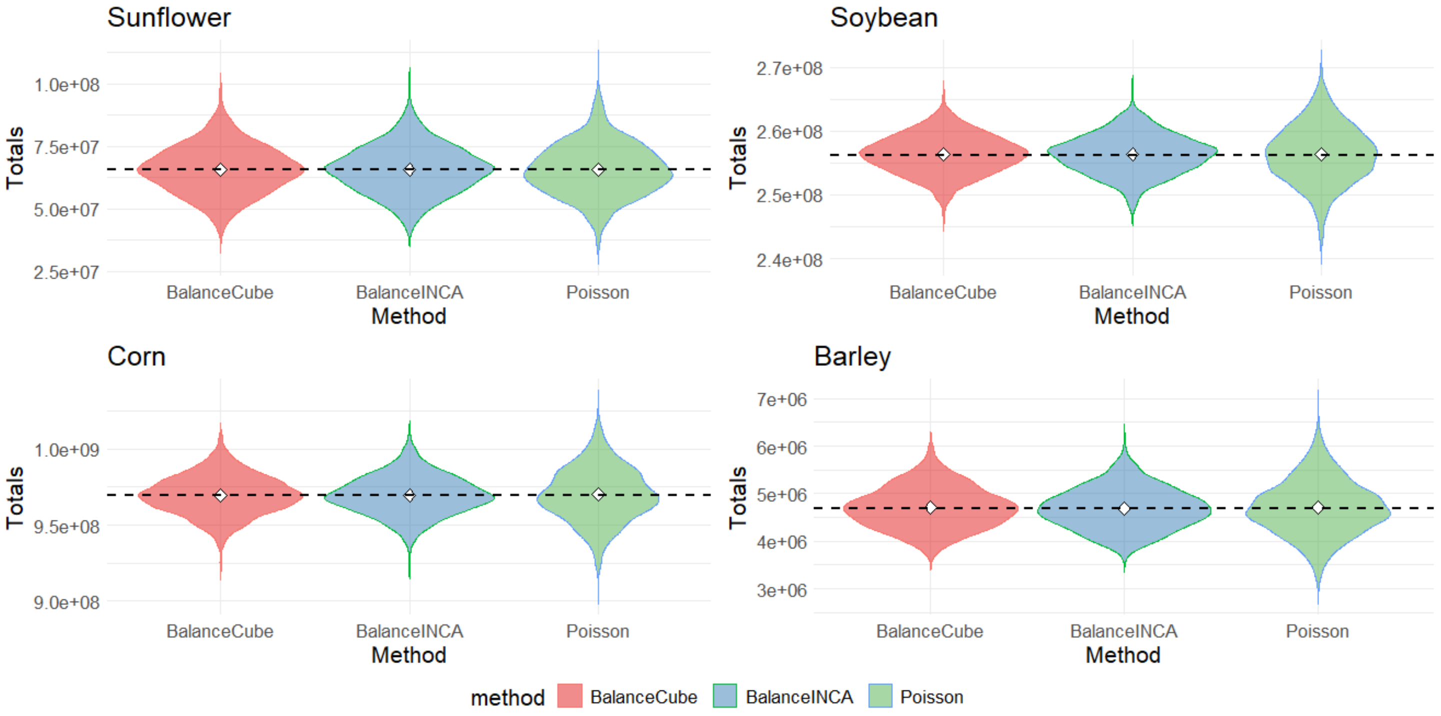
Case study – Simulation

- R = 1,000 simulations.
- Calculate the MPPS inclusion probabilities for each crop.
- Apply three different sampling methods:
  1. PO Sampling (current).
  2. Balance Sampling – Cube method.
  3. Balance Sampling – INCA method.
- Compute and compare for 3 sampling methods:
  - Percentage relative bias: compute the Monte Carlo expectations.
  - Relative efficiency: compute the Monte Carlo variance across the 1,000 replications of the point estimates (PE).

Case study – Percentage Relative Bias

$$100 * \frac{Average\ of\ PE_k - True\ Total}{True\ Total}$$, K= Cube, INCA, PO

Study Variables	Cube	INCA	PO
Sunflower	-0.11	**-0.09**	-0.16
Barley	0.24	-0.28	**0.18**
Soybean	0.07	**0.07**	0.10
Corn	0.01	**-0.01**	0.05



Case study – Relative Efficiency

- The relative efficiency (RE) = $\frac{Var_k(\hat{T}_{Balanced})}{Var(\hat{T}_{PO})}, k = Cube, INCA.$
- Values < 1 means MPPS Balance sampling with both Cube and INCA methods are better.

Study Variables	**RE Cube**	**RE INCA**
Sunflower	0.59	**0.56**
Barley	0.92	**0.92**
Soybean	**0.97**	0.97
Corn	0.96	**0.96**

Conclusion

- Cube method and INCA perform better than Poisson sampling when applying for MPPS inclusion probabilities.
- INCA method performs slightly better than cube method.
- Future research will focus on
  - Hypothesis test on the relative bias and efficiency among three sampling methods.
  - Apply INCA and cube methods on other crops.

References

- Bailey, J.T. and Kott, P.S. (1997). "An application of multiple list frame sampling for multi-purpose surveys". ASA Proceedings of the Section on Survey Research Methods.
- Brewer, K.R.W. (1963). "A Model of Systematic Sampling with Unequal Probabilities". Australian Journal of Statistics, Volume 5, Issue 1: 5-13.
- Brewer, K.R.W. (1999) "Design-based or prediction-based inference? Stratified random vs stratified balanced sampling". International Statistical Review, 67(1): 35–47.
- Cheng, Y., Smith, L., Bailey, J., and Young, L.J. (2023). "On Farm Grain Stocks Sample Methodology Review". United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) RDD Research Report Number RDD-23-01. https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/QAS%20sample%20methodology%20review%20final%20report.pdf
- Deville, J.C. and Tille, Y. (2004). Efficient balanced sampling: the cube method. Biometrika 91, 893–912.

References (cond.)

- Deville, J.C. and Tille, Y. (2005). “Variance approximation under balanced sampling”. Journal of Statistical Planning and Inference, 128(2): 569–591.
- Field Offices and Headquarters Units (2018). “Policy and Standards Memoranda: No. PSM-ASMS-12”. National Agricultural Statistics Service (http://nassportal/NASSdocs/Documents/PSM-ASMS-12.pdf).
- Kott, P.S., Amrhein, J.F., and Hicks, S.D. (1998). “Sampling and estimation from multiple list frames”. Survey Methodology, 24(1).
- Kott, P.S. and Bailey, J.T. (2000). “The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling”. ASA Proceedings of the Section on Survey Research Methods.
- Sartore, L., Toppin, K., Young, L.J., and Spiegelman, C. (2019). “Developing Integer Calibration Weights for Census of Agriculture”. Journal of Agricultural, Biological, and Environmental Statistics, Vol. 24, No. 1, pp. 26-48.

Thank you!

Contact information:

yang.cheng@usda.gov