# Inclusion Probability Proportional to Size Sampling: A Nonlinear Programming Approach to Ensure a Nonnegative and Stable Variance Estimator

**Sun-Woong Kim**
**Steven G. Heeringa**
**Peter W. Solenberger**

**Dongguk University**
**&**
**Survey Research Center**
**University of Michigan**

# Overview

- Inclusion Probability Proportional to Size (IPPS) Sampling
- Horvitz-Thompson Estimator and Variance Estimator
- Some Desirable Properties: Non-negativity and Stability of Variance Estimator
- Some Methods by Earlier Workers
- Nonlinear Programming (NLP) Approaches
- Implementation of NLP Approaches
- An Example
- Discussion

# Inclusion Probability Proportional to Size (IPPS) Sampling

## IPPS sampling design

$$D = (S, P),$$

where $S$: a collection of possible samples
$P$: a positive function on $S$

with the following properties:

$$\sum_{s \in S} P(s) = 1$$

$$\sum_{s \in S,\ i \in s} P(s) = np_i\,,\ i = 1, \cdots, N$$

where $s$ is a sample, $P(s)$ is the selection probability of each sample and $p_i$ is the relative size of each unit

## Inclusion probabilities

✓ The first-order inclusion probability $\pi_i$

: Probability that the $i$th unit is in the sample $s$

That is,

$$\sum_{s \in S,\ i \in s} P(s) = np_i = \pi_i$$

✓ The second-order inclusion probability $\pi_{ij}$

: Probability that the $i$th and $j$th units are both in the sample $s$

That is,

$$\sum_{s \in S,\ i, j \in s} P(s) = \pi_{ij}$$

## Horvitz-Thompson Estimator and Variance Estimator

- **H-T (1952) estimator of the population total $Y$:**

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

where $y_i$ is the value of characteristic of the unit $i$

- **Variance estimators:**

✓ Sen, Yates and Grundy (1953)'s form :

$$\mathrm{v}_1(\hat{Y}_{HT}) = \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

✓ A different form :

$$\mathrm{v}_2(\hat{Y}_{HT}) = \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2 \sum_{i=1}^{n} \sum_{j>i}^{n} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j \pi_{ij}} y_i y_j$$

## Some Desirable Properties: Non-negativity and Stability of Variance Estimator

✓ According to a sampling design, the terms

$$\left(\pi_i \pi_j - \pi_{ij}\right)$$

often vary widely and sometimes cause the variance estimator to be negative and unstable

✓ Achieving non-negativity and stability of Sen-Yates-Grundy's variance estimator may be essential in creating a sampling design

✓ In addition, the second-order inclusion probability $\pi_{ij}$ must be larger than zero with respect to unbiased variance estimation

# Some Methods by Earlier Workers

## Jessen (1969)

Four methods of selecting probability non-replacement samples described and examined their properties

Some of them partially achieve those desired properties and would not be appropriate to use in practice

## Nigam et al. (1984)

Suggested a IPPS sampling scheme using binary block designs that are sort of experimental designs

For the cases where $n = 2$ as well as $n > 2$, it involves considerable trial and error to find a sampling design, although it achieves the desired properties, that is,

$$0.4\pi_i\pi_j < \pi_{ij} \le \pi_i\pi_j$$

# Nonlinear Programming (NLP) Approaches

## Approach I

First, in order to minimize Sen-Yates-Grundy variance estimator, construct the following nonlinear objective function:

$$\textbf{Min} \sum_{j}^{N} \sum_{j>i}^{N} \frac{\pi_i \pi_j}{\pi_{ij}},$$

which is equivalent to

$$\textbf{Min} \sum_{j}^{N} \sum_{j>i}^{N} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}$$

Second, add the bounded constraints

$$c\pi_i \pi_j < \pi_{ij} \le \pi_i \pi_j, \; 0 < c < 1, \; j > i = 1, 2, \cdots, N$$

and IPPS constraints

$$\sum_{j \ne i}^{N} \pi_{ij} = (n-1)\pi_i, \; i = 1, 2, \cdots, N$$

Third, run NLP after deciding $c$, which indicates the level of stability of variance estimator, and find a set of $\pi_{ij}$ that is a solution to the NLP problem

## Approaches II and III

Orginally developed by Kim, Heeringa and Solenberger (2003) for minimizing variance

Finding a set of $\pi_{ij}$ to optimize

$$\textbf{Min}\sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j-\pi_{ij})^2$$

**or**

$$\textbf{Max}\sum_{i=1}^{N}\sum_{j>i}^{N}\pi_{ij}$$

subject to the same constraints with Approach I

## Remark)

**The constraint $c\pi_i\pi_j < \pi_{ij} \leq \pi_i\pi_j$ is differently expressed as**

$$c < \delta_{ij} \leq 1,$$

**where** $\delta_{ij} = \dfrac{\pi_{ij}}{\pi_i\pi_j}$

**Then obtaining a sampling design providing MINMAX $\delta_{ij}$ among possible solutions by NLP may be preferable since it would remain more stable for the variance estimator**

# Implementation of NLP Approaches

- ✓ **SAS/OR NLP Procedure is available to optimize those non-linear objective functions under the certain constraints**

- ✓ **Repeating the steps in Approach I by some reasonable rules yields a set of $\pi_{ij}$ achieving MINMAX $\delta_{ij}$**

- ✓ **The value of $c$ less than 0.5 may be favorable because NLP is unlikely to be feasible for the higher values**

# An Example

## Table 1. Yates and Grundy (1953)

**Three populations with $N=4$, $n=2$**

| Unit | $i$: | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Relative Sizes | $p_i$: | 0.1 | 0.2 | 0.3 | 0.4 |
| Population A | $y_i$: | 0.5 | 1.2 | 2.1 | 3.2 |
| | $y_i/p_i$ | 5 | 6 | 7 | 8 |
| Population B | $y_i$: | 0.8 | 1.4 | 1.8 | 2.0 |
| | $y_i/p_i$ | 8 | 7 | 6 | 5 |
| Population C | $y_i$: | 0.2 | 0.6 | 0.9 | 0.8 |
| | $y_i/p_i$ | 2 | 3 | 3 | 2 |

**Table 2. The second-order inclusion probabilities obtained by different sampling schemes**

| Units $(i, j)$ | $\pi_{ij}$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | Method | | | | | | |
| | J2 | J3 | J4 | N | A1 | A2 | A3 |
| 1, 2 | 0.200 | 0.066 | 0.010 | 0.040 | 0.037 | 0.036 | 0.036 |
| 1, 3 | 0.000 | 0.067 | 0.050 | 0.060 | 0.055 | 0.055 | 0.055 |
| 1, 4 | 0.000 | 0.067 | 0.140 | 0.100 | 0.109 | 0.109 | 0.109 |
| 2, 3 | 0.000 | 0.066 | 0.140 | 0.100 | 0.109 | 0.109 | 0.109 |
| 2, 4 | 0.200 | 0.267 | 0.250 | 0.260 | 0.255 | 0.255 | 0.255 |
| 3, 4 | 0.600 | 0.467 | 0.410 | 0.440 | 0.437 | 0.436 | 0.436 |

**J2: Jessen's method 2**
**J3: Jessen's method 3**
**J4: Jessen's method 4**
**N: Nigam et al.'s method**
**A1: Approach I**
**A2: Approach II**
**A3: Approach III**

**Table 3. $\delta_{ij}$ obtained by different sampling schemes**

| Units $(i, j)$ | $\delta_{ij}$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | Method | | | | | | |
| | J2 | J3 | J4 | N | A1 | A2 | A3 |
| 1, 2 | 2.500 | 0.825 | 0.125 | 0.500 | 0.456 | 0.454 | 0.454 |
| 1, 3 | 0.000 | 0.558 | 0.417 | 0.500 | 0.454 | 0.454 | 0.454 |
| 1, 4 | 0.000 | 0.419 | 0.875 | 0.625 | 0.681 | 0.683 | 0.683 |
| 2, 3 | 0.000 | 0.275 | 0.583 | 0.417 | 0.454 | 0.455 | 0.455 |
| 2, 4 | 0.625 | 0.834 | 0.781 | 0.813 | 0.795 | 0.795 | 0.795 |
| 3, 4 | 1.250 | 0.973 | 0.854 | 0.917 | 0.909 | 0.909 | 0.909 |

**Note. The values in the thick borders indicate MIN $\delta_{ij}$, while those in A1, A2, A3 present MINMAX $\delta_{ij}$**

**J2: Jessen's method 2**
**J3: Jessen's method 3**
**J4: Jessen's method 4**
**N: Nigam et al.'s method**
**A1: Approach I**
**A2: Approach II**
**A3: Approach III**

## Table 4. Comparison of stabilities of variance estimators

| Pop | $CV\left(\hat{V}\left(\hat{Y}\right)\right)$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **PPS** | **J2** | **J3** | **J4** | **N** | **A1** | **A2** | **A3** |
| **A** | **1.600** | **NA** | **2.121** | **1.341** | **1.349** | **1.244** | **1.242** | **1.242** |
| **B** | **1.600** | **NA** | **2.121** | **1.341** | **1.349** | **1.244** | **1.242** | **1.242** |
| **C** | **1.000** | **NA** | **1.467** | **2.627** | **1.392** | **1.481** | **1.483** | **1.483** |
| **Average** | **1.400** | **NA** | **1.903** | **1.769** | **1.363** | **1.323** | **1.322** | **1.322** |
| **Relative Stability** | **100** | **NA** | **74** | **79** | **103** | **106** | **106** | **106** |

**Note. 'NA' indicates 'Not Available' due to some zero $\pi_{ij}$**

**PPS: Probability proportional to size sampling with replacement**
**J2: Jessen's method 2**
**J3: Jessen's method 3**
**J4: Jessen's method 4**
**N: Nigam et al.'s method**
**A1: Approach I**
**A2: Approach II**
**A3: Approach III**

## Discussion

- ✓ Surely achieved the desired properties such as non-negativity and stability in variance estimation by using NLP approaches

- ✓ Very flexible for $n = 2$ as well as $n > 2$

- ✓ NLP approaches easy to carry out by using some publicly available software

- ✓ But there may be a tradeoff between variance and variance estimator since the objective function is an increasing function of 'c' However, the variance for a sampling design having MINMAX $\delta_{ij}$ may be lower than in probability proportional to size(PPS) sampling

- ✓ Helpful to do some studies for a lot of populations
- ✓ Developing a software application formed from the suggested NLP approaches and SAS/OR NLP Procedure and checking the efficiencies of some NLP algorithms provided in SAS/OR highly recommended

In our written paper, more will be coming out!