A Modified Two-Stage Sampling Scheme with Integrated Second Stage Sample

Chang-Tai Chao

Department of Statistics
National Cheng Kung University

November 28, 2019

Outline

Outline

Agricultural Council Projects

- Statistic Office/統計室
  - Annual Technology Projects since 2013 (Collaborate with Prof. Ma)
  - Primary Farm Household Income Survey/主力農家所得調查
  - Agricultural Household Survey/農家戶口調查
  - Annual number of peasant households
- Animal Protection Section/動物保護科
  - Regular biennial projects
    - 全國家犬貓數量調查
    - 全國遊蕩犬數量調查
  - Other projects
    - 寵物飼養態度調查
    - 遊蕩犬分布地圖

Census v.s. Sampling Survey

- Census
  - All population units will be observed
  - Enormous resource is required.
  - It is time-consuming to process the data.
  - Result is often not necessarily to be true.
- Sampling Survey
  - Usually provides better population information than census at lower cost.
  - Can collect/process the data more quickly, so result can come out more quickly as well.
  - Estimates based on sampling are often more accurate than what based on census.
  - Might provide better results.
  - Error can be described in a probability format.

$\longrightarrow$

Census of Agriculture, Forestry, Fishery and Animal Husbandry

- Every five years
- Provide information about the size of operator characteristics, production practices and gross income of the agriculture sector.
- Evaluate and revise the current policy accordingly.
- Provide the sampling population/population structure for various agricultural sampling survey

Expensive, time-consuming and tedious

Primary Farm Household Income Survey

- Annual sampling survey (Statistic Office Agriculture Council)
- Collect and process the data in a timely manner with much lower cost

Primary Farm Household Income Survey

Based on 2010 Census data:

- Target population : Primary Farm Households
  - 50 million NTD $>$ annual agricultural gross income(初級農產品收入) $>$ 200,000 NTD
  - At least one household member under the age of 65 is currently engaged in the agriculture work
- Population quantity of interest : Average annual household income
- Population Size : 150,456
- Sample size : 1,000

Primary Farm Household Income Survey

- Estimation precision
- Estimation of subpopulation of interest
    - Production type
    - Production scale

Primary Farm Household Income Survey

Production Type

- Crop Farms
    - Rice
    - Vegetables
    - Fruits
    - Coarse Grain and Special Crops
    - Other Crop
- Livestock Farms
    - Hog Farms
    - Chicken Farms
    - Other Livestock Farms

Primary Farm Household Income Survey

Production Scale

- Crop Farms
  - Cultivated area
- Livestock Farms
  - Year-end Feeding number ??
  - Agricultural gross income

Primary Farm Household Income Survey

- Sampling Design: Stratified sampling
- Stratification: production type by scale
    - Type : Categorical variable
    - Scale : Continuous
        - Equal stratum size
        - **Optimal Boundary Algorithm**
- Allocation
    - Proportional Allocation
    - **Neyman Allocation**

Result

- Maximum relative estimation error under 95% confidence level
  - SRSWOR : 15.48%
  - Stratified design with equal stratum size : 5.12%
  - Proposed stratified design : 4.22%
- The proposed sampling design has been practiced since 2013.

Obstacles

- About 280 out of 319 townships would be intersected
- Most of the agriculture households are located at the rural or remote area.
- The dispersion of the sampled units cause certain difficulty.
- The travel cost may be intolerable under a limited budget.

Possible Solution

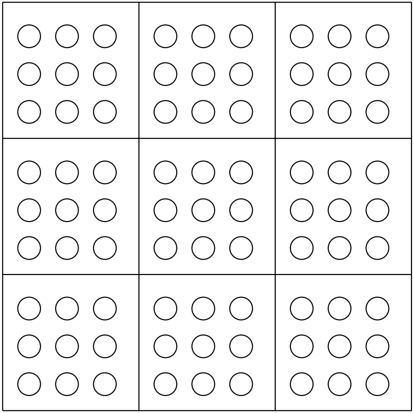
Cluster Sampling Design

- Pros: save sampling cost
- Cons:
  - less estimation precision
  - practically impossible to guarantee a proper sample/sample size of secondary units (farm households) to estimate the subpopulation of interest.

Outline

Modified Two-Stage Sampling

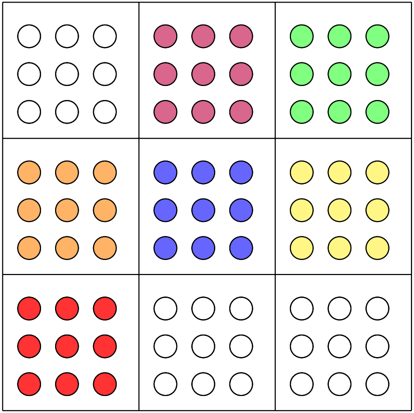
9 primary units, 81 secondary units



,

Modified Two-Stage Sampling

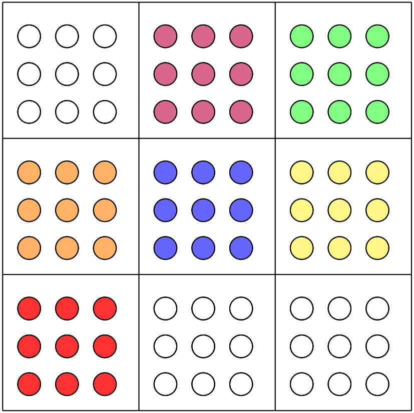
9 primary units, 81 secondary units

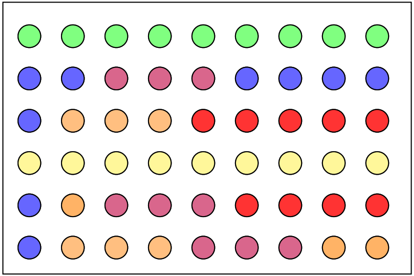


,

Modified Two-Stage Sampling

9 primary units, 81 secondary units





,

Outline

Sampling Design

- First-stage sampling design: simple random sampling without replacement.
- Second-stage sampling design: simple random sampling without replacement.
- Evaluate the feasibility of the proposed design.

Estimations

- Arithmetic average: $\hat{\mu}_{1.1} = \frac{1}{m}\sum_{i\in S_1}\sum_{j\in S_i} y_{ij}$
- Horvitz Thompson-ssu: $\hat{\mu}_{1.2} = \frac{1}{M}\frac{N}{n}\sum_{i\in S_1}\sum_{j\in S_2}\frac{y_{ij}}{\pi_{ij}}$
- Probability proportional to size: $\hat{\mu}_{1.3} = \frac{1}{M}\frac{N}{n}\frac{1}{m}\sum_{i\in S_1} m_i\frac{\hat{y}_i}{p_i}$
- Horvitz Thompson-psu: $\hat{\mu}_{1.4} = \frac{1}{M}\frac{N}{n}\sum_{i\in S_1'}\frac{\hat{y}_i}{\pi_i}$
- Ratio Type Estimator: $\hat{\mu}_{1.5} = \frac{\sum_{i\in S_1}\hat{y}_i}{\sum_{i\in S_1} M_i}$

$$\pi_{ij} = \frac{m}{K},\ p_i = \frac{M_i}{K}, \hat{y}_i = \frac{M_i}{m_i}\sum_{j\in S_i} y_{ij},\ \pi_i = 1 - \frac{\binom{K-M_i}{m}}{\binom{K}{m}}$$

Simulation data

- Population size of ssu: $M = 30,116$
- Population size of psu: $N = 300$
- PSU sample size: $n = 200$
- SSU sample size: $m = 300 - 7,000$

Comparable Classical Designs

- Simple random sampling : Cost/Number of PSU selected
- Classical two-stage sampling: Estimation precision.

Comparison of MSE for different estimators

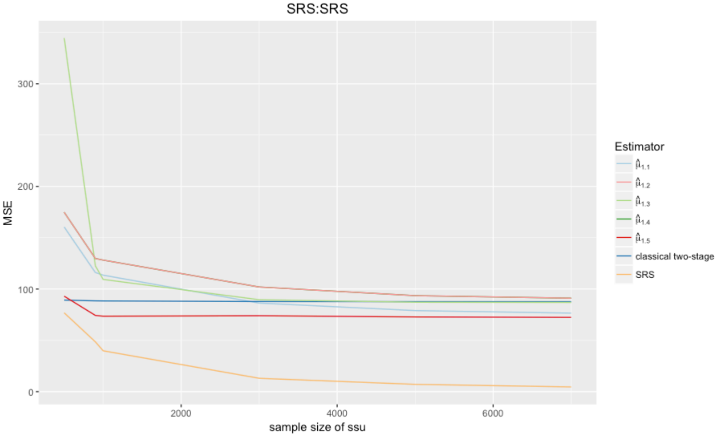


Figure: The mse of the proposed estimators (SRS/SRS) and the baselines with different ssu sample sizes.

Number of PSU's intersected

- 200 when SSU sample size $\cong$ 300
- 300 (the population size) when SSU sample size $\cong$ 1000

Outline

Sampling Design

- First-stage sampling design: simple random sampling without replacement
- Second-stage sampling design: stratified sampling (Neyman allocation)

Estimators

$\hat{\tau}'$ is the usual unbiased estimator under stratified random sampling design for τ', the total of the first-stage sample.

$$\hat{\tau}' = \sum_{i \in S_1} \sum_{h=1}^{H} \frac{M'_h}{m_h} \sum_{j \in S_{hi}} y_{hij}$$

- $$\hat{\mu}_{2.1} = \frac{1}{M} \frac{N}{n} \hat{\tau}'$$
- $$\hat{\mu}_{2.2} = \frac{1}{K} \hat{\tau}'$$

Simulation Results

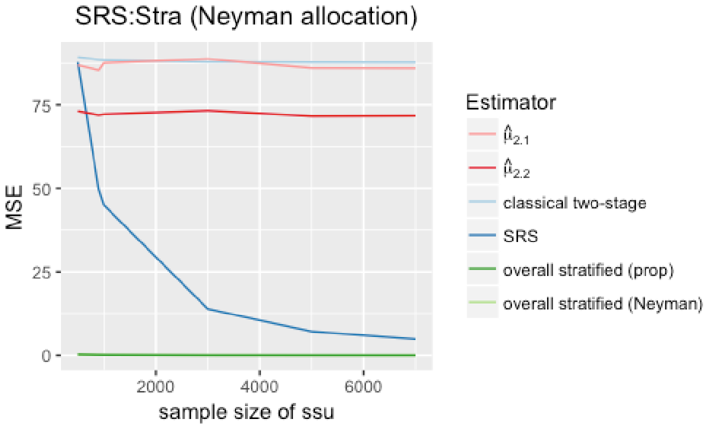


Figure: The mse of the proposed estimators (SRS/Stratified) and the baselines with different ssu sizes.

Outline

Motivation

- The variance of estimator is composed of the variability of:
  1. between PSU (first stage)
  2. within PSU (second stage)
- The between primary unit variance is usually of the greater portion comparing to the within PSU variance.

First-Stage Sampling Design

Two sampling designs are used to select a better first-stage sample to improve the estimation mean square error.

- Stratified sampling design
  - proportional allocation
- Systematic sampling design
  - Assigning equal inclusion probability to each primary unit.

The primary units are selected with equal inclusion probability, so that the unbiased estimator can be constructed with a relatively easier manner.

Second Stage Sampling Design

Stratified sampling design with Neyman allocation

$\hat{\tau}'$ is the usual unbiased estimator under stratified random sampling design for τ', the total of the first-stage sample.

$$\hat{\tau}' = \sum_{i \in S_1} \sum_{h=1}^{H} \frac{M'_h}{m_h} \sum_{j \in S_{hi}} y_{hij}$$

- $$\hat{\mu}_{k.1} = \frac{1}{M} \frac{N}{n} \hat{\tau}'$$
- $$\hat{\mu}_{k.2} = \frac{1}{K} \hat{\tau}'$$

$$k = \begin{cases} 3: & \text{stratified sampling in the first stage} \\ 4: & \text{systematic sampling in the first stage} \end{cases}$$

MSE v.s. m (ssu sample size)

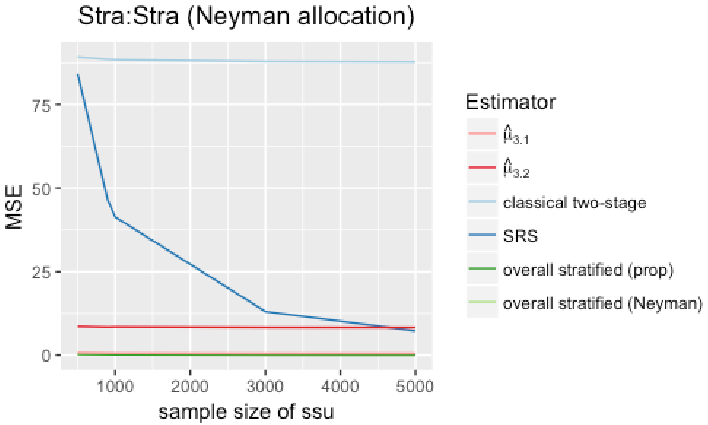


Figure: The mse of the proposed estimators (Stratified/Stratified) and the baselines for different ssu sizes.

MSE v.s. m (ssu sample size)

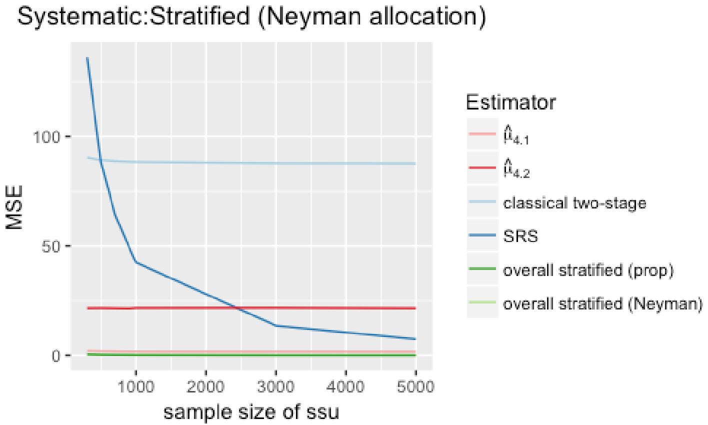


Figure: The mse of the proposed estimators (Systematic/Stratified) and the baselines for different ssu sizes.

Outline

2015 Taiwanese Agriculture Census Data

- Target population: primary farm households (200,000 NTD $<$ annual agriculture gross income $<$ 50 million NTD) and (least one household member under the age of 65 is currently engaged in the agriculture work.)
  - Population size of the township: 319
  - Population size of the farm household: 217,747
- Primary variable of interest: agriculture gross income

Designs

- Proposed Two-Stage design
  - Primary unit: Township
  - Secondary unit: Farm household
  - Primary unit sample size $n = 200$
  - Secondary unit sample size $m = 1600$
- MSE of comparable designs
  - Original stratified design: 343.7165
    - Number of townships intersected: 280
  - Classical Two-Stage design: 12297.24
  - Simple random sampling: 18519.09

Simulation results

O: 343.72, C: 12297.24, S: 18519.09

Table: Comparison of the MSE of the proposed estimators

Estimator	MSE
$\hat{\mu}_{1.5}$	6886.08
$\hat{\mu}_{2.1}$	3737.19
$\hat{\mu}_{2.2}$	1607.04
$\hat{\mu}_{3.1}$	849.18
$\hat{\mu}_{3.2}$	1103.94
$\hat{\mu}_{4.1}$	678.91
$\hat{\mu}_{4.2}$	1025.07

Outline

Summary

- The proposed modified two-stage sampling design is suggested in contemplation of striking balance between the estimation performance and survey cost.
- The performance of this modified two stage sampling design can be significantly improved when a proper sampling design is applied in the first stage.

Future Research

- Closed forms of the associated variance and variance estimation
- Bootstrap/Jackknife
- Estimation under a general probability first-stage design

Notations

- N: the number of primary units in the population
- M_i: the number of secondary units in the ith primary unit
- $M = \sum_{i=1}^{N} M_i$: the total number of secondary units in the population
- $h = 1, 2, ..., H$: the hth stratum of the primary units
- M'_h: the population size of the first-stage sample in the hth stratum
- y_{ij}: the value of variable of interest for the jth secondary unit in the ith primary unit
- $y_i = \sum_{j=1}^{M_i}$: the total of value of secondary unit in the ith primary unit

Notations-Conti

- $\tau = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$: the population total
- $\mu_i = y_i / M_i$: the mean per secondary unit in the ith primary unit
- $\mu_1 = \tau / N$: the population mean per primary unit
- $\mu = \tau / M$: the overall population mean
- S_1: the set of the selected primary units
- S_i: the set of the selected secondary units in the ith primary unit.
- $K = \sum_{i \in S_1} M_i$: the total number of secondary units in the selected primary units

Thank you.

Background Knowledge - Census v.s. Sampling

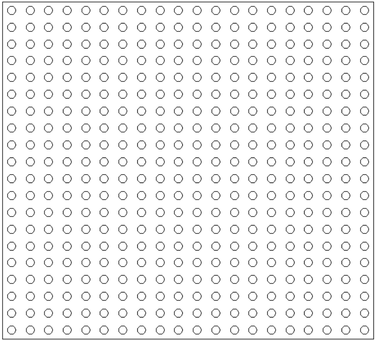


Figure: Population

Background Knowledge

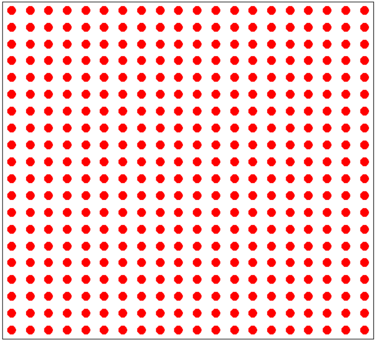


Figure: Census

Background Knowledge

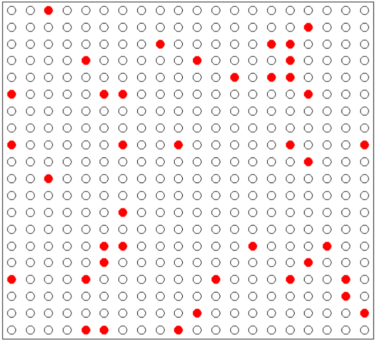


Figure: Sampling Survey

Background Knowledge - Stratified Sampling

The population is partitioned into strata/subpopulations.

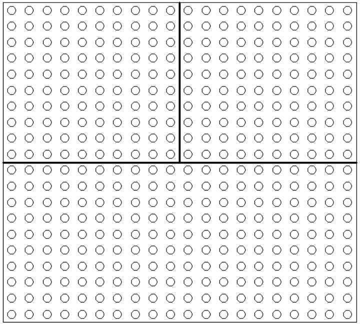


Figure: Stratified Sampling

Background Knowledge

Within-stratum sampling selections are independent.

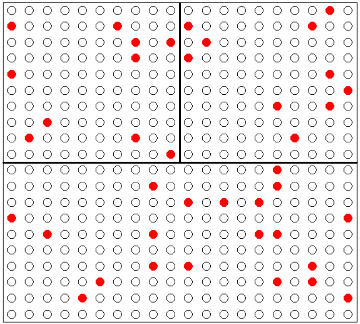


Figure: Stratified Sampling

Background Knowledge - Optimal Stratification Boundary

- Lavallée-Hidiroglou Algorithm and Random Search method
  - Kozak (2004)
  - Stratification variable = Survey variable
- Log-linear model between variable of interest (gross income) and stratification variable (cultivated land area)
  - Rivest (2002)

$\leftarrow$

Background Knowledge - Cluster Sampling

- The population is partitioned into N primary sampling units (PSU), each PSU is a collection of m_i secondary sampling units (SSU).
- A set of PSU's is selected at the first stage sampling.

  One-stage design : All the SSU's in the selected PSU's are selected into the sample.

  Two-stage design : A set of SSU's is selected within each selected PSU's.

  Multi-stage design :

Background Knowledge - Cluster Sampling

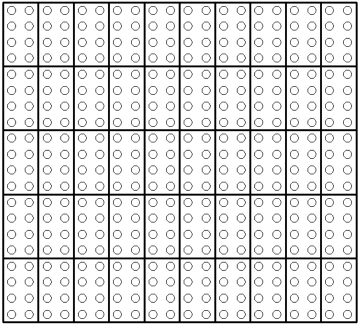


Figure: Cluster Sampling, $N = 50$, $m_i = 8, \forall\, i$

Background Knowledge - One-stage cluster design

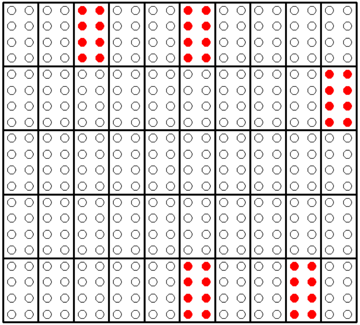


Figure: One-stage

Background Knowledge - Two-stage cluster sampling

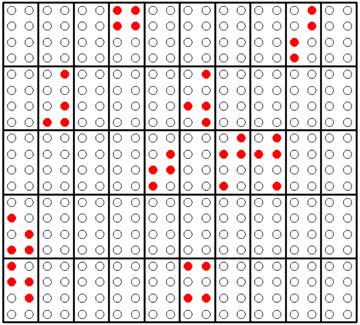


Figure: Two-stage