Estimation Methods for Dual Frame Sample of Cell and Landline Numbers

Mingue Park

(Korea University)

2012. 6. 15

WAPOR Meeting at University of Hong Kong

Outlines

- Introduction
- Definition of Dual Frame Survey
- Estimation Methods for Dual frame Sample for Classical Surveys
- Estimator for Dual Frame Sample of Cell & Landline Numbers
- Discussion

1. Introduction

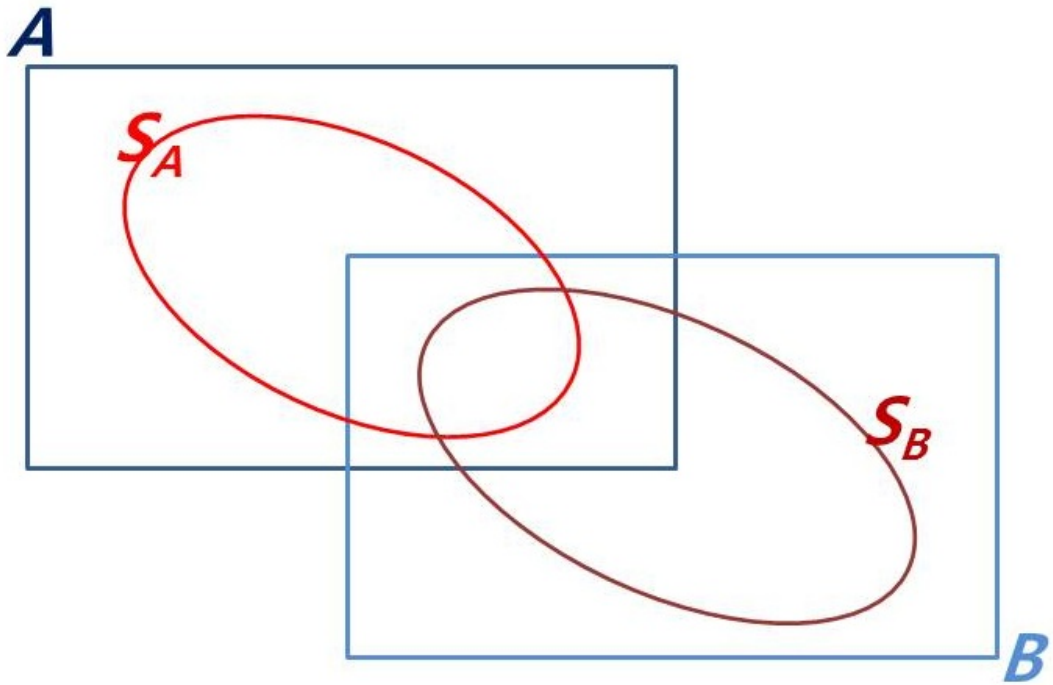
- Rapid increase of cell phone users and cell phone only households possibly causes a significant coverage bias if the conventional landline survey is only considered.
- Dual frame of cell and landline numbers survey is getting popular for many telephone surveys.
- Dual frame survey may reduce the coverage bias of the single frame survey.
- Appropriate nonresponse adjusted estimator is required to reduce the risk in using dual frame survey.
- Clarify the definition of dual frame survey and suggest a possible estimator for dual frame sample of Cell & Landline Numbers.

2. Definition of Dual Frame Survey

- Definition: A survey based on samples selected from two *potentially overlapping* frames.
  - Two frames cover the population $U = A \cup B$.
  - Samples are drawn *independently* from two frames.
  - *Two independent samples* are combined to produce estimates of population parameters such as mean or total.
- Dual(multiple) mode survey
  - Use a single frame that cover the population U properly.
  - Sample is drawn from the frame.
  - Two or multiple modes are used to collect data.
  - Use the conventional estimation strategies to produce population parameters.

3. Estimation Methods for Classical Dual Frame Sample

- Due to its characteristics, estimation of the population parameters using dual frame surveyed data has been a challenging problem for survey statisticians.



3. Estimation Methods for Dual Frame Sample

Notations and assumptions

- Population

$$U = A \cup B,\ a = A \cap B^c,\ b = A^c \cap B,\ ab = A \cap B$$

$$|U| = N,\ |A| = N_A,\ |B| = N_B,\ |a| = N_a,\ |b| = N_b,\ |ab| = N_{ab}$$

- Sample

$$S_a = S_A \cap a,\ S_b = S_B \cap b,\ S'_{ab} = S_A \cap ab,\ S''_{ab} = S_B \cap ab$$

$$|S_A| = n_A,\ |S_B| = n_B,\ |S_a| = n_a,\ |S_b| = n_b,\ |S'_{ab}| = n'_{ab},\ |S''_{ab}| = n''_{ab}$$

- n_A, n_B, n_a, n_b, n'_{ab}, n''_{ab} and corresponding sample total y_A, y_B, y_a, y_b, y'_{ab}, y''_{ab} are known.
- N_A and N_B are known

3. Estimation Methods for Dual Frame Sample

Dual Frame Estimator with simple random samples

- Hartley (1962)

$$\hat{Y}_H = f_A^{-1} y_a + p f_A^{-1} y'_{ab} + (1-p) f_B^{-1} y''_{ab} + f_B^{-1} y_b,$$

where

$$f_A = \frac{n_A}{N_A}, \; f_B = \frac{n_B}{N_B},$$

and p is a constant to be chosen to minimize $Var\left(\hat{Y}_H\right)$.

  - Unbiased

3. Estimation Methods for Dual Frame Sample

Dual Frame Estimator with simple random samples

- Lund (1968)

$$\hat{Y}_L = f_A^{-1} y_a + \left[p f_A^{-1} n'_{ab} + (1-p) f_B^{-1} n''_{ab} \right] \bar{y}_{ab} + f_B^{-1} y_b,$$

where

$$\bar{y}_{ab} = \frac{y_{ab}}{n_{ab}} = \frac{y'_{ab} + y''_{ab}}{n'_{ab} + n'_{ab}}.$$

  - Unbiased
  - $Var\left(\hat{Y}_L\right) \leq Var\left(\hat{Y}_H\right)$

3. Estimation Methods for Dual Frame Sample

Dual Frame Estimator with simple random samples

- Fuller and Burmeister (1972)

$$\hat{Y}_{FB} = (N_A - \hat{N}_{ab})\bar{y}_a + \hat{N}_{ab}\bar{y}_{ab} + (N_B - \hat{N}_{ab})\bar{y}_b,$$

where

$$\bar{y}_a = \frac{y_a}{n_a},\ \bar{y}_b = \frac{y_b}{n_b},$$

and $\hat{N}_{ab}$ is the smallest root of

$$(n_A + n_B)x^2 - (n_A N_B + n_B N_A + n''_{ab} N_A + n''_{ab} N_B)x + n_{ab} N_A N_B = 0.$$

  - Asymptotically unbiased & $Var\left(\hat{Y}_{FB}\right) \leq Var\left(\hat{Y}_L\right)$
  - Based on likelihood method in which parameters are $\bar{Y}_a,\ \bar{Y}_b,\ \bar{Y}_{ab},\ N_{ab}$

3. Estimation Methods for Dual Frame Sample

Dual Frame Estimator with simple random samples

- Bankier (1986)

$$\hat{Y}_s = f_A^{-1} y_a + (f_A + f_B)^{-1} \bar{y}_{ab} + f_B^{-1} y_b$$

- Bankier also provided a raking ratio estimator with initial estimator $\hat{Y}_s$ where the marginal distributions of A and B are used for calibration.
  - Skinner (1991) shows the asymptotic efficiency of the raking ratio estimator.

3. Estimation Methods for Dual Frame Sample

Dual Frame Estimator

- Similar types of estimators were suggested by Lohr and Rao (2006) and Skinner and Rao (1996)
- All considered estimators are easily extended to more general sampling designs with simple manipulation such as by taking $f^{-1} = \pi^{-1}$, where π is an inclusion probability.

Can these estimators be used to analyze the data obtained through cell and landline dual telephone surveys? NO! Need to handle nonresponse

4. A Possible Estimator

Assumptions

- RDD is used for both landline & cell phone survey.
- Exist screened frames for both numbers.
- For the landline RDD, region is used for stratification.
- For landline RDD, either all household members in the selected household are measured or sampling fraction of household members across the households is a constant.
- Post-stratified estimation is used to handle nonresponse.
  - Post-stratified estimator using demographic information is often used to handle nonresponse in Korea

4. A Possible Estimator

$$\hat{Y} = \hat{Y}_{c,post} + \hat{Y}_{l,post} + \hat{Y}_{cl,post},$$

$$\hat{Y}_{cl,post} = p\hat{Y}'_{cl,post} + (1-p)\hat{Y}''_{cl,post},$$

$$p = \frac{\hat{V}\left(\hat{Y}''_{cl,post}\right)}{\hat{V}\left(\hat{Y}'_{cl,post}\right) + \hat{V}\left(\hat{Y}''_{cl,post}\right)}$$

$$\hat{Y}_{c,post} = \sum_{g=1}^{G} N_g \overline{yl}_{c_{rg}}, \; \hat{Y}_{l,post} = \sum_{h=1}^{H}\sum_{g=1}^{G} N_{hg} \overline{yl}_{l_{r_{hg}}}$$

$$\hat{Y}'_{cl,post} = \sum_{g=1}^{G} N_g \overline{yl}_{cl_{rg}}, \; \hat{Y}''_{l,post} = \sum_{h=1}^{H}\sum_{g=1}^{G} N_{hg} \overline{yl}_{cl_{r_{hg}}}$$

4. A Possible Estimator

$$\hat{V}\left(\hat{Y}'_{cl,post}\right) = N^2 \frac{1-n/N}{n} \sum_{g=1}^{G} \frac{n_g}{n}(1-\delta_g) s^2_{yl_{cl\,rg}} + N^2 \sum_{g=1}^{G} \left(\frac{n_g}{n}\right)^2 \frac{1-f_g}{m_g} s^2_{yl_{cl\,rg}},$$

$$\hat{V}\left(\hat{Y}''_{cl,post}\right) = \sum_{h=1}^{H} \hat{V}_h,$$

$$\hat{V}_h = N_h^2 \frac{1-n_h/N_h}{n_h} \sum_{g=1}^{G} \frac{n_{hg}}{n_h}(1-\delta_{hg}) s^2_{yl_{cl\,r_{hg}}} + N^2 \sum_{g=1}^{G} \left(\frac{n_{hg}}{n_h}\right)^2 \frac{1-f_{hg}}{m_{hg}} s^2_{yl_{cl\,r_{hg}}},$$

$$\delta_g = \frac{1-n_g/n}{m_g} \frac{n}{n-1}, \delta_{hg} = \frac{1-n_{hg}/n_h}{m_{hg}} \frac{n_h}{n_h-1},$$

$f_g = m_g/n_g, \quad f_{hg} = m_{hg}/n_{hg}.$

4. A Possible Estimator

Properties

- Sampling design and response distribution are considered to define the variance estimators.
- The propose estimator is asymptotically optimal if
  - propensity score is constant for all element in the same post-stratum
  - independent response mechanism
- Model for y is not considered.
- A variance of $\hat{Y}$ can be obtained by

$$\hat{V}\left(\hat{Y}_{c,post}\right) + \hat{V}\left(\hat{Y}_{l,post}\right) + p^2\hat{V}\left(\hat{Y}'_{cl,post}\right) + (1-p)^2\hat{V}\left(\hat{Y}''_{cl,post}\right)$$

5. Discussion

- Unless appropriate treatments on nonresponse error were given, dual (cell and landline) frame survey used to reduce coverage error may increase whole nonsampling error.
- To take care of the nonresponse error, post stratified estimator is considered.
- Proposed estimator is asymptotically optimal if the assumed response distribution is correct
- The response distribution assumed may be inappropriate, especially for cell phone survey because various sources of nonresponse are expected and their behaviors on the survey questions are also suspected quite different.

5. Discussion

- For the statistical validity of the estimator, thorough investigation on the nonresponse mechanism should be preceded.
- Model for the variables of interest can be considered for quota samples of cell & landline numbers.

Q/A