An Enhanced Approach for Solving the Overlap Problem in Dual Frame RDD Surveys

So-Hyung Park, Sun-Woong Kim, Michael Traugott, Sang-Kyung Lee

Dongguk University, University of Michigan, Hyundae Research Institute

Outline

- **Background**
- **Two-Stage Procedure for Respondent Selection**
- **Conducting Dual Frame RDD Survey in Korea**
- **Examining Initial Weighting Strategies**
- **Exploring Ratio Mean and Variance**
- **Conclusions**

Background

- Dual frame RDD survey designs of landline and cell phone numbers have been popular in many countries
- They eliminate the coverage bias due to cell-only populations
- But they are subject to the overlap and overrepresentation problem due to population elements selected from both frames
- One of the solutions is to use a compensatory weight for unequal probabilities of selection
- We may choose one of the following initial weighting strategies

Brick et al. (2006)

- Their dual frame study used a sample of households with no sampling of household member

- The weights were adjusted to account for households that had multiple chances of being sampled because they had more than one telephone number.

- In the landline sample, the divisor (the number of landlines) was three if there were three or more lines (less than 3 percent had more than three landline numbers). In the cell sample the divisor could be as large as four because about 3 percent of cell sample households had four or more cell numbers.

Dutwin et al. (2008)

- They assigned a weight equal to the number of adults in the household (capped at 3), multiplied by the reciprocal of the number of telephones in the household (capped at a minimum of .33)

Best (2010)

- The probability that a person is sampled from the landline frame is the landline sampling fraction, multiplied by the number of landline telephones in the household that are used to receive calls, divided by the number of adults in the household.

- The probability that a person is sampled by cell phone is the cell phone sampling fraction if the person has a cell phone, or it is zero if the person does not have a cell phone.

- The weight equals to the reciprocal of the above probability

- He was concerned only with relative probabilities and didn’t compute the actual selection probability for each household

BRFSS (2010)

- Initial weights are the number of adults in the household, multiplied by the reciprocal of the number of residential telephone numbers in the household

Principal Problems

- Those weighting strategies are often inappropriate or ambiguous because of the lack of exact information on actual use of landline or cell phones by a respondent

- Since a cell phone is a personal or shared device and a landline phone can be a personal device, we need to consider respondent selection and estimation procedures different from the conventional methods

Two-Stage Procedure for Respondent Selection

- Illustrated by Lepkowski and Kim (2005), but little attention has been paid to it
- Different from a conventional within-household selection
- Randomly choose one eligible person who uses the phone number selected
- Avoids complicated estimation procedures used to combine results from overlapping frame (e.g., Lepkowski and Groves, 1986)

Two-Stage Procedure for Respondent Selection (Cont.)

- **Steps for Respondent Selection**

1. With assumption that both landline and cell numbers are 1) for the household, 2) shared, or 3) personal numbers, a few questions are used to identify such status from the informant for each number selected from the RDD frame.

2. If the phone number is for the household or a shared number, one eligible person using the number is randomly chosen and asked to provide the information on other phone numbers that could have been used to reach the person. If the phone number is for a single person, the person is interviewed.

Two-Stage Procedure for Respondent Selection (Cont.)

- Questions for Informants
  - How many adults, including yourself, use this phone number?
  - Please tell me their relationships (e.g., “husband, daughter, and son”)
    1. _____
    2. _____
    3. _____
    .
    .
    .
  - Random selection of respondent

Two-Stage Procedure for Respondent Selection (Cont.)

- Questions for Respondents
  - How many phone numbers, including this number, do you use? Include both landline and cell numbers.
  - (For each number) How many people, including yourself, use that phone number for personal calls?

Conducting Dual Frame RDD Survey in Korea

- Population: Adults 18 or over
- Sample Design: List-assisted RDD for a landline sample and RDD based on 10,000-blocks for a cell sample

  Note that unlike the USA, cell phone numbers do not have area code
- Sample Size: 1,508 (Landline 899, Cell 609)
- Survey Period: November 1 to December 27, 2010
- Data Collection: Computer-assisted telephone interviewing (CATI)

Conducting Dual Frame RDD Survey in Korea (Cont.)

- Telephone Availability Status

	Frequency	Percent
Both	1,215	80.6%
Cell only	184	12.2%
Landline only	109	7.2%
TOTAL	1,508	100.0%

Conducting Dual Frame RDD Survey in Korea (Cont.)

- Based on the number of people (including under 18) using the same number

1) Provided by Informants

Type of number	Frame	
	Landline	Cell
Household	59.3%	0.3%
Personal	11.1%	87.2%
Personal and household (1 person household)	16.9%	12.5%
Shared	12.7%	0.0%
Total	100.0%	100.0%

Conducting Dual Frame RDD Survey in Korea (Cont.)

- Based on the number of adults using the same number

2) Provided by Informants

Type of number	Frame	
	Landline	Cell
Household	33.6%	0.3%
Personal	15.7%	87.2%
Personal and household (1 person household)	17.2%	12.5%
Shared	33.5%	0.0%
Total	100.0%	100.0%

Conducting Dual Frame RDD Survey in Korea (Cont.)

- Based on the number of adults using the same number

3) Provided by Respondents

Type of number	Frame	
	Landline	Cell
Household	40.7%	0.2%
Personal	33.5%	87.3%
Personal and household (1 person household)	17.5%	12.5%
Shared	8.3%	0%
Total	100.0%	100.0%

Examining Initial Weighting Strategies

- Notation

1. Type of phone: i

   Landline: $i = 1$

   Cell: $i = 2$

2. Frame sizes: N_i

   Landline frame : $N_1 = \sum_{h=1}^{H} N_{1h}$, h=stratum

   Cell frame : N_2 (No strata)

3. Sample sizes: n_i

   Landline frame : $n_1 = \sum_{h=1}^{H} n_{1h}$

   Cell frame : n_2 (No strata)

Examining Initial Weighting Strategies(Cont.)

4. Selection probabilities of RDD numbers

Landline: $p_{1h} = \frac{n_{1h}}{N_{1h}}$

Cell: $p_2 = \frac{n_2}{N_2}$

5. Indicator variable

δ_{ik}: 'use' of phone i by respondent k

6. Phone ownership

α_{ik}: Number of phone i's to be reached to the respondent k

7. Eligible persons

β_{ijk}: Number of adults who use jth phone i with respondent k

Examining Initial Weighting Strategies(Cont.)

8. Selection probabilities of respondents

$$(a)\begin{cases} \pi_{1k} = \delta_{1k} p_{1h} \ (k \in h) \\ \pi_{2k} = \delta_{2k} p_2 \end{cases}$$, if phone ownership and eligibles are not considered

$$(b)\begin{cases} \pi_{1k} = \delta_{1k} \alpha_{1k} p_{1h} \ (k \in h) \\ \pi_{2k} = \delta_{2k} \alpha_{2k} p_2 \end{cases}$$, if phone ownership only is considered

$$(c)\begin{cases} \pi_{1k} = \sum_{j=1}^{\alpha_{ik}} \frac{\delta_{1k} p_{1h}}{\beta_{1jk}} \ (k \in h) \\ \pi_{2k} = \sum_{j=1}^{\alpha_{ik}} \frac{\delta_{2k} p_2}{\beta_{2jk}} \end{cases}$$, if both phone ownership and eligibles are considered

Examining Initial Weighting Strategies(Cont.)

- Weights for respondent *k* using from phone *i*

Regardless of which frame is used,

1) Landline only person: $W_{ik} = 1/\pi_{1k}$

2) Cell only person: $W_{ik} = 1/\pi_{2k}$

3) Landline and cell person :

$$W_{ik} = \frac{1}{\pi_{1k} + \pi_{2k}}$$, if overlap is not considered

$$W_{ik} = \frac{1}{\pi_{1k} + \pi_{2k} - \pi_{1k}\pi_{2k}}$$, if overlap is considered

Note. There are six types of weights according to (a), (b) and (c).

Examining Initial Weighting Strategies(Cont.)

- Coefficient of variation for weights

1) Landline or Cell RDD frame

$$cv(\mathrm{w_i}) = \frac{se(\mathrm{w_i})}{\mathrm{w_i}} = \frac{\sqrt{v(\mathrm{w_i})}}{\mathrm{w_i}} \text{ , i = 1,2}$$

$$\text{where } \mathrm{w_1} = \sum_{\mathrm{h=1}}^{\mathrm{H}} \sum_{\mathrm{k=1}}^{\mathrm{n_{1h}}} \mathrm{w_{1hk}} \text{ , } \mathrm{w_2} = \sum_{\mathrm{k=1}}^{\mathrm{n_2}} \mathrm{w_{2_k}}$$

$$v(w_1) \doteq \sum_{h=1}^{H} n_{1h} \sum_{k=1}^{n_{1h}} (w_{1hk} - \bar{w}_{1h})^2 / (n_{1h} - 1)$$

$$v(w_2) \doteq n_2 \sum_{k=1}^{n_C} (w_{2k} - \bar{w}_2)^2 / (n_2 - 1)$$

Examining Initial Weighting Strategies(Cont.)

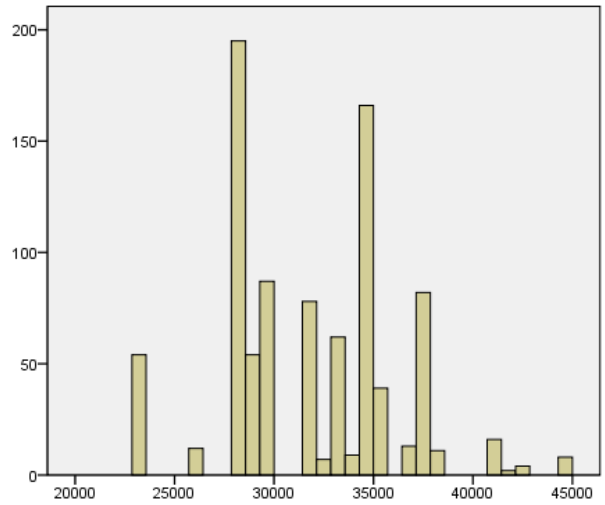
2) Dual frame RDD

$$cv(\mathrm{w}) = \frac{se(\mathrm{w})}{\mathrm{w}} = \frac{\sqrt{v(\mathrm{w})}}{\mathrm{w}}$$

where $\mathrm{w} = \mathrm{w}_1 + \mathrm{w}_2$

Examining Initial Weighting Strategies(Cont.)

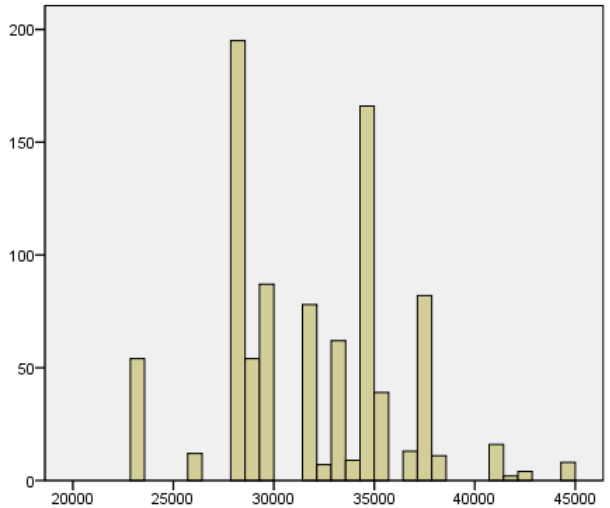
- Distributions of weights in landline frame



W_{ik} by (a)

cv=0.002

(1) Not considering overlap



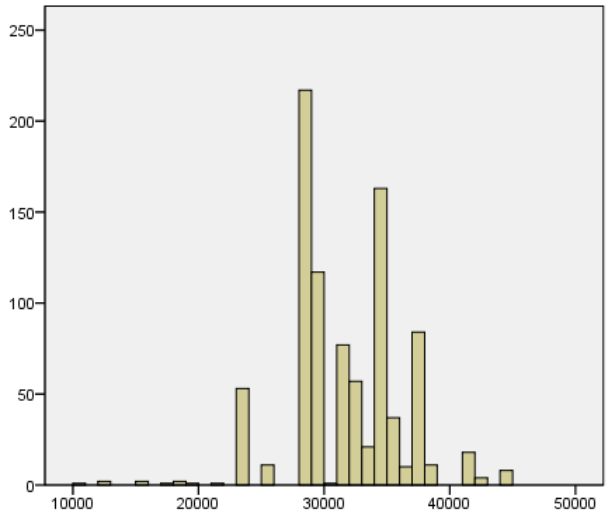
W_{ik} by (a)

cv=0.002

(2) Considering overlap

Examining Initial Weighting Strategies(Cont.)

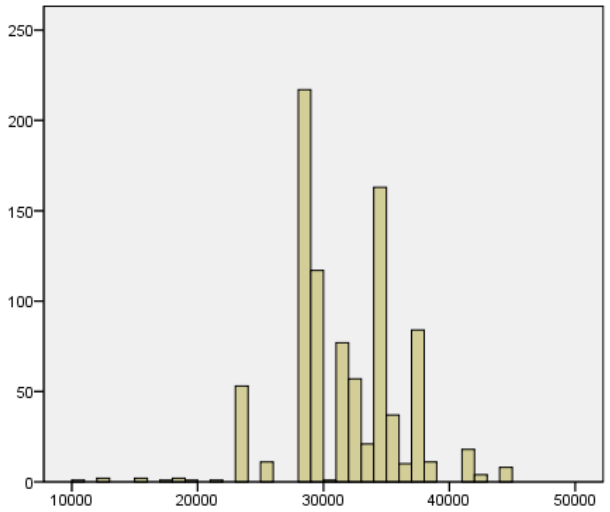
- Distributions of weights in landline frame



W_{ik} by (b)

cv=0.003

(1) Not considering overlap



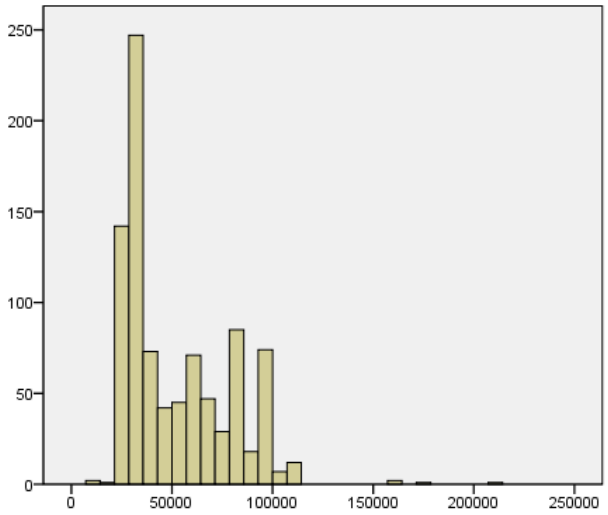
W_{ik} by (b)

cv=0.003

(2) Considering overlap

Examining Initial Weighting Strategies(Cont.)

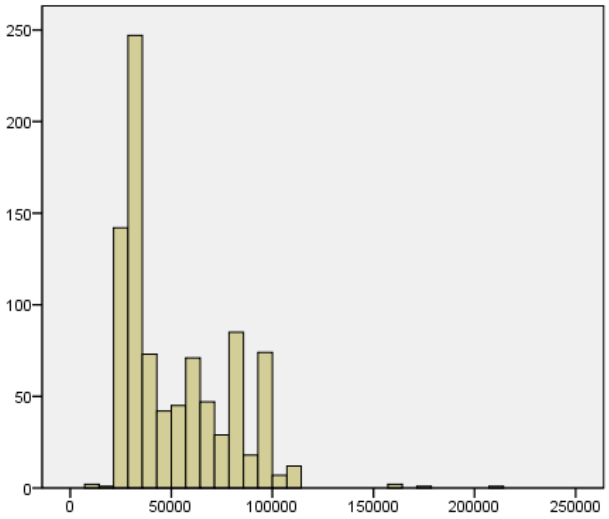
- Distributions of weights in landline frame



W_{ik} by (c)

cv=0.016

(1) Not considering overlap



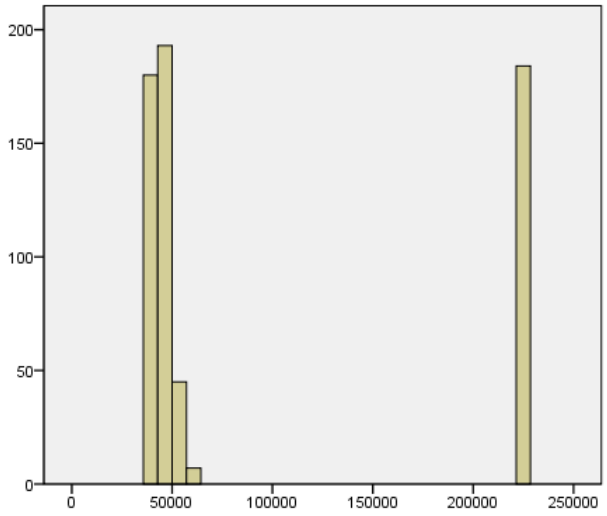
W_{ik} by (c)

cv=0.016

(2) Considering overlap

Examining Initial Weighting Strategies(Cont.)

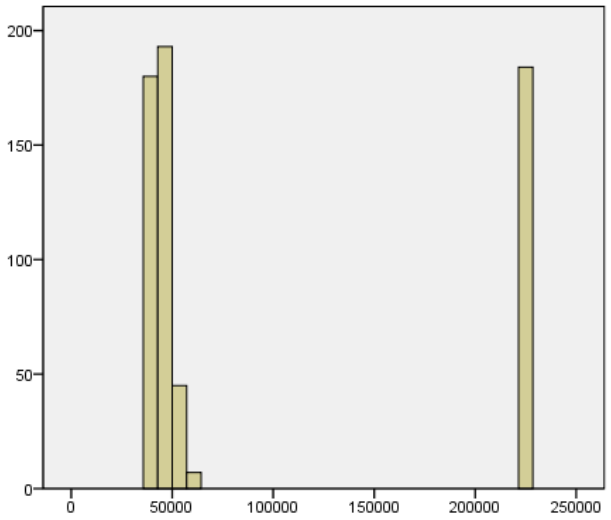
- Distributions of weights in cell frame



W_{ik} by (a)

cv=0.035

(1) Not considering overlap



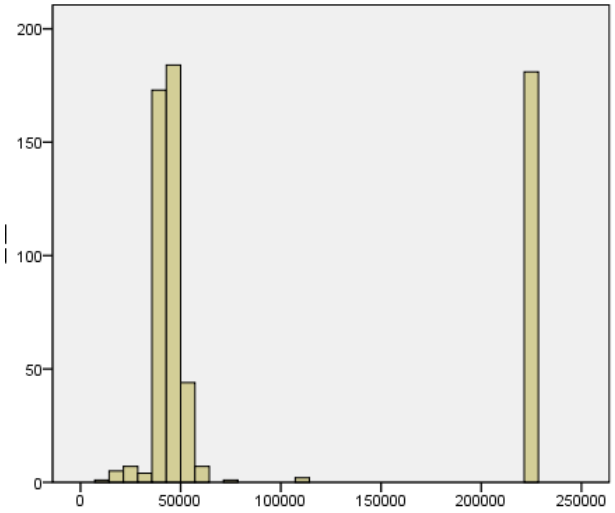
W_{ik} by (a)

cv=0.035

(2) Considering overlap

Examining Initial Weighting Strategies(Cont.)

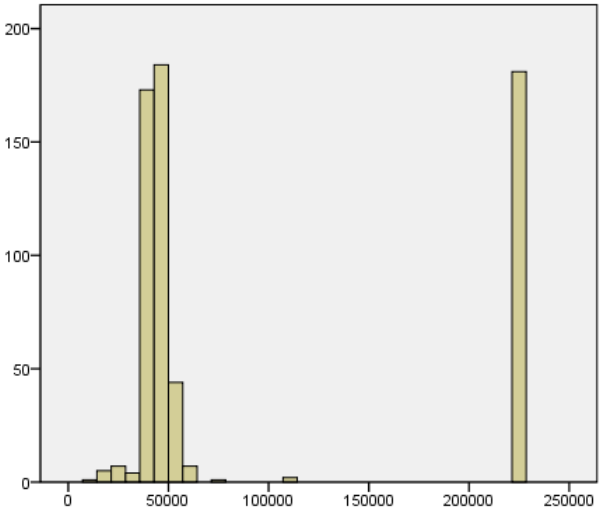
- Distributions of weights in cell frame



W_{ik} by (b)

cv=0.035

(1) Not considering overlap



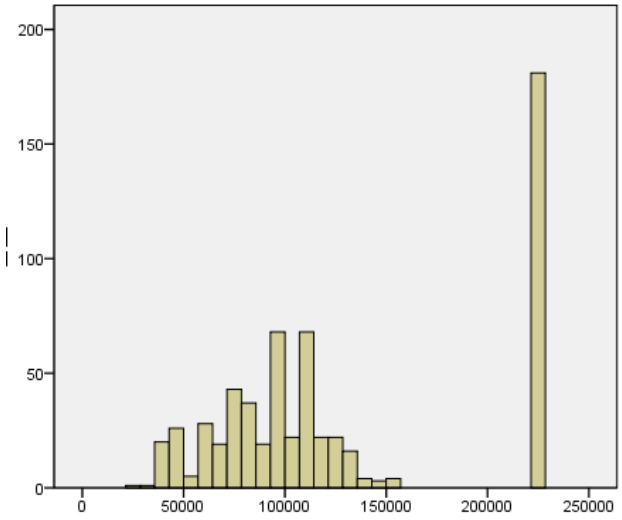
W_{ik} by (b)

cv=0.035

(2) Considering overlap

Examining Initial Weighting Strategies(Cont.)

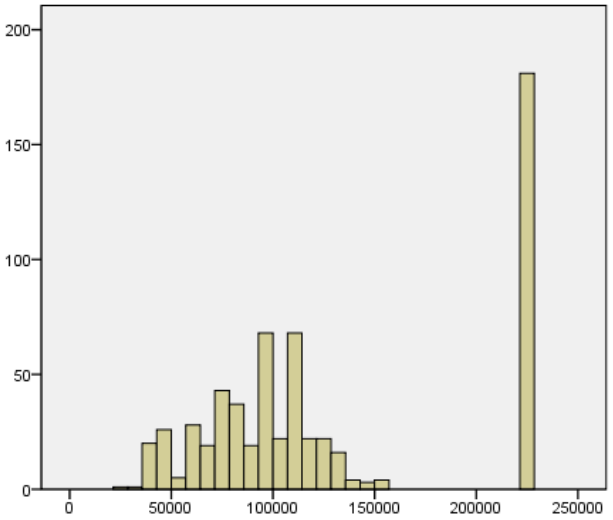
- Distributions of weights in cell frame



W_{ik} by (c)

cv=0.021

(1) Not considering overlap



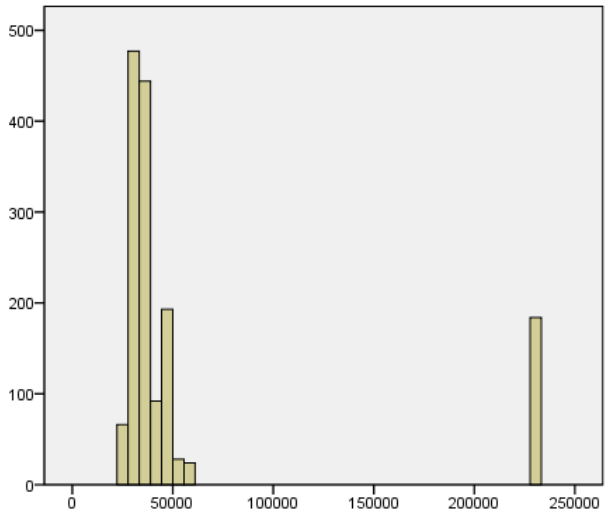
W_{ik} by (c)

cv=0.021

(2) Considering overlap

Examining Initial Weighting Strategies(Cont.)

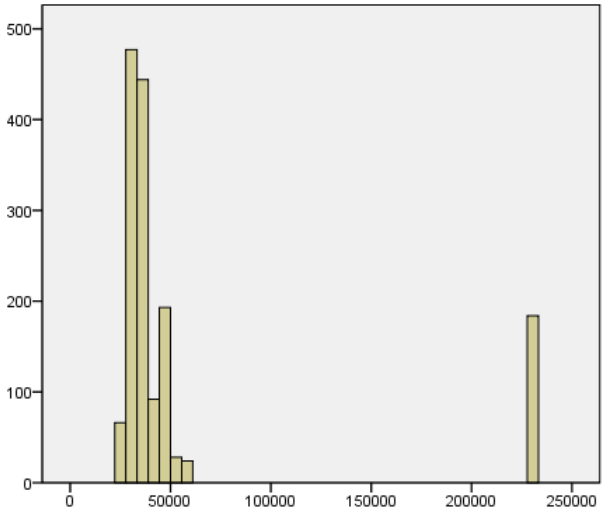
- Distributions of weights in both landline and cell frame



W_{ik} by (a)

cv=0.024

(1) Not considering overlap



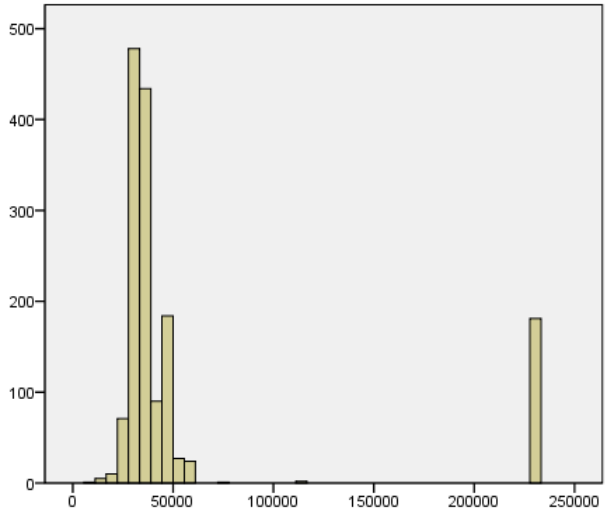
W_{ik} by (a)

cv=0.024

(2) Considering overlap

Examining Initial Weighting Strategies(Cont.)

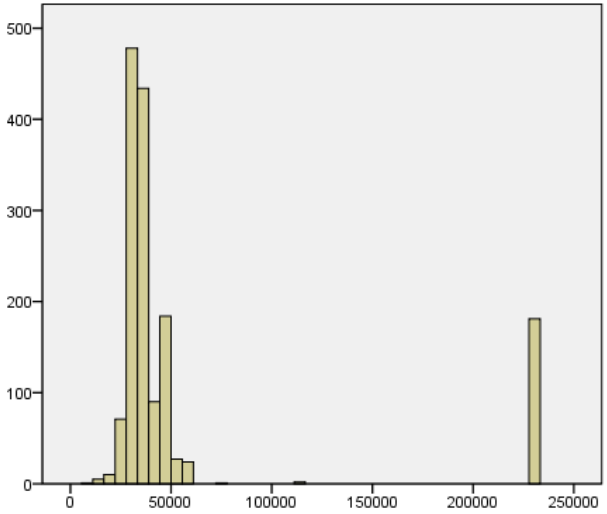
- Distributions of weights in both landline and cell frame



W_{ik} by (b)

cv=0.024

(1) Not considering overlap



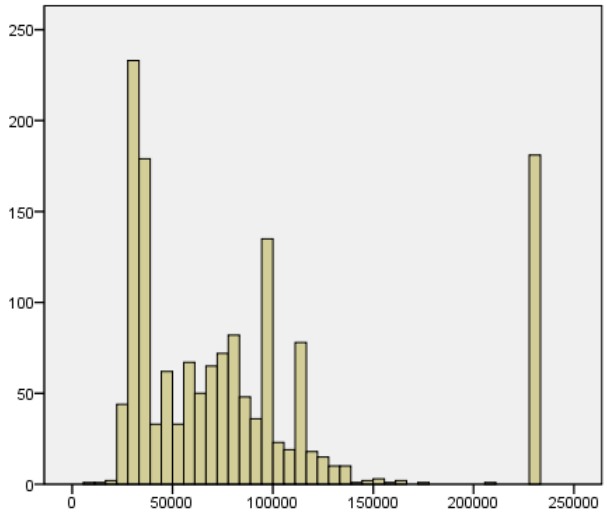
W_{ik} by (b)

cv=0.024

(2) Considering overlap

Examining Initial Weighting Strategies(Cont.)

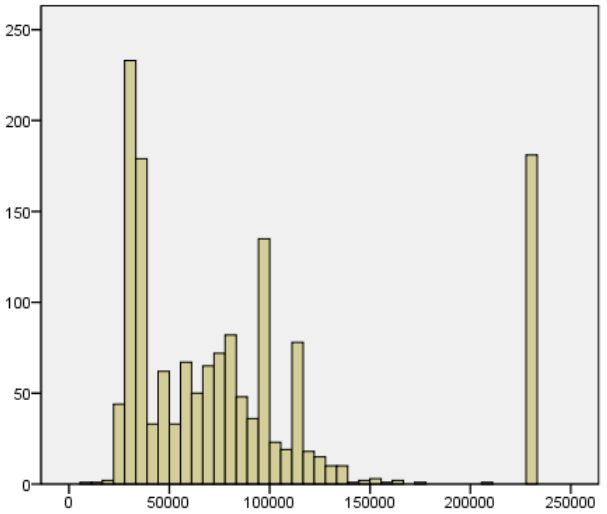
- Distributions of weights in both landline and cell frame



W_{ik} by (c)

cv=0.014

(1) Not considering overlap



W_{ik} by (c)

cv=0.014

(2) Considering overlap

Examining Initial Weighting Strategies(Cont.)

- Summary
  - The effect of the term $\pi_{1k}\pi_{2k}$ is negligible, when the overlap in dual frame is considered.
  - The distributions of weights according to either phone ownership or eligibles vary
  - All coefficients of variation are small enough to ensure that the bias of the ratio mean considered next is not appreciable

Exploring the Ratio Mean and Variance

- Ratio mean in dual frame

$$\bar{y}_{DW} = \frac{u_1 + u_2}{w_1 + w_2}, \text{where}$$

$$u_1 = \sum_{h=1}^{H} \sum_{k=1}^{n_{1h}} w_{1hk} y_{1hk}, \; u_2 = \sum_{k=1}^{n_2} w_{2k} y_{2k}$$

Exploring the Ratio Mean and Variance (Cont.)

- Estimated variance in dual frame

$$v(\bar{y}_{DW}) \doteq [v(u_1 + u_2) + \bar{y}_{DW}{}^2 v(w_1 + w_2) - 2\bar{y}_{DW} c(u_1 + u_2, w_1 + w_2)]/(w_1 + w_2)^2$$

$$= [v(u_1) + V(u_2) + \bar{y}_{DW}{}^2\{v(w_1) + V(w_2)\} - 2\bar{y}_{DW}\{c(u_1, w_1) + c(u_2, w_2)\}]/(w_1 + w_2)^2$$

where

$$v(u_1) \doteq \sum_{h=1}^{H} n_{1h} \sum_{k=1}^{n_{1h}} (u_{1hk} - \bar{u}_{1h})^2/(n_{1h} - 1)$$

$$v(u_2) \doteq n_2 \sum_{k=1}^{n_C} (u_{2k} - \bar{u}_2)^2/(n_2 - 1)$$

Exploring the Ratio Mean and Variance (Cont.)

$$c(u_1, w_1) \doteq \sum_{h=1}^{H} n_{1h} \sum_{k=1}^{n_{1h}} (u_{1hk} - \bar{u}_{1h})(w_{1k} - \bar{w}_{1h}) / (n_{1h} - 1)$$

$$c(u_2, w_2) \doteq n_2 \sum_{k=1}^{n_C} (u_{2k} - \bar{u}_2)(w_{2k} - \bar{w}_2) / (n_2 - 1)$$

$$\bar{u}_{1h} = \frac{\sum_{k=1}^{n_{1h}} w_{1hk}\, y_{1hk}}{n_{1h}},\ \bar{u}_2 = \frac{\sum_{k=1}^{n_2} w_{2k} y_{2k}}{n_2},\ \bar{w}_{1h} = \frac{\sum_{i=1}^{n_{1h}} w_{1hk}}{n_{1h}},\ \bar{w}_2 = \frac{\sum_{k=1}^{n_2} w_{2k}}{n_2}$$

Exploring the Ratio Mean and Variance (Cont.)

- Design effect (deff) in dual frame

$$v(\bar{y}) \doteq s^2_{\bar{y}_{Dw}}/n$$

$$s^2_{\bar{y}_{Dw}} = \frac{n}{n-1}\frac{\sum_{i=1}^{2}\sum_{k=1}^{n_i} w_{ik}(y_{ik}-\bar{y}_{DW})^2}{\sum_{i=1}^{2}\sum_{k=1}^{n_i} w_{ik}}, n = \sum_{i=1}^{2} n_i$$

$$deff(\bar{y}_{Dw}) = v(\bar{y}_{Dw})/v(\bar{y})$$

Exploring the Ratio Mean and Variance (Cont.)

- Application

y_{h1k}(or y_{2k}) : Employment status (Yes, No)

	(1) Wik by (a)	(2) Wik by (a)	(1) Wik by (b)	(2) Wik by (b)	(1) Wik by (c)	(2) Wik by (c)
Ratio Mean $\bar{y}_{Dw}$	0.5191	0.5191	0.5241	0.5241	0.3618	0.3618
Design effect $deff(\bar{y}_{Dw})$	2.0958	2.0958	2.1486	2.1486	1.1003	1.1003

(1) Not considering overlap

(2) Considering overlap

Conclusions

- We successfully completed a two-stage procedure for solving the overlap problem in a national dual frame survey
- We examined 6 person-level initial weighting strategies based on data on the phone ownership and eligibles from the survey
- The distributions of weights according to phone ownership or eligibles vary
- The bias of the ratio mean in dual frame may be not appreciable
- There would be substantial decrease of variance due to the smaller coefficients of variation for weights based on both phone ownership and eligibles in dual frame

Thank you

Contact at sohyung@dongguk.edu