ANALYSING LANDLINE AND CELL-PHONE SURVEYS IN HEALTH STUDIES

María del Mar Arcos.
Faculty of Medicine. University of Granada. Spain.

ABSTRACT
This work is a review of several estimation methods for dual-frame designs in the particular context of health telephone surveys. Due to the recent increase of the number of people who has no landline phone but only mobile phone, it is very important nowadays that telephone surveys incorporate landline together with cell-phone samples. Otherwise, large bias may appear in the estimations. Given that the two frames made of landline owners and cell-phone owners intersect, some adapted estimation methods have to be used and there are several in the literature. The paper presents the different existing methods and compares these methods on a real health survey concerning the attitude of the Andalusian population regarding the public health system. The results suggest that the use of multiple frames might be useful in other health surveys where good estimates are wanted for both the whole population and particular subgroups at the same time.
KEYWORDS: Multiple frames, coverage bias, health surveys.
MSC: 62D05.

RESUMEN
Este trabajo es una revisión de varios métodos de estimación para marcos duales en el contexto particular de encuestas telefónicas salud. Debido al reciente aumento del número de personas que no tienen teléfono fijo, y sólo tienen teléfono móvil, es muy importante hoy en día que las encuestas telefónicas incorporen teléfono fijo, junto con muestras de teléfonos móviles. De lo contrario, un gran sesgo puede darse en las estimaciones. Dado que los dos marcos hechos de los propietarios de teléfonos fijos y los propietarios de teléfonos celulares se solapan, tienen que ser utilizados métodos de estimación adaptados y hay varios en la literatura. El artículo presenta los diferentes métodos existentes y compara estos métodos en una encuesta real para la salud en relación con la actitud de la población andaluza sobre el sistema de salud pública. Los resultados sugieren que el uso de múltiples marcos podría ser útil en encuestas de salud proporcionando buenas estimaciones tanto para toda la población como para subgrupos determinados.

1. INTRODUCTION

From 2000 to the present,there has been a steady increase in the use of telephone surveys, which have replaced all other data collection methods (the majority of which were face-to-face interviews). The telephone survey presents numerous advantages compared to a face-to-face one. In some subject areas face-to-face surveys have been completely ousted by telephone interviewing. Moreover, studies have reported improved results from phone surveys compared with face-to-face interviews. However, telephone surveys also present some drawbacks with regard to coverage, due to the absence of a telephone in some households and the generalized use of mobile phones, which are sometimes replacing fixed (land) lines entirely . The potential for coverage error as a result of the exponential growth of the cell-phone-only population has led to the development of dual-frame surveys. In these designs, a traditional sample from the landline frame is supplemented with an independent sample from the banks of numbers designated for cell-phones. By drawing samples from both cell phones and landline phones instead of from a single frame, it is possible to reduce survey costs, improve the coverage of the overall sample and potentially even increase response rates, depending on the specific survey being conducted. Some applications of dual frame techniques in health surveys can be seen in [6] and [7].

In this work, we look at a unified approach to estimation in multiple frames. We investigate how the approach performs with data from the Survey of Opinions and Attitudes of the Andalusian Population regarding the public health system, a dual frame survey looking at the opinion of Andalusian population about its health system and the introduction of a new financial model, the pharmaceutical copayment.

2. SURVEY OF OPINIONS AND ATTITUDES OF THE ANDALUSIAN POPULATION REGARDING THE PUBLIC HEALTH SYSTEM (OPIAH) 2013

The 2013 survey of Opinions and attitudes of the Andalusian population regard s the health system (OPIAH) is a population-based survey conducted by a public scientific research institute specialising in the social sciences. Its aim is to reflect the opinions of the Andalusian population with regard to various aspects of the public health system and specially about the controversial pharmaceutical copayment. Nowadays, the Andalusian health system is a universal-coverage system where the right to health is associated with the condition of citizen. The state is in charge of financing the system through taxes, so that the medical assistance is guaranteed for all the population. However, in the past few years due to the economic crisis and the progressive aging of the population (the consumption of health career sources is much higher in elderly people), the government has introduced a new measure: the pharmaceutical copayment. This copay means that when a doctor prescribes a treatment the patient will have to face a part of the costs while the main part of it will still relay on the state. In Andalusia, the proportion of survey subjects only reachable by landline communication has decreased to below 10%. In economic good times, and due to rising numbers of internet connections, the proportion of people only reachable by cell phone also declined. However, in recent years this proportion has risen to around 20%. The number of people not reachable by phone now only represent a residual percentage of the population (less than 2%). In this survey, on decided to carry out telephone interviews with adults using both landlines and cell phones. Taking into account the time and budget available, 2402 interviews were performed by qualified interviewers, specially trained in survey techniques. The number of interviews to be conducted via landline and via cell phone was determined by calculating the optimum proportion (in the sense of minimum variance) for each type of telephone, taking into account costs and the percentage of possession of each type of device (following [3]). As a result, the sample sizes ascertained were 1919 for landlines and 483 for cell phones. The base weights are the ratio of the number of telephone numbers in the frame to the number sampled. The weights were further adjusted to account for people who had multiple chances of being sampled because they had more than one telephone number.

3. ESTIMATION IN LANDLINE AND CELL-PHONE SURVEYS

Consider a finite set of N population units identified by the integers, $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$, and let A and B be two sampling-frames, both can be incomplete, but it is assumed that together they cover the entire finite population. Let $\mathcal{A}$ be the set of population units in frame A and $\mathcal{B}$ the set of population units in frame B. The population of interest, $\mathcal{U}$, may be divided into three mutually exclusive domains, $a = \mathcal{A} \cap \mathcal{B}^c, b = \mathcal{A}^c \cap \mathcal{B}$ and $ab = \mathcal{A} \cap \mathcal{B}$. Because the population units in the overlap domain ab can be sampled in either survey or both surveys, it is convenient to create a duplicate domain $ba = \mathcal{B} \cap \mathcal{A}$, which is identical to $ab = \mathcal{A} \cap \mathcal{B}$, to denote the domain in the overlapping area coming from frame B. Let N, N_A, N_B, N_a, N_b, N_{ab}, N_{ba} be the number of population units in $\mathcal{U}$, $\mathcal{A}$, $\mathcal{B}$, a, b, ab, ba, respectively. It follows that $N_A = N_a + N_{ab}$, $N_B = N_b + N_{ba}$ and $N = N_a + N_b + N_{ab} = N_a + N_b + N_{ba}$.

Let y be a variable of interest in the population and y_k its value on unit k, for $k = 1, \ldots, N$. The entire set of population y values is our finite population F. The objective is to estimate the finite population total $Y = \sum_{k=1}^{N} y_k$ of y, that can be written as

$$Y = Y_a + \eta Y_{ab} + (1 - \eta) Y_{ba} + Y_b, \tag{1}$$

where $0 \leq \eta \leq 1$, and $Y_a = \sum_{k \in a} y_k$, $Y_{ab} = \sum_{k \in ab} y_k$, $Y_{ba} = \sum_{k \in ba} y_k$ and $Y_b = \sum_{k \in b} y_k$. Two probability samples s_A and s_B are drawn independently from frame A and frame B of sizes n_A and n_B, respectively. Each design induces first-order inclusion probabilities π_{Ak} and π_{Bk}, respectively, and sampling weights $d_{Ak} = 1/\pi_{Ak}$ and $d_{Bk} = 1/\pi_{Bk}$. Units in s_A can be divided as $s_A = s_a \cup s_{ab}$, where $s_a = s_A \cap a$ and $s_{ab} = s_A \cap (ab)$. Similarly, $s_B = s_b \cup s_{ba}$, where $s_b = s_B \cap b$ and $s_{ba} = s_B \cap (ba)$. Note that s_{ab} and s_{ba} are both from the same domain ab, but s_{ab} is part of the frame A sample and s_{ba} is part of the frame B sample.

A multiplicity adjusted estimator of Y is given by

$$\hat{Y}_H = \hat{Y}_a + \eta \hat{Y}_{ab} + (1 - \eta) \hat{Y}_{ba} + \hat{Y}_b, \tag{2}$$

where $\hat{Y}_a = \sum_{k \in s_a} d_{Ak} y_k$ is the Horvitz-Thompson estimator for the total of domain a and similarly for the other domains. [3] proposes to choose η in order to minimize the variance of the estimator. If we let

$$d_k^\circ = \begin{cases} d_{Ak} & \text{if } k \in s_a \\ \eta d_{Ak} & \text{if } k \in s_{ab} \\ (1-\eta) d_{Bk} & \text{if } k \in s_{ba} \\ d_{Bk} & \text{if } k \in s_b \end{cases} \tag{3}$$

then $\hat{Y}_H = \sum_{k \in s} d_k^\circ y_k$. Since each domain is estimated by its expansion estimator, $\hat{Y}_H$ is an unbiased estimator of Y for a given η. Since frames A and B are sampled independently, the variance of $\hat{Y}_H$ is given by

$$V(\hat{Y}_H) = V(\hat{Y}_a + \eta \hat{Y}_{ab}) + V((1-\eta)\hat{Y}_{ba} + \hat{Y}_b), \tag{4}$$

where the first component of the right hand side is computed under $p_A(\cdot)$ (the sampling design in frame A) and the second one under $p_B(\cdot)$, and both are always understood conditional on the finite population F.

Choice of a value for η has attracted much attention in literature; the value of η that minimizes the variance in (4) depends on unknown population variances and covariances and, when estimated from the data, it depends on the values of the variable of interest. This implies a need to recompute weights for every variable of interest y, which will be inconvenient in practice for statistical agencies conducting surveys with numerous variables and lead to inconsistencies in the estimates (see [5]).

[2] proposed modifying Hartley's estimator by incorporating additional information regarding estimation of the overlap domain. The resulting estimator is:

$$\hat{Y}_{FB} = (\hat{Y}_a + \beta_1 \hat{Y}_{ab} + (1-\beta_1)\hat{Y}_{ba} + \hat{Y}_b + \beta_2(\hat{N}_{ab} - \hat{N}_{ba})) \tag{5}$$

where $\hat{N}_{ab}$ is the Horvitz-Thompson estimator of N_{ab}, that is $\hat{N}_{ab} = \sum_{k \in s_A} d_k^\circ \delta_k(ab)$, $\hat{N}_{ba} = \sum_{k \in s_B} d_k^\circ \delta_k(ab)$ and $\delta_k(ab) = 1$ if $k \in ab$ and 0 otherwise. β_1 and β_2 are selected to minimize the variance. In this case, and as with Hartley's estimator, a new set of weights must be calculated for each response variable, leading to the inconsistency of the estimator. Optimum values depend on covariances among the Horvitz-Thompson estimators and it is also possible to obtain values of β_1 outside $[0, 1]$.

The estimator developed by [2] incorporates information regarding the estimation of N_{ab} to improve over $\hat{Y}_H$, but has the drawback of not being a linear combination of y values, unless in particular cases like when using simple random sampling. [11] propose a modification of the estimator proposed by [2] for simple random sampling to handle complex designs. They introduce a pseudo maximum likelihood (PML) estimator that does not achieve optimality like the FB estimator, but it can be written as a linear combination of the observations and the same set of weights can be used for all variables of interest.

When inclusion probabilities in domain ab are known for both frames, and not just for the frame from which the unit was selected, *single-frame* methods can be used that combine the observations into a single dataset and adjust the weights in the intersection domain for multiplicity. In particular, observations from frame A and frame B are combined and the two samples drawn independently from A and B are considered as a single stratified sample over the three domains a, b and ab. To adjust for multiplicity, the weights are defined as follows for all units in frame A and in frame B,

$$d_k^\star = \begin{cases} d_{Ak} & \text{if } k \in s_a \\ (1/d_{Ak} + 1/d_{Bk})^{-1} & \text{if } k \in s_{ab} \cup s_{ba} \\ d_{Bk} & \text{if } k \in s_b \end{cases} . \tag{6}$$

The estimator is given by the expression (see ([4])):

$$\hat{Y}_{KA} = \sum_{k \in s} d_k^\star y_k. \tag{7}$$

Its variance is given by $V(\hat{Y}_{KA}) = V(\sum_{k \in s_A} d_k^\star y_k) + V(\sum_{k \in s_B} d_k^\star y_k)$, where the first component of the right hand side is computed under $p_A(\cdot)$ and the second one under $p_B(\cdot)$. If N_A and N_B were known, the single-frame estimator $\hat{Y}_{KA}$ could be adjusted using raking ratio estimation [10].

[8] used calibration procedures for estimation from dual frame sampling assuming that some kind of auxiliary information is available. For example, assuming that there are p auxiliary variables, $\underline{x}_k = (x_{1k}, ..., x_{pk})$

is the value taken by such auxiliary variables on unit k. It is assumed that the vector of population totals of the auxiliary variables, $\underline{t}_x = \sum_{k\in U} \underline{x}_k$ is also known. In this context, the dual frame calibration estimator can be defined as follows,

$$\hat{Y}_{Cali}^{DF} = \sum_{k\in s} d_k^{DF} y_k \tag{8}$$

where weights d_k^{DF} are chosen to be as close as possible to basic design weights and, at the same time, satisfy benchmark constraints on the auxiliary variables, i.e. they are such that

$$\min_{d_k^{DF}} \sum_{k\in s} G(d_k^{DF}, d_k^{\circ}), \quad \text{subject to} \quad \sum_{k\in s} d_k^{DF} \underline{x}_k = \underline{t}_x,$$

with $G(\cdot,\cdot)$ a given distance measure.

4. RESULTS FOR THE OPIAH SURVEY

To examine the performance of the dual-frame estimation methods in practice, we applied them to the dataset from the OPIAH survey. Several main variables are included in this study, related to the payment or not of medicines of the user. Table 1 shows the point and 95% confidence level estimation of proportion of people that belives that copay is a measure necessary for the maintenance of public health. The used estimators are: Kalton Anderson estimator (KA) ([4]), ranking ratio estimator (RR), Hartley estimator (HAR) and Skinner and Rao estimator (PML).

The confidence intervals computed are based on the pivotal method. This method yields a confidence interval for the population total as follows: $\hat{Y} \mp z_{\alpha/2}\sqrt{\hat{V}(\hat{Y})}$ where $z_{\alpha/2}$ is the critical value of a standard normal distribution.

The variance estimator for the Hartley estimator can be see in [3]. For the calculation of an unbiased estimator for the variance of the single-frame estimator KA and for the variance of the PML estimator, we adopted the approach proposed by [9], which provides two consistent estimators of variances. The variance for the single-frame RR estimator is determined using the residuals technique and Deville's method (see [1]).

Variance estimation methods exposed so far depend on each specific estimator. Instead, one can consider jackknife, which can be used to estimate variances irrespective of the type of estimator. Our results are similar and are not included in Table 1.

Table 1: Point and 95% confidence level estimation of proportions people ....

	Estimator	Lower limit	Upper limit	length
HAR	43.44	37.94	48.93	10.99
PML	45.66	39.06	52.26	13.20
KA	46.51	41.00	52.02	11.02
RR	41.31	37.57	45.06	7.49

From this study we obtained the following findings:

- There are important differences between the estimates produced with each dual-frame method.
- Among all the estimation strategies, the ranking ratio method performs best, and produces the smallest confidence interval

From the numerical results our recommendation is to use the RR estimator when the population size of both frames are known. In addition, [10] and [11] show that raking may be beneficial in reducing nonresponse biases.

Finally, let us note that the results obtained in applying these methods in the OPIAH survey indicate that the pharmaceutical copayment is a change with low acceptance trough users and that currently only 37-45% of those surveyed state that the copayment is actually needed in Andalusian health system.

RECEIVED: SEPTEMBER, 2015.
REVISED: JANUARY, 2016.

REFERENCES

[1] DEVILLE, J. C. [1993]: **Estimation de la variance pour les enquêtes en deux phases** Manuscript, INSEE, Paris.

[2] FULLER, W. A. AND BURMEISTER, L. F. [1972]: Estimators for samples selected from two overlapping frames **Proceedings of social science section of The American Statistical Asociation**.

[3] HARTLEY, H. O. [1962]: Multiple frame surveys In **Proceedings of the Social Statistics Section, American Statistical Association**, pages 203–206.

[4] KALTON, G. AND ANDERSON, D. W. [1986]: Sampling rare populations **Journal of the Royal Statistical Society. Series A (General)**, 149:65–82.

[5] LOHR, S. L. [2009]: Multiple-frame surveys **Handbook of Statistics**, 29:71–88.

[6] LU, B., SAHR, T., IACHAN, R., DENKER, M., DUFFY, T., AND WESTON, D. [2013]: Design and analysis of dual-frame telephone surveys for health policy research **World Medical and Health Policy**, 5:217–232.

[7] METCALF, P. AND SCOTT, A. [2009]: Using multiple frames in health surveys **Statistics in Medicine**, 28, 10,:1512–1523.

[8] RANALLI, M., ARCOS, A., RUEDA, M., AND TEODORO, A. [2015]: Calibration estimation in dual-frame surveys **Statistical Methods and Applications**, First online: 01 September 2015:1–29.

[9] RAO, J. N. K. AND SKINNER, C. J. [1996]: Estimation in dual frame surveys with complex designs In **Proceedings of the Survey Method Section, Statistical Society of Canada**, pages 63–68.

[10] SKINNER, C. J. [1991]: On the efficiency of raking ratio estimation for multiple frame surveys **Journal of the American Statistical Association**, 86:779–784.

[11] SKINNER, C. J. AND RAO, J. N. K. [1996]: Estimation in dual frame surveys with complex designs **Journal of the American Statistical Association**, 91:349–356.