Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Statistical foundations of cell-phone surveys

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2009
•
VOLUME 35
•
NUMBER 1

Canada

by Kirk M. Wolter, Phil Smith and Stephen J. Blumberg

December 2010

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 203-215
Statistics Canada, Catalogue No. 12-001-X

Statistical foundations of cell-phone surveys

Kirk M. Wolter, Phil Smith and Stephen J. Blumberg [1]

Abstract

The size of the cell-phone-only population in the USA has increased rapidly in recent years and, correspondingly, researchers have begun to experiment with sampling and interviewing of cell-phone subscribers. We discuss statistical issues involved in the sampling design and estimation phases of cell-phone studies. This work is presented primarily in the context of a nonoverlapping dual-frame survey in which one frame and sample are employed for the landline population and a second frame and sample are employed for the cell-phone-only population. Additional considerations necessary for overlapping dual-frame surveys (where the cell-phone frame and sample include some of the landline population) are also discussed. We illustrate the methods using the design of the National Immunization Survey (NIS), which monitors the vaccination rates of children age 19-35 months and teens age 13-17 years. The NIS is a nationwide telephone survey, followed by a provider record check, conducted by the Centers for Disease Control and Prevention.

Key Words: Cell-phone study; Random digit dialing; Dual-frame survey; Network sampling; Indirect sampling; Linking rules; Weighting of survey data; National Immunization Survey.

1. Introduction

The number of persons with cell phones in the USA has increased rapidly in recent years, and the percent of adults living in households with cell phones is expected to soon exceed the percent living in households with landlines (CTIA 2008; Blumberg and Luke 2008; Arthur 2007; Ehlen and Ehlen 2007). Correspondingly, survey researchers have begun to experiment with the sampling and interviewing of cell-phone subscribers (Lavrakas, Shuttles, Steeh and Fienberg 2007). This article is about the issues of statistical design and estimation that arise in cell-phone surveys. It emphasizes theoretically rigorous but practical solutions to the emergent problems survey researchers are facing in cell-phone surveys today.

Standard telephone surveys driven by random-digit-dialing (RDD) sampling only cover the population of households that have at least one working landline telephone actually used for voice communications. In an RDD survey, one assumes that the landline telephone is a household appliance and that all persons in the population are attached to one and only one household. Thus, one can sample people indirectly by sampling their telephone numbers and proceed from there to use reasonably standard and well-known methods of estimation.

The cell-phone survey brings a paradigm shift and new challenges. Most people think of the cell phone as a personal appliance, not a household device. Some people do share a cell phone, including 10-20 percent of cell-phone-only adults (Carley-Baxter, Peytchev and Lynberg 2008), but many do not, and thus it cannot be assumed that all residents of a household can be reached through the same cell-phone line. Some residents of a household can be reached through more than one cell-phone line. Some residents can be reached only by a cell-phone line while others can be reached through both cell and landline telephones. Thus, in the cell-phone survey, the household may no longer provide the same unifying organization that it does in standard telephone surveys.

To address the growing risk of bias (due to under-coverage) in telephone surveys, one can consider dual-frame telephone survey designs that include both an RDD sample of landline telephones and a sample of cell-phone lines. The telephone numbers on the two sampling frames are non-overlapping, but the corresponding people and households that may be the objects of the survey are partially overlapping.

A rigorous theory of estimation for such telephone survey designs has been lacking, although some initial descriptions of weighting have been advanced by Brick, Dipko, Presser, Tucker and Yuan (2006), Brick, Edwards and Lee (2007), and Frankel, Battaglia, Link and Mokdad (2007). In this article, we provide a general theory of unbiased estimation for population totals in the context of dual-frame telephone survey designs and derive the corresponding survey weights. We show what information must be collected in the survey itself to enable the calculation of the sampling weights.

To introduce ideas, we let A signify the portion of the overall population of interest accessible through the landline sampling frame, let B denote the portion accessible through the cell-phone sampling frame, and let C denote the portion not accessible through either frame (the *phoneless population* and other relatively small components of the

[1] Kirk M. Wolter, NORC and the University of Chicago. E-mail: wolter-kirk@norc.org; Phil Smith, National Center for Immunization and Respiratory Diseases; Stephen J. Blumberg, National Center for Health Statistics.

total population). We let a be the subpopulation in A not accessible through cell-phone lines (the *landline-only population*), let b be the subpopulation in B not accessible through landlines (the *cell-phone-only population*), and let ab be the subpopulation accessible through both landlines and cell-phone lines (the *mixed population*). We will sharpen this notation in succeeding sections.

Whether or not a unit in the population of interest is accessible through landlines or cell-phone lines is itself a complex matter. Throughout this article, when we say that a unit is accessible through landlines, we shall mean that there is both physical access to one or more landlines (usually residential landlines only) and a respondent would actually answer the landline if it rang for voice communications. Many adults today maintain a landline telephone strictly for computer communications and utilize a cell phone for all voice communications. By our definition, such adults are not considered to have landline access and instead are considered to be in the cell-phone-only population. Similarly, when we say that a unit is accessible through cell-phone lines, we shall mean that there is both physical access to a cell phone and intent to answer the cell phone if it rang. All other units in the population of interest that are not accessible through either landlines or cell-phone lines are considered phoneless. Current evidence suggests, although no one knows for sure, that about 20 to 30 percent of adults are domain b, 5 to 10 percent are in domain C, and the balance are spread across domains a and ab.

What we know so far from the cell-phone surveys we and others have conducted is that the data collection is relatively expensive, with average-interviewer-hours-per-completed case running around three times the average for standard RDD surveys. The higher cost is brought, in part, by the legal requirement (in the US, the Telephone Consumer Protection Act) of manually dialing the selected cell-phones. Response rates are somewhat lower than those achieved in RDD surveys. Interview length may be problematic, with some respondents less willing to submit to a lengthy interview by cell phone than by landline phone. Privacy issues may constrain the cell-phone interview, if the respondent is not in a private place at the time of the interview. The cell-phone user's propensity to respond may vary monotonically with his or her level of use of the cell phone, with the heavy user more willing to answer the phone than the lighter or occasional user. Most breakoffs occur during the opening seconds of the interview attempt. Because cell-phone surveys are relatively new, people are not used to being called and the interviewer has mere seconds to sell the survey. On the other hand, we find many cell-phone respondents to be quite cooperative once their attention has been held through the survey's introductory script.

Due to all of these circumstances in the environment, we currently view the cell-phone sample as a relatively small supplementary sample, with the main sample continuing to be a larger RDD sample of landlines. The cell-phone sample is intended to round out the coverage of the population of interest. In the future, as the environment matures and if costs come down, it may be possible to shift towards a more balanced approach with similarly sized landline and cell-phone samples, or even to a state where the cell-phone sample begins to dominate and the landline sample is used as a supplement to round out coverage.

In Section 2, we introduce the topic of *networks* of *sampling units*, *reporting units*, and *estimation units* and show how cell-phone surveys equate to a sampling of networks. Section 3 introduces various key concepts that will be needed as we discuss survey estimation, among them being the idea of a *link* (or edge) between the *nodes* (or vertices) in the network. Section 4 describes the duality that exists between the populations corresponding to the different types of nodes. Our approach will remind some readers of Lavallée's (2007) methods for indirect sampling. The heart of the paper is Section 5, which sets forth unbiased estimators of population totals for cell-phone surveys and for corresponding dual-frame telephone survey designs. Section 6 gives an example, illustrating implications of the new methods of estimation for an existing telephone survey regarding the vaccination coverage of young children and teenagers. We close in Section 7 with a brief summary.

Throughout the article, we emphasize the development of rigorous but practical design and estimation procedures for population B. The methods of RDD surveys, *i.e.*, the methods for population A, are well known and, to a degree, have been used for decades; for a recent review of these methods see Wolter, Chowdhury and Kelly (2008).

2. Networks of units and the response protocol

In general, at least three types of units arise in the context of a cell-phone survey, as follows:

- Sampling units (SU)
- Reporting units (RU)
- Estimation units (EU).

The SU is the unit of sampling in the survey. In actual practice, telephone numbers may be sampled directly from cell-phone frames, or they may be sampled in stages, with perhaps exchanges or banks of numbers serving as the primary sampling units and numbers themselves being selected in one or more stages of subsampling within the primary units. To keep the discussion simple, in this article we will present the telephone number itself as the SU.

The actual target of the survey interview and the unit of analysis is what we shall call the EU. Some surveys focus on the collection and analysis of data on households or families, in which case the household or family is the EU. Other surveys focus on person level data, where the eligible persons may be children under age 18, adults age 18+, or some demographic segment of the population, such as Hispanic females aged 0-34. Still other surveys focus on both household- and person-level data, in which case the survey involves at least two types of EUs and two levels of analysis.

The adult is the respondent or RU in telephone surveys. The EU may or may not have the capacity to respond directly for itself, and instead an RU responds on its behalf. If the EU is an adult, then the same adult or even a different adult may serve as the corresponding RU. If the EU is a household, family, consumer unit, or child, then one or more adults may serve as the corresponding RU. The response protocol, specified by the survey methodologist, actually determines which RUs are permitted to respond for which EUs. In a typical survey, one respondent adult (or RU) would be contacted by telephone and interviewed for each SU selected into the sample.

SUs, RUs, and EUs may bear different relationships to one another in a cell-phone survey. Figure 1 gives nine networks that illustrate some of the types of relationships that are possible. In the first network, one SU is linked to one RU, which in turn responds for one EU. This arrangement could occur if one adult uses one telephone line, and the adult in turn reports for the household or for him or herself or for one child. In the second network, one SU is linked to two RUs, each of which can respond for the EU. This arrangement would occur, for example, if two adults shared the same telephone line and each was permitted by survey protocol to respond for the household. The fifth network could occur if two adults each had their own telephone line not shared with the other adult, while each adult in the pair is allowed by survey protocol to respond for each of two children.

More complicated networks are possible and surely must exist in the world. For example, the eighth network shows an arrangement of three adults sharing two telephone lines. The first of the lines is shared by all three adults, while the second line is only used by the third adult. The first of the adults is permitted by survey protocol to respond for two EUs, such as the adult's biological children; the second adult is not permitted to respond for any EUs; and the third adult is permitted to respond only for a third EU that is not reportable by the first two adults.

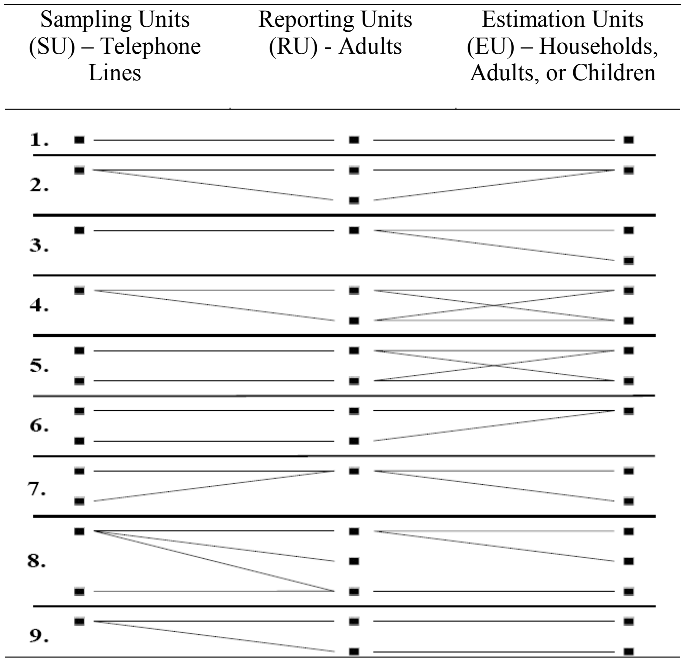
Sampling Units (SU) – Telephone Lines
Reporting Units (RU) - Adults
Estimation Units (EU) – Households, Adults, or Children
1.
2.
3.
4.
5.
6.
7.
8.
9.

Figure 1 Examples of networks in a cell-phone survey

3. Links between units in the network

A *link* is a salient relationship between two nodes in the network. In the context of Figure 1, the links are represented by the line segments that join the different nodes. To provide a foundation for survey estimation, we need to explore links between (i) RUs and SUs, (ii) EUs and RUs, (iii) and EUs and SUs.

3.1 Link of RU and SU

Two concepts are central to creating a link between an RU and an SU, namely, the concepts of (a) an *Active Personal Cell Number* (*APCN*) and (b) *usual access* to the cell-phone line.

An APCN is a telephone line that is in service at the time of the cell-phone survey and can ring through to an eligible adult who uses the cell phone, at least partially, for personal matters. In other words, an APCN meets three tests:

- It is in service
- It connects to an eligible adult respondent
- It is not used exclusively for business purposes.

We say that a given adult has usual access to a given APCN if and only if the individual has

- Regular,
- Substantial, and
- Ongoing use of the cell-phone line.

Each APCN has one or more regular adult users, and each individual user has usual access to one or more cell phones. In many cases, there is a unique one-to-one relationship between the cell-phone line and the adult user. In some cases, there is a one-to-many relationship between the cell-phone line and its users.

We treat a given SU and a given RU as linked if and only if the SU is an APCN and the RU has usual access to the SU. A cell-phone survey must work with and recognize the links that exist between the population of SUs and the population of RUs.

3.2 Link of EU and RU

A given EU is linked to one or more RUs via natural relationships that exist in the world, such as those created by family or place of residence. For example, an adult respondent may respond to the survey interview on behalf of his or her household, family, or consumer unit. He or she may respond for him or herself, for a dependent child under age 18, or for his or her own parent or sibling.

All surveys require a response protocol that defines which adult respondents are to respond for which EUs. The protocol is selected by the survey methodologist in light of feasibility, cost, and accuracy-of-reporting concerns. It is this protocol that establishes the links between EUs and RUs.

3.3 Link of EU and SU

The foregoing links between RUs and SUs and between EUs and RUs determine the links between EUs and SUs. We say a given EU is linked to a given SU if and only if the EU is linked to at least one RU that in turn is linked to the SU.

Some notation will become useful in our work in the following sections. Let j denote a given EU in the population of interest and let i be a given SU in the population. Then define the indicator or link variables

$$\begin{aligned} \ell_{ij} &= 1, \quad \text{if the } j^{\text{th}} \text{ EU is linked to the } i^{\text{th}} \text{ SU} \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

4. Duality between the populations of SUs and EUs

To begin the process of determining an unbiased estimation procedure for cell-phone surveys, we establish that a duality exists between the population of SUs or cell phones (henceforth denoted by U^{SB}) and the population of EUs that are linked to cell phones (denoted by U^{EB}). The goal of a cell-phone survey is to make inferences concerning U^{EB}, but we will soon see that this goal is equivalent to making certain inferences concerning U^{SB} (in this notation, the first superscript designates the type of unit while the superscript B refers to the cell-phone sampling frame. Later we will use the superscript A to signify the landline sampling frame).

In the EU domain, a population total of interest is given by

$$Y^{\mathrm{EB}} = \sum_{j \in U^{\mathrm{EB}}} Y_j,$$

where the Y-variable on the right-hand side is a questionnaire item or other recoded or derived variable attached to the units in the population U^{EB}. Similarly, in the SU domain, a population total is defined by

$$X^{\mathrm{SB}} = \sum_{i \in U^{\mathrm{SB}}} X_i,$$

where the X-variable on the right-hand side is any fixed characteristic attached to the units in the population U^{SB}.

While the interest of the survey analyst centers on the total from the population of EUs (and on other parameters of this population), one can obtain a corresponding parameter in the SU domain by writing

$$Y^{\mathrm{EB}} = \sum_{j \in U^{\mathrm{EB}}} Y_j = \sum_{j \in U^{\mathrm{EB}}} \sum_{i \in U^{\mathrm{SB}}} \frac{Y_j \ell_{ij}}{\sum_{i' \in U^{\mathrm{SB}}} \ell_{i'j}} = \sum_{i \in U^{\mathrm{SB}}} X_i = X^{\mathrm{SB}}, \quad (1)$$

where the X-variable is now defined specifically by

$$X_i = \sum_{j \in U^{\mathrm{EB}}} \frac{Y_j \ell_{ij}}{\sum_{i' \in U^{\mathrm{SB}}} \ell_{i'j}}. \quad (2)$$

From (1), one can see the correspondence between estimation in the SU domain and estimation in the EU domain. The total X^{SB}, with X_i defined as in (2), is equivalent to the total of interest Y^{EB}, and thus the problem of estimation of Y^{EB} is equivalent to the problem of estimation of X^{SB}.

We note that (2) arises in substantially the same form in the theory of indirect sampling. See Lavallée (2007), Theorem 4.1. In indirect sampling, SUs are linked to naturally defined clusters of EUs; if a given SU is selected into the sample, the survey data are collected for all EUs in the linked clusters. The analogy here is that the clusters are defined by the RUs that respond to the cell-phone interview attempt, and survey data are collected from the respondent for all EUs to which he or she is linked. The current situation is such that the cluster is defined by the SU-RU pair. An identifiability problem arises in this regard that does not occur in general in indirect sampling, and we elaborate on this matter in Section 5.5.

In (2), we effectively allocate an equal share of Y_j to each SU i to which it is linked. We could, alternatively, achieve the same ends by allocating Y_j to its linked SUs in proportion to some other known measure of the intensity of the relationship between j and i. Although one could conceive of an optimal allocation of Y_j to its linked SUs, as in Deville and Lavallée (2006), such an allocation may be difficult to execute or may not be of great import in large scale practical settings.

5. Estimation

As mentioned in the introduction, some EUs will be linked exclusively to cell phones, some will be linked exclusively to landlines, and some will be linked to both landlines and cell phones. Phoneless EUs, if any, will not be linked to cell phones or to landlines. To provide notation for this environment, let U^{E} be the overall population of EUs of interest, and let U^{S} be the overall population of SUs. Let U^{EA} be the elements of U^{E} that are linked to landlines, let U^{EB} be the elements that are linked to cell-phone lines, let U^{Ea} be the elements that are linked only to landlines, let U^{Eb} be the elements that are linked only to cell-phone lines, let U^{Eab} be the elements that are linked to both landlines and cell-phone lines, and let U^{EC} be the elements that are phoneless. Note that $U^{\mathrm{E}} = U^{\mathrm{EA}} \cup U^{\mathrm{EB}} \cup U^{\mathrm{EC}}$, $U^{\mathrm{EA}} = U^{\mathrm{Ea}} \cup U^{\mathrm{Eab}}$, and $U^{\mathrm{EB}} = U^{\mathrm{Eab}} \cup U^{\mathrm{Eb}}$, where U^{Ea}, U^{Eab}, and U^{Eb} are disjoint sets. Also, let U^{SA} be the population of landlines, such that $U^{\mathrm{S}} = U^{\mathrm{SA}} \cup U^{\mathrm{SB}}$. Landlines and cell-phone lines reflect disjoint subsets of the overall population of SUs.

In the following Sections 5.1 and 5.2, we discuss unbiased estimation for the subpopulation, say $U^{\mathrm{ET}} = U^{\mathrm{EA}} \cup U^{\mathrm{EB}}$, that is linked to at least one telephone of any kind. We use the super-script T to designate this telephone subpopulation. Subsequently, in Section 5.4, we briefly discuss coverage of the phoneless population.

For EUs in U^{E}, define the indicator variables

$$\delta_j = 1, \text{ if none of the RUs linked to } j \text{ have access to landline service, while at least one of these RUs has usual access to cell-phone service}$$
$$= 0, \text{ otherwise}$$

$$\phi_j = 1, \text{ if none of the RUs linked to } j \text{ have usual access to cell-telephone service, while at least one of these RUs has access to landline service}$$
$$= 0, \text{ otherwise.}$$

The δ-variable is an indicator of cell-phone-only status and the ϕ-variable is an indicator of landline-only status.

Then the population total of interest may be decomposed as

$$Y^{\mathrm{ET}} = Y^{\mathrm{EA}} + Y^{\mathrm{Eb}}, \tag{3}$$

where

$$Y^{\mathrm{Eb}} = \sum_{j \in U^{\mathrm{ET}}} \delta_j Y_j$$

is the total of the cell-phone-only domain, and

$$Y^{\mathrm{EA}} = \sum_{j \in U^{\mathrm{ET}}} (1 - \delta_j)\, Y_j$$

is the total of the complement of this domain, including EUs that are linked exclusively to landlines and mixed EUs that are linked to both landlines and cell phones. The total of EUs may also be written as

$$Y^{\mathrm{ET}} = Y^{\mathrm{Ea}} + Y^{\mathrm{Eab}} + Y^{\mathrm{Eb}}, \tag{4}$$

where

$$Y^{\mathrm{Ea}} = \sum_{j \in U^{\mathrm{ET}}} \phi_j Y_j$$

is the total of the landline-only population, and

$$Y^{\mathrm{Eab}} = \sum_{j \in U^{\mathrm{ET}}} (1 - \delta_j)\,(1 - \phi_j)\, Y_j$$

is the total of the mixed population that has a combination of landline and cell-phone access. Finally, the population total may be written as

$$Y^{\mathrm{ET}} = Y^{\mathrm{Ea}} + Y^{\mathrm{EB}}, \tag{5}$$

where

$$Y^{\mathrm{EB}} = \sum_{j \in U^{\mathrm{ET}}} (1 - \phi_j) Y_j$$

is the total of the complement (in the telephone population) of the landline-only population.

We view (3) and, to some extent, (4) as the decompositions of current practical interest and importance in telephone surveys in the USA and, in what follows, we present methods of estimation for each. Because of the current high relative cost of cell-phone interviews, surveys based on decomposition (5) would not be cost effective. It would almost always be better to represent the domain U^{Eab} using a sample of landlines than using a sample of cell phones. If the relative cost of cell-phone interviewing shifts downward in the future, decomposition (5) could become economically viable. It may also be viable for surveys in other countries where the cost structure is more favorable to cell-phone interviews.

5.1 Case of nonoverlapping domains

In this section, we will use a sample of cell-phone lines for purposes of estimation for the cell-phone-only population U^{Eb} and a sample of landlines for estimation for the entire landline population U^{EA}. We observe that it is not possible to directly select a sample of cell-phone-only lines, because cell-phone-only status is not available on the sampling frame but rather is determined in the survey screening interview. To operationalize this design, one would screen-out cell-phone respondents who classify themselves in the mixed domain and terminate the interview, continuing the interview only for cell-phone-only respondents.

Let s^{SB} denote a probability sample of SUs (cell-phone lines) selected from the population U^{SB}, and let $\{W_i^{\mathrm{SB}}\}$ denote the set of base sampling weights such that

$$\hat{X}^{\mathrm{SB}} = \sum_{i \in s^{\mathrm{SB}}} W_i^{\mathrm{SB}} X_i$$

is an unbiased estimator of the population total X^{SB}, where X_i is a characteristic of the i^{th} unit in the population. Assuming simple random sampling without replacement within strata, the base weights are of the form

$$W_i^{\mathrm{SB}} = N_h / n_h, \tag{6}$$

where h signifies the sampling stratum in which the i^{th} SU is selected, N_h is the number of SUs on the sampling frame in stratum h, and n_h is the sample size in stratum h. Typically, the cell-phone sampling frame would include all telephone numbers within the exchanges assigned by the telephone system to cell phones. Simple random sampling would be the most common method of sample selection from such exchanges. There is little information available on the cell-phone sampling frame to enable stratification of the sample, except for the coarse geographic information embodied within the area code.

Let s^{EB} be the corresponding sample of EUs, *i.e.*, $s^{\mathrm{EB}} = \{j \in U^{\mathrm{EB}} \mid j$ is linked to at least one SU i in $s^{\mathrm{SB}}\}$. We will use this sample to estimate the domain total of EUs that are linked only to a cell phone, Y^{Eb}. From (1) and (2), we can readily see that the unbiased estimator of the domain total is given by

$$\begin{aligned}\hat{Y}^{\mathrm{Eb}} &= \sum_{i \in s^{\mathrm{SB}}} W_i^{\mathrm{SB}} \left\{ \sum_{j \in U^{\mathrm{EB}}} \delta_j Y_j \ell_{ij} \Big/ \sum_{i' \in U^{\mathrm{SB}}} \ell_{i'j} \right\} \\ &= \sum_{j \in s^{\mathrm{EB}}} \delta_j Y_j W_j^{\mathrm{EB}},\end{aligned} \tag{7}$$

where the EU level sampling weights are defined by

$$W_j^{\mathrm{EB}} = \sum_{i \in s^{\mathrm{SB}}} W_i^{\mathrm{SB}} \ell_{ij} \Big/ \sum_{i' \in U^{\mathrm{SB}}} \ell_{i'j} \quad \text{for } j \in s^{\mathrm{EB}}. \tag{8}$$

Again, see Lavallée (2007) for expression of these weights in the context of indirect sampling.

Before leaving domain b, we observe in passing that it is possible to subsample the EUs and collect the survey information only for the subsample instead of enumerating all EUs linked to the sample RUs. If the statistician would choose some form of subsampling, perhaps to control sample size or cost, then an additional weighting factor would appear in the weights in (8). Such subsampling is referred to as two-stage indirect sampling in Lavallée (2007, Section 5.1).

Turning to domain A, let s^{SA} denote a standard RDD sample of landline telephones, let s^{EA} be the implied sample of EUs, *i.e.*, $s^{\mathrm{EA}} = \{j \in U^{\mathrm{EA}} \mid j$ is linked to at least one SU i in $s^{\mathrm{SA}}\}$, and let

$$\hat{Y}^{\mathrm{EA}} = \sum_{j \in s^{\mathrm{EA}}} W_j^{\mathrm{EA}} Y_j \tag{9}$$

be the standard unbiased estimator of the population total. For brevity, we shall not derive the standard sampling weights here; for more information about these weights, see Wolter *et al.* (2008).

From (7) and (9), the unbiased estimator of the population total of the EUs is given by

$$\hat{Y}^{\mathrm{ET}} = \hat{Y}^{\mathrm{EA}} + \hat{Y}^{\mathrm{Eb}} \tag{10}$$

and the weights needed to support this estimator are $\{W_j^{\mathrm{EA}}\}$ and $\{W_j^{\mathrm{EB}}\}$.

5.2 Case of overlapping domains

We now proceed with estimation starting from the decomposition (4). This means that in the cell-phone sample we will interview not only the cell-phone-only population, but also the mixed population (*i.e.*, those that use both landline and cell telephones). The estimator of the population total of interest is now of the form

$$\hat{Y}^{\mathrm{ET}} = \hat{Y}^{\mathrm{Ea}} + \hat{Y}^{\mathrm{Eab}} + \hat{Y}^{\mathrm{Eb}}, \tag{11}$$

where

$$\hat{Y}^{\mathrm{Ea}} = \sum_{j \in s^{\mathrm{EA}}} W_j^{\mathrm{EA}} \phi_j Y_j$$

is the estimator for the landline-only domain derived from the landline sample, $\hat{Y}^{\mathrm{Eb}}$ is defined in (7) and is the estimator for the cell-phone-only domain derived from the cell-phone sample, and $\hat{Y}^{\mathrm{Eab}}$ is an estimator of the mixed domain obtained from both samples. The estimator of the mixed domain is

$$\begin{aligned}\hat{Y}^{\mathrm{Eab}} &= \lambda \sum_{j \in s^{\mathrm{EA}}} W_j^{\mathrm{EA}} (1 - \phi_j) Y_j \\ &\quad + (1 - \lambda) \sum_{j \in s^{\mathrm{EB}}} W_j^{\mathrm{EB}} (1 - \delta_j) Y_j.\end{aligned} \tag{12}$$

The weights need to support estimator (11) are $\{W_j^{\mathrm{EA}}\}$ and $\{W_j^{\mathrm{EB}}\}$.

See Hartley (1962) for discussion of the mixing parameter λ in a dual-frame survey, focusing on considerations of sampling variability. Turning to considerations of bias, Brick *et al.* (2006) report that the propensity to respond to a cell-phone survey may be positively related to the frequency of use of the cell phone. Thus, the two pieces on the right side of (12) may be subject to a differential nonresponse bias not removed by the standard weighting-class methods. In the mixed population, infrequent users of the cell phone may be less likely to respond if surveyed in the cell-phone sample than if surveyed in the landline sample. If these adults would be substantially different from other adults in the mixed population with respect to the key characteristics under study in the survey, then (12) and also (11) could be subject to a nonreponse bias.

5.3 Variance estimation

To make inferences from the sample to the overall population, we require an estimator of the variance of the estimated total. First, consider the case of nonoverlapping domains. By working in the SU population, we can employ methods of variance estimation appropriate to the survey design. From (7), the estimated total for the cell-phone only domain may be written by

$$\hat{Y}^{\text{Eb}} = \sum_{i \in s^{\text{SB}}} W_i^{\text{SB}} X_i,$$

where

$$X_i = \sum_{j \in U^{\text{EB}}} \delta_j Y_j \ell_{ij} \Big/ \sum_{i' \in U^{\text{SB}}} \ell_{i'j}. \tag{13}$$

Assuming simple random sampling, the unbiased estimator of the variance of the estimated total is given by

$$v(\hat{Y}^{\text{Eb}}) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} s_{xh}^2,$$

where

$$s_{xh}^2 = \frac{1}{n_h - 1} \sum_{i \in s_h^{\text{SB}}} \left(X_i - \frac{1}{n_h} \sum_{i' \in s_h^{\text{SB}}} X_{i'} \right)^2.$$

If we would ignore the finite population correction factor, which would be possible in almost any real telephone survey, the variance estimator becomes

$$v(\hat{Y}^{\text{Eb}}) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i \in s_h^{\text{SB}}} \left(W_i^{\text{SB}} X_i - \frac{1}{n_h} \sum_{i' \in s_h^{\text{SB}}} W_{i'}^{\text{SB}} X_{i'} \right)^2. \tag{14}$$

Now let $v(\hat{Y}^{\text{EA}})$ be an estimator of the variance of $\hat{Y}^{\text{EA}}$ for the RDD sample of landlines. Such estimators are well known and we do not review them here; see for example, Wolter *et al.* (2008). Because sampling is independent in the landline and cell-phone sampling frames, the unbiased estimator of the variance of the estimated total for the entire telephone population becomes

$$v(\hat{Y}^{\text{ET}}) = v(\hat{Y}^{\text{EA}}) + v(\hat{Y}^{\text{Eb}}). \tag{15}$$

To facilitate the following developments, we let $\hat{V}^{\text{EB}}[\delta Y]$ be another symbol to represent the estimator of variance in (14). This notation will emphasize the fact that the estimator of variance is based on the X_i variable in (13) defined in terms of the characteristic $\delta_j Y_j$, which is the characteristic of interest for cell-phone-only EUs. Also, let the symbol $\hat{V}^{\text{EA}}[Y]$ be the estimator $v(\hat{Y}^{\text{EA}})$ defined in terms of the characteristic Y_j. With this notation, (15) becomes $v(\hat{Y}^{\text{ET}}) = \hat{V}^{\text{EA}}[Y] + \hat{V}^{\text{EB}}[\delta Y]$.

Second, consider variance estimation for the case of overlapping domains. The estimator of the total of the telephone population is now $\hat{Y}^{\text{ET}}$ in (11). For fixed λ, the unbiased estimator of variance is clearly seen from the work done in (14) and (15). It is

$$\begin{aligned} v(\hat{Y}^{\text{ET}}) = {} & \hat{V}^{\text{EA}}[\phi Y + \lambda(1-\phi) Y] \\ & + \hat{V}^{\text{EB}}[\delta Y + (1-\lambda)(1-\delta) Y]. \end{aligned} \tag{16}$$

The first term on the right side of (16) is the variance estimator for the RDD sample of landlines applied to the composite characteristic $\phi_j Y_j + \lambda(1-\phi_j) Y_j$, which is the characteristic for landline-only EUs plus a λ-portion of the characteristic for mixed EUs. The second term on the right side of (16) is the variance estimator for the cell-phone sample applied to the composite characteristic $\delta_j Y_j + (1-\lambda)(1-\delta_j) Y_j$, which is the characteristic for cell-phone-only EUs plus a $(1-\lambda)$-portion of the characteristic for mixed EUs.

Estimators of covariance matrices can be built up from expressions like (15) and (16), facilitating statistical inference concerning other population parameters of interest.

5.4 Adjustments of the sampling weights

The sampling weights may be adjusted because of nonresponse or a planned calibration to known control totals.

Thus far, we have not addressed the various types of missing data that may occur in a cell-phone survey. We will focus on deriving adjustments for missing data that arise during the cell-phone interviews, assuming that standard adjustments for missingness in the landline sample have already been incorporated in the $\{W_j^{\text{EA}}\}$ weights.

Missing data can arise due to three factors: (i) *nonresolution* of the SU; (ii) an *incomplete screening interview* of the RU; and (iii) an *incomplete main interview* of the RU. In this article, we adopt the convention that the resolution step refers to the classification of the SU as an ACPN or something else, such as a disconnected line or a dedicated business line; nonresolved SUs and SUs resolved as non-ACPNs do not continue with the interview. The screening step refers to a brief preliminary interview intended to ascertain telephone status and to determine any demographic or other eligibility characteristics of any EUs linked to the RU; RUs for which the screening interview is incomplete or for which the screening interview is complete but no eligible EUs are linked to the RU do not continue with the interview. If the survey protocol calls for including only cell-phone-only EUs, as in Section 5.1, then the interview would terminate at this point for any mixed EUs. On the other hand, if the survey protocol calls for including both cell-phone-only and mixed EUs, as in Section 5.2, then the interview would continue for all such EUs. The interview step refers to the collection of the main survey items that form the substance of the survey for each of the eligible EUs linked to the RU. The survey methodologist must institute a definition of what constitutes a completed interview. In particular, the methodologist must decide whether *breakoffs* (an interview attempt that is completed for some but not all of the eligible EUs linked to the RU) are to be treated as a completed interview or not. Some other authors may organize the steps in the survey response process somewhat differently than the convention adopted here.

Adjustments to the sampling weights can be made for nonresolution and screener nonresponse, assuming a missing-at-random model for the response mechanism. These two adjustments must be made at the SU level. Let $\{s_\alpha^{\mathrm{SB}}\}$ be a partition of the cell-phone sample into user-specified weighting cells α, and let the base sampling weights from (6) now be denoted by W_{1i}^{SB}, where the subscript 1 has been added simply to signify the first step in a multi-step adjustment process. Telephone area codes, rate centers, and census environmental variables at the county or area code level can be used to form the weighting cells; otherwise, little covariate information is available concerning cell-phone numbers. The cell-specific resolution completion rates are defined by

$$R_{1\alpha} = \frac{\sum_{i' \in s_\alpha^{\mathrm{SB}}} r_{1i'} W_{1i'}^{\mathrm{SB}}}{\sum_{i' \in s_\alpha^{\mathrm{SB}}} W_{1i'}^{\mathrm{SB}}},$$

where r_{1i} is a resolution indicator variable (= 1, if resolved, = 0, if not resolved), and the nonresolution adjusted weights are $W_{2i}^{\mathrm{SB}} = r_{1i} W_{1i}^{\mathrm{SB}} / R_{1\alpha}$ for $i \in s_\alpha^{\mathrm{SB}}$.

Let e_{1i} be an indicator of whether i is a resolved APCN (= 1, if resolved APCN, = 0, otherwise), and let $\{s_\beta^{\mathrm{SB}}\}_{\beta=1}^{B}$ be a partition of the cell-phone sample into user-specified weighting cells, which could be the same as or different than the foregoing partition. Then, the cell-specific screener completion rates are

$$R_{2\beta} = \frac{\sum_{i' \in s_\beta^{\mathrm{SB}}} r_{2i'} e_{1i'} W_{2i'}^{\mathrm{SB}}}{\sum_{i' \in s_\beta^{\mathrm{SB}}} e_{1i'} W_{2i'}^{\mathrm{SB}}},$$

where r_{2i} is a screener indicator variable (= 1, if screener completed, = 0, if screener not completed), and the screener-nonresponse adjusted weights are $W_{3i}^{\mathrm{SB}} = r_{2i} e_{1i} W_{2i}^{\mathrm{SB}} / R_{2\beta}$ for $i \in s_\beta^{\mathrm{SB}}$. Note that the appropriate sum of the weights is preserved at each step of the adjustment process.

Next, an adjustment to the sampling weights must be made for interview nonresponse. Depending on how breakoffs are classified by the survey methodologist, there may be two cases to consider: (i) the RU completes or fails to complete the interview for all of its linked and eligible EUs en masse, or (ii) the RU selectively completes or fails to complete the interview on an EU by EU basis. If breakoffs would be classified as incomplete interviews, then only Case i would apply. Let e_{2i} be an indicator of whether the RU is screened and is linked to at least one EU that is eligible for the interview (= 1, if screened and eligible, = 0, otherwise), and let r_{3i} be the interview indicator variable (= 1, if the interview is complete, = 0, otherwise).

For Case i, the weight adjustment can be made at the SU level and is given by $W_{4i}^{\mathrm{SB}} = r_{3i} e_{2i} W_{3i}^{\mathrm{SB}} / R_{3\gamma}$ for $i \in s_\gamma^{\mathrm{SB}}$, where $R_{3\gamma}$ is the weighted interview completion rate computed within user-specified weighting cells γ. Again, options for constructing weighting cells are limited in a cell-phone survey; they may be specified in terms of the information available at the previous weighting steps or any information collected in the screening interview. The weighted interview completion rate is

$$R_{3\gamma} = \frac{\sum_{i' \in s_\gamma^{\mathrm{SB}}} r_{3i'} e_{2i'} W_{3i'}^{\mathrm{SB}}}{\sum_{i' \in s_\gamma^{\mathrm{SB}}} e_{2i'} W_{3i'}^{\mathrm{SB}}}.$$

The estimated total for the cell-phone-only domain may now be expressed by

$$\hat{Y}^{\mathrm{Eb}} = \sum_{j \in s^{\mathrm{EB}}} \delta_j Y_j W_{4j}^{\mathrm{EB}}, \tag{17}$$

where

$$W_{4j}^{\mathrm{EB}} = \sum_{i \in s^{\mathrm{SB}}} W_{4i}^{\mathrm{SB}} \ell_{ij} \Big/ \sum_{i' \in U^{\mathrm{SB}}} \ell_{i'j}$$

and s^{EB} is the set of eligible EUs reported in the screening interviews. The weight is zero for any eligible EUs in s^{EB} for which the RU failed to complete the main interview. The estimated total for the mixed domain, if called for by the survey protocol, is defined similarly by

$$\hat{Y}^{\text{Eab}} = \lambda \sum_{j\in s^{\text{EA}}} W_j^{\text{EA}}(1-\varphi_j)Y_j + (1-\lambda)\sum_{j\in s^{\text{EB}}} W_{4j}^{\text{EB}}(1-\delta_j)Y_j.$$

For Case ii, the noninterview adjustment must be made at the EU level. The EUs are treated as spawned cases and a decision is made for each one as to whether it has a completed interview or not. The estimated total for the cell-phone-only domain is (17), where the weight is now defined by

$$W_{4j}^{\text{EB}} = r_{3j}\, e_{2j}\, W_{3j}^{\text{EB}} / R_{3\gamma} \quad \text{for} \quad j \in s^{\text{EB}},$$

$$W_{3j}^{\text{EB}} = \sum_{i\in s_3^{\text{SB}}} W_{3i}^{\text{SB}} \ell_{ij} \Big/ \sum_{i'\in U^{\text{SB}}} \ell_{i'j},$$

and

$$R_{3\gamma} = \frac{\sum_{j'\in s_\gamma^{\text{EB}}} r_{3j'} W_{3j'}^{\text{EB}}}{\sum_{j'\in s_\gamma^{\text{EB}}} W_{3j'}^{\text{EB}}}.$$

Here, the weighting cells, γ, are defined in terms of characteristics of the EUs as determined from the screening interview and other sources.

For either Case i or ii, to facilitate computations, take W_{4j}^{EA} to be defined and equal to zero for EUs in the cell-phone sample, and take W_{4j}^{EB} to be equal to zero for EUs in the landline sample. If the survey protocol is as in Section 5.1, then we conclude that the survey weights for estimating the population total of interest are defined by

$$W_j = W_{4j}^{\text{EA}} + W_{4j}^{\text{EB}} \delta_j \tag{18}$$

for $j \in s^{\text{ET}}$, where $s^{\text{ET}} \in s^{\text{EA}} \cup s^{\text{EB}}$. Otherwise, if the survey protocol is as in Section 5.2, then we conclude that the survey weights are defined by

$$W_j = W_{4j}^{\text{EA}}\{\phi_j + \lambda(1-\phi_j)\} + W_{4j}^{\text{EB}}\{\delta_j + (1-\lambda)(1-\delta_j)\} \tag{19}$$

for $j \in s^{\text{ET}}$.

The nonresponse-adjusted weights from (18) or (19) may be calibrated (Deville and Särndal 1992) to external control totals within socio-economic or geographic cells for the population of EUs, using poststratification, raking, or GREG (generalized regression estimation) techniques. If accurate sources are available, control totals may be established and calibration may be conducted separately for domains *A* and *b* or for domains *a*, *ab*, and *b*. If control totals are not available by telephone status, then calibration must use control totals for the entire population regardless of telephone status.

To illustrate these ideas, we briefly examine the GREG estimator. Let us suppose that we have available a $1 \times p$ auxiliary variable $\mathbf{Z}_j$ for the observed, eligible EUs for which the control totals $\mathbf{Z}^{\text{ET}} = \sum_{j\in U^{\text{ET}}} \mathbf{Z}_j$ are known. For example, the *z*-variable may arise from a fully saturated model in terms of explanatory variables age, race, and sex. Let s_4^{ET} be the set of EUs with a completed main interview and let $n_4^{\text{ET}} = \#(s_4^{\text{ET}})$ be the number of eligible EUs reported in the completed interviews obtained within the consolidated telephone sample. Stack the *y*-values, *z*-values, and weights into the matrices $\mathbf{Y} = (Y_1, ..., Y_{n_4^{\text{ET}}})', \mathbf{Z} = (\mathbf{Z}'_1, ..., \mathbf{Z}'_{n_4^{\text{ET}}})'$, and $\mathbf{W} = \text{diag}(W_1, ..., W_{n_4^{\text{ET}}})'$. Then the GREG estimator (Cassel, Särndal, and Wretman 1976) of the total of the telephone population of interest takes the familiar form

$$\tilde{Y}^{\text{ET}} = \hat{Y}^{\text{ET}} + (\mathbf{Z}^{\text{ET}} - \hat{\mathbf{Z}}^{\text{ET}})\,\hat{\beta} = \sum_{j\in s_4^{\text{ET}}} W_j\, g_j\, Y_j,$$

where the estimated coefficients are given by $\hat{\beta} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}\mathbf{Y}$, $\hat{Y}^{\text{ET}} = \sum_{j\in s_4^{\text{ET}}} W_j Y_j$, $\hat{\mathbf{Z}}^{\text{ET}} = \sum_{j\in s_4^{\text{ET}}} W_j \mathbf{Z}_j$, and $g_j = 1 + (\mathbf{Z}^{\text{ET}} - \hat{\mathbf{Z}}^{\text{ET}})\mathbf{Z}'_j$. Lavallée (2007, Chapter 7) derives the Taylor series estimator of the variance of the GREG estimator in an indirect sampling context. Also see Wolter (2007, Chapter 6) for estimation of the variance of the GREG estimator.

Before leaving the topic of calibration, we note that we have largely left aside the small phoneless population, which fundamentally is impossible to sample in a telephone survey. Yet, in all likelihood, the overall population total $Y^{\text{E}} = Y^{\text{ET}} + Y^{\text{EC}}$ will be the parameter of interest, not the total of the telephone population Y^{ET}, and the known control totals used in calibration may be totals for the overall population $\mathbf{Z}^{\text{E}} = \mathbf{Z}^{\text{ET}} + \mathbf{Z}^{\text{EC}}$, not totals for the telephone population $\mathbf{Z}^{\text{ET}}$. To include the phoneless population, we may consider use of a revised GREG estimator with $g_j = 1 + (\mathbf{Z}^{\text{E}} - \hat{\mathbf{Z}}^{\text{ET}})\mathbf{Z}'_j$. This revision takes the same model for the phoneless population as for the telephone population. See Keeter (1995) and Chowdhury, Montgomery and Smith (2008) for other considerations in the calibration of weights for the phoneless population.

5.5 Identifiability assumptions

The foregoing theory assumes fundamentally that if SU *i* is selected into the sample of cell-phone lines, then X_i defined in (2) is observable in the cell-phone interview. Yet the 9th network (and also the 8th) in Figure 1 illustrates a potential problem for the theory. For this network, two RUs are linked to one SU, and in turn each RU is linked to only one EU. To continue this illustration, we suppose that these two EUs are not linked to any other RUs in the population. At the time of the survey interview, only one of the RUs will typically be reached and interviewed (unless the survey protocol would specifically mandate that an interview be attempted with each RU linked to the selected SU). The respondent RU will report for its linked EU, but by the very nature of this network, the respondent cannot report for the EU that is linked to the companion RU who shares the sample cell-phone line. Thus, there is at least one EU that is linked to the SU that cannot be observed, *i.e.*, data cannot be collected in the cell-phone interview. Thus, we say X_i is *not identifiable*. The situation regarding the reportability of the two EUs would be reversed if the cell-phone interview attempt would have rung through to the companion RU.

To maintain the unbiasedness of the estimator of the population total, the X_i must be identifiable for every respondent SU selected into the sample of cell-phone lines. We need to make one of two assumptions. First, we could assume the problem away by acting as if networks like numbers 8 and 9 either do not exist or are trivial in number.

Secondly, the more realistic case would be to assume an extra randomization step, namely, that the interview call attempt to the given SU has reached a randomly selected RU linked to the SU. This randomization could be viewed as conceptual (that is, occurring naturally and not directed by the survey methodologist). To be formal and rigorous, one would need to collect information on the number of RUs linked to the SU and the probability that the cell-phone call attempt would ring through to the respondent RU. The probability would be approximated by the respondent's self-report of his or her share of use of the cell phone. If only one RU is linked to the SU, then this probability is 1.0 and clearly this simple value would not need to be collected in the interview once it is reported that there is only one RU. If two or more RUs are linked to the SU, then the probability or share to be collected is denoted by τ_{ik} for RUs indexed by *k*, where $\sum_{k\in U_i^{\text{RB}}} \tau_{ik} = 1$ and U_i^{RB} is the set of RUs that are linked to the i^{th} SU. With this additional information in hand, an unbiased estimator of

$$X_i = \sum_{j\in U^{\text{EB}}} \frac{\delta_j\, Y_j\, \ell_{ij}}{\sum_{i'\in U^{\text{SB}}} \ell_{i'j}}$$

is given by

$$\hat{X}_i = \sum_{j\in U^{\text{EB}}} \frac{1}{\sum_{i'\in U^{\text{SB}}} \ell_{i'j}} \sum_{k\in U_i^{\text{RB}}} \alpha_{ik} \frac{\delta_j\, Y_j\, \ell_{ij}\, \ell_{ikj}}{\tau_{ik} \sum_{k'\in U_i^{\text{RB}}} \ell_{ik'j}}, \qquad (20)$$

where α_{ik} is an indicator variable signifying whether the k^{th} RU was the realized respondent or not for the i^{th} SU in s^{SB} and

$$\begin{aligned} \ell_{ikj} &= 1, \quad \text{if SU } i \text{ is linked to RU } k \text{ which} \\ &\qquad\ \ \text{in turn is linked to EU } j \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

The data are now identified and one can plug (20) into (7), giving the revised estimator

$$\hat{Y}^{\text{Eb}} = \sum_{j\in s^{\text{EB}}} \delta_j\, Y_j\, W_{0j}^{\text{EB}} \qquad (21)$$

with revised weights

$$W_{0j}^{\text{EB}} = \sum_{i\in s^{\text{SB}}} W_i^{\text{SB}} \frac{1}{\sum_{i'\in U^{\text{SB}}} \ell_{i'j}} \sum_{k\in U_i^{\text{RB}}} \alpha_{ik} \frac{\ell_{ij}\ell_{ikj}}{\tau_{ik} \sum_{k'\in U_i^{\text{RB}}} \ell_{ik'j}}. \qquad (22)$$

As an approximation, one could take the RUs to be equal users of the cell phone, in which case τ_{ik} would simply be the reciprocal of the number of RUs linked to the SU *i* for all RUs *k*. Adjustments for nonresponse and calibration to control totals would proceed as before.

Alternatively, the survey methodologist could call for a real randomization step, which would require that the interviewer make a roster of the RUs linked to the SU and select one at random, or a pseudo randomization step using the last birthday method. Such methods are probably not feasible at this time, due to the difficulty of gaining cooperation in cell-phone interviews.

5.6 Implications for data collection

Certain information must be collected in the survey interview in order to support the calculation of the estimators discussed here.

To support the use of δ_j, the cell-phone survey must collect information to establish whether any of the RUs linked to the EU have access to a landline telephone. The respondent RU must report this information both for himself or herself and for other RUs that may be linked to the EU.

To support the use of ϕ_j, the landline survey must collect information to establish whether any of the RUs linked to the EU have regular access to a cell phone. The respondent RU must report this information both for himself or herself and for other RUs that may be linked to the EU. This report may be quite straightforward in the event that the response protocol only links EUs to RUs within the same household. For more complicated response protocols, the report could be difficult to obtain.

To support the use of $\sum_{i'\in U^{\text{SB}}} \ell_{i'j}$ in calculating the survey weights, the survey must collect information to establish how many SUs in the population are linked to the reported EU *j*. The respondent RU must be able to report the number of cell phones, including their own, that ring to an RU who is linked to the given EU.

If the estimator given in (21) and (22) would be used in order to identify all of the EUs, then additional information must be collected in the interview. The respondent RU must know and report the number of RUs, including themselves, that are linked to both the selected SU and the reported EU.

The respondent RU must also know and report their share of use of the cell phone on which the interview is completed or be able to say that use is approximately equal.

6. Example: The National Immunization Survey (NIS)

We illustrate the information that must be collected in the survey interview using the NIS, a survey of parents of children age 19-35 months and of teens age 13-17 years sponsored by the Centers for Disease Control and Prevention (CDC) for the purpose of monitoring vaccination coverage rates (*i.e.*, the proportion of children who are up-to-date with respect to the recommended vaccination schedule) in the USA. Data collection in the NIS occurs in two phases: an RDD telephone survey of households with landline telephones that have children or teens in the eligible age range, followed by a survey mailed to the vaccination providers of the age-eligible children. The sampling frame for the telephone survey phase of the NIS consists of all landline telephone numbers in 1+ banks in the USA. Cellular telephone numbers in dedicated cellular banks are currently not included in the NIS sampling frame. When a household with an age-eligible child is identified in the telephone survey, the interview is conducted with the adult in the household who is identified as the most knowledgeable about the vaccination status of the child (nearly always the mother or father). During the telephone interview, data are collected for each age-eligible child in the household, including the demographic characteristics of the child, demographic characteristics of the child's mother, and socio-economic characteristics of the child's household. At the end of the telephone interview, consent is asked to contact the child's vaccination providers. If consent is given, all vaccination providers named by the telephone interview respondent are contacted by mail to obtain the child's provider-reported vaccination history, which is used in statistical analysis to evaluate vaccination status. Smith, Hoaglin, Battaglia, Khare and Barker (2005) provide a detailed description of the statistical methods used by the NIS.

Because of the growth of the cell-phone-only population, the proportion of the NIS target population that is covered by the landline sampling frame has decreased in recent years. Using data from the National Health Interview Survey, Khare, Singleton, Wouhib and Jain (2008) estimate that about 18 percent of eligible children and 10 percent of eligible teens may be missing from the NIS sampling frame. To address the increase in cell-phone-only households in the NIS target population, cell-phone interviews could be added to the NIS.

For the NIS, the telephone number is the SU, the knowledgeable mother or father is the RU, and the age-eligible child is the EU. For the landline RDD or A sample, the parent is a resident of the household to which the sample landline number is assigned, while for the cell-phone or B sample, the parent has regular access to the cell phone to which the sample telephone number is assigned. Children are not subsampled in the NIS, but rather the knowledgeable parent reports for all of their age-eligible children who live in their home (but not for any children who may live elsewhere). These elements of the survey protocol establish the links between RUs and SUs and between EUs and RUs.

One comprehensive NIS design is to conduct estimation by way of nonoverlapping domains and decomposition (3). That is, the A sample is used to represent all children linked to a landline household and the B sample is used to represent all children linked to a cell-phone-only parent. We considered and rejected decompositions (4) and (5) due to considerations of cost and the potential for differential nonresponse bias in estimation for the mixed population.

To implement the estimator in (10), we determine whether the A-sample child is landline-only through use of the following three questionnaire items:

A1. Next I have some questions about cell phones in your household. In total, how many working cell phones do you and your household members have available for personal use? Please don't count cell phones that are used exclusively for business purposes.

A2. How many [of these] cell phones do [LIST ALL ELIGIBLE CHILDREN]'s parents and guardians usually use?

A3. Of all the telephone calls that you and your family receive, are nearly all received on cell phones, nearly all received on regular phones, or some received on cell phones and some received on regular phones? (IF ASKED ABOUT INCLUDING BUSINESS CALLS: Please do not include any business-related calls in your answer).

For the cell-phone or B sample, we establish whether the child is cell-phone-only using the following two questions.

B1. Do you have a landline in your household? (INTERVIEWER PROBE IF YES: Please do not include modem only lines, fax only lines, lines used just for a home security system, beepers, pagers, or the cell phone).

B2. Thinking just about the landline home phone, not your cell phone, if that telephone rang and someone was home, under normal circumstances how likely is it that it would be answered? Would you say extremely likely, somewhat likely, somewhat unlikely, or not at all likely?

We would use Question B2, due to Cantor, Brownlee, Zukin and Boyle (2008), to determine whether the landline is actually used for voice communications and thus whether the respondent is in the *ab* or *b* domain.

Also for the B sample, to determine the number of cell phones in the population that are linked to a given age-eligible child, we would use the following two questions:

B3. Next, I have some questions about cell phones in your household. In total, how many working cell phones do you and your household members have available for personal use? Please do not count cell phones that are used exclusively for business purposes, and please include the number we called.

B4. How many of these cell phones do [LIST CHILDREN]'s parents and guardians usually use? Please include the number we called.

Responses to questions A1-A3 and B1-B4 permit the calculation of survey weights and implementation of the unbiased estimator of the population total given in (10).

7. Summary

In this article, we used some theory of indirect sampling and network sampling to demonstrate a statistical framework for the design and analysis of cell-phone surveys. We exhibited an unbiased estimator of the population total with respect to estimation units linked to sampling units. By implication, this theory gives a means of constructing estimators of other population parameters that can be expressed as functions of totals. We illustrated the issues using the NIS, a telephone survey about young children and teens.

Information from the survey interviews is needed to classify estimation units into the cell-phone-only domain, the landline-only domain, or the mixed domain. Reporting error could result in misclassifications and undermine the unbiasedness of the estimator, as could survey nonresponse in the cell-phone and landline interviews.

Acknowledgements

The authors thank associate editor for helpful comments.

References

Arthur, A. (2007). The birth of a cellular nation. *The Source*. Mediamark Research Inc. Available from: http://www.mediamark.com/mri/TheSource/sorc2007_09.htm, 3.

Blumberg, S.J., and Luke, J.W. (2008). Wireless substitution: Early release of estimates from the National Health Interview Survey. National Center for Health Statistics. Available from: http://www.cdc.gov/nchs/nhis.htm.

Brick, J.M., Dipko, S., Presser, S., Tucker, C. and Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.

Brick, J.M., Edwards, W.S. and Lee, S. (2007). Sampling telephone numbers and adults, interview length, and weighting in the California Health Interview Survey cell phone pilot study. *Public Opinion Quarterly*, 71, 793-813.

Cantor, J., Brownlee, S., Zukin, C. and Boyle, J. (2008). Do We Need to Worry About Wireless Substitution in Public Opinion Polls about Health Reform. Presentation at the AcademyHealth 25th Annual Research Meeting, Washington, DC.

Carley-Baxter, L., Peytchev, A. and Lynberg, M. (2008). Comparison of cell phone and landline surveys: A design perspective. Paper presented at the annual meeting of the American Association for Public Opinion Research, New Orleans, LA.

Cassel, C.-M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Chowdhury, S., Montgomery, R. and Smith, P.J. (2008). Adjustment for noncoverage of nonlandline telephone households in and RDD Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexanderia, VA.

CTIA (2008). Wireless Quick Facts. Available from http://www.ctia.org/advocacy/research/index.cfm/AID/10323.

Deville, J.-C., and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, 32, 165-176.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Ehlen, J., and Ehlen, P. (2007). Cellular-only substitution in the United States as lifestyle adoption: Implications for telephone survey coverage. *Public Opinion Quarterly*, 71, 717-733.

Frankel, M., Battaglia, M., Link, M. and Mokdad, A. (2007). Integrating cell phone numbers into Random Digit-Dialed (RDD) landline surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, (Alexandria, VA), 3793-3800.

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

Keeter, S. (1995). Estimating non-coverage bias from a phone survey. *Public Opinion Quarterly*, 59, 196-217.

Khare, M., Singleton, J.A., Wouhib, A. and Jain, N. (2008). Assessment of Potential Bias in the National Immunization Survey (NIS) from the Increasing Prevalence of Households Without Landline Telephones. Presented at the National Immunization Conference, Centers for Disease Control and Prevention.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer Science+Business Media, LLC.

Lavrakas, P.J., Shuttles, C.D., Steeh, C. and Fienberg, H. (2007). The state of surveying cell phone numbers in the United States: 2007 and Beyond. *Public Opinion Quarterly*, 71, 840-854.

Smith, P.J., Hoaglin, D.C., Battaglia, M.P., Khare, M. and Barker, L.E. (2005). Statistical methodology of the National Immunization Survey, 1994-2002. National Center for Health Statistics, Hyattsville, MD. *Vital and Health Statistics*, Series 2, 138.

Wolter, K.M. (2007). *Introduction to Variance Estimation*, *Second Edition*. New York: Springer-Verlag.

Wolter, K.M., Chowdhury, S. and Kelly, J. (2008). Design, conduct, and analysis of random digit dialing surveys. In *Handbook of Statistics*: *Sample Surveys*, *Theory*, *Methods and Inference*, (Eds., D. Pfeffermann and C.R. Rao), Elsevier, Oxford, UK.