Estimation in a mixed-mode, web and face-to-face, survey

Kaur Lumiste[1]

[1]University of Tartu, e-mail: kaur.lumiste@ut.ee

Abstract

Growing survey costs and falling response rates are problems for many survey companies and national statistics agencies. One heavily researched possible cure for this is to combine survey modes - mixed-mode design.

In September 2012 European Social Survey in Estonia, Slovenia and UK are planning an experiment with mixed-mode designs. Web and telephone survey modes are considered in conjunction with the usual face-to-face interview mode. The aim is to test for mode effects, influences in respondents' answers caused by the mixed-mode design, and develop means and protocols to avoid them.

In Estonia the experiment involves a web survey mode in conjunction with face-to-face interviews. At first an invitation is sent to sampled persons inviting them to fill the survey online. If a person shows no signs of activity, even after two reminders, then an interviewer is given a task to survey that person. The design gives us three random subgroups of the sample: people who fill the survey online, people who answer in the face-to-face interview and non-response subgroup. These subgroups tend to be different from one-another and, with auxiliary information, they can be used for better estimation.

Current paper gives a short overview on the preliminary studies made on estimation in this experimental mixed-mode design.

Keywords: Mixed-mode survey, web survey, face-to-face interview, estimation

1 Introduction

All survey designs pursue the somewhat incompatible objectives of reducing error and limiting costs. Survey companies and national statistics agencies are trying to find ways of making surveys more cost effective while not giving away precision. One heavily researched option is to combine survey modes (e.g. internet and postal survey or telephone and CAPI). Mixed-mode designs have special appeal for reducing coverage and non-response error, while also bringing costs down (Dillman & Messer, 2010). Ensuring that all members of a population have a known, nonzero chance of being sampled is very difficult, if not possible with certain designs. For example a web survey leaves out respondents who do not have access to internet, so a mixed-mode design should be considered.

Mixed-mode designs may reduce non-response error. People who are unwilling to participate in a telephone survey may be willing to respond by mail or over the internet. Groves & Kahn (1979) and Millar, Dillman, & O'Neill (2009) showed that some people prefer certain modes for being surveyed, while objecting to others. More importantly, those preferring different modes could differ from one another, as Link & Mokdad (2006) found that respondents to telephone and mail versions of the survey differed on demographic characteristics including gender, age, and income.

But all this does not come without drawbacks, using multiple modes may introduce mode effects, thereby increasing measurement error. For example Hochstim (1967) showed that personal interviews produced more "excellent" answers (40%) to a simple question, "Do you consider your health to be excellent, good, fair, or poor" than did mail surveys (30%). Interviewer presence encourages respondents to give answers consistent with social norms, a behaviour known as social desirability bias. Also different designs require different question wording, for example paper and web questionnaires may use check-all-that-apply questions, while telephone surveys use forced-choice items offering respondents a "yes/no" choice for each item. For a more complete list of mode effects the reader is referred to Dillman & Messer (2010).

In September 2012 three participating countries of the European Social Survey (ESS) will conduct a mixed-mode experiment in the background of data collection for ESS round 6. The experiment aims to test the feasibility of using other survey modes in conjunction with the face-to-face interviews used so far. Currently telephone and internet surveys are being considered. The experiment will be conducted simultaneously in Estonia, Great Britain and Slovenia, and Estonia will test CAPI in conjunction with the web survey method. First, sampled persons are invited to fill the ESS questionnaire online. If the invitation is ignored, as well as the two reminders, an interviewer is sent for a face-to-face interview.

The experiment's design divides sampled persons into two subgroups (with random sizes) - people who answer online and those who did not. The grouping is not completely random (like simple random sampling without replacement) since respondents' mode preference may be dependant on some demographic characteristics, as mentioned earlier. The subgroup of people who did not answer online is again divided into two groups - respondents by face-to-face interview, and non-respondents.

This paper presents preliminary studies on a possible estimation method in case of this special experimental case. First, two different estimators for the population total can be defined using these two groups of respondents and auxiliary information, and then also a linear combination of the two estimators is proposed.

2 Estimation in mixed-mode surveys

2.1 Preliminaries

Let $U = (1, 2, \dots, N)$ denote a finite population of N units. Let a random vector (design vector) $\mathbf{I} = (I_1, I_2, \dots, I_N)$ describe the sampling process on U and I_i is the sample inclusion indicator for unit $i \in U$. The probability sampling design generates for element i a known inclusion probability, $E(I_i) = \pi_i > 0$, and a corresponding sampling design weight $a_i = 1/\pi_i$. In case of non-response data can only be collected from a sample subgroup $r \subseteq s$. The study variable y is recorded for all $i \in r$ and our objective is to estimate the population total $Y = \sum_U y_i$. The basic design unbiased estimator of Y from a full sample s is $\hat{t}_{\mathrm{HT}} = \sum_s a_i y_i$, the Horwitz-Thopmson (HT) estimator.

Auxiliary information has become more and more crucial in effective estimation and dealing with non-response. The auxiliary vector value $\mathbf{x}_i : J \times 1$ is assumed available for every element $i \in s$ (or every $i \in U$ if it is compiled from comprehensive registers) and J is the number of auxiliary variables available.

2.2 Mixed-mode

Since data from the respondents will be collected in two parts, let us define two random vectors:

$$\mathbf{W} = (W_1, W_2, \dots, W_n \,|s),$$

where $W_i = 1$ if unit i in sample s answers in the web mode and $W_i = 0$ otherwise, and

$$\mathbf{F} = (F_1, F_2, \dots, F_n \,|s),$$

where $F_i = 1$ if unit i in sample s answers in the face-to-face mode and $F_i = 0$ otherwise. Note that vector $\mathbf{W} + \mathbf{F}$ indicates the units that belong to set r. As the data collection begins the sample s is divided into two subsets - sampled persons who choose to answer online, $r_{web} = \{i \mid W_i = 1, s\}$, and remaining sampled persons i.e. $s_{ftf} = s - r_{web} = \{i \mid W_i = 0, s\}$.

We can now define $P(W_i = 1 \mid s) = p_i$, which is the probability that unit i will answer the survey via the internet, and an unbiased estimate for the population total Y can be found:

$$\hat{t}_{web} = \sum_{r_{web}} \frac{y_i}{\pi_i p_i},$$

since $\pi_i p_i = P(I_i = 1) \cdot P(W_i = 1 \mid s) = P(i \in s) \cdot P(i \in r_{web} \mid s) = P(i \in r_{web})$.

The probabilities p_i have to be estimated and this can be done using auxiliary information, but we will come back to this later in section 2.3.

Since sampled units, who do not answer online, are approached for a face-to-face interview, the probability of that happening is $1 - p_i$. We now define $P(F_i = 1 \mid s_{ftf}) = q_i$, which is the probability of person i answering to the survey in a face-to-face interview under the condition that he belongs to s_{ftf}. The following unbiased estimator can be constructed:

$$\hat{t}_{ftf} = \sum_{r_{ftf}} \frac{y_i}{\pi_i (1 - p_i) q_i}.$$

We now have two different estimates for Y and we can get a more efficient estimator by linearly combining them

$$\hat{t} = \alpha \hat{t}_{web} + (1 - \alpha) \hat{t}_{ftf}$$

where $\alpha \in (0,1)$ and can be found by minimizing $Var(\hat{t})$. In practice it would be very rare, but for simplicity let us assume that $\hat{t}_{web}$ and $\hat{t}_{ftf}$ are independent. Then the optimal α is

$$\alpha^* = \frac{Var(\hat{t}_{ftf})}{Var(\hat{t}_{web}) + Var(\hat{t}_{ftf})}.$$

2.3 Estimating mode participation probabilities

As mentioned earlier, research has shown that mode preference can be dependant on demographic characteristics like age and gender. Usually these variables can be retrieved from population registries for all sampled elements and can be taken as auxiliary variables. Särndal (2011) uses auxiliary information to estimate response probabilities $\theta_i = P(i \in r \mid s)$, but we adapt it to estimate mode participation probabilities p_i and q_i. In general case for estimating θ_i we need two conditions:

1. The estimates $\hat{\theta}_i$ for θ_i are linearly dependant on auxiliary variables $\mathbf{x}_i$, meaning that there is a constant vector $\boldsymbol{\lambda}$ so that

$$\hat{\theta}_i = \boldsymbol{\lambda}'\mathbf{x}_i. \tag{0.1}$$

2. The estimates $\hat{\theta}_i$ satisfy restrictions in the response/non-response case for I_i being here the response indicator:

$$\sum_s a_i(I_i - \hat{\theta}_i)\mathbf{x}_i = 0 \text{ or}$$

$$\sum_r a_i\mathbf{x}_i = \sum_s a_i\hat{\theta}_i\mathbf{x}_i. \tag{0.2}$$

With these restrictions $\boldsymbol{\lambda}$ can be found by substituting (0.1) into (0.2) so that we get

$$\sum_r a_i\mathbf{x}_i' = \boldsymbol{\lambda}' \sum_s a_i\mathbf{x}_i\mathbf{x}_i'.$$

By extracting $\boldsymbol{\lambda}'$and using it in (0.1), we get an estimate for the answering probability

$$\hat{\theta}_i = \left(\sum_r a_i\,\mathbf{x}_i\right)'\left(\sum_s a_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\mathbf{x}_i.$$

For mode participation probabilities p_i and q_i, given the auxiliary vector $\mathbf{x}_i$, the estimators take the following form:

$$\hat{p}_i = \left(\sum_{r_{web}} a_i\,\mathbf{x}_i\right)'\left(\sum_s a_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\mathbf{x}_i \text{ and } \hat{q}_i = \left(\sum_{r_{ftf}} a_i\,\mathbf{x}_i\right)'\left(\sum_{s_{ftf}} a_i\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\mathbf{x}_i.$$

3 Conclusions

Estonia's ESS team will conduct an experiment with combining web survey mode with CAPI. Estimators for such a design are studied and the preliminary results presented.

Further research aims to study the properties of these estimators, find the optimal α if the two population totals are correlated and test the estimators in a simulation study.

References

Dillman, D. A. & Messer, B. L. (2010). Mixed-Mode Survey. In: P. Marsden and J. Wrigth, ed. 2010 *Handbook of Survey Methodology*. Bingley: Emerald Publishing Limited, 551-574.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews.* NewYork: Acadmic Press.

Hochstim, J. R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976-989.

Link, M.W., & Mokdad, A. (2006). Can web and mail survey modes improve participation in an RDD-bsaed national health surveillance? *Journal of Official Statistics*, 22, 293-312.

Millar, M., Dillman, D. A. & O'Neill, A. C. (2009). *Are mode preferences real?* Technical Report 09-003, Social and Economic Sciences Research Center, Washington State University, Pullman, WA.

Särndal, C.-E., (2011). The 2010 Morris Hansen Lecture. Dealing with Survey Nonresponse in Data Collection, in Estimation. *Journal of Official Statistics* 28. 1-21.