Estimation of a Proportion: Sub Sampling the Non Response Stratum

Carlos. N. Bouza

Facultad de Matemática y Computación, Universidad de La Habana
[2] Division of Engineering and Applied Science, California Institute of Technology, Pasadena, United States
Email: bouza@matcom.uh.cu

Abstract. We consider the estimation of a proportion when there are non-responses. They may be motivated by the fact that a "yes", as a response, identifies belonging to a sensitive group. We suggest taking a sample among the non-respondents and using a randomized response technique in the second interview. This approach motivates sampling models which are developed in this paper. The performance of the proposed model is evaluated using real world data.

Keywords: Randomized responses, non-respondent stratum, sub sampling

1 Introduction

The usual theory of survey sampling is developed assuming that the finite population $U = \{u_1, \cdots, u_N\}$ is composed of individuals that can be perfectly identified. A sample s of size $n \leq N$ is selected. The variable of interest Y is measured in each selected unit. Real life surveys should deal with the existence of missing observations. There are three solutions to cope with this fact: to ignore the non-respondents, to sub sample the non-respondents or to impute the missing values, while to ignore the non-responses is a dangerous decision and to sub sample is a conservative and costly solution. Once sub-sampling is chosen, to grant the cooperation of the interviewed is very important. We then consider using a randomized response procedure for obtaining information from the re-interviewed no-respondents, which seems to be a solution for granting cooperation and true answers.

The use of randomized response (RR) was introduced by Warner [5]. It provides the opportunity of reducing response biases as well as non-responses. Thereof, this technique protects the privacy of the respondents by preventing their belonging to a stigmatized group from being detected. We consider that, as a RR technique is applied, a cooperative attitude of the interviewed will be present. Different extensions have been developed, see Singh [4] and Bouza [6], for detailed discussions.

Let the finite population under study be $U = \{u_1, \cdots, u_N\}$. A sample s is selected from U and it is divided into $s_1 + s_2$. The individuals in s_1 give a response at the first visit while those in s_2 do not respond. Based on the non-respondent stratum model, U is divided into two strata

$$U_1 = \{u \in U \mid u \text{ responds at the first visit}\}$$

$$U_2 = \{u \in U \mid u \text{ does not respond at the first visit}\}$$

The cause of the non-response can be traced to the intention of not being identified as a member of a stigmatized group, say belonging to A. Hence we can also stratify U accordingly into

$$U_A = \{u \in U \mid u \in A\}$$

$$U_{A^*} = \{u \in U \mid u \notin A\}$$

It is expected that the individuals in U_A tend not to respond or tend to give false reports.

2 Randomized Responses Procedures

Suppose that an inquiry is developed to estimate the proportion of individuals belonging to A. If $|B|$ denotes the number of units in B then

$$\theta_A = |U_A| / |U| = N_A / N$$

is the parameter to be estimated. The usual approach is to ask a selected individual if he/she belongs to A (to carry the stigma) and we deal with the Bernoulli random variable

$$X_{A(i)} = \begin{cases} 1 & \text{if } u_j \text{ carries the stigma} \\ 0 & \text{otherwise} \end{cases}$$

Hence $E\left[X_{A(i)} \right] = \theta_A$.

Warner [5] proposed to provide a random mechanism to the interviewed, which should develop an experiment that selects the statement

1. I belong to A with probability $p \neq 0.5$.
2. I do not belong to A with probability $1-p$.

The evaluated variable is

$$Y_i = \begin{cases} 1 & \text{if } u_i \text{ responds Yes} \\ 0 & \text{if } u_i \text{ responds No} \end{cases}$$

The individual is not told which statement is been evaluated. The random sample permits to evaluate the number of Yes

$$.n_Y = \sum_{i-1}^{n} Y_i$$

Warner [5] derived that

$$\theta_{s(A)} = \frac{p-1}{2p-1} + \frac{n_Y / n}{2p-1}$$

is unbiased for θ_A and that its variance is

$$V(\theta_{s(A)}) = \frac{\theta_A(1-\theta_A)}{n} + \frac{0,25\left[\left(2p-1\right)^{-2} - 1 \right]}{n}$$

This technique is known as the related question method.

A variation of this method was recently proposed by Christofides [1]. He proposed the use of a random mechanism that generates a random integer $e \in \{1, \cdots, L\}$. It assigns a probability $P\left(e = j\right) = p_j$. $j = 1, \cdots L$ to each integer. The interviewed uses the mechanism, selects randomly an e and reports $d_i = L + 1 - e$ if he/she belongs to A and zero otherwise. The distributional problem is described by noting that the response is

$$X_i = \begin{cases} L + 1 - e & \text{if } u_i \in A \\ 0 & \text{if } u_i \notin A^* \end{cases}$$

As

$$P\left(X_i = L + 1 \right) = \theta$$
$$P\left(X_i = 0 \right) = 1 - \theta$$

The probability that a the random report is equal to k is

$$P\left(d_i = k\right) = \theta p_k + (1-\theta) p_{L+1-k}, \quad k = 1, \cdots, L$$

Then the expected value of a report is

$$E\left(d_i\right) = \mu_e + \theta(L + 1 - 2\mu_e) = \mu_d.$$

Because the other term is multiplied by zero. The random mechanism generates the variable e and therefore its distribution is known including its expectation and variance

$$E\left(e_i\right) = \sum_{k=1}^{L} kp_k = \mu_e$$
$$V\left(e_i\right) = \sum_{k=1}^{L} (k-\mu_e)^2 p_k = \sigma_e^{\ 2}$$

Its variance is determined as

$$E(d_i - \mu_d)^2 = V\left(e\right) + \theta(1-\theta)[L+1-2\mu_e]^2$$

The mean of the responses is given by

$$.d_s = \sum_{i=1}^{n} d_i / n$$

Using the relationship fixed in μ_d a naive estimator is given by

$$\theta_{s(A)} = \frac{d_s - \mu_e}{L+1-2\mu_e}$$

Therefore the mechanism should satisfy that $\mu_e \neq \left(L+1\right)/2$, and so the unbiasedness of the estimator follows that

$$E[\theta_{s(A)}] = \frac{\mu_d - \mu_e}{L+1-2\mu_e} = \theta_A$$

because only d_s is random. Therefore the error is calculated by deriving from $V\left(d_s\right)$ as

$$V[\theta_{s(A)}] = \frac{V(d_s)}{(L+1-2\mu_e)^2} = \frac{\theta_A(1-\theta_A)}{n} + \frac{\sigma_e^2}{n(L+1-2\mu_e)^2}$$

where the first term is the error of the usual estimator of a proportion and the second term is the increase in inaccuracy due to the use of the RR technique of Christofides [1].

A comparison of both methods indicates that Warner's proposal should be preferred whenever the following inequality holds

$$\frac{0.25\left[\left(2p-1\right)^{-2} - 1\right]}{n} < \frac{\sigma_e^2}{n(L+1-2\mu_e)^2}$$

Then the sampler can design the RR device to ensure that it is more accurate than the technique of the related response.

3 The Non Response Problem

As quoted previously we will analyze the nonresponse problem. The sampler selects the sample S without knowing that U is stratified into

$$U_1 = \{u \in U \mid u \text{ responds at the first visit}\}$$
$$U_2 = \{u \in U \mid u \text{ does not respond at the first visit}\}$$

The probability of selecting a $u \in U_t$ is

$$\left|U_t\right| = N_t / N = W_t, \quad t = 1,2.$$

Therefore $S = S_1 \cup S_2$, $S_t \subset U_t$, $\left|S_t\right| = n_t$, $t = 1,2$ and $n_1 + n_2 = n$.

A second visit is planned supposing that the individual selected in a subsample $S_2' \subset S_2$ will respond. The refusal to answer may be because a certain violation of the privacy is present by answering 'I belong to A'. Therefore to use a randomized response scheme should provide full response in the subsample.

The most popular method is to subsample the non-respondents using the rule of Hansen Hurwitz [3], see Cochran [2]. They suggested determining the subsample size

$$\left|S_2'\right| = n_2' < \left|S_2\right| = n_2$$

by introducing a constant $\lambda > 1$ and calculating

$$n_2' = n_2 / \lambda$$

Take the total sum of yes in S_1 as

$$.n_{Y(1)} = \sum_{i=1}^{n} I\left(S_1\right) Y_i$$

$I\left(S_1\right)$ is a Bernoulli random variable that takes the value 1 if $i \in U_1$. Hence $E\left[I\left(S_1\right)\right] = W_1$ and $E\left[n_{Y(1)} \middle| S\right] = \dfrac{\left|U_1 \cap U_A\right|}{N} = n\theta_{A(1)}$.

which is the proportion of individuals with the stigma in U_1.

Note that

$$.n_1 = \sum_{i=1}^{n} I\left(S_1\right)$$

has the expectation $E\left[n_1\right] = nW_1$.

Similarly using

$$.n_{Y(2)} = \sum_{i=1}^{n}\left[1 - I\left(S_1\right)\right] Y_i$$

we have that $E\left[1 - I\left(S_1\right)\right] = W_2$ and $E\left[n_{Y(2)} \middle| S\right] = \dfrac{\left|U_2 \cap U_A\right|}{N} = n\theta_{(2)A}$

As $.n_{Y(2)}$ is not known we resample S_2 and obtain the number of persons in A included in the subsample S_2'

$$n_{Y(2)}' = \sum_{i=1}^{n_2} I(S_2' \middle| S_2) Y_i$$

Its conditional expectation is

$$E\left(n_{Y(2)}' \middle| S_2\right) = n_2' \frac{n_{S_2(A)}}{n_2} = n_2' \theta_{S_2(A)}$$

For the usual sub sampling approach,

$$\theta_{s(A|NR)} = \frac{n_1 \theta_{S_1(A)} + n_2 \theta_{S_2'(A)}}{n}$$

where $\theta_{S_2'(A)} = \dfrac{n_{Y(2)}'}{n_2'}$ is unbiased, and so the estimators are unbiased for the expectation within each stratum.

It is easily derived from the results reported in the literature for the case of the sample means that

$$E\left\{V\left[\theta_{s(A|NR)} \middle| S\right]\right\} = \frac{\theta_{(A)}(1-\theta_{(A)})}{n} + \frac{(\lambda - 1) W_2 \theta_{A(2)}(1-\theta_{A(2)})}{n}$$

is the error of this estimator.

Our proposal is to use a RR method for obtaining the responses of the non-respondents in the first visit.

If we use Warner's RR technique the estimator of $\theta_{A(2)}$ is

$$\theta(W)_{S_2'(A)} = \frac{p-1}{2p-1} + \frac{n_{Y(2)}' / n_2'}{2p-1}$$

which is unbiased. Hence

$$\theta(W)_{s(A|NR)} = \frac{n_1 \theta_{S_1(A)} + n_2 \theta(W)_{S_2'(A)}}{n}$$

is unbiased too. Rewriting it as

$$\theta(W)_{s(A|NR)} = \frac{n_1\theta_{S_1(A)} + n_2\theta_{S_2(A)} + n_2\left[\theta(W)_{S_2'(A)} - \theta(W)_{S_2(A)}\right]}{n}$$

this result is easily derived. The first term is the estimator of the proportion of individuals that belong to U_A and the expectation of the other term is zero. Considering the variance of the first term,

$$V\left[\frac{n_1\theta_{S_1(A)} + n_2\theta_{S_2(A)}}{n}\right] = \frac{\theta_A(1-\theta_A)}{n}$$

The variance of the second term is the variance of Warner's related question method. Noting that the expected value of the cross product is zero and taking the equality

$$\frac{n_2\left[\theta(W)_{S_2^{'}(A)} - \theta(W)_{S_2(A)}\right]}{n} = \frac{n_2\left[\theta(W)_{S_2^{'}(A)} - \theta_{A(2)}\right]}{n} - \frac{n_2\left[\theta_{S_2(A)} - \theta_{A(2)}\right]}{n}$$

we have that

$$V\left[\frac{n_2\left[\theta(W)_{S_2'(A)} - \theta(W)_{S_2(A)}\right]}{n}\Big|S_2\right] = \left(\frac{n_2}{n}\right)^2\left[E\left[\left[\theta(W)_{S_2'(A)} - \theta_{A(2)}\right]^2\Big|S_2\right]\right] - E\left[\left[\theta_{S_2(A)} - \theta_{A(2)}\right]^2\Big|S\right]^2$$

The first term at the right hand side of the equation is the variance of the estimator under Warner's model. As

$$E\left(\theta(W)_{S_2(A)} - \theta_{A(2)}\right)^2 = \frac{\theta_{A(2)}(1-\theta_{A(2)})}{n_2} + \frac{0.25}{n_2}\left[(2p-1)^{-2} - 1\right]$$

$$V\left[\frac{n_2\left[\theta(W)_{S_2'(A)} - \theta(W)_{S_2(A)}\right]}{n}\Big|S_2\right] = \left(\frac{n_2}{n}\right)^2\left[\theta_{A(2)}(1-\theta_{A(2)})\left(\frac{1}{n_2'} - \frac{1}{n_2}\right) + \frac{0.25}{n_2}\left[(2p-1)^{-2} - 1\right]\right]$$

Substituting with $n_2' = n_2/\lambda$, we have that the factor of $\theta_{A(2)}(1-\theta_{A(2)}$ is $(\lambda-1)/n_2$. Simplifying with $\left(n_2/n\right)^2$ we have that this variance is equal to $n_2\left\{\theta_{A(2)}(1-\theta_{A(2)}(\lambda-1) + 0.25\left[\left(2p-1\right)^{-2} - 1\right]\right\}/n^2$.

Substituting this results in the variance of the proposed estimator we have that

$$E\left[V\left[\theta\left(W\right)_{(A|NR)}\right]\right] = n^{-1}\left[\theta_{(A)1}1-\theta_{(A)}) + W_2\theta_{A(2)}(1-\theta_{(A)}) + 0{,}25\left[\left(2p-1\right)^{-2} - 1\right]\right]$$

The third term represents the increment in the variability of double sampling estimator for non-response when Warner's RR is used.

Using the method of Christofides [1] the corresponding random mechanism is applied to the persons in the NR subsample. The mean of the responses of the re-interviewed is given by

$$d_{S_2'} = \sum_{i=1}^{n_2'} d_i / n_2'$$

and the estimator of the proposition in the NR stratum is

$$\theta(C)_{s_2'(A)} = \frac{d_{s_2'} - \mu_e}{L+1-2\mu_e}$$

with $\mu_e \neq \left(L+1\right)/2$. The unbiasedness of it follows conditionally and then is unbiased

$$\theta(C)_{s(A|NR)} = \frac{n_1\theta_{S_1(A)} + n_2\theta(C)_{S_2'(A)}}{n}$$

The conditional error of the estimator of the proportion of individuals in $A \cap U_2$ is calculated once we derive $V_{Cs'} = V(d_{s_2'})$. It is easily derived from the results discussed in the previous section

$$V[\theta(C)_{s_2'(A)}] = \frac{V(d_{s_2'})}{(L+1-2\mu_e)^2} = \frac{\theta_{(A|NR)}(1-\theta_{(A|NR)})}{n_2'} + \frac{\sigma_e^2}{n_2'(L+1-2\mu_e)^2}$$

Using the procedures applied for obtaining the variance under Warner' s design we have that

$$E[V(\theta(C)_{(A|NR)})] = \frac{\theta(A)[1-\theta(A)]}{n} + W_2 \frac{\theta(A(2))[1-\theta(A(2))]}{n} + \frac{\sigma_e^2}{n(L+1-2\mu_e)^2}$$

4 Comparisons

Analytically we obtain the preference to Warner's based strategy when

$$p < \frac{2\sigma_e^2}{(L+1-2\mu_e)^2} + \frac{1}{2}$$

Therefore the sampler can model the preference for a certain RR procedure by tuning p, μ_e and σ_e. In practice the confidence for a certain procedure relies on subjective issues that make the interviewed feel that to tell the truth does not jeopardize his/her privacy of belonging to a sensitive group. We used an inquiry developed among workers of a Hotel Chain devoted mainly to foreign guests. Some of the questions were related with their morality. They were:

A_0 = Have you been sanctioned during the last 12 months?

A_1 = Have you committed a bribery during the last 6 months?

If the response to A_1 is Yes then respond?

A_2 = I have subtracted food?

A_3 = I have subtracted cleaning products or instruments?

The proportion of workers $\theta_{(A_0)}$ was known. A ballot was carried out with a set of questions that were placed in the box of the random sample of workers. Those who did not respond were considered in U_2. A meeting with them was made and the RR based ballot was given to them together with the two devices. Warner procedure was used with $p = 0,25$. The set used for Christofides procedure was $\{1,\cdots,10\}$ and $p_j = 0,1$, The direct question was also made. A comparison between the proportion of A_0 $\theta_{(A_0)}$ and the estimation made by each procedure $\theta_{s(A_0|NR)}$ is given in Table 1. The computed difference of the estimation obtained from the non-respondents and from the population proportion (first column) indicates that the proportion is severely underestimated when the direct procedure is used. The RR methods are closer to the population parameter. The second column gives more evidence on these facts. The estimation of the proportion of the non-respondents is considerably larger than when a RR procedure is used for obtaining the responses. This results support that they are confident that the RR does not jeopardize their identity.

Table 1. Results obtained with A_0

Response Procedures	$\theta_{s(A_0)} - \theta_{s(A_0\|NR)}$	$\theta_{s2(A_0\|NR)}$
Direct	-0.115	0.082
Warner	-0.0031	0.268
Christofides	0.070	0.193

We have not population knowledge for the rest of the questions but the computation of $\theta_{s_2(AT|NR)}$, $T = 0,1,\cdots,4$ is given in Table 2. Note that the use of the direct question gives a considerably large underestimation of the proportion. Note that if the interviewed are more confident, they will declare the truth more frequently. Hence a larger proportion should denote more confidence, indicating that the non-respondents considered Warner's RR procedure more jeopardy protective.

Table 2. Estimation of the proportions

Response Procedures	$\theta_{S(A_0\|NR)}$	$\theta_{S(A_1\|NR)}$	$\theta_{S(A_2\|NR)}$	$\theta_{S(A_3\|NR)}$	$\theta_{S(A_4\|NR)}$
Direct	0.082	0.158	0.106	0.092	0.002
Warner	0.268	0.331	0.233	0.271	0.196
Christofides	0.193	0.267	0.209	0.202	0.085

Acknowledgements: The work of this paper was carried out within the plan of a CITMA project.

References

1. T. C. Christofides, "A generalized response technique". Metrika, Vol. 57, No. 2, pp. 195-200, 2003.
2. W. G. Cochran, "Sampling Techniques", John Wiley & Sons; 3rd edition, January 1, 1977.
3. M.H. Hansen and W. N. Hurwitz, "The problem of non-response in sample surveys" J. Amer. Stat. Ass. Vol 41, No. 3, 517-529, 1946.
4. S. Singh, "Randomized response model". Metrika, Vol. 56, No. 1, pp. 131-142, 2002.
5. S. L.Warner, "Randomized response : a survey technique for eliminating evasive answer bias", J. Amer. Stat. Assoc. Vol. 60, No. 1, pp63-69, 1965.
6. C. N. Bouza, "A Review of Randomized Responses Procedures: the Qualitative Variable Case", Investigación Operacional, Vol. 30, No. 3, pp 240-247, 2010.

