Survey Nonresponse Adjustments in Methods for the Treatment of Sensitive Questions

Leonardo Trujillo[1,3], **Luz Mery Gonzalez**[2,3]
1, 2 Department of Statistics, National University of Colombia, Bogota.
3 Corresponding authors: Email: ltrujilloo@unal.edu.co, lgonzalezg@unal.edu.co

Abstract

Randomized response and item count techniques have originally been designed as statistical methods to reduce nonresponse as well as not reliable answers from the respondents. However, understanding how these methods work by the respondents could be a very difficult task and still sometimes it is possible to obtain a big proportion of people refusing to collaborate in the fear of the protection of their confidentiality. In a particular survey, the elements in the sample are exposed to a specific set of survey operations. In the case of randomized response techniques, an additional stage of introducing the methods to both interviewers and respondents and convincing them how the confidentiality of the former is protected is necessary. In this paper, a simple but powerful method is formulated in order to describe the unknown response mechanism for the sample as accurately as possible. This corresponds to the very well-known response homogeneity group. However, as far as the authors know these models have not been applied to randomized response techniques. Data are assumed as missing at random within sample subgroups, conditionally on the sample. This is a substantial improvement on assuming that data are missing at random throughout the population. At the end, a simulation study shows the effects of ignoring the nonresponse on the bias and the variance of the estimators obtained under the application of these techniques and in particular to the seminal Warner randomized response model and some item count techniques available in the literature.

Keywords: Design based inference, item count techniques, randomized response techniques, survey sampling, two phase estimators.

1 Introduction

Sometimes due to embarrassment; fear of having any personal consequences as receiving fines, punishment or simply because people does not want to reveal their intimacy, the respondents in a survey can refuse to participate. On the other hand, some people answering the survey could give false answers for some specific type of questions because they do not want to reveal the truth even in surveys from national statistical offices. For researchers and in particular for statisticians, the first problem is known as a nonresponse error in the survey and the second one is known as a bias in the response. Accessing information regarding a sensitive characteristic in the population induces these two particular problems: nonresponse and non truthful answers. The two sources of error frequently appear to be a problem when the characteristic of interest being estimated corresponds to sensitive questions related to phenomena such as abortion, contraceptive methods, domestic violence, emigration status of a person, euthanasia, fraud and plagiarism, health problems, illegal crops, income, opinion about high ranked staff in a company or authorities, racism, sexual preferences, tax evasion, use of illegal drugs, among many others.

Randomized Response Techniques (RRTs) and other techniques using indirect questions are useful in order to get a trustful answer but also keeping the confidentiality of the respondents. Alternative techniques being studied recently are the item count techniques (ICTs, Droitcour et al. (1991); Blair and Imai (2010); Imai (2011); Hussain, Shah and Shabbir (2012)) or the three card method (Chaudhuri, 2011).

The aim of this paper is to implement randomized response and item count techniques for finite population sampling in order to estimate the total of individuals in the population having a particular sensitive characteristic. The estimators for the population total as their corresponding variance estimators are obtained in this paper for any (complex) sampling design. Also, survey nonresponse adjustments are proposed for all the proposed estimators.

2 Randomized Response Techniques

Randomized Response Techniques(RRTs) and other techniques using indirect questions are useful in order to get a trustful answer but also keeping the confidentiality of the respondents. Warner (1965); Christofides (2003); Huang (2004); Kim and Warde (2004); Soberanis and Miranda (2011) are few examples refering to these techniques. The first of these techniques was proposed by Warner (1965). Warner s method proceeds as follows: suppose all the people in a population belong either to group A or group A^c and you want to estimate the proportion of people belonging to group A using a sensitive question in your survey. A simple random sample with replacement of size n is taken from the population in a way such as every person responds only one question about belonging to group A with probability P or a question about belonging to group A^c with probability $1 - P$. The interviewee responds "yes" or "no" with the interviewer not knowing which of the two questions he/she is answering.

Let π the unknown probability of belonging to group A in the population and let P the probability of a person answering the question referring of belonging to group A. Warner (1965) shows how just knowing P and not the actual response of every individual in the survey, it is possible to get an estimation of π. Sarndal, Swensson and Wretman (1992) extends this classical model of Warner (1965) to be applied under any (complex) sampling design. Two indicator variables are defined, y_k if the k-th individual possesses the sensitive attribute A and x_k if the k-th individual in the survey responds with a "yes". For every element k, it follows that $p(x_k = 1) = y_k P + (1 - y_k)(1 - P)$ and then, assuming $p \neq \frac{1}{2}$, it follows that $\hat{y}_k = \frac{x_k+P-1}{2P-1}$. For every k, $\hat{y}_k$ is unbiased for y_k with respect to the random mechanism with $Var_{RC}(\hat{y}_k) = \frac{P(1-P)}{(2P-1)^2} = V_0$. In order to estimate t_y unbiasedly with complete response, Sarndal, et al. (1992) shows that $\hat{t}_{RR} = \sum\limits_{s} \frac{\hat{y}_k}{\pi_k}$ and also its corresponding variance.

3 Nonresponse Adjustments for Randomized Response Methods

Randomized response techniques have been originally proposed as useful methods to reduce nonresponse and measurement errors. However, the comprehension from interviewers and interviewees in how these methods protect the privacy of the respondents require a strong sensitization and training of them, and it is even possible that a big proportion of people refuse to answer. In this paper, we formulate a simple model corresponding to the homogeneous group model of response. However, as far as we know this kind of models have never been applied for randomized response techniques. Data are assumed as "missing at random" among population subgroups in the obtained sample. This is much better, than just assuming "missing at random" data throughout all the population. Suppose the sample is selected under a sampling design $p(.)$ with inclusion probabilities π_k and π_{kl}. The sample S is partitioned in H_s groups s_h having the same probability of response.

For every s and $h = 1, 2, \ldots, H_S$

$$\begin{aligned} p(k \in r|s) &= \pi_{k|s} = \theta_{hs} \quad \text{assuming } k \in S_h \\ p(k, l \in r|s) &= \pi_{kl|s} = p(k \in r|s)p(l \in r|s) \\ &= \begin{cases} \theta_{hs}\theta_{h's} & \text{if } k \in s_h,\ l \in s_{h'} \quad h \neq h' \\ \theta^2_{h_s} & \text{if } k, l \in s_h \quad k \neq l \end{cases} \end{aligned}$$

Among many other examples, the probability of answering a sensitive question about abortion could be positively correlated with socioeconomic strata. After considering the nonresponse mechanism, higher stratum could be unrepresented generating bias in the estimators. That is why, it is necessary to account for nonresponse. Also, the potential use of weapons as self-defense could be higher in some population groups but also their nonresponse patterns for this particular issue.

Let r_h the group of respondents from s_h; $n_h = \#(s_h)$,$m_h = \#(r_h)$ with

$$r = \bigcup_{h=1}^{H_s} r_h \qquad m_r = \sum_{h=1}^{H_s} m_h$$

The aim is to estimate $t_y = \sum_h y_k$ being the total of individuals in the population having the sensitive feature. Once the sample has been chosen and the values m_h have been observed, we define the vector $bm = (m_1, m_2, \cdots, m_{H_s})$ with

$$\begin{aligned} \pi_{k|s,bm} &= p(k \in r|s, bm) = \frac{m_h}{n_h} = f_h = \hat{\theta}_h \\ \pi_{kl|s,bm} &= p(k, l \in r|s, bm) = \begin{cases} \frac{m_h}{n_h}\frac{m_h-1}{n_h-1} & \text{if } k, l \in s_h \quad k \neq l \\ \frac{m_h}{n_h}\frac{m_{h'}}{n_{h'}} & \text{if } k \in s_h,\ l \in s_{h'} \quad h \neq h' \end{cases} \end{aligned}$$

Then, under the presence of nonresponse and considering an arbitrary sample design $p(.)$ on the first phase, a stratified simple random sampling for the second phase sample of respondents and applying the Warner randomized response model:

$$\hat{t}_{RR\pi^*} = \sum_r \frac{\hat{y}_k}{\pi^*_k} = \frac{1}{2P-1} \sum_{h=1}^{H_s} f_h^{-1} \sum_{r_h} \frac{\check{x}_k + P - 1}{\pi_k}$$

being unbiased with its corresponding variance given by

$$Var(\hat{t}_{RR\pi^*}) = \sum\sum_U \Delta_{kl}\check{\hat{y}}_k\check{\hat{y}}_l + E_pE_{bm}\left(\sum_{h=1}^{H_s} n_h^2\frac{1-f_h}{m_h}S^2_{\check{\hat{y}}s_h}|s\right) + N\frac{P(1-P)}{(2P-1)^2}$$

with $S^2_{\check{\hat{y}}s_h}$ the variance of $\check{\hat{y}}_k = \frac{\hat{y}_k}{\pi_k}$ in s_h, $E_p(.)$ the expected value with respect to the sampling design and $E_{bm}(.|s)$ the expected value with respect to the distribution of bm given s.

4 Imai's Classic Item Count Technique

An alternative method that has recently caught the attention from researchers is the item count technique (ICT). Suppose a simple random sample with n respondents is obtained from the population. Also, suppose that there are J control items and one sensitive question. Let T_k be a random variable assigned to the k-th element where T_k = 0 means the k-th respondent that belongs to the control group is answering a partial list of J items in the control group whereas T_k = 1 means that the k-th respondent that belongs to the treatment group is answering the whole list of $J+1$ items including the sensitive question and the control questions. Suppose also that a latent indicator variable Z_{tkj} is assigned to every element in the population being equal to 1 if the answer is affirmative for the j-th item, $j = 1, \cdots, J$, and 0 otherwise. This variable depends on the assigned group for the k-th element, $t = 0, 1$. For example, if Z_{1kj} = 1 this indicates that the latent answer of the k-th element to the j-th control item is affirmative under the treatment condition. Analogously, we have the corresponding interpretations for Z_{1kj} = 0, Z_{0kj} = 1 o Z_{0kj} = 0. In particular, for $Z_{1k,J+1}$; this variable indicates the latent answer of the k-th element to the sensitive question under the treatment condition. Finally, $Z_{0k,J+1}$ has an undefined value since the questionnaire that is applied to the set of elements in the control group do not include the sensitive question.

Since the method works asking the total number of items that apply in the particular list assigned to the k-th element and not to the evaluation of every item individually, the possible answers can be defined as $Y_{1k} = \sum_{j=1}^{J+1} Z_{1kj}$ and $Y_{0k} = \sum_{j=1}^{J} Z_{0kj}$. We finally denote Y_{tk} as the random variable for total count of items that apply to the individual k depending on the assigned group (control or treatment) and y_k its corresponding observed value after the application of the survey.

Let $\pi = P(Z_{1k,J+1} = 1)$ the unknown probability of the element k having the sensitive characteristic and let $\theta_j = P(Z_{1kj} = 1)$ the assumed known probability of the element k having the nonsensitive characteristic j, $j = 1, ..., J$. According with the assumptions above, $\theta_j = P(Z_{1kj} = 1) = P(Z_{0kj} = 1)$

Analogous to the classic Imai technique, where the first sample is obtained under a simple random sampling design, the technique proceeds as follows in the general case: firstly, a first phase sample s_a is obtained according to an arbitrary sample design $p(s_a)$ of size n. From this sample s_a, a subsample s_t is obtained according to an arbitrary sample design $p(s_t|s_a)$ of size n_1 with its associated inclusion probabilities of first and second order π_{ak} and π_{akl}, respectively. The n_1 individuals in this two-phase treatment sample respond to a questionnaire with $J+1$ questions where the first J questions are the control items and the remaining question corresponds to the sensitive question. The other $n_0 = n - n_1$ individuals in the two-phase control (complement) sample $s_t^c = s_c$ are asked about only J non-sensitive questions and they only report how many of the questions apply for them but not which ones. We will denote $\pi_{k|s_a} = \sum\limits_{s_t \ni k} p(s_t|s_a) = P(T_k = 1)$ and $\pi_{k|s_a}^c = 1 - \pi_{k|s_a} = P(T_k = 0)$ as the corresponding first order inclusion probabilities for the treatment and control group in the second phase, respectively. Then, our proposed estimator for Imai s item count technique for the total population of individuals having the sensitive characteristic in a finite population is given by

$$\hat{t}_{\pi^*} = \sum_{s_a} T_k \frac{Y_{1k}}{\pi_{ak}\pi_{k|s_a}} - \sum_{s_a}(1 - T_k)\frac{Y_{0k}}{\pi_{ak}\pi_{k|s_a}^c} = \sum_{s_t} \frac{y_k}{\pi_{ak}\pi_{k|s_a}} - \sum_{s_c} \frac{y_k}{\pi_{ak}\pi_{k|s_a}^c}$$

After applying conditional variances under a two phase sampling design and under three basic assumptions at Imai (2011) paper, the total expression for the variance of our proposed estimator for the Imai's technique for finite populations equals to

$$V(\hat{t}_{\pi^*}) = \pi^2 \sum\sum_U \frac{\Delta_{akl}}{\pi_{ak}\pi_{al}} + \left[\sum_{j=1}^J \theta_j(1-\theta_j)\right] \sum_U \frac{1}{\pi_{ak}\pi_{k|s_a}\pi_{k|s_a}^c} + \pi(1-\pi)\sum_U \frac{1}{\pi_{ak}\pi_{k|s_a}} +$$
$$E_{p_a}\left(\sum\sum_{s_a} \frac{\Delta_{kl|s_a}}{\pi_{ak}\pi_{al}} \left[\frac{\pi}{\pi_{k|s_a}} + \left\{\sum_{j=1}^J \theta_j\right\}\left\{\frac{1}{\pi_{k|s_a}} + \frac{1}{\pi_{k|s_a}^c}\right\}\right] \left[\frac{\pi}{\pi_{l|s_a}} + \left\{\sum_{j=1}^J \theta_j\right\}\left\{\frac{1}{\pi_{l|s_a}} + \frac{1}{\pi_{l|s_a}^c}\right\}\right]\right)$$

5 Nonresponse Adjustments for the Item Count Technique

Following the same homogeneous group model in Section 3, the estimator for the total of people having the sensitive characteristic under the item count technique is given by the unbiased estimator:

$$\hat{t}_{RR\pi^*} = \sum_{m_t} \frac{\sum_{j=1}^{J+1} Z_{1kj}}{\pi_{ak}\pi_{k|s_a}\pi_{1k|s,s_a}} - \sum_{m_c} \frac{\sum_{j=1}^{J} Z_{0kj}}{\pi_{ak}\pi_{k|s_a}^c\pi_{0k|s,s_a}}$$

with its corresponding variance

$$Var(\hat{t}_{RR\pi^*} - t) = \pi^2 \sum\sum_U \frac{\Delta_{akl}}{\pi_{ak}\pi_{al}} +$$
$$E_a\left(\left[\sum\sum_{s_a} \frac{\Delta_{kl/s_a}}{\pi_{ak}\pi_{al}} \left(\frac{\pi + \sum_{j=1}^J \theta_j - \pi\pi_{k|s_a}}{\pi_{k|s_a}[1 - \pi_{k|s_a}]}\right) \left(\frac{\pi + \sum_{j=1}^J \theta_j - \pi\pi_{l|s_a}}{\pi_{l|s_a}[1 - \pi_{l|s_a}]}\right)\right]\right) +$$
$$E_a\left(\sum\sum_{s_a} \frac{\pi_{kl/s_a}\Delta_{1kl|s,s_a}\left(\pi + \sum_{j=1}^J \theta_j\right)^2}{\pi_{ak}\pi_{k|s_a}\pi_{1k|s,s_a}\pi_{al}\pi_{l|s_a}\pi_{1l|s,s_a}} + \sum\sum_{s_a} \frac{(1 - \pi_{k|s_a} - \pi_{l|s_a} + \pi_{kl/s_a})\Delta_{0kl|s,s_a}\left(\sum_{j=1}^J \theta_j\right)^2}{\pi_{ak}\pi_{k|s_a}^c\pi_{0k|s,s_a}\pi_{al}\pi_{l|s_a}^c\pi_{0l|s,s_a}}\right) +$$
$$\sum_U \frac{1}{\pi_{ak}}\left(\sum_{j=1}^J \theta_j(1-\theta_j)\right)\left(\frac{1}{\pi_{k|s_a}\pi_{1k|s,s_a}} + \frac{1}{\pi_{k|s_a}^c\pi_{0k|s,s_a}}\right) + \sum_U \frac{\pi(1-\pi)}{\pi_{ak}\pi_{k|s_a}\pi_{1k|s,s_a}}$$

6 Simulation Study and Areas of Further Work

In this paper, once we have obtained the finite population expressions for both methods (randomized response and item count techniques), it follows an expression for the estimation of the population total and its corresponding variance under the presence of nonresponse. In both cases, the total variance can be written in terms of three components of variance: a first component regarding to the variance of the estimator under the technique being applied, a second term considering the sampling design being applied and a third extra term due to the presence of nonresponse.

A simulation study in 1,000 fictitious populations prove that in both cases, the additional third term is not necessarily negligible and sometimes can represent up to the 30% of the variance. Then, not taking into account the presence of nonresponse can underestimate the real variance of the unbiased proposed estimator adjusted for nonresponse. Simulations were done considering a stratified simple random sampling design but we have the feeling that under more complex survey designs (e.g. unequal inclusion probabilities) the scenarios could be actually worse and this is left as an area of further work. Also, new item count techniques as the one proposed by Hussain et al. (2012) are currently studied by the authors but not presented here.

7 References

1. Blair G. and Imai K. (2010). Statistical analysis of list experiments. *Political Analysis*, 20(1), pp. 47 - 77.
2. Chaudhuri A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys*. London, UK: CRC Press.
3. Christofides T.C. (2003). Randomized response in stratified sampling. *Journal of Statistical Planning and Inference*, 128 (1): 303 - 310.
4. Droitcour J., Caspar R.A., Hubbard M.L., Parsley T.L., Visscher W. and Ezzati, T.M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. In: Biemer, P.P. et al. (eds). *Measurement Errors in Surveys*, pp. 185 - 210. New York: John Wiley and Sons.
5. Huang, K.C. (2004). A survey technique for estimating the proportion and sensitivity in a dichotomous finite population. *Statistica Neerlandica*, 58, pp. 75-82.
6. Hussain, Z., Shah, E.A. and Shabbir, J. (2012). An alternative item count technique in sensitive surveys. *Revista Colombiana de Estadistica*, 35 (1), pp. 39-54.
7. Imai, K. (2011). Multivariate regression analysis for the item count technique. Journal of the American Statistical Association, 106 (494), pp. 407 - 416.
8. Kim, J.M. and Warde, W.D. (2004). A stratified Warner randomized response model. Journal of Statistical Planning and Inference, 120, pp. 155 - 165.
9. Sarndal, C.E.; Sweenson, B. and Wretman J. (1992), Model Assisted Survey Sampling, Springer - Verlag, New York.
10. Soberanis V. and Miranda V. (2011), The generalized logistic regression estimator in a finite population sampling without replacement setting with randomized response. Revista Colombiana de Estadistica, 34(3), pp. 451-460.
11. Warner, S.L. (1965), Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60, pp. 63-69.