THE RANDOMIZED RESPONSE PROCEDURE OF HUANG : A RANKED SET SAMPLING LOOK

Carlos Bouza-Herrera
Universidad de La Habana

ABSTRACT.
In this paper we extend the scrambled response method of Huang et al (2010) to the case in which Ranked Set Sample is used. The superiority of this method is derived in terms of the diminution in the variance . The auxiliary variable is used for ranking.

KEYWORDS: Warner randomized response, scrambling, ranked set sampling gain in accuracy

MSC: 62D05

RESUMEN
En este trabajo extendemos el métodos del enmascaramiento de Huang et al (2010) al caso en que se utiliza el muestreo por rangos ordenados. La superioridad de este método es derivada en términos de la disminución de la varianza. La variable auxiliar es usada para ordenar.

1. INTRODUCTION

The use of random response (RR) was introduced by Warner (1965). It provides the opportunity of reducing response biases due to dishonest answers to sensitive questioning. Thereof this technique protects the privacy of the respondent by granting that his belonging to a stigmatized group cannot be detected. Sensitive study variables are often dealt in survey research such as proportion of adulterated milk packs of a company, proportion of illicit drugs usage. The more sensitive a question, the greater is the number of non-response or dishonest responses.
We consider the case in which we have non-responses. A sample among the non-respondents is to be taken and to this reduced group a RR technique is applied for obtaining a cooperative attitude of the interviewed when they are convinced that they will not be identified as a member of an stigmatized group.
Let the finite population under study be $U=\{u_1,..,u_N\}$. A sample S is selected from U and it is divided in S_1+S_2. The individual in S_1 give a response at the first visit while those in S_2 do not respond. The usual model fixes that 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 in the intention of not being identified as a member of a stigmatized group, say belonging to A. Hence we can stratify U accordingly by
$U_A=\{u\in U \mid u\in A\}$
$U_{A^*}=\{u\in U \mid u\notin A\}$
The interest of the inquiry is to estimate the proportion of individuals carrying the stigma, say belonging to A.
If $\mid B \mid$ denotes the number of units in B then
$\theta_A = \mid U_A \mid / \mid U \mid = N_A/N$
Another set of RR procedure is implemented by using scrambling variables.
Ranked set sampling is a sampling procedure, which is a cost effective when compared to the commonly used simple random sampling in situations where visual ordering of a set of units can be easily done, but exact measurement of units is difficult and expensive. McIntyre (1952) proposed the sample mean based on RSS as an estimator of the population mean. He found that the estimator based on RSS is more efficient than SRS. Takahasi and Wakimoto (1968) provided the necessary mathematical theory of RSS. Al-Saleh and Al-Kadiri (2000) suggested double ranked set sampling method (DRSS) for estimating the population mean, and they showed that the ranking at the second stage is easier than the ranking at the first stage. Al-Saleh and Al-Omari (2002) suggested multistage ranked set sampling (MSRSS) method to increase the efficiency when estimating the population mean for specific value of the sample size. Jemain and Al-Omari (2006a, 2006b) considered

double percentile ranked set sampling (DPRSS) and multistage median ranked set sampling (MMRSS) methods respectively for estimating the population mean. They found that DPRSS and MMRSS are more efficient than the commonly used SRS for the same sample size. Jemain et al. (2008) investigated balanced groups ranked set sampling method for estimating the population mean. Jemain et al. (2007) suggested multistage extreme ranked set sampling method for estimating the population mean. Al-Hadrami and Al-Omari (2006) considered the Bayesian inference of the variance of the normal distribution using moving extreme ranked set sampling. For more details about RSS see Bouza and Al-Omari (2014).

2. RANKED SET METHODOLOGY AND THE BALANCED DESIGN

.

The balanced RSS, see Chen et al. (2004) procedure is described as follows:

Step 1. Draw randomly *m* sets using SRSWR, each of size *m* from the population

Step 2. In each set rank the measurements with a cost free method.

Step 3.Then, from the first set the element with the smallest rank is chosen for the actual measurement. From the second set the element with the second smallest rank is chosen.

The process is continued by selecting the ith order statistics of the ith random sample until the element with the largest rank from the mth set is chosen.

The scheme yields the following data $\{X_{i\,m}\}$ 1,2,...,where $[i{:}m]$ is the ith order statistics of the ith random sample of size m, and it is denoted by the ith judgment order statistics. It can be noted that the selected elements are independent order statistics but not identically distributed.

Usually, from both practical and theoretical issues m is kept small. A sample of size n is obtained by repeating the procedure r times up to obtaining n=mr observations. $\{X_{i\,m\,j\,r}\}$ where $i\ m\ j$ the ith judgment order statistics in the jth cycle, which is the ith order statistics of the ith random sample of size m in the jth cycle. It should be noted that all of order statistics are mutually independent.

When the ranking variable is a scrambling variable the rank may be made using the variables introduced by the RR procedure. The inquiry may be developed as follows. The respondents generate the scrambling variable and they communicate the values and then are ranked. Now ξ is the scrambled variable and the distribution is completely known by the sampler.

3. THE RANDOMIZED RESPONSE PROCEDURE OF HUANG

Let Y be a sensitive variable evaluated in a finite population $U=\{u_1,\ldots,u_N\}$. The individual u_i , with a value of Y that carries a stigma , will tend to give incorrect information or to refuse to answer. It is well known that when dealing with sensitive questions we should face the need to reduce the refusals to respond and the response bias.

The use of randomized responses (RR) models is still placing challenges both for the theory of survey sampling as well as for its application. Different extensions of RR have been introduced, see Chaudhuri-Mukerjee R (1988) for the earlier proposals. The interest in RR methods to estimate the mean of a sensitive quantitative variable is increasing. RR methods for scrambling the responses have been proposed by different authors. See for example Eichhorn-Hayre (1983), Gupta et al. (2002), Ryu et al. (2005).

Let us consider a sensitive quantitative variable Y and a known probability distribution P is determined for selecting a value $Z \in Z^*=\{Z^*_1,\ldots,Z^*_k\}$ independent of Y . As Z^* is determined by the statistician both the mean and variance of Z, μ_Z and σ^2_Z , are known a priori. Each selected person $i \in s$, $i=1,\ldots,n$ responds with a coded value.

The RR proposed by Huang et al (2010) is to use a pair of randomization mechanisms for collecting responses in each of the independent samples of size n_i (i=1,2), drawn from using SRSWR . The respondent $j \in s_i$ selects randomly between reporting the true response Y or the scrambled response

$$Z^*_{ij} = S_{ij}Y_j + D_{ij}$$

For reporting the interviewee performs the experiments Bernoulli A, B y C, such that E(S)=T, E(C)=W and the report is modeled by

$$Z^*_{ij} = (1 - C_j)Y_j + C_j\left(S_{ij}Y_j + D_{ij}\right)$$

The expected value and variance of the report are

$$E(Z^*_{ij}) = (1 - W)\mu_Y + W(\mu_Y + \bar{D}_i)$$

$$V_{Hi}^2 = V(Z_{ij}^* \ ) = \sigma_Y^2 + W(\mu_Y^2 + \sigma_Y^2)\mathcal{S}_i^2 + W(1-W)\overline{D}_i^2 + W\delta_i^2$$

An unbiased estimator of the mean of Y is easily derived . It is

$$\hat{\mu}_H = \frac{\overline{D}_1\bar{Z}_2^* - \overline{D}_2\bar{Z}_1^*}{\overline{D}_1 - \overline{D}_2}, \bar{Z}_i^* = \frac{\sum_{j=1}^{n_i} Z_{ij}^*}{n_i}, i = 1,2$$

Its variance is given by

$$V(\hat{\mu}_H \ ) = \frac{\overline{D}_1^2 V_{H2}^2}{n_2(\overline{D}_1 - \overline{D}_2)^2} + \frac{\overline{D}_2^2 V_{H1}^2}{n_1(\overline{D}_1 - \overline{D}_2)^2},$$

The estimation of the sensitive is derived from the same experiment by computing

$$\hat{\mu}_H^W = \frac{\bar{Z}_2^* - \bar{Z}_1^*}{\overline{D}_1 - \overline{D}_2}$$

Its variance is

$$V(\hat{\mu}_H^W) = \frac{V_{H2}^2}{n_2(\overline{D}_1 - \overline{D}_2)^2} + \frac{V_{H1}^2}{n_1(\overline{D}_1 - \overline{D}_2)^2}$$

Let us consider the RR procedure developed by Huang et al (2010). The proposal is to use a pair of randomization mechanisms for collecting responses in each of the independent samples of size n_i (i=1,2), drawn from using SRSWR . The respondent $j \in s_i$ selects randomly between reporting the true response Y or the scrambled response

$$Z_{ij}^* = S_{ij}Y_j + D_{ij}$$

For reporting the interviewee performs the experiments Bernoulli A, B y C, such that E(S)=T, E(C)=W and the report is modeled by

$$Z_{ij}^* = (1 - C_j)Y_j + C_j\left(S_{ij}Y_j + D_{ij}\right)$$

The expected value and variance of the report are

$$E(Z_{ij}^*) = (1-W)\mu_Y + W(\mu_Y + \overline{D}_i)$$

$$V_{Hi}^2 = V(Z_{ij}^* \ ) = \sigma_Y^2 + W(\mu_Y^2 + \sigma_Y^2)\mathcal{S}_i^2 + W(1-W)\overline{D}_i^2 + W\delta_i^2$$

An unbiased estimator of the mean of Y is easily derived . It is

$$\hat{\mu}_H = \frac{\overline{D}_1\bar{Z}_2^* - \overline{D}_2\bar{Z}_1^*}{\overline{D}_1 - \overline{D}_2}, \bar{Z}_i^* = \frac{\sum_{j=1}^{n_i} Z_{ij}^*}{n_i}, i = 1,2$$

Its variance is given by

$$V(\hat{\mu}_H \ ) = \frac{\overline{D}_1^2 V_{H2}^2}{n_2(\overline{D}_1 - \overline{D}_2)^2} + \frac{\overline{D}_2^2 V_{H1}^2}{n_1(\overline{D}_1 - \overline{D}_2)^2},$$

The estimation of the sensitive is derived from the same experiment by computing

$$\hat{\mu}_H^W = \frac{\bar{Z}_2^* - \bar{Z}_1^*}{\overline{D}_1 - \overline{D}_2}$$

Its variance is

$$V(\hat{\mu}_H^W) = \frac{V_{H2}^2}{n_2(\overline{D}_1 - \overline{D}_2)^2} + \frac{V_{H1}^2}{n_1(\overline{D}_1 - \overline{D}_2)^2}$$

4. RSS FOR THE RR OF HUANG

Let us develop the RSS model for the RR procedure proposed by Huang et al (2010). The sampler ranks the units selected. As in the SRSWR case we assume that the respondent $j \in s_{ijt}$ selects randomly between reporting the true response Y or the scrambled response

$$Z_{i(j)t}^* = S_{ijt}Y_{i(j)t} + D_{ijt}$$

The report is modeled by:

$$Z_{i(j)t}^* = (1 - C_{ijt})Y_{i(j)t} + C_{ijt}\left(S_{ijt}Y_{i(j)t} + D_{ijt}\right)$$

The expected value and variance of the report are

$$E(Z_{i(j)t}^*) = (1-W)\mu_{Y_{(j)}} + W(\mu_{Y_{(j)}} + \overline{D}_i)$$

$$V_{Hi(j)}^2 = V(Z_{i(j)t}^* \ ) = \sigma_{Y_{(j)}}^2 + W\left(\mu_{Y_{(j)}}^2 + \sigma_{Y_{(j)}}^2\right)\vartheta_i^2 + W(1-W)\bar{D}_i^2 + W\delta_i^2$$
$$= \sigma_Y^2 - \Delta_{(j)}^* + W\left(\mu_{Y_{(j)}}^2 + \sigma_Y^2 - \Delta_{(j)}^*\right)\vartheta_i^2 + W(1-W)\bar{D}_i^2 + W\delta_i^2$$
$$= \sigma_Y^2\left(1 + W\vartheta_i^2\right) + W\vartheta_i^2\mu_{Y_{(j)}}^2 + W(1-W)\bar{D}_i^2 + W\delta_i^2 - \Delta_{(j)}^*(1 + W\vartheta_i^2), \ \Delta_{(j)}^*$$
$$= \left(\mu_{Y_{(j)}} - \mu_Y \ \right)^2, j = 1,..,m_i; i = 1,2$$

Take

$$\hat{\mu}_{HRSSi} = \frac{\sum_{t=1}^{r_i} \ \sum_{j=1}^{m_i} \ Z_{i(j)t}^*}{m_i r_i} = \frac{\sum_{t=1}^{r_i} \ \sum_{j=1}^{m_i}(1-C_{ijt})Y_{i(j)t}}{m_i r_i} + \frac{\sum_{t=1}^{r_i} \ \sum_{j=1}^{m_i} \ C_{ijt}\left(S_{ijt}Y_{i(j)t} + D_{ijt}\right)}{m_i r_i}, i = 1,2$$

Its expectation is

$$E(\hat{\mu}_{HRSSi}) = \frac{(1-W)\sum_{j=1}^{m_i}\mu_{Y_{(j)}}}{m_i} + \frac{W\sum_{j=1}^{m_i} \ \mu_{Y_{(j)}}}{m_i} + W\bar{D}_i = \mu_Y + W\bar{D}_i, i = 1,2$$

Then, we have the unbiasedness of

$$\hat{\mu}_{HRSS} = \frac{\bar{D}_1\hat{\mu}_{HRSS2} - \bar{D}_2\hat{\mu}_{HRSS1}}{\bar{D}_1 - \bar{D}_2}$$

Due to the independence

$$V(\hat{\mu}_{HRSS}) = \frac{\bar{D}_1^2 V(\hat{\mu}_{HRSS2}) + \bar{D}_2^2 V(\hat{\mu}_{HRSS1})}{(\bar{D}_1 - \bar{D}_2)^2}$$

where

$$V(\hat{\mu}_{HRSSi}) = \frac{\sigma_Y^2\left(1 + W\vartheta_i^2\right)}{n_i} - \frac{(1 + W\vartheta_i^2)\sum_{j=1}^{m_i} \ \Delta_{(j)}^*}{n_i m_i} + \frac{W\vartheta_i^2\sum_{j=1}^{m_i} \ \mu_{Y_{(j)}}^2}{n_i m_i} + \frac{W(1-W)\bar{D}_i^2 + W\delta_i^2}{n_i}, \ \Delta_{(j)}^*$$
$$= \left(\mu_{Y_{(j)}} - \mu_Y \ \right)^2, j = 1,..,m_i; i = 1,2$$

Comparing this variances with the corresponding to the SRSWR model we have

$$V(\hat{\mu}_{HRSSi}) - V_i^{*2} = W\vartheta_i^2\left(\frac{\sum_{j=1}^{m_i} \ \mu_{Y_{(j)}}^2}{n_i m_i} - \mu_Y^2\right) - \frac{(1+W\vartheta_i^2)\sum_{j=1}^{m_i} \ \Delta_{(j)}^*}{m_i \ n_i} = -\frac{\sum_{j=1}^{m_i} \ \Delta_{(j)}^*}{m_i \ n_i}, \text{ i=1,2}$$

Hence

$$G_H(SRSWR, RSS) = V(\hat{\mu}_{HRSS}) - \ V(\hat{\mu}_H \ ) = \sum_{i=1}^{2} \ \frac{\bar{D}_i^2}{(\bar{D}_1 - \bar{D}_2)^2}\frac{\sum_{j=1}^{m_i} \ \Delta_{(j)}^*}{m_i \ n_i}$$

Is the gain in accuracy due to using RSS. As it is positive RSS should be preferred

In the RSS case the sensitive is estimated unbiasedly by

$$\hat{\mu}_{HRSS}^W = \frac{\hat{\mu}_{HRSS2} \ - \hat{\mu}_{HRSS1}}{\bar{D}_1 - \bar{D}_2}$$

Its variance is

$$V(\hat{\mu}_H^W) = \frac{V(\hat{\mu}_{HRSS2})}{(\bar{D}_1 - \bar{D}_2)^2} + \frac{V(\hat{\mu}_{HRSS1})}{(\bar{D}_1 - \bar{D}_2)^2}$$

From the above given results we have that the gain in accuracy of the proposed estimator is

$$G_H^W(SRSWR, RSS) = \frac{1}{(\bar{D}_1 - \bar{D}_2)^2}\sum_{i=1}^{2} \ \frac{\sum_{j=1}^{m_i} \ \Delta_{(j)}^*}{m_i \ n_i}.$$

Acknowledgments. The author thanks the two anonymous referees for their valuable comments. This paper was benefited by the support of projects PNCB of CITMA and CAPES-MES 209-13

RECEIVED DECEMBER, 2014
REVISED MAY, 2015

REFERENCES

[1] AL-HADHRAMI, S.A. & AL-OMARI, A.I.(2006): Bayesian inference on the variance of normal distribution using moving extremes ranked set sampling. ***Journal of Modern Applied Statistical Methods***, 8, 273-281.
[2] AL-OMARI, A.I and BOUZA, C. N. (2014): Review Of Ranked Set Sampling: Modifications and Applications. **Revista Investigación Operacional,** 3, 215-240**.**
[3] AL-SALEH, M.F. & AL-KADIRI, M. (2000): Double ranked set sampling. ***Statistics & Probability Letters***, 48: 205–212.
[4] BAR-LEV, S. K., E. BOBOVITCH and B. BOUKAI (2004): A note on randomized response models for quantitative data. **Metrika** , 60: 255—260.
[5] BOUZA, C. N. (2005): Sampling Using Ranked Sets:Concepts, Results and Perspectives**. Revista Investigación Operacional.** 26, 3,275-293
[6] CHAUDHURI, A. (2004): Christofides' randomized response technique in complex sample surveys, **Metrika,** 60, 223–228.
[7] CHEN, Z., Z. BAI AND K. B. SINHA.(2004**): Ranked Set Sampling**. Springer, Berlin.
[8] EICHHORN B. H. and HAYRE L. S.(1983): Scrambled Randomized Response Methods for Obtaining Sensitive Quantitative Data. **J of StatisticalPlanning and Inference,** 7, . 307-716.
[9] GUPTA, S., GUPTA, B., and SINGH, S. (2002), Estimation of Sensitivity level of personal interview survey questions. **Journal of Statistical Planning and inference**, 100, 239-247.
[10] HUANG, K. C**.** (2010) Unbiased estimators of mean, variance and sensitivity level for quantitative characteristics in finite population sampling. **Metrika** 71, 341–352.
[11] JEMAIN, A.A. & AL-OMARI, A.I. (2006a):. Double percentile ranked set samples for estimating the population mean. **Advances and Applications in Statistics**, 6, 261-276.
[12] JEMAIN, A.A., & AL-OMARI, A.I. (2006b): Double quartile ranked set samples. **Pakistan Journal of Statistic*s***, 22, 217-228.
[13] JEMAIN, A.A., AL-OMARI, A.I. & IBRAHIM, K. (2007a): Multistage median ranked set sampling for estimating the population median. **Journal of Mathematics and Statistics**, 3, 58-64.
[14] JEMAIN, A.A., AL-OMARI, A.I. & IBRAHIM, K. (2008): Some variations of ranked set sampling. **Electronic Journal of Applied Statistical Analysis**, 1, 1-15.
[15] MCINTYRE, G. A. (1952): A Method of Unbiased Selective Sampling Using Ranked Set. **Australian, J. Agricultural Research**, 3, 385-390.
[16] RYU, J. -B., KIM, J. -M., HEO, T. -Y. and PARK, C. G.(2005): On stratified Randomized response sampling. **Model Assisted Statistics and Application**,. 1, 31–36.
[17] WARNER, S. L. (1965): Randomized response: a survey technique for eliminating evasive answer Bias. **Journal of American Statistical Association**, 60, 63–69.
[18] TAKAHASi K. and WAKIMOTO, K. (1968): On unbiased estimates of the population mean based on sample stratified by means of ordering. **Annals of the Inst. of Statistical Mathematics**. 20, 1-31