RANDOMIZED ESTIMATION A PROPORTION USING RANKED SET SAMPLING AND WARNER'S PROCEDURE

Agustin Santiago*, Jose M. Sautto*, Carlos N. Bouza**
*Universidad Autónoma de Guerrero
**Universidad de LA Habana

ABSTRACT
This paper is devoted to the analysis of the estimation of the proportion of a sensitive variable. The use of a randomized response (rr) procedure gives confidence to the interviewed that his privacy is protected. We consider that a simple random sampling with replacement design is used for selecting a sample. The behavior of the rr procedure, when ranked set sampling is the design used, is developed under three different ranking criteria. The usual gain in accuracy associated with the use of ranked set sampling is exhibited only by one of the designs. The behavior of the models is illustrated using data provided by a study of samples of persons infected with AIDS.

KEYWORDS: Order statistics, ranked set sampling, randomized response

MSC: 62D05

RESUMEN
Este paper se dedica al análisis de la estimación de la proporción de una variable sensible. El uso de un procedimiento de respuestas aleatorizadas (rr) brinda confianza al entrevistado de que su privacidad está protegida. Consideramos que se utiliza muestreo simple aleatorio para selección la muestra. . El comportamiento de procedimiento de rr, al usar como diseño el muestreo por conjuntos ordenados, lo que es desarrollado para tres criterios diferentes de ordenación. La usual ganancia en precisión es exhibida solo para uno de los diseños. El comportamiento de los modelos es ilustrado usando datos provenientes de un estudio con muestras de en personas infectadas con AIDS.

PALABRAS CLAVE: estadísticos de orden, muestreo por conjuntos ordenados, respuestas aleatorizadas

1. INTRODUCTION

Commonly it is required to obtain information on sensitive attributes and a sample is selected for interviewing a sample of persons. Collecting trustworthy responses on sensitive issues through direct questioning in personal interviews using techniques is not often successful because they do not protect the respondents' privacy. Therefore in practice the data collected on sensitive features are affected by the existence of respondent bias.

Randomized response models are used to decrease the both non responses and answer bias and to provide privacy protection to the respondents.

Warner (1965) proposed the randomized response (RR) method as a mean of avoiding response bias. The initial model looked for the estimation of the proportion of persons with the stigma. The model used a randomized trial. The seminal paper of Warner (1965) has 50 years of being created and still different contributions are being generated. The models are generally based on the selection a sample using simple random sampling with replacement.

Consider a population U of size N with two strata U_A and U_{A*}. Therefore to conduct an inquiry is serious issue, to belong to U_A is stigmatizing. Hence the respondents will tend using of random response (RR). It provides the opportunity of reducing response biases due to dishonest answers to sensitive questioning. Therefore this technique protects the privacy of the respondent by granting that his belonging to a stigmatized group cannot be detected. The interest of the inquiry is to estimate the proportion of individuals carrying a stigma, identified with the belonging to A. If $|A|$ denotes the number of units with the stigma and we are interested in estimating the probability $P(A) = |U_A| /|U| = N_A/N$.

The RR technique has been successfully applied in many areas and different modifications and extensions on this method have proposed the literature on sampling. It is still receiving attention from the researchers, see for example Gupta et al. (2002) and Ryu et al. (2005), Sahai (2006).

A challenging sampling design is ranked set sampling (rss). It was suggested by McIntyre (1952) and appears as a more efficient then srswr. Takahasi-Wakimoto (1968) and Dell-Clutter (1972) gave a mathematical support to rss and the list of new results is growing rapidly. See Patil (2002) for a review on the theme. Recently Chen et al. (2006) suggested a randomized response model for ordered categorical variables. They used an ordinal logistic regression for ranking. We present these results in section 2. Considering that a sensitive variable is evaluated, we consider the use of RR for collecting the information. We develop an extension of the RSS estimator of Chen et al. (2006) using Warner's model. The proposal is presented in section 3. The derived variance of the proposed estimator is larger than the variance of Chen's proposal. Considering that a sensitive question is evaluated we suspect that its use will reduce answer biases. Section 4 develops a study using real life data. The experiments sustained our suspecting. The answers to the direct question of the interviewed produced estimations more different than the real one. The proposed estimator was closer. This facts support the recommendation of using it for obtaining a gain in accuracy with respect to the usual simple random sampling with replacement (srswr) model.

2. RANKING ORDERED CATEGORICAL VARIABLES

The proposal of Chen et al. (2006) for ordered categorical variables allows to use RSS. They used a set of explanatory variables $Z = (Z_1, \dots, Z_k)$ for fitting a logistic regression. Take the variable of interest X_j in an item where

$$X_j = i \text{ if item } j \text{ is classified in the class } C(i)$$

Hence the probability distribution function is the multinomial $M(1, P_1, \dots, P_q)$, $P_i = Prob\{X = i\}, i = 1, \dots, q$. Initially a random sample is selected and in each sample item are measured **Z** and **X***. The ordinal logistic regression (ORL) is fitted to the data using a statistical package. Considering
$c_i = P(\text{classifying an item in a category 1 to i}) = \sum_{t \le i} P_t,\ i = 1, \dots, q$
The logit function is $logit(c_i) = log\left(\frac{c_i}{1-c_i}\right) = L_i$. Using the collected data the fitted logit model is the proportional odds model

$$L_i = \alpha_i + \boldsymbol{\beta^T z}, \qquad i = 1, \dots, q$$

The model's probability of classifying a particular item r in the *i-th* category is denoted π_{ri} and its cumulative probability by c_{ri}. The model fitted produces the corresponding estimates $\hat{\pi}_{ri}$ and $\hat{c}_{ri}$.
The procedure proposed by Chen et al (2006) considers the selection of a random sample of size *m* using srswr. The class of the *r-th* judgmental order statistic for X is denoted by $X_{(r)}$. The ranking is made as follows

Chen et al (2006) Ranking procedure for ordinal variables

Step 1 Use the fitted model and compute $\{\hat{\pi}_{ri,}\ \hat{c}_{ri}\}, i = 1, .., q, r = 1, .., m$

Step 2 Classify item r in the category h such that $\hat{\pi}_{hi} = Max\{\hat{\pi}_{ri}, i = 1, .., q\}, r = 1, \dots, m.$

Step 3 $Rank(r) > Rank\ (r *)$ if r is assigned to $C(i)$ and $r *$ to $C(j)$ being $j < i$.

Step 4 An item in $C(i)$ is ranked using the computed $\hat{c}_{ri}'s$: $Rank(r) > Rank(r *)$ if $\hat{c}_{ri} < \hat{c}_{r^*i}$.

The procedure is repeated n_r times for each $X(r) - class, t = 1, .., m$. For the experiment j the item with $rank(j)$ is interviewed. The RSS sample sets is

$$\begin{array}{ccc} X_{(1)1} \cdots & X_{(1)t} \cdots & X_{(1)n_1} \\ \vdots & \vdots & \vdots \\ X_{(r)1} \cdots & X_{(r)t} & X_{(r)n_t} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ X_{(m)1} \cdots & X_{(m)t} & X_{(m)n_m} \end{array}$$

The $n_t's$ are not necessarily equal. The use of an equal number of experiments yields a balanced RSS sampling design; in other case it is unbalanced.
The *r-th row* is a sample from the stratum defined by the *r-th* order statistic. The probability of mass function is $p_{(r)i}, i = 1, .., q$.
Let us consider the particular case in which the interviewed persons are questioned to declare the belonging to a certain group *A*. The response can be modeled as

$$I[X_{(r)j}] = \begin{cases} 1 & \text{if a YES is the answer} \\ 0, & \text{otherwise} \end{cases}$$

We are interested in estimating $\theta(A)$, the proportion of persons belonging to A in the population. $\theta(A)$ may be estimated using the RSS proposed by Chen et al (2004) by

$$p_c = \frac{\sum_{r=1}^{m} \frac{1}{n_r} \sum_{t=1}^{n_r} I\left[X_{(r)j}\right]}{m}$$

Now we have $p_{(r)A} = \mu(r)$ and $m\mu = mP(A) = \mu_{(1)} + \cdots + \mu_{(m)}$. Hence

$$E(p_c) = \frac{\sum_{r=1}^{m} \frac{1}{n_r} \sum_{t=1}^{n_r} p_{(r)A}}{m} = \theta(A)$$

It has been derived; see that the variance of the statistics of order r is

$$\sigma_{(r)}^2 = \sigma^2 - (\mu_{(r)} - \mu)^2$$

Therefore we may consider that

$$V\left(I\left[X_{(r)j}\right]\right) = p_{(r)A}\left(1 - p_{(r)A}\right) = \theta(A)\left(1 - \theta(A)\right) - (p_{(r)A} - \theta(A))^2$$

and, as result, taking $\vartheta = \sum_{t=1}^{n_r} \frac{1}{n_r}$

$$V(p_c) = \sum_{r=1}^{m} \frac{p_{(r)A}\left(1 - p_{(r)A}\right)}{m^2} \sum_{t=1}^{n_r} \frac{1}{n_r} = \frac{\theta(A)(1 - \theta(A))}{m} - \vartheta \sum_{r=1}^{m} \frac{(p_{(r)A} - \theta(A))^2}{m^2}$$

The second sum is positive and represent the gain in accuracy due to the use of the proposal of Chen et al (2006) with respect to the use of srswr.

The optimal choice of the sample sizes is given by the expression

$$n_{r(opt)} = n \frac{\sqrt{p_{(r)A}\left(1 - p_{(r)A}\right)}}{\sum_{r=1}^{m} \frac{\sqrt{p_{(r)A}\left(1 - p_{(r)A}\right)}}{m}}, \qquad n = \sum_{r=1}^{m} n_r$$

It establishes that the order statistics with larger standard deviation should have larger samples sizes. That is, the order statistics with smaller gains in accuracy measured by $(p_{(r)A} - \theta(A))^2$.

We will consider the case in which A is a sensitive group and evaluate the behavior of this sampling design when a randomized response mechanism is introduced for obtaining the responses.

3. A RANDOMIZED RESPONSE STRATEGY

The probability of carrying a stigma $\theta(A)$ is the parameter to be estimated. The usual approach is to ask to a selected individual if he/she belongs to A (to carry the stigma). Warner (1965) proposed to provide a random mechanism to the interviewed who develops an experiment that selects the between the statements:

1. I belong to A, with probability $p \neq 0{,}5$ and
2. I do not belong to A, with probability $1 - p$.

The evaluated variable is

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

The individual does not reveal which statement is evaluating. The random sample permits to evaluate the number of Yes

$$n_Y = \sum_{I=1}^{n} Y_i$$

Commonly each respondent in the sample is asked to select a card form a deck after shuffling. The deck has a proportion p of cards with the statement 1. After de selection the respondent answers Yes of No without revealing the selected statement. This technique is known as the related question method. Warner (1965) derived that

$$p_W = \frac{\frac{n_y}{n}}{2p - 1} + \frac{p - 1}{2p - 1}$$

is the maximum likelihood estimator of θ(A). It is unbiased and its existence is supported by the use of $p \neq 0{,}5$. Its variance is

$$V(p_w) = \frac{\theta(A)(1 - \theta(A))}{n} + \frac{p(1 - p)}{n(2p - 1)^2}$$

The second term in the above expression is the increase in the variance due to the introduction of the randomized mechanism.

Let us consider the use of this RR model when RSS is used.

After conforming the RSS sample using Chen et al. (2006) procedure the interviewed uses the RR mechanism for selecting the statement to be evaluated. The response obtained will be again

$$I\left[X_{(r)j}\right] = \begin{cases} 1 \text{ if a YES is the answer} \\ 0 \qquad\qquad \text{otherwise} \end{cases}$$

but

$$Prob\left(I\left[X_{(r)j}\right] = 1\right) = pp_{(r)A} + (1-p)(1-p_{(r)A})$$

Now the estimator of the probability of carrying the stigma for the sample of the class of the *r–th* order statistics is,

$$\hat{p}_{W(r)A} = \frac{\sum_{t=1}^{n_r} I\left[X_{(r)j}\right]}{n_r(2p-1)} - \frac{1-p}{n_r(2p-1)}$$

A naïve estimator based for the RSS sample is

$$\hat{p}_{cW} = \frac{\sum_{r=1}^{m} \hat{p}_{W(r)A}}{m} = \sum\nolimits_{r=1}^{m} \left(\frac{\sum_{t=1}^{n_r} I\left[X_{(r)j}\right]}{mn_r(2p-1)} - \frac{1-p}{mn_r(2p-1)} \right)$$

Its unbiasedness follows form the fact that, for any $r = 1, \ldots, m$,

$$E\left(\hat{p}_{W(r)A}\right) = \frac{n_r\left[pp_{(r)A} + (1-p)(1-p_{(r)A})\right]}{n_r(2p-1)} - \frac{1-p}{n_r(2p-1)} = p_{(r)A}$$

The variance of the estimator is readily obtained as

$$V(\hat{p}_{cW}) = \frac{\sum_{r=1}^{m} V\left(\hat{p}_{W(r)A}\right)}{m^2} = \sum\nolimits_{r=1}^{m} \frac{p_{W(r)A}(1-p_{W(r)A})}{mn_r} + \vartheta \frac{p(1-p)}{m(2p-1)^2} - \vartheta \sum\nolimits_{r=1}^{m} \frac{(p_{W(r)A} - \theta(A))^2}{m^2}$$

The second term represents and increment in the variance due to the use of the randomization procedure. In practice the non sampling error produced by providing incorrect answers, for avoiding to be stigmatized, is present when direct questions are made.

We performed a large study for evaluating the behavior of the proposal when managers are interviewed for establishing the quality of the protection of the environment by their enterprises.

4. EVALUATION OF THE PERFORMANCE OF $\hat{p}_{cW}$

To test the model proposed we interviewed the directors of different enterprises that produce highly contaminated garbage. They were asked to report if they send contaminated garbage to the municipal sites. They gave the report. Afterwards they were provided of the set of cards where the 60% of the cards fixed the selection of the sensitive question, *p=0.60*, "The enterprise contaminates the environment".

The cards were shuffled by the interviewed for reporting "yes or no".

The characterization of leaching of elements from solid waste compost was made by evaluating samples of grab. We consider that it provided the real result. That is, a "Yes or No" was produced by analyzing the grab. The grab samples were prepared from multiple grab samples using coning and quartering methods. The compost was collected from composting facilities which was screened reducing the particles mechanically six times separated in a trammel and passed through a fine. The type of grab came from aliment, metallurgical, textile and chemical factories. The grab sample procedure is described in Tisdell and Breslin (1995).

We considered *3* different sets of variables for fitting the logistic regression. The measurement of contamination in the air and the rivers, of the basin used for sending the residuals of the industries, produced the explanatory variables. The reports of the closest monitoring station were used for measuring them in a large research conducted for detecting the highly contaminating enterprises. Table 1 gives a description of them. An inspection to the enterprises established if they were contaminating the environment. The inquiry took place a year after the auditing performed. The objective of it was to check if they changed their status. Presumably the managers will avoid declaring the incompetence to solve the problems detected previously.

Table 1. Logistic regression models used for estimating the proportion of contaminating enterprises.

Model	Explanatory variables
WQ: main metallic contaminators in the river	Percentage of plumb, chrome and nickel
AQ: main contaminators of the quality of the air	Percentage of Sulphuric acid and carbon dioxide
GQ: main metallic contaminators in the river and main contaminators of the quality of the air	Percentage of plumb, chrome, nickel, Sulphuric acid, and carbon dioxide

The population a census was performed. The collected population data were sampled. Three sampling fractions were used *f=0.05, 0.10* and *0.20.* The evaluation of the behavior of the estimators was made by selecting *1000* samples using each sample fraction.

Table 2. Average of *1000* proportion estimates for *m=2, n_r=10*

Model	Aliment factories.		Metallurgical factories		Textile factories		Chemical factories	
True proportion	0,87		0,78		0,90		0,85	
Model	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$
WQ	0,74	0,89	0,65	0,72	0,45	0,92	0,55	0,86
AQ	0,75	0,87	0,51	0,73	0,68	0,90	0,67	0,82
GQ	0,76	0,84	0,62	0,71	0,77	0,89	0,63	0,79

Table 2 presents the average of the proportions computed with the two estimators for an overall sample size *n=20* with *m=2* and constant value of the *n_r`s*. It is clear that the managers cheated. The direct responses produced and under estimation of the true proportion. The use of the RR allows obtaining a closer estimation.

Table 3. Average of *1000* proportion estimates for *m=4, n_r=4*

Model	Aliment factories.		Metallurgical factories		Textile factories		Chemical factories	
True proportion	0,87		0,78		0,90		0,85	
Model	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$	$\hat{p}_c$	$\hat{p}_{cW}$
WQ	0,75	0,86	0,62	0,70	0,41	0,92	0,57	0,81
AQ	0,74	0,87	0,55	0,71	0,65	0,93	0,69	0,81
GQ	0,72	0,88	0,64	0,70	0,74	0,91	0,62	0,76

Table 3 is devoted to present the average of the proportions computed with the two estimators with *m=4, n_r=4*. The comparison of them leads to a similar conclusion. Note that it seems to be better to us e use of the RR allows to obtain a closer estimation.

Table 4 Computed ε_u. For $u = c, cW$ and for *m=2, n_r=10*

Model	Aliment factories		Metallurgical factories		Textile factories		Chemical factories	
Model	ε_c	ε_{cW}	ε_c	ε_{cW}	ε_c	ε_{cW}	ε_c	ε_{cW}
WQ	1,87	0,81	1,85	0,89,	1,90	0,91	1,87	0,81
AQ	1,92	0,80	1,91	0,88	1,,96	0,92	1,92	0,80
GQ	1,91	0,75	1,91	0,91	1,93	0,92	1,91	0,75

The accuracy of the estimators was analyzed computing.

$$\varepsilon_u = \sum_{h=1}^{1000} \frac{|\hat{p}_u - \theta(A)|_u}{1000\theta(A)}, \qquad u = c, cW$$

The results are given en Table 4 and 5. The direct question is considerably more inaccurate than the randomized one.

Table 5 Computed ε_u. For $u = c, cW$ and for *m=4, n_r=5*

Model	Aliment factories.		Metallurgical factories		Textile factories		Chemical factories	
Model	ε_c	ε_{cW}	ε_c	ε_{cW}	ε_c	ε_{cW}	ε_c	ε_{cW}
WQ	1,91	0,87	1,92	0,89	1,93	0,90	1,91	0,87
AQ	1,91	0,88	1,93	0,89	1,93	0,91	1,91	0,88
GQ	1,92	0,87	1,93	0,90	1,94	0,93	1,92	0,87

Note that the estimator based on the randomized response procedure performs better for the smaller value of m.

5. **CONCLUSIONS**

From the results obtained it can be seen that the estimates of the proportion of individuals with stigma, based on the randomized response procedure, are closer to the true proportion than the direct estimate obtained from the responses of the staff of the Business. In all the factories there are dishonest answers, although we can identify those that were more dishonest in their answers. First appears the textile factory, followed by chemical and metallurgical factories. The factory that had the highest proportion of honest answers was food.

Now, with respect to the estimators of randomized response, where there is greater precision are the food and textile factories, although in the latter there is an overestimation of the proportion. As expected, when the sample size n_r is smaller, the accuracy of the estimates is worse and, if the sample size is large and m is smaller, the accuracy of the estimators is better.

RECEIVED: MAY, 2018
REVISED: SEPTEMBER, 2018

REFERENCES

[1] BOUZA, C. N. (2010): A Review of Randomized Responses Procedures: the Qualitative Variable Case. **Investigación Operacional**, 30, 240-247.

[2] DELL G.P. and J.L. CLUTTER (1972): Ranked Set Sampling Theory With Order Statistics Background. Biometrics. 28, 545-553.

[3] CHRISTOFIDES, T. C. (2003): A generalized response technique. **Metrika** 57, 195-200.

[4] 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.

[5] MCINTYRE, G. A. (1952): A Method For Unbiased Selective Sampling Using Ranked Sets. **Australian J. of Agriculture Res.** 3, 385-390.

[6] PATIL, G.P (2002): Ranked Set Sampling. In **Encyclopedia of Enviromentrics**. (A.H. El-Shaarawi and W.W. Pieegoshed, (Editors). 3,1684-1690. Wiley, Chichester

[7] 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.

[8] SAHA, A. (2006): A Generalized Two-Stage Randomized Response Procedure In Complex Sample Surveys. **Aust. N. Z. J. Stat**. 48, 429–443

[9] SINGH, S. (2002): Randomized response model. **Metrika**, 56, 131-142.

[10] TAKAHASHI K. and WAKIMOTO, K. (1968): On Unbiased Estimates of the Population Mean Based on Sample Stratified by Means of Ordering. **Annals Inst. of Statistical Mathematics**. 20, 1-31.

[11] TISSDEL, S.E. and BRESLIN V. T. (1995): Characterization of Leaching of Element from Municipal Solid Waste **Compost. J. of Environmental Quality**. 24, 827-833

[12] WARNER, S. L. (1965): Randomized Response: A Survey Technique For Eliminating Evasive Answer Bias,.J. **Amer. Statist. Assoc**., 60, 63-69.