International Journal of Applied Mathematics
& Statistical Sciences (IJAMSS)
ISSN(P): 2319-3972; ISSN(E): 2319-3980
Vol. 6, Issue 2, Feb-Mar 2017; 55-62
© IASET

RANKED SET SAMPLING STRATEGIES FOR THE ELIMINATED SCRAMBLING VARIANCE RESPONSE MODELS

CARLOS N. BOUZA-HERRERA[1], AGUSTIN SANTIAGO[2] & JOSE M. SAUTTO[3]

[1] Universidad De La Habana, Cuba

[2, 3] Universidad Autónoma De Guerrero, México

ABSTRACT

Commonly in surveys respondents carrying a stigma refuse responding or will give false reports. Warner (1965) introduced the technique of Randomized Response (RR). Many RR models wen are based on the use of a scrambled response technique. In this paper we develop an RSS extension for the Scrambling method proposed by Hussain (2012).

KEYWORDS: Randomized Response, Ranked Set Sampling

1. INTRODUCTION

Take a population U of identifiable units. The researcher selects a sample form U and obtains response to Y and aims estimating the population mean of the variable of Y. Consider that the response to Y is sensitive. Respondents carrying a stigma will refuse responding or will give false reports. This situation is often encountered in survey research when information on induced abortions, drug abuse, and family income, accepting briberies, etc., is inquired. Warner (1965) introduced the technique of Randomized Response (RR). The theory of estimating the mean of a sensitive quantitative variable Y is commonly developed using the scrambled response technique.

Gupta-Thornton (2002), (GT), proposed a RR procedure that provides more confidence to the respondents. Their model was improved by Hussain (2012).

Ranked Set Sampling is challenging common sample designs as it generally generates leads togains in accuracy with respect to simple random sampling with replacement (SRSWR). It was proposed by McIntyre in 1952 and has been extended to more sophisticated problems, see Chen et al. (2004). Some recent results are Al-Saleh and Al Omari (2002), Al-Nasser (2007), Bouza (2010). In this paper we develop an RSS extension of the model of Hussain (2012).

2. RR PROCEDURE

The need of obtaining true responses through the use of RR is a recurrent theme in applications. We consider a population of persons $U=(1,\cdots,i,\cdots,N)$. Take Y_i as the value of a variable of Y and Y with possible stigmatizing values. An estimate is required of$\mu_Y=\frac{\sum_{i=1}^{N}Y_i}{N}$.

Gupta and Thornton (2002) considered the sampling design to be random sampling with replacement (SRSWR), and proposed a RR procedure based on a two-step randomization mechanism. In addition to the sensitive variable $Y$the statistician determines a probabilitydensityfunction$f(x)$and a non-sensitive variable X is generated. Hence are known$E(X)=\mu_X\in\Re$and $V(X)=\sigma_X^2\in\Re^+$.

The sampler also fixes a randomizer that generates independent Bernoulli distributed $\beta$with $E(\beta) = T$. In the first stage, the interviewee generates a value of X and on the second stage generates a value of β.

$$\beta = \begin{cases} 1 \text{ implies that he or she reports the true value of } Y \\ 0 \text{ implies that he or she reports } Z = X + Y \end{cases}$$

Therefore, the report is the random variable

$$B_i = \beta Y_i + (1 - \beta)Z_i, i = 1, \cdots, n.$$

The expectation and variance of the report are

$$E_M(\beta_i) = TY_i + (1 - T)(X_i + Y_i), i = 1, \cdots, n$$

$$V(\beta_i) = E(B_i^2) - (T\mu_Y + (1 - T)\mu_X)^2$$

As is used a simple random sampling with replacement (SRSWR)

$$E\big(E_M(\beta_i)\big) = T\mu_Y + (1 - T)(\mu_X + \mu_Y), i = 1, \cdots, n,$$

Then an unbiased estimator is

$$\hat{\mu}_Y = \bar{B} - (1 - T)\mu_Y = \frac{1}{n}\sum_{i=1}^{n} B_i - (1 - T)\mu_Y,$$

With variance

$$V(\hat{\mu}_Y) = \frac{\sigma_Y^2}{n} + \frac{(1-T)\left(\sigma_X^2 + T\mu_X^2\right)}{n},$$

See Gupta-Thorton (2002).

Hussain (2012) proposed a new model. It is based on the selection of two responses from each respondent. Each response was used for computing an estimation. They are correlated but have equal variances. The procedure is as follows:

Revised Gupta & Thornton RR (H-GT)

Fix a Randomization mechanism (RM) that generates independent Bernoulli variables β with probability T

Fix a mechanism that generates a random variable X with density $f(x)$

The respondent "i" is requested to use $f(x)$ and he/she generates two values of X , $X_{ij}, j = 1,2$

The respondent uses RM for selecting between:

(i) Reporting the true response on the sensitive variable $Y$with probability T.

(ii) Reporting$Z_{ij} = X_i + Y_{ij}$ with probability $1 - T, j = 1,2$.

Now each respondent´s reports are modeled by

$$R_{ij} = \beta_j Y_j + \left(1 - \beta_j\right)Z_{ij}, i = 1, \cdots, n; j = 1, 2.$$

We will now consider the case $E\left(\beta_j\right) = T, j = 1,2$.

Two estimators are computed$\hat{\mu}_{Y+} = \frac{1}{n}\sum_{i=1}^{n} R_{i1} + (1-T)\mu_X$and $\hat{\mu}_{Y-} = \frac{1}{n}\sum_{i=1}^{n} R_{i2} - (1-T)\mu_X$.

Both estimators are unbiased and Hussein (2012) proposed to use the combined estimator

$$\hat{\mu}_{YW} = W\hat{\mu}_{Y+} + (1-W)\hat{\mu}_{Y-}, W \in]0,1[$$

with variance

$$V(\hat{\mu}_{YW}) = W^2 V(\hat{\mu}_{Y+}) + (1-W)^2 V(\hat{\mu}_{Y-}) + 2W(1-W)Cov(\hat{\mu}_{Y+}, \hat{\mu}_{Y-})$$

It is readily obtained that both variances are equal to

$$V(\hat{\mu}_X) = \frac{\sigma_Y^2}{n} + \frac{(1-T)(\sigma_X^2 + T\mu_X^2)}{n}$$

On the other hand

$$Cov(\hat{\mu}_{Y+}, \hat{\mu}_{Y-}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i^*=1}^{n} Cov(R_{i1}, R_{i^*2})$$

$$= \frac{1}{n^2}\left(\sum_{i \neq i^*}^{n} Cov(R_{i1}, R_{i^*2}) + \sum_{i^*=1}^{n} Cov\left(R_{i1}, R_{i\ \ 2}\right)\right) = \frac{1}{n^2}\sum_{i^*=1}^{n} Cov\left(R_{i1}, R_{i\ \ 2}\right)$$

As

$$Cov(R_{i1}, R_{i2}) = E(R_{i1}, R_{i2}) - E(R_{i1})E(R_{i2}) = \sigma_Y^2 + (1-T)(\sigma_X^2 + T\mu_X^2)$$

$$Cov(\hat{\mu}_{Y+}, \hat{\mu}_{Y-}) = \frac{\sigma_Y^2}{n} + \frac{(1-T)(\sigma_X^2 + T\mu_X^2)}{n}$$

Then using H-GT for obtaining responses from a sample selected with SRSWR an unbiased estimator of μ_Y is

$$\hat{\mu}_{YW} = W\hat{\mu}_{Y+} + (1-W)\hat{\mu}_{Y-}, W \in]0,1[.$$

Its variance is given by

$$V(\hat{\mu}_{YW}) = \frac{\sigma_Y^2}{n}$$

Hence this RR procedure does not increase the sampling error by scrambling the variable in the randomization response procedure.

3. RSS ALTERNATIVES

We will consider the use of RSS. It consists in the selection of m independent samples of size m using SRSWR. Dell-Clutter (1972) established that in such case we obtain not the i-th OS but the '*ith* – judgmental' one. See Stokes (1977) and Patil et al. (1995) for studies on the use of concomitant variables for ranking the sampled units. The application of RR is an important theme, see an example in Bouza (2016).

Denote by $Y_{(t:1)}, \cdots, Y_{(t:t)}, \cdots, Y_{(t:m)}$ the corresponding order statistics (OS) of the ranked sample $s_{t.}$ The OS´s $Y_{(t:1)}, \cdots, Y_{(t:t)}, \cdots, Y_{(t:m)}$are measured by the sampler. The process is repeated $k = 1, \cdots, r$ times (cycles). Take $Y_{(t:t)k}$ as the OS measured in the cycle k in sample t. The sample mean of the RSS design is

$$\bar{y}_{rss} = \frac{1}{m\,r}\sum_{t=1}^{m}\sum_{k=1}^{r} Y(t:t)k$$

As

$E(Y(t{:}\,t)k) = \mu_{Y(t)}, t = 1, \cdots, m; k = 1, \cdots, r$

And

$\mu_Y = \frac{1}{m} \sum_{k=1}^{r} \mu_{Y(t)}$

We have the unbiasedness of$\bar{y}_{rss}$. As the samples are independent

$V(\bar{y}_{rss}) = \frac{1}{m^2 r} \sum_{t=1}^{m} \sigma_{y(t)}^2$

From the relation

$\sigma_{y(t)}^2 = \sigma_Y^2 - \Delta_{Y(t)}^2, \Delta_{Y(t)} = \mu_{Y(t)} - \mu_Y$

We derive

$V(\bar{y}_{rss}) = \frac{\sigma_Y^2}{mr} - \frac{1}{m^2 r} \sum_{t=1}^{m} \Delta_{Y(t)}^2$

Ranking the selected individual should be made cheaply. That is the case in many applications. Take the use of medical records, subjective predictions of Y, etc. Take the ranking variable as a concomitant variable A. In practice we prefer that A be correlated with Y.

Now the report is the OS

$R_{(t:t)k} = \beta Y_{(t:t)k} + (1 - \beta) Z_{(t:t)k}, t = 1, \cdots, m; k = 1, \cdots, r$

Where

$Z_{(t:t)k} = Y_{(t:t)k} + X_{tk}$

$X_{tk}$is the result of generating a value of X using $f(x)$ by the respondent ranked t in the RSS-sample t in the cycle k.

From this reasoning we derive the following result:

Proposition: If we use GT for obtaining responses from an RSS selected from a finite population using an auxiliary variable A correlated with Y. Considering the $n = mr$ reports:

$R_{(t:t)k} = \beta Y_{(t:t)k} + (1 - \beta) Z_{(t:t)k}, t = 1, \cdots, m; k = 1, \cdots, r$

1) An unbiased estimator of μ_Y is

$\hat{\mu}_{Y(rss)} = \bar{R}_{rss} - (1 - T)\mu_X = \frac{1}{m\, r} \sum_{t=1}^{m} \sum_{k=1}^{r} R_{(t:t)k} - (1 - T)\mu_X.$

2) Its variance is $V\left(\hat{\mu}_{Y(rss)}\right) = \frac{\sigma_Y^2}{n} + \frac{(1-T)\left(\sigma_X^2 + T\mu_X^2\right)}{n} - \frac{1}{m^2 r} \sum_{t=1}^{m} \Delta_{Y(t)}^2$

Proof:

As

$E\left(R_{(t:t)k}\right) = T\mu_Y + (1 - T)(\mu_X + \mu_Y) = T\mu_Y + (1 - T)\mu_X, t = 1, \cdots, m; k = 1, \cdots, r$

$E\left(\hat{\mu}_{Y(rss)}\right) = \mu_Y$and the unbiasedness of the estimator is derived.

On the other hand

$$V\left(R_{(t:t)k}\right) = E\left(R^2{}_{(t:t)k}\right) - (T\mu_Y + (1-T)\mu_X)^2$$

The first term is

$$E\left(R^2{}_{(t:t)k}\right) = E\left(\beta^2 Y_{(t:t)k} + (1-\beta)^2 Z^2_{(t:t)k} + 2\beta(1-\beta)Y_{(t:t)k}Z_{(t:t)k}\right)$$

$$= T\left(\mu_Y^2 + \sigma^2_{Y(t)}\right) + (1-T)\left(\mu_X^2 + \sigma^2_{X(t)} + \mu_Y^2 + \sigma_Y^2\right), t = 1, \cdots, m; k = 1, \cdots, r$$

Therefore

$$V\left(R_{(t:t)k}\right) = \left(\mu_Y^2 + \sigma^2_{Y(t)}\right) + (1-T)\left(\mu_X^2 + \sigma^2_{X(t)} + \mu_Y^2 + \sigma_Y^2 - 2\mu_X\mu_Y\right) - (T\mu_Y + (1-T)\mu_X)^2 = \sigma^2_{Y(t)} + (1-T)(\sigma_X^2 + T\mu_X)$$

and

$$V\left(\hat{\mu}_{Y(rss)}\right) = \frac{\sigma_Y^2}{n} + \frac{(1-T)\left(\sigma_X^2 + T\mu_X^2\right)}{n} - \frac{1}{m^2 r}\sum_{t=1}^{m} \Delta^2_{Y(t)}$$

The last term in $V\left(\hat{\mu}_{Y(rss)}\right)$ is the gain in accuracy due to the use of RSS.

In some applications the members of the k-th selected sample may be convinced to share the values of the generated values of X. Then the ranking is made on $X = A$ and we deal with OS´s of X. If X is not correlated with Y we have that the unbiasedness of $\hat{\mu}_{Y(rss)}$ holds and

$$V\left(\hat{\mu}_{Y(rss)}|X\right) = \frac{\sigma_Y^2}{mr} + \frac{(1-T)\left(\sigma_X^2 - \frac{1}{m^2 r}\sum_{t=1}^{m} \Delta^2_{X(t)} + T\mu_X^2\right)}{n}, \Delta_{X(T)} = \mu_{X(t)} - \mu_X$$

This result can be considered as a Corollary to the previous proposition

Corollary 4: Under the conditions of Proposition 3, if $A = X$ is used for ranking, we have to:

1. If X is Uncorrelated with Y

$$R_{(t:t|X)k} = \beta Y_{(t:t)k} + (1-\beta)\left(X_{tk} + X_{(t:t)k}\right), t = 1, \cdots, m; k = 1, \cdots, r$$

1.1). An Unbiased Estimator Of μ_YIs

$$\hat{\mu}_{Y(rss|X)} = \frac{1}{m\, r}\sum_{t=1}^{m}\sum_{k=1}^{r} R_{(t:t|Y)k} - (1-T)\mu_X.$$

1.2). Its Variance Is

$$V\left(\hat{\mu}_{Y(rss|X)}\right) = \frac{\sigma_Y^2}{n} + \frac{(1-T)\left(\sigma_X^2 + T\mu_X^2 - \frac{1}{m}\sum_{t=1}^{m} \Delta^2_{Y(t)}\right)}{n}$$

2. If X is Correlated with Y

$$R_{(t:t|X)k} = \beta^2 Y_{(t:t)k} + (1-\beta)\left(X_{(t:t)k} + Y_{(t:t)k}\right), t = 1, \cdots, m; k = 1, \cdots, r$$

2.1). An Unbiased Estimator Of μ_Y Is

$$\hat{\mu}_{Y(rss|X)} = \frac{1}{m\,r}\sum_{t=1}^{m}\sum_{k=1}^{r} R_{(t:t|X)k} - (1-T)\mu_X.$$

2.2). Its Variance Is

$$V\left(\hat{\mu}_{Y(rss|X)}\right) = \frac{\sigma_Y^2}{n} + \frac{(1-T)\left(\sigma_X^2 + T\mu_X^2\right)}{n} - \left(\frac{1}{nr}\sum_{t=1}^{m}\Delta_{Y(t)}^2 + (1-T)\left(\frac{1}{r}\sum_{t=1}^{m}\Delta_{Y(t)}^2\right)\right)$$

These results suggest that using a probability function with large values of $\left|\Delta_{Y(t)}\right|$ increases the gain in accuracy.

In the RSS extension of H-GT we have the reports

$$R_{(t:t)k1} = \beta Y_{(t:t)k} + (1-\beta)Z_{(t:t)k,} R_{(t:t)k2} = \beta Y_{(t:t)k} - (1-\beta)Z_{(t:t)k}$$

The corresponding unbiased RSS-estimators are

$$\hat{\mu}_{Y(rss+)} = \frac{1}{mr}\sum_{t=1}^{m}\sum_{k=1}^{r} R_{(t:t)k1} - (1-T)\mu_X,\ \hat{\mu}_{Y(rss-)} = \frac{1}{mr}\sum_{t=1}^{m}\sum_{k=1}^{r} R_{(t:t)k2} + (1-T),$$

The unbiased estimator of μ_Y is

$$\hat{\mu}_{Y(rssW)} = W\hat{\mu}_{Y(rss+)} + (1-W)\hat{\mu}_{Y(rss-)}$$

We have two possible cases when the ranking variable is considered.

Case 1 $A \neq X$

$$V\left(\hat{\mu}_{Y(rssW)}\right) = \frac{\sigma_Y^2}{n} - \frac{1}{nm}\sum_{t=1}^{m}\Delta_{Y(t)}^2.$$

Case 2 $A = X$ and Y

$$V\left(\hat{\mu}_{Y(rssW|X)}\right) = \frac{\sigma_Y^2}{n}$$

The last result holds because in that case the ranking is random.

CONCLUSIONS

These result indicates that, using RSS is recommended only if we have additional information that allows obtaining a non-random ranking of the sensitive variable.

REFERENCES

1. Al-Nasser, D. A. (2007). L-Ranked set sampling: a Generalization procedure for robust visual sampling. *Communications in Stat.: Simulation and Computation.*, 33-43.
2. Al-Saleh, M. a.-O. (2002). Multistage ranked set sampling. *Journal of Statistical Planning and Inference*, 273-286.
3. Bouza, C. (2010). Ranked set sampling using auxiliary variables of a randomized response procedure for estimating the mean of a sensitive quantitative character. *Journal of Modern Applied Statistical Methods*, 248-254.
4. Chen, Z. B. (2004). *Ranked Set Sampling: Theory and Applications.* New York: Springer Science+Business Media.

5. Greenberg, B. G. (1971). Applications of RR technique in obtaining quantitative data. *Journal of the American Statistical Association*, 243-250.

6. Gupta, S. a. (2002). Circumventing social desirability response bias in personal interview surveys. *American Journal of Mathematical and Management Sciences*, 369-383.

7. Hussain, Z. &. (2007). Estimation of mean of a sensitive quantitative variable. *Journal of Statistical Research*, 83-92.

8. Hussain, Z. (2012). Improvement of the Gupta and Thornton Scrambling Model through Double Use of Randomization Device. *International Journal of Academic Research in Business and Social Sciences*, 91-97.

9. Hussain, Z. a. (2007). Estimation of mean of a sensitive quantitative variable. *Journal of Statistical Research*, 83-92.

10. Hussain, Z. a. (2011). Improved estimation of mean in Randomized response models. *Hacettepe Journal of Mathematics and Statistics*, 91-104.

11. Warner, S. L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 63-69.