RANK SET SAMPLING IN SITUATIONS OF NON-RESPONSE, WHILE CONSIDERING THE PROBLEMS OF ALLOCATION

M. Iqbal Jeelani[1]*, Carlos, N. Bouza**, S.E.H.Rizvi*, S.A.Mir**, Manish.Kr. Sharma*, Faheem Jeelani*
*Agricultural Statistics, SKUAST- India
**Facultad de Matemática y Computación, Universidad de La Habana, Cuba

ABSTRACT
In this article the problem of allocation in Ranked Set Stratified Sampling in situations of non-response has been considered. We considered the fixed cost and variance, based on various combinations of response rates and inverse ratio of sub sample among the non-respondents. A numerical example is given to illustrate the results.

KEYWORDS: Ranked set sampling, stratification, no responses

MSC: 62D05

RESUMEN
En este trabajo consideramos el problema de la afijación en el Muestreo Estratificado Por Conjuntos Ordenados (Ranked Set Sampling). Consideramos en situaciones donde hay no respuestas, bajo el costo fijo y la varianza, basado en varias combinaciones de tasas de respuesta y la razón inversa de la sub muestra entre los no respondieres. Un ejemplo numérico se brinda para ilustrar los resultados.

PALABRAS CLAVE: Muestreo por rangos ordenados , estratificación, no respuestas

1. INTRODUCTION

In most of the surveys, it is usually observed that the data may not be obtained in the first attempt, sometimes due to the absence and sometimes due to the refusal of the respondent. The first step to deal with the problem of Non-Response was made by Hansen and Hurwitz (1946), in which they divided the population into two groups viz respondents and non-respondents. Another method to obtain unbiased estimators from the information collected from the respondent in the first attempt only was proposed by Politz and Simmons (1949). Kish and Lansing (1954) proposed the adding of a sample of non-responding units from previous surveys for obtaining information about the non-respondents.
In this article we have considered the problem of allocations in rank set stratified sampling under combinations of response rates and inverse ratio of sub sample class among non-respondents keeping cost and variance fixed using Hansen and Hurwitz (1946) procedure.

2. HANSEN AND HURWITZ PROCEDURE OF NON-RESPONSE UNDER STRATIFIED SAMPLING

Let us consider a finite population $U = \{u_1, \ldots \ldots u_N\}$composed of individuals that can be identified, let a sample s of size $n \leq N$ is selected. The variable of interest Y is measured in a selected sample. When some of the units in the sample do not give a response then the existence of non-response does not permit us to compute sample mean, which means that the population U is divided into two strata: U_1 units which give a response at first visit and U_2 which contains the non-responding individuals (Hansen &Hurwitz, 1946).

1 Email: jeelani.miqbal@gmail.com

Now let us consider a population consisting of N units divided into L strata. Let the size of i^{th} stratum be $N_{i,}(1,2,\ldots.L)$.We select a sample of size n from the population in such a way that n_i units are selected from the i^{th} stratum. Thus, we have

$$\sum_{i=1}^{L} n_i = n \tag{2.1}$$

Let the non-responses occur in each stratum, then we select a sample of size m_i units out of $n_{i2,}$ then on-responding units in the i^{th} stratum such that $n_{i2} = K_i\, m_i,\ K_i \geq 1$using Hansen and Hurwitz procedure and the information are observed on all the m_i units by interview method. The Hansen-Hurwitz estimator of population mean for the i^{th}stratum under stratified simple random sampling will be;

$$\bar{y}_i^* = \frac{n_{i1}\bar{y}_{0i1}+n_{i2}\bar{y}_{0mi}}{n_i}, (1,2,\ldots.L) \tag{2.2}$$

Where$\bar{y}_{0i1}$ and$\bar{y}_{0mi}$are the sample means based on n_{i1} responding units and m_i non-responding units in the i^{th} stratum. Combining the estimators over all strata we get the estimator of population meanY_{ssrsi}, given by

$$Y_{ssrsi} = \sum_{i=1}^{k} p_i\bar{y}_i^* \qquad ; \text{where } p_i = \frac{N_i}{N} \tag{2.3}$$

Obviously, we have $E[\bar{Y}_{ssrsi}] = Y_{ssrs}$, then the variance of $\bar{Y}_{ssrsi}$ is given by;

$$V[\bar{Y}_{ssrsi}] = \sum_{i=1}^{L}\left(\frac{1}{n_i}-\frac{1}{N_i}\right)p_i^2S_{0i}^2 + \sum_{i=1}^{L}\frac{(K_i-1)}{n_i}W_{i2}p_i^2S_{0i2}^2 \tag{2.4}$$

where$W_{i2} = \frac{N_{i2}}{N_i}, S_{0i}^2$, $K_i = {}^{n_{i2}}/_{m_i}$ (W_{i2} is non-response rate of the i^{th} stratum, K_i is inverse ratio of sub sample class among the non-respondents and S_{0i}^2 is mean squares of entire group and S_{0i2}^2 is the mean square of the non-response group respectively in the i^{th} stratum.

3. RANK SET STRATIFIED SAMPLING IN NON-RESPONSE SITUATIONS AND ALLOCATION PROCEDURE

McIntyre (1952) first proposed rank set sampling and claimed that rank set sampling produces more accurate estimators of the sample mean than the usual simple random sampling (SRS) design. The mathematical support to his claim was given by Takahasi and Wakimoto (1968), Dell and Clutter (1972), Chen (2002), Bouza (2013), Jeelani *et al.* (2014a), (2014b), Jeelani *et al.* (2015), etc..

In this section we consider the use rank set sampling scheme for selecting the sub-sample. Rank set sampling procedure is used for sub-sampling s_2. Take a sub sample $s'_{2(rss)}$ from s using Rank set sampling procedure. That is we select n'_2 independent samples of size ${}^{n_2}/_{K}$ using simple random sampling. The units are ranked accordingly with the variable closely related with the variable of interest Y.

Let there be n'_2 independent samples$Y_{11}, Y_{12} \ldots, Y_{1n'_2};\ Y_{21}, Y_{22} \ldots, Y_{2n'_2}; \ldots; Y_{n'_21}, Y_{n'_22} \ldots, Y_{n'_2n'_2}$ they are ranked and we obtain$Y_{1:1}, Y_{2:1} \ldots, Y_{n'_2:1};\ Y_{1:2}, Y_{2:2} \ldots, Y_{n'_2:2}; \ldots; Y_{1:n'_2}, Y_{2:n'_2} \ldots, Y_{n'_2:n'_2}$The estimate of μ_2 is made by using the estimator:

$$\bar{y}'_{2(rss)} = \frac{\sum_{j=1}^{n'_2} y_{(j:j)}}{n'_2} \tag{3.1}$$

$$E\left(\bar{y}'_{2(rss)}\right) = E\left(\frac{\sum_{j=1}^{n'_2} E\left(y_{(j:j)}\right)}{n'_2}\right) = E[\bar{y}_2] = \mu_2 \tag{3.2}$$

The rank set sampling counter part of $\bar{\bar{y}}$ in simple random sampling is

$$\bar{\bar{y}} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}'_2 = w_1\bar{y}_1 + w_2\bar{y}'_{2(rss)} \tag{3.3}$$

It can be represented by

$$\bar{\bar{y}}_{(rss)} = (w_1\bar{y}_1 + w_2\bar{y}_2) + w_2(\bar{y}'_{2(rss)} - \bar{y}_2) \tag{3.4}$$

Its conditional variance is

$$V\left(\bar{\bar{y}}_{(rss)}|s\right) = \frac{\sigma^2}{n} + w_2^2V\ (\bar{y}'_{2(rss)} - \bar{y}_2|s) \tag{3.5}$$

Explicit expression of second term in the R.H.S. it is:

$$V(w_2\left(\bar{y}'_{2(rss)} - \bar{y}_2\right)|s) = w_2^2E\left(\bar{y}'_{2(rss)} - \mu_2\right) - (\bar{y}_2 - \mu_2)|s)^2$$

$$= w_2^2\left[E\left(\bar{y}'_{2(rss)} - \mu_2\right)|s)^2 + E(\bar{y}_2 - \mu_2)|s)^2 - 2E\left(\left(\bar{y}'_{2(rss)} - \mu_2\right)\left(\bar{y}_2 - \mu_{2)}\right)\Big|s\right)\right] \tag{3.6}$$

Using the Ist, 2nd and 3rd terms in equation (3.6) the counter part of variance of SRS under non response in case of RSS is given below as proved by Bouza (2002).

$$V\left(w_2\left(\bar{y}'_{2(rss)} - \bar{y}_2\right)|s\right) = w_2^2\left(\frac{\sigma^2_{2(rss)}}{n'_2} - \frac{\sigma^2_2}{n_2}\right) = w_2^2\left(\frac{\sigma^2}{n'_2} - \frac{\sigma^2_2}{n_2} - \frac{\sum_{j=1}^{n'_2}\Delta^2_{(j)}}{n'_2}\right) \quad (3.7)$$

Substituting $n'_2 = {n_2}/{K}$, then we have;

$$EV\left(\bar{\bar{y}}_{(rss)}\right) = \frac{\sigma^2}{n} + \frac{W_2(K-1)\sigma^2_2}{n} - W_2E\left(\frac{K\sum_{j=1}^{n'_2}\Delta^2_{(j)}}{n}\right) \quad (3.8)$$

Using rank set stratified sampling procedure proposed by Jeelani *et al.* (2014b) the variance of the estimator of population $\bar{\bar{y}}_{(rssh)}$ based on Hansen and Hurwitz (1946) is given by:

$$V\left(\bar{\bar{y}}_{(rssh)}\right) = \sum_{h=1}^{L}\left(\frac{1}{n_h} - \frac{1}{N_h}\right)p_h^2S_{0h}^2 + \sum_{h=1}^{L}\frac{(K_h-1)}{n_h}W_{h2} - S^2_{0h2}\left(\frac{Kh\sum_{h=1}^{n'_2}\Delta^2_{(h)}}{n}\right) \quad (3.9)$$

where, $h = 1,2,3,4, \ldots, L$, $W_{h2} = \frac{N_{h2}}{N_h}$, S^2_{0h} and S^2_{0h2} are the non-response rates, mean squares of entire group and non-response group respectively in the h^{th} stratum and K_h is inverse ratio of sub-sample among non-response.

Let n_{h1} and n_{h2} are the number of units in response and non response group in the h^{th} stratum, then

$$E\left({n_{h1}}/{N_{h1}}\right) = E\left({n_{h2}}/{N_{h2}}\right) = K_h\, E\left({m_h}/{N_{h2}}\right) \quad (3.10)$$

Since $K_h = {n_{h2}}/{m_h}$ also $(K_h > 1)$

Assuming the information on all the units of sub sample from non-response group of the h^{th} stratum is already available, then the value of n_h and K_h should be chosen in such a way so that maximum precision will be attained for fixed cost. Since the simplest cost function that can be taken is

$$C = a + \sum_{h=1}^{L} n_h c_h \quad (3.11)$$

Where the overhead cost a is the constant and c_h is the average cost of surveying one unit in the h^{th} stratum and n_h is the number of units in the sample., which depends upon the nature and size of units in the stratum.

Utilizing the equation (3.11) the cost function in case of rank set stratified sampling will take the form given below:

$$C = \sum_{h=1}^{L} C_{ho}n_h + \sum_{h=1}^{L} C_{h1}n_{h1} + \sum_{h=1}^{L} C_{h2}m_h \quad (3.12)$$

where C_{ho} is cost of making the Ist attempt.

C_{h1} is the cost of getting information per unit from group.

C_{h2} is the cost of getting information from non-response group.

The expected value of the n_{h1} is $W_{h1}n_h$ and for m_h is ${W_{h2}n_h}/{K_h}$, then the total expected cost in case of Rank set stratified sampling will be ;

$$C = \sum_{h=1}^{L} C_{ho}n_h + \sum_{h=1}^{L} W_{h1}C_{h1}n_{h1} + \sum_{h=1}^{L} {C_{h2}W_{h2}n_h}/{K_h} \quad (3.13)$$

Now let us consider a cost function using the Lagrange's multiplier which will determine the optimum values of number of units in the sample and the inverse ratio of the sub sample class among non-respondents given in the equation (3.13), then we have;

$$\psi = V\left(\bar{\bar{y}}_{(rssh)}\right) + \eta(C) \quad (3.14)$$

where, η is the unknown constant and

$$V\left(\bar{\bar{y}}_{(rssh)}\right) = \sum_{h=1}^{L}\left(\frac{1}{n_h} - \frac{1}{N_h}\right)p_h^2S_{0h}^2 + \sum_{h=1}^{3}\frac{(K_h-1)}{n_h}W_{h2} - S^2_{0h2}\left(\frac{K_h\sum_{h=1}^{n'_2}\Delta^2_{(h)}}{n}\right)$$

Now differentiating the ψ with respect to inverse ratio of sub sample class among non-respondents, we have;

$$n_h = {k_hp_hs_h}/{\sqrt{\eta C_{h2}}} \quad (3.15)$$

where, $p_h = {N_h}/{N}$

Now differentiating ψ with respect to the total number of units in the sample we have;

$$\frac{\partial\psi}{\partial n_h} = \frac{(1+(k_h-1)W_{h2})p_h^2S_h^2}{n_h^2} + \eta\left(C_{h0} + C_{h1}W_{h1} + \frac{C_{h2}W_{h2}}{K_h}\right) = 0 \quad (3.16)$$

Utilizing the equation (3.17) and eliminating η in the above equation we have ;

$$k_h = \sqrt{C_{h2}W_{h1} / C_{h0} + C_{h1}W_{h1}} \tag{3.17}$$

It can be seen from equation (3.17) that the value of inverse ratio of sub sample class among non-respondents increases as the cost of getting information among non-respondents increases.

Now if we fix the total cost then we may have;

$$C_0 = \sum_{h=1}^{L}\left(C_{h0} + C_{h1}W_{h1} + \frac{C_{h2}W_{h2}}{K_h}\right)n_h \tag{3.18}$$

Since $k_h = \sqrt{C_{h2}W_{h1} / C_{h0} + C_{h1}W_{h1}}$ and $n_h = k_h p_h s_h / \sqrt{\eta C_{h2}}$ then we have;

$$1 / \sqrt{\eta} = C_0 / \sum_{h=1}^{L}\left(\sqrt{(C_{h0} + C_{h1}W_{h1}) + W_{h1}} + \sqrt{C_{h2}W_{h2}}\right)p_h s_h$$

(3.19)

Then the total number of units in the sample for fixed cost will be given as follows;

$$n_h = \sqrt{(W_{h1}/(C_{h1}h)}\, p_h s_h C_0 / \sum_{h=1}^{L}\sqrt{(C_{h0} + C_{h1}W_{h1})W_{h1}} + \sqrt{C_{h2}W_{h2}})\, p_h s_h \tag{3.20}$$

Now if we fix the precision then we may have;

$$1 / \sqrt{\eta} = \sum_{h=1}^{L}\left(\sqrt{(C_{h0} + C_{h1}W_{h1}) + W_{h1}} + \sqrt{C_{h2}W_{h2}}\right)p_h s_h / V\left(\bar{\bar{y}}_{(rssh)0}\right) + \frac{1}{N}\sum_{h=1}^{L} p_h S_h^2 \tag{3.21}$$

$$V\left(\bar{\bar{y}}_{(rssh)0}\right) = \sum_{h=1}^{L}\left(\frac{1+(K_h-1)}{n_h}W_{h2}\right)p_h^2 S_{0h}^2 \frac{1}{N}\sum_{h=1}^{L} p_h S_h^2 - \left(\frac{Kh\sum_{h=1}^{n_2'}\Delta_{(h)}^2}{n}\right) \tag{3.22}$$

(we obtain the equation (3.21) by substituting inverse ratio from equation (3.17).

Then the total number of units selected with minimum cost for a fixed precision is given by;

$$n_h = \left(\sqrt{W_{h1}/(C_{h0} + C_{h1}W_{h1}}\, p_h S_h\right)\sum_{h=1}^{L}\left(\sqrt{(C_{h0} + C_{h1}W_{h1})} + W_{h2}\sqrt{C_{h2}}\right)p_h s_h / \left(\bar{\bar{y}}_{(rssh)0}\right) + \frac{1}{N}\sum_{h=1}^{L} p_h S_h^2$$

(3.23)

4. NUMERICAL ILLUSTRATION

For numerical illustration apple data is considered. The data refers to yield of 672 orchards in metric tons taken from district Baramulla in Kashmir valley, India. For the purpose of illustration, we have randomly divided the 420 orchards into five strata. The summary details of cost of making the I[st] attempt, cost of getting information per unit from response group and finally cost of getting information from non-response group are given Table: 1 below and for calculation of variance a specified precision of 2.438 and specified cost of 3600 has been taken in this example. It can be seen that for a fixed precision the sample size and the cost expected decreases with increase in the response rate, also for fixed cost sample size increases with the increase in response rates and variances decreases with the increases in response rates, which are given in Table:2.

Table. 1: Details of cost

	C_{h2}	C_{h1}	C_{h0}	S_h	p_h
Stratum 1	50.00	30.00	5.00	6.42	0.53
Stratum 2	60.00	36.00	10.00	8.32	0.70
Stratum 3	55.00	27.00	8.00	7.10	0.67
Stratum 4	66.00	42.00	15.00	10.13	0.90
Stratum 5	72.00	34.00	12.00	9.33	0.89

Table. 2: Response rates, Inverse ratio's, expected cost among Strata for fixed cost and specified precision

		$W_{h1} = 0.40$				
		Stratum 1	Stratum 2	Stratum 3	Stratum 4	Stratum 5
	K_h	1.15	1.17	1.23	1.2	1.25

Variance 3.471	Fixed Cost $C_0 = 3600$	n_h	10	23	36	15	18
Expected Cost 4687.44	Fixed Precision $V(\bar{\bar{y}}_{(rssh)0}) = 2.438$	n_h	13	30	55	26	29
			$W_{h1} = 0.60$				
		K_h	1.24	1.28	1.4	1.36	1.47
Variance 3.128	Fixed Cost $C_0 = 3600$	n_h	14	27	61	18	23
Expected Cost 4258.90	Fixed Precision $V(\bar{\bar{y}}_{(rssh)0}) = 2.438$	n_h	16	33	65	28	31
			$W_{h1} = 0.14$				
		K_h	1.33	1.38	1.47	1.41	1.58
Variance 2.861	Fixed Cost $C_0 = 3600$	n_h	21	27	31	20	34
Expected Cost 3861.77	Fixed Precision $V(\bar{\bar{y}}_{(rssh)0}) = 2.438$	n_h	25	36	43	28	37
			$W_{h1} = 0.80$				
		K_h	1.49	1.54	1.63	1.57	1.74
Variance 2.520	Fixed Cost $C_0 = 3600$	n_h	15	21	25	14	28
Expected Cost 3749.11	Fixed Precision $V(\bar{\bar{y}}_{(rssh)0}) = 2.438$	n_h	19	30	37	22	31
			$W_{h1} = 0.90$				
		K_h	1.61	1.66	1.75	1.69	1.86
Variance 2.152	Fixed Cost $C_0 = 3600$	n_h	23	29	33	22	36
Expected Cost 3693.44	Fixed Precision $V(\bar{\bar{y}}_{(rssh)0}) = 2.438$	n_h	27	38	45	30	39

5. CONCLUSIONS

It is concluded that under different combinations of responses rates and inverse ratio of sub sample under non-respondents ,new allocation schemes for fixed cost and precision clearly shows that inverse ratio of sub sample among non-respondents is independent of total cost and fixed variance, also it is can be said that this inverse ratio is the key function of cost of making the Ist attempt, cost of getting information per unit from Response group and finally cost of getting information from non-response group.

RECEIVED: DECEMBER, 2016
REVISED: MARCH, 2017

Acknowledgements: The authors thanks two anonymous referees for their helpful comments. The results of this paper are connected with the CYTED project BigData and Decision Support Systems in Agriculture.

REFERENCES

[1]. BOUZA, C.N. (2002): Ranked set sub-sampling the non response strata for estimating the difference of means. **Biometrical. J.,** 44, 903-915.
[2]. BOUZA, C.N. (2013): .**Handling Missing Data in Ranked Set Sampling**. Springer Briefs in Statistics, Springer, Berlin.
[3]. CHEN, Z. (2002): .Ranked-set sampling with regression-type estimators. **J. Statist. Plann. and Infer**., 92, 181-192.
[4]. DELL, T.R. and CLUTTER, J.L. (1972): Ranked set sampling theory with order statistics. Background. Biometrics, 28, 545-555.
[5]. HANSEN, M.H. and HURWITZ, W.N. (1946): The problem of non-response in sample surveys. **Jour. Amer. Stat. Assoc.**, 41, 517-529.
[6]. JEELANI, M.I., MIR, S.A., KHAN,I.,NAZIR,N and JEELANI,F.(2015): Rank Set Sampling in Improving the Estimates of Simple Regression Model. **Pakistan Journal of Statistics and Operation Research**. 11 , 39-45
[7]. JEELANI..I, MIR.S.A.,KHAN.I.,NAZIR.N., and JEELANI.F.(2014a): Non-Response Problems in Rank set sampling. **Pakistan Journal of Statistics** 30, 555-562
[8]. JEELANI, M.I., MIR,S.A., MAQBOOL,S., KHAN, I., SINGH,K.N., ZAFFER, G., NAZIR, N. and JEELANI, F.(2014b): Role of Rank Set Sampling in Improving the Estimates of Population Mean under Stratification. **Amer. J. Math. and Statist**., 4, 46-49.
[9]. KISH, L. and LANSING, J.B. (1954): .Response errors in estimating the value of homes. **J.Amer. Stat. Assoc**., 49, 520-538.
[10]. MCINTYRE, G.A. (1952): A Method for unbiased selective sampling, using ranked sets. **Aust. J. Agric. Res**., 3, 385-390.
[11]. MUTTLAK, H.A. (1996): Median Ranked set sampling. **J. of Applied Statistical Science**, 6, 91-98.
[12]. POLITZ, A.N. and SIMMONS, W.R. (1949): .An attempt to get the ‘not at homes into the sample without call backs. **J. Amer. Stat. Assoc**. 44, 9-31.