Jour. Ind. Soc. Ag. Statistics
47(2), 1995 : 134–141

Sampling Without Replacement in Qualitative Randomized Response Model

Rajendra Singh and O.P. Kathuria*
Indian Veterinary Research Institute, Izatnagar-243122

(Received : April, 1989)

Summary

It is shown that the probability of 'yes' response is same for simple random sampling with replacement (SRSWR) as well as without replacement (SRSWOR). Some estimators have been derived for SRSWOR and their variances obtained using randomized response models for binary and discrete data. Estimators developed under SRSWOR are more efficient than estimators under SRSWR irrespective of randomized response model used, provided N is finite. The mean square error of estimators of randomized response model under SRSWOR and SRSWR are compared with the MSE of conventional estimator under various assumptions about the underlying population. This study established the supremacy of unrelated question randomized response model under SRSWOR over open interview with nominal untruthful reporting of order 5% under the same scheme.

Key words : Sensitive attribute; Randomised response; Unbiased estimate.

Introduction

Warner [4] developed a model for estimating the proportion of individuals possessing a sensitive attribute without requiring the individual respondent to report to the interviewer whether or not he possesses the sensitive attribute. The technique consists in presenting a random device to the respondent, (say) a spinner with a face mark such that the spinner points to the letter A with probability P and to letter $\bar{A}$ with probability $(1 - P)$, $P \neq \frac{1}{2}$ where A represents the sensitive attribute and $\bar{A}$ the non-sensitive attribute (complement of A). A simple random sample of n individuals is drawn with replacement. The respondent is asked to spin the spinner unobserved by the interviewer and report only 'yes' or 'No', whether or not the spinner points to the letter representing the group to which the interviewer belongs. The response to either question will divide the sample space into two mutually exclusive and complementary classes.

* Indian Agricultural Statistics Research Institute, New Delhi-110012

Remark : The classes of estimators proposed by Biradar and Singh [1], Singh and Kataria [3], Singh [5] and Singh and Upadhyaya [4] are not shown to be the special cases of the propsed general class of estimators as Srivastava [7] has shown that these classes of estimators are not improvements over the original one.

ACKNOWLEDGEMENTS

The authors are grateful to the referee for valuable, fruitful and constructive comments which resulted in substantial improvement of the original manuscript. They are also thankful to Prof. Ravindra Singh for help during the revision of the manuscript.

REFERENCES

[1] Biradar, R.S. and Singh, H.P., 1992. A class of estimators for finite population correlation coefficient using auxiliary information. *Jour. Ind. Soc. Ag. Statistics*, 44(3), 271-285.

[2] Gujarati, D., 1978. Basic Econometrics (International Students Edition), Mcgraw-Hill International Book Company, Tokyo

[3] Singh, H.P., 1988. An improved class of estimators of population mean using auxiliary information. *Jour. Ind. Soc. Ag. Statistics*, 40 (2), 96-104.

[4] Singh, H.P. and Upadhyaya, L.N., 1986. On a class of estimators of the population mean in sampling using auxiliary information. *Jour. Ind. Soc. Ag. Statistics*, 38, (1), 100-104.

[5] Singh, S. and Kataria, P., 1990. An estimator of finite population variance. *Jour. Ind. Soc. Ag. Statistics*, 42, 186-188.

[6] Srivastava, S.K., 1980) : A class of estimators using auxiliary information in sample surveys. *Canadian J. Statist.*, 8 (2), 253-254.

[7] Srivastava, S.K., 1992. A note on improving classes of estimators in survey sampling. *Jour. Ind. Soc. Ag. Statistics*, 44 (3) , 267-270.

[8] Srivastava, S.K. and Jhajj, H.S., 1980. A class of estimators using auxiliary information for estimating finite population variance. *Sankhyā*, C, 42, 87-96.

[9] Srivastava, S.K. and Jhajj, H.S., 1981. A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika*, 68, 341-343

[10] Srivastava, S.K. and Jhajj, H.S., 1983. Class of estimators of mean and variance using auxiliary information when correlation coefficient is also known. *Biom. J.*, 25, 4, 401-409.

[11] Srivastava, S.K. and Jhajj, H.S., 1986. On the estimation of finite population correlation coefficient. *Jour. Ind. Soc. Ag. Statistics*, 38, (1), 82-91.

Further refinements in the model and in choice of unrelated questions were suggested by Greenberg *et al* [1], Moors [3], Horvitz *et al* [2] and others but selection of respondents in all cases was done by SRSWR.

As is well known for a finite population, SRSWOR is always superior to SRSWR, hence the objective of this paper is to prove that probability of 'Yes' response is same for SRSWR as well as SRSWOR and to develop estimators for the randomized response model of Warner [4] and unrelated question randomized response model of Greenberg *et al* [1] under SRSWOR for binary and discrete data and to compare their efficiencies with respect to SRSWR.

The mean square errors of these estimators have also been worked out to see the effect of false reporting.

2. *Probability of 'Yes' Response in Sampling Without Replacement*

The following theorem is proved :

Theorem : The probability of 'Yes' response is same in SRSWR as well as SRSWOR.

Proof. Let the proportion of idividuals possessing the sensitive character A be π and the proportion of individuals not possessing the sensitive character $(\bar{A})$ be $(1-\pi)$. The randomized instrument used here is a deck of cards. On each card in Warner model is recorded one of the following two statements which occur with relative frequencies P and 1–P respectively.

1. I belong to group A
2. I belong to group $\bar{A}$

Let the population consist of N_1 individuals possessing character A and N_2 individuals possessing character $\bar{A}$ where $N_1 + N_2 = N$.

A Simple random sample of K individuals is drawn from the population without replacement and each one is asked to select a card at random from a well shuffled deck of cards unseen by the interviewer and reply 'Yes' or 'No' depending upon whether or not he belongs to the respective group indicated on the card. Assume that $(k+1)$th individual is drawn.

Then the total number of ways to draw $(k+1)$ individuals out of N individuals considering order is equal to $N(N-1)(N-2)\ldots(N-K)$. We know that K individuals drawn first do not disclose their identity and so we do not know to which group they belong. Therefore, the number of ways in which the $(K+1)$th individual drawn belongs to group A when order is also considered is equal to $N_1(N-1)\ldots(N-K)$. Similarly, number of ways in which the $(K+1)$th individual drawn belongs to group $\bar{A}$ when order is also considered is equal to $N_2 (N-1)(N-2)\ldots(N-K)$. If λ be the probability of 'Yes' response, then under Warner model the probability that the i–th person selected at $(K+1)$th draw says 'Yes' is

$$\lambda = P\,\frac{N_1 (N-1)(N-2)\ldots(N-K)}{N(N-1)(N-2)\ldots(N-K)} + (1-P)\,\frac{N_2 (N-1)(N-2)\ldots(N-K)}{N(N-1)\ldots(N-K)}$$

$$= \frac{N_1}{N}\,P + (1-P)\,\frac{N_2}{N}$$

$$= P\pi + (1-P)(1-\pi) \text{ where } \pi = \frac{N_1}{N} \qquad \text{(I)}$$

In unrelated question randomized response model, consider a situation where the following three statements are stored in the randomized device with known frequencies P, P_1 and P_2 respectively where $P_1 + P_2 = (1-P)$:

1. I belong to group A
2. say "Yes"
3. say "No"

The randomised device in this case may be thought to comprise of a box containing red, white and blue balls; a red ball chosen with probability P requires the sensitive question to be answered while choice of a white or blue balls with probabilities P_1 and P_2 respectively refers to an instruction to reply 'Yes' or 'No'. The non-sensitive question is thus built into the randomisation device itself with corresponding 'Yes' and 'No' responses being given in statements 2 and 3 above, the expected value of Yes response being equal to $P_1/(P_1+P_2)$.

Using the same notations, logic and derivations as stated above, the probability that the i–th person selected at $(k+1)$th draw says "Yes" is

$$\lambda' = P\,\frac{N_1 (N-1)(N-2)\ldots(N-K)}{N(N-1)(N-2)\ldots(N-K)} + P_1$$

$$= P\,\frac{N_1}{N} + P_1$$

$$= P\pi + P_1 \qquad \text{(II)}$$

This proves that probability of 'Yes' response is same in sampling with replacement and sampling without replacement. The number of 'Yes' in the population is $N\lambda'$ and number of 'No' is $N(1-\lambda')$.

3. *Mean and variance of 'Yes' response :*

Let a SRSWOR of size n be drawn and let n_1 be the number of 'Yes' responses in the sample. Then the probability distribution of n_1 'Yes' responses is :

$$P(n_1) = \frac{\binom{N\lambda}{n_1}\binom{N(1-\lambda)}{(n-n_1)}}{\binom{N}{n}}$$

and

$$E(n_1) = \sum_{n_1=0}^{n} n_1 P(n_1)$$

$$= n\lambda \sum_{n_1=1}^{n} \frac{\binom{N\lambda-1}{n_1-1}\binom{N-(1-\lambda)}{n-n_1}}{\binom{N-1}{n-1}} \cdot$$

$$= n\lambda$$

An unbiased estimate of proportion π in the population is given by the proportion in the sample. Therefore, using [I] under Warner's model, we get

$$\hat{\pi}_w P + (1-\hat{\pi}_w)(1-P) = \frac{n_1}{n}$$

or

$$\hat{\pi}_w = \frac{1}{(2P-1)}[n_1 n^{-1} - (1-P)] \text{ if } P \neq \frac{1}{2} \qquad \text{(III)}$$

Using (II) under unrelated question randomized response model, we get

$$\hat{\pi}_u P + P_1 = \frac{n_1}{n}$$

or

$$\hat{\pi}_u = \frac{n_1 n^{-1} - P_1}{P} \qquad \text{(IV)}$$

where the suffixed 'w' and 'u' have been used to denote the estimators under Warner and unrelated question models respectively.

Variance of $\hat{\pi}$

It may be seen that

$$\text{Var } (\hat{\pi}_w)_{wor} = \frac{1}{(2P-1)^2}\frac{\text{Var}(n_1)}{n^2}$$

$$= \frac{N-n}{N-1}\frac{1}{(2P-1)^2}\frac{\lambda(1-\lambda)}{n} \text{ if } P \neq \frac{1}{2} \qquad \text{(V)}$$

$$\text{and} \qquad \text{Var}\ (\hat{\pi}_u)_{wor} = \frac{N-n}{N-1}\,\frac{1}{P^2}\,\frac{\lambda'(1-\lambda')}{n} \tag{VI}$$

where the suffix 'wor' is used to denote the variance under sampling without replacement.

But we know that the variances in case of Warner and unrelated question models under sampling with replacement 'wr' are :

$$\text{Var}\ (\hat{\pi}_w)_{wr} = \frac{1}{2P-1)^2}\,\frac{\lambda(1-\lambda)}{n} \ \text{if}\ P \neq \frac{1}{2} \tag{VII}$$

$$\text{Var}\ (\hat{\pi}_u)_{wr} = \frac{1}{P^2}\,\frac{\lambda'(1-\lambda')}{n} \tag{VIII}$$

Now comparing (V) with (VII) and (VI) with (VIII) we get

$$\text{Var}\ (\hat{\pi}_w)_{wor} = \frac{N-n}{N-1}\ \text{Var}\ (\hat{\pi}_w)_{wr} \tag{IX}$$

$$\text{and} \qquad \text{Var}\ (\hat{\pi}_u)_{wor} = \frac{N-n}{N-1}\ .\ \text{Var}\ (\hat{\pi}_{uw})_{wr} \tag{X}$$

A n>1 any sampling scheme, therefore

$$\left.\begin{array}{rl} & \text{Var}\ (\hat{\pi}_w)_{wor} < \text{Var}\ (\hat{\pi}_w)_{wr} \\ \text{and} & \text{Var}\ (\hat{\pi}_u)_{wor} < \text{Var}\ (\hat{\pi}_u)_{wr} \end{array}\right| \tag{XI}$$

This shows that sampling without replacement is always superior to sampling with replacement and the relative efficiencfy of SRSWOR as compared to SRSWR is equal to $\frac{N-1}{N-n}$ irrespective of randomized response model used and is same as in the open interview.

As N tends to be large, the distribution of n_1,

$$P(n_1) = \frac{\binom{N\lambda}{n_1}\binom{N(1-\lambda)}{(n-n_1)}}{\binom{N}{n}}$$

approches binomial distribution

$$\binom{n}{n_1}\lambda^{n_1}(1-\lambda)^{n-n_1}\ ,\ n_1 = 0, 1, 2, \ldots n.$$

$$\text{and} \qquad \text{Var}\ (\hat{\pi})_{wor} = \text{Var}\ (\hat{\pi})_{wr} \tag{XII}$$

4. *Comparison between randomized response model and open interview under untruthful reporting*

Let the individuals belonging to group A tell the truth with probability T_a and those belonging to group $\bar{A}$ tell the truth with probability T_b under SRSWOR.

If $y_i = 1$ or 0 according as ith member of the sample reports that he is in group A or not, the conventional estimate of true population proportion is

$$(\hat{\pi})_{wor} = \sum_{i=1}^{n} y_i/n \qquad \text{(XIII)}$$

Then expected value, response bias, and variance of this regular estimate are given by

$$E(\hat{\pi})_{wor} = \pi T_a + (1-\pi)(1-T_b) \qquad \text{(XIV)}$$

$$\begin{aligned} \text{Bias } (\hat{\pi})_{wor} &= E(\hat{\pi})_{wor} - \pi \\ &= \pi(T_a + T_b - 2) + (1 - T_b) \qquad \text{(XV)} \end{aligned}$$

and $$\text{Var}(\hat{\pi})_{wor} = \frac{N-n}{N-1} \; \frac{[\pi T_a + (1-\pi)(1-T_b)]\,[\,(1-\pi \; T_a - (1-\pi)(1-T_b)\,]}{n} \qquad \text{(XVI)}$$

Table 1 compares the mean squre errors (the variance plus the square of the bias) of Warner's and unrelated question randomized response model with regular method of estimation in SRSWOR under the assumption that the interviewed individual tells the truth in the radnomized method but only tells the truth in the non-random method with probabilities given by T_a and T_b. Table 1 is constructed under the assumption that P in each case is low enough to induce full cooperation in the randomzied approach.

It is important to note that except for the cases where the bias of the regular estimate is 0 or negligible, there appears to be sizeable potential gain through the randomized response. The Potential gain of randomized response technique is even larger for larger samples. Comparison of Table 1 with the values calculated by Warner ([1] p.67, Table 1) reveals that use of SRSWOR in randomized response provides greater gains than SRSWR when randomized response techniques are compared with open interview. This gain increases further when unrelated question randomized response model with known proportion of individual in non-sensitive group is used. The gain due to uncorrelated question randomized response model over regular estimator is significant even for nominal untruthful reporting of order 5% by individuals in sensitive group in an open interview and that too for even a small value of P = 0.6.

Table 1. Comparison of randomized and regular estimates for true probability of $\pi = 0.6$, $n = 1000$ and $N = 15000$

Regular Estimates		$\frac{\text{Mean Square Error Under Warner's Model}}{\text{Mean Square Error of Regular estimator}}$				$\frac{\text{Mean Square Error Under Unrelated Model}}{\text{Mean Square Error of Regular estimator}}$			
Probability of 'truth' T_a	T_b	P = 0.6	P = 0.7	P = 0.8	P = 0.9	P = 0.6 P_1 = 0.3	P = 0.7 P_1 = 0.25	P = 0.8 P = 0.15	P = 0.9 P_1 = 0.07
0.95	1.00	5.1600	0.9553	0.5659	0.3147	0.5154	0.3731	0.3012	0.0429
0.90	1.00	1.5200	0.3782	0.1667	0.0927	0.1518	0.1099	0.0887	0.0715
0.70	1.00	0.1785	0.0330	0.0196	0.0109	0.0178	0.0129	0.0104	0.0084
0.50	1.00	0.0646	0.0120	0.0071	0.0039	0.0065	0.0047	0.0038	0.0030
1.00	0.95	9.3962	1.7396	1.0306	0.5731	0.9386	0.6794	0.5484	0.4423
1.00	0.90	3.2089	0.5941	0.3519	0.1957	0.3205	0.2320	0.1873	0.1510
1.00	0.70	0.3993	0.0739	0.0438	0.0244	0.0399	0.0289	0.0233	0.0188
1.00	0.50	0.1451	0.0269	0.0159	0.0088	0.0145	0.0105	0.0085	0.0068
0.95	0.95	17.8797	3.3102	1.9610	1.0906	1.7860	1.2998	1.0436	0.8415
0.90	0.90	9.2843	1.7189	1.0183	0.5663	0.9274	0.6713	0.5419	0.4370
0.70	0.70	1.5200	0.2814	0.1667	0.0927	0.1518	0.1099	0.0887	0.0715
0.50	0.50	0.5692	0.1054	0.0624	0.0347	0.0569	0.0412	0.0332	0.0268

REFERENCES

[1] Greenberg, B.G., Kuebler, Roy R. Jr., Abernathy, James R. and Hortvitz Daniel G., 1971. Application of the randomized response technique in obtaining quantitative data. *J. Amer. Statist. Assoc.* **66.** 243-250.

[2] Horvitz, Daniel G., Greenberg, B.G. and Abernathy, J.R., 1975. Recent development in randomized response designs. A survey of statistical Design and linear model. Edited by J.N. Srivastava, North Holland Publishing Company, Amsterdam, 271-285.

[3] Moors, J.J.A., 1971. Optimization of the unrelated question randomized response model. *J. Amer. Statist. Assoc.* **66,** 627-629.

[4] Warner, Stanley L., 1965. Randomized response : A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* **60,** 63-69.

Jour. Ind. Soc. Ag. Statistics
47(2), 1995 : 142–150

A Class of Estimators for Mean of Symmetrical Population when the Variance is not known

R. Karan Singh and S.M.H. Zaidi
Lucknow University, Lucknow

(Received : December, 1991)

SUMMARY

A class of estimators of population mean (μ) when the variance (σ^2) is unknown, is proposed in case of symmetrical populations. Bias and mean squared error are found for the class. Various estimators are shown to belong to the class and sub-class of optimum estimators in the sense of having minimum mean squared error is found.

Key words : Class of estimators, Coefficient of variation, Mean square error, Unknown variance.

Introduction

Utilising known square of coefficient of variation $C^2\left(=\frac{\sigma^2}{\mu^2}\right)$, Searles [2] proposed an improved estimator of population mean μ; but when C^2 is unknown, the problem of estimation consists of estimators using the estimates of C^2 given by

$$\hat{C}^2 = \frac{s^2}{\bar{y}^2} \text{ or } \hat{C}^2 = \frac{s^2}{\bar{y}^2}\left(1-\frac{s^2}{n\bar{y}^2}\right)^{-1}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2$ for the values $y_1, y_2, ..., y_n$ of a random sample of size n.

In this paper, with $u = \frac{s^2}{n\bar{y}^2}$, the following class of estimators are proposed for population mean μ

$$t = f\left(\bar{y}, \frac{s^2}{n\bar{y}^2}\right) = f(\bar{y}, u)$$

where $f(\bar{y}, u)$ satisfying the validity conditions of Taylor's series expansion, is the function of $(\bar{y}, u)$ such that $f(\mu, 0) = \mu$, first order partial derivative