Optimal Allocation in Stratified Randomized Response Model

Javid Shabbir
Department of Statistics, Quaid-i-Azam University
Islamabad 45320, Pakistan

Sat Gupta
Department of Mathematical Sciences, University of North Carolina at Greensboro,
383 Bryan Building Greensboro, NC 27402, USA

Abstract

A Warner (1965) randomized response model based on stratification is used to determine the allocation of samples. Both linear and log-linear cost functions are discussed under uni and double stratification. It observed that by using a log-linear cost function, one can get better allocations.

Key Words: Randomized response, linear and log-linear functions, cost function, stratification.

Corresponding address: Department of Statistics, Quaid-i-Azam University, Islamabad 45320, Pakistan

Introduction

Various randomized response techniques have developed to obtain truthful answers since the pioneering work by Warne (1965). The significant contribution is by Greenberg *et al*. (1969), Moors (1971), Mangat and Singh (1990), Mangat (1994) and Singh *et al*. (2000). Hong *et al*. (1994) used the proportional allocation for obtaining the response more accurately. Recently Kim and Warde (2003) discussed the Warner's model and used the optimum allocation in stratification. It is no doubt that optimal allocation gives better results as compared to proportional allocation, as also discussed by Cochran (1977). Kim and Warde (2003) mainly focused of getting the truthful response. They did not obtain the allocation of samples and cost function. Khare (1987) discussed the allocation of samples in the presence of non-response. In this paper our emphasize is to look the allocation of samples in relation to cost function in randomized response model. We used both linear and loglinear cost functions and also study is extended to double stratification.

Warner's Stratified Model

Let y_h and x_h ($h = 1$, 2,….., L) be the response and auxiliary variables of stratum *h*, having strata, *L*. A population of size $N = \sum_{h=1}^{L} N_h$ is divided into various strata according to some particular characteristics that may be age, height or martial status *etc*. or some other suitable similar characteristics. A sample of size $n=\sum_{h=1}^{L} n_h$ is obtained from each stratum to measure the response *y*. Each selected respondent from each stratum is instructed to use the randomized device before giving the response in the form of yes or no. We assumed that the number of observations in each stratum is known. The Warner (1965) randomized device consists of sensitive and its negative statements with probability P_h and $(1-P_h)$ respectively. The probability of yes answer under the assumption that response will be true is given by

$$G_h = P_h\pi_h + (1-P_h)(1-\pi_h), \qquad h = 1, 2, \ldots, L ,$$

where G_h is the proportion of yes answer in stratum *h*, P_h is the probability that the respondent in the sample of stratum *h* has a sensitive question card and π_h is the proportion of respondent with yes answer in stratum *h*. The maximum likelihood estimate of π_h is given by $\hat{\pi}_h = \dfrac{\hat{G}_h - (1-P_h)}{2P_h - 1}$, $P_h \neq 0.5$ where $\hat{G}_h$ is the proportion of yes answer in the sample of stratum *h*. Since $\hat{G}_h$ follows the binomial distribution $B(n_h,\ G_h)$ and selection is made independently in different strata. Therefore maximum likelihood estimate of π is $\hat{\pi} = \sum_{h=1}^{L} W_h\hat{\pi}_h$, where $W_h = N_h / N$ is the stratum weight. The $\hat{\pi}$ is an unbiased estimate of π *i.e.* $E(\hat{\pi}) = E[\sum_{h=1}^{L} W_h\hat{\pi}_h] = \sum_{h=1}^{L} W_h E(\hat{\pi}_h) = \sum_{h=1}^{L} W_h\pi_h = \pi$ and its variance is given by

$$Var(\hat{\pi}) = Var[\sum_{h=1}^{L} W_h\hat{\pi}_h] = \sum_{h=1}^{L} W_h^2 Var(\hat{\pi}_h),$$

$$Var(\hat{\pi}) = \sum_{h=1}^{L} \frac{W_h^2}{n_h}[\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}] \tag{2.1}$$

Cost Functions

We discuss the linear and loglinear cost functions in stratified randomized response

Linear Cost Function

Define:

$$C = c_0 + \sum_{h=1}^{L} C_h n_h^{\delta}, \ \delta > 0, \tag{3.1}$$

where *C* is the total cost, c_0 is the fixed cost, C_h is the cost per unit in stratum *h* and δ is a constant. We select a sample of size n_h to minimize the $V(\hat{\pi}) = V$ (say) for a specified cost or to minimize the cost for a specified variance.

For obtaining n_h, we define a function ψ by using the equations (2.1) and (3.1) as

$$\psi = \sum_{h=1}^{L} \frac{W_h^2}{n_h}[\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}] + \lambda[\sum_{h=1}^{L} C_h n_h^{\delta} - (C - c_0)] \qquad (3.2)$$

where λ is the lagrangian multiplier

Now solving (3.2) for. n_h, we get

$$n_h = \left[W_h^2 \left\{ \frac{\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}}{\delta\lambda} \right\} \right]^{1/(\delta+1)} . \qquad (3.3)$$

Taking summation of (3.3) and $\sum_{h=1}^{L} n_h = n$, we get

$$\frac{n_h}{n} = \frac{[W_h^2\{\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}\}/C_h]^{1/(\delta+1)}}{\sum_{h=1}^{L}[W_h^2\{\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}\}/C_h]^{1/(\delta+1)}} . \qquad (3.4)$$

Case 1:

If cost is fixed then by (3.2) and (3.4), we have

$$n = \frac{(C-c_0)^{1/\delta} \sum_{h=1}^{L}[W_h^2\{\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}\}/C_h]^{1/(\delta+1)}}{[\sum_{h=1}^{L}[W_h^2\{\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}\}(C_h)^{1/\delta}]^{\delta/(\delta+1)}]^{1/\delta}} . \qquad (3.5)$$

Case 2:

If variance is fixed then by using (2.1) and (3.4), we have

$$n = [\sum_{h=1}^{L}[W_h^2\{\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}\}C_h^{1/\delta}]^{\delta/(\delta+1)}]$$
$$\times\ [\sum_{h=1}^{L}[W_h^2\{\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}\}/C_h]^{1/(\delta+1)}]/V . \qquad (3.6)$$

Using (3.6), we get optimum variance

$$Var(\hat{\pi})_{opt} = [\sum_{h=1}^{L}[W_h^2\{\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}\}C_h^{1/\delta}]^{\delta/(\delta+1)}]$$

$$\times \; [\sum_{h=1}^{L}[W_h^2\{\pi_h(1-\pi_h)+\frac{P_h(1-P_h)}{(2P_h-1)^2}\}/C_h]^{1/(\delta+1)}]/n\,. \tag{3.7}$$

Log-linear Cost Function

Define:

$$C = c_0 + \sum_{h=1}^{L} C_h \log n_h \quad \delta > 0\,, \tag{3.8}$$

By (2.1) and (3.8), we have

$$V = \sum_{h=1}^{L}\frac{W_h^2}{n_h^2}\{\pi_h(1-\pi_h)+\frac{P_h(1-P_h)}{(2P_h-1)^2}\} + \lambda[\sum_{h=1}^{L} C_h \log n_h - (C-c_0)]\,. \tag{3.9}$$

Solving (3.9) for n_h, we get

$$\frac{n_h}{n} = \frac{W_h^2\{\pi_h(1-\pi_h)+\frac{P_h(1-P_h)}{(2P_h-1)^2}\}/C_h}{\sum_{h=1}^{L}W_h^2\{\pi_h(1-\pi_h)+\frac{P_h(1-P_h)}{(2P_h-1)^2}\}/C_h}\,. \tag{3.10}$$

Case 1:

If cost is fixed then by (3.8) and (3.10), we have

$$n = \exp\left[\frac{(C-c_0)-\sum_{h=1}^{L} C_h \log\left\{W_h^2\{\pi_h(\pi_h-1)+\frac{P_h(1-P_h}{(2P_h-1)^2}\}/C_h\right\}}{\sum_{h=1}^{L} C_h}\right]$$

$$\times \left[\sum_{h=1}^{L} W_h^2\{\pi_h(\pi_h-1)\frac{P_h(1-P_h)}{(2P_h-1)^2}\}/C_h\right] \tag{3.11}$$

Case 2:

If variance is fixed, then by (2.1) and (3.10), we have

$$n = [\sum_{h=1}^{L} C_h][\sum_{h=1}^{L} W_h^2\{\pi_h(1-\pi_h)+\frac{P_h(1-P_h)}{(2P_h-1)^2}\}/C_h]/V \tag{3.12}$$

From (3.12), the optimum variance is

$$V(\hat{\pi})_{opt} = [\sum_{h=1}^{L} C_h][\sum_{h=1}^{L} W_h^2\{\pi_h(1-\pi_h)+\frac{P_h(1-P_h)}{(2P_h-1)^2}\}/C_h]/n\,. \tag{3.13}$$

Numerical Illustration

The following data is used to determine the sample size and expected cost. We obtain the followings (i), (ii) and (iii), linear cost functions by substitution of δ = 2, 1, 0.5 in (3.1) and (iv) is the log-linear cost function.

(i) $C = c_0 + \sum_{h=1}^{L} C_h n_h^2$, (ii) $C = c_0 + \sum_{h=1}^{L} C_h n_h$,

(iii) $C = c_0 + \sum_{h=1}^{L} C_h \sqrt{n_h}$ and (iv) $C = c_0 + \sum_{h=1}^{L} C_h \log(n_h)$.

Table1: Artificial data for two groups

Stratum	π_h	P_h	W_h	C_h
1 2	0.4 0.6	0.6 0.7	0.3 0.7	4 9
1 2	0.2 0.4	0.6 0.7	0.3 0.7	9 4

Table 2: Fixed cost

Cases	$(C - c_0) = 95$ **units**			$(C - c_0) = 100$ **units**		
	n	n_1	n_2	n	n_1	n_2
(i)	6	3	3	6	3	3
(ii)	15	9	7	17	6	11
(iii)	115	67	48	142	45	96
(iv)	3396	2120	1276	6332	1549	4783

Table 3: Fixed variance

Cases	$V = 1$				$V = 0.175$			
	n	n_1	n_2	**E. cost**	n	n_1	n_2	**E. cost**
(i)	3	2	1	22	15	6	9	662
(ii)	3	2	1	17	16	6	10	90
(iii)	3	2	1	14	16	5	11	34
(iv)	3	2	1	3	3	1	2	2

From Tables 2 and 3, it is observed that the log-linear cost function is better as compared to ordinary cost function. The expected cost is decreasing with increase of π_h, P_h and W_h.

Double Stratification

To estimate the proportion of yes answers from respondent *y*, it is reasonable to stratify the population on the basis of auxiliary variable *x*, but when such information on *x* is lacking or stratum weights, W_h are unknown, then we use the double sampling technique, (see Cochran, 1977; Rao, 1973).

(i) A sample of size n' inexpensively is selected from *N* with replacement to observe the auxiliary variable *x*.

(ii) The collected samples are arranged into L strata on the basis of x. Let n'_h be the number of units falling in stratum h, *i.e.* $n' = \sum_{h=1}^{L} n'_h$.

(iii) A sub-sample of size $n_h \le n'_h (= \nu_h n'_h)$, $0 < \nu_h \le 1$, where ν_h is known for each strata, is selected with replacement and the information on y_h in stratum h is collected from respondents using the randomized device to estimate the proportion π. Also $n'_h / n' = w_h$ is an unbiased estimate of $N_h / N = W_h$.

As $\hat{\pi}$ is an unbiased estimate of π, therefore its variance ignoring the finite population correction (*fpc*) is given by

$$Var(\hat{\pi}) = \frac{1}{n'}[\pi(1-\pi) + \frac{P(1-P)}{(2P-1)^2}] + \sum_{h=1}^{L} \frac{W_h^2}{n'}[\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}](\frac{1}{\nu_h} - 1),$$

(5.1) where $P = \sum_{h=1}^{L} P_h$, and $\pi = \sum_{h=1}^{L} \pi_h$.

Linear Cost Function

Define:

$$C = c_0 + c'(n')^{\delta} + \sum_{h=1}^{L} C_h n_h , \qquad (5.2)$$

where c' be the cost of classification per unit, c_0 is the fixed cost and C_h be the cost of measuring unit in stratum h.

Since n_h is random variable therefore the expected cost is.

$$E(C) = C^* = c_0 + c'(n')^{\delta} + n' \sum_{h=1}^{L} C_h \nu_h W_h . \qquad (5.3)$$

Here we choose n' and ν_h to minimize $Var(\hat{\pi}) = V$ (say) for a specified expected cost or to minimize the expected cost for specified variance (see Cochran, 1977).

Now we define a function ψ by using a constant λ as By (5.1) and (5.3), we have

$$\psi = \frac{1}{n'}[\pi(1-\pi) + \frac{P(1-P)}{(2P-1)^2}] + \sum_{h=1}^{L} \frac{W_h^2}{n'}[\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}](\frac{1}{\nu_h} - 1)$$
$$+ \lambda[c'(n')^{\delta} + n' \sum_{h=1}^{L} C_h \nu_h W_h - (C^* - c_0)]. \qquad (5.4)$$

$$\psi = \frac{1}{n'}\pi_B^2 + \sum_{h=1}^{L} \frac{W_h}{n'\nu_h}[\pi_h(1-\pi_h) + \frac{P_h(1-P_h)}{(2P_h-1)^2}]$$

$$+\lambda[c'n'^{\delta}+n'\sum_{h=1}^{L}C_h\nu_hW_h-(C^*-c_0)]\,, \tag{5.5}$$

where $\pi_B^2=[\pi(1-\pi)+\frac{P(1-P)}{(2P-1)^2}]-\sum_{h=1}^{L}W_h[\pi_h(1-\pi_h)+\frac{P_h(1-P_h)}{(2P_h-1)^2}]$.

Now solving (5.5) for n' and ν_h, we get

$$n'=[\frac{\pi_B^2}{\delta\lambda c'}]^{1/(\delta+1)} \tag{5.6}$$

and $$\nu_h=\frac{[\pi_h(1-\pi_h)+\frac{P_h(1-P_h)}{(2P_h-1)^2}]^{1/2}[\delta\lambda c']^{1/(\delta+1)}}{(\pi_B^2)^{1/(\delta+1)}(\lambda C_h)^{1/2}} \tag{5.7}$$

Case 1:

If cost is fixed, then substituting of n' and ν_h from (5.6) and (5.7) in (5.3), we get

$$(A_1A_2^{\delta})^{1/(\delta+1)}+A_3\sqrt{A_2}-(C^*-c_0)=0\,, \tag{5.8}$$

where

$$A_1=[(\frac{\pi_B^2}{\delta})^{\delta}c']^{1/(\delta+1)},\quad A_2=1/\lambda \quad \text{and}$$

$$A_3=\sum_{h=1}^{L}\sqrt{C_h}W_h[\pi_h(1-\pi_h)+\frac{P_h(1-P_h)}{(2P_h-1)^2}]^{1/2}$$

Case 2:

If variance is fixed, then substituting the values of n' and ν_h from (5.6) and (5.7) in (5.1), we get

$$V=\delta A_1(\lambda)^{1/(\delta+1)}+A_3\sqrt{\lambda}\,, \tag{5.9}$$

Log-linear Cost Function

Define:

$$C^*=c_0+c'\log n'+n'\sum_{h=1}^{L}C_h\nu_hW_h\,. \tag{5.10}$$

By (5.1) and (5.12), we have

$$\psi=\frac{1}{n'}[\pi(1-\pi)+\frac{P(1-P)}{(2P-1)^2}]+\sum_{h=1}^{L}\frac{W_h^2}{n'}[\pi_h(1-\pi_h)+\frac{P_h(1-P_h)}{(2P_h-1)^2}](\frac{1}{\nu_h}-1)$$

$$+\lambda[c'\log n'+n'\sum_{h=1}^{L}C_h\nu_hW_h-(C^*-c_0)]\,. \tag{5.11}$$

Solving (5.11) for n' and ν_h, we have

$$n'=\pi_B^2A_2/c' \text{ and } \nu_h=\frac{[\pi_h(1-\pi_h)+\frac{P_h(1-P_h)}{(2P_h-1)^2}c']^{1/2}}{\sqrt{A_2}\pi_B^2\sqrt{C_h}}\,.$$

Case 1:

If cost is fixed, then substituting n' and v_h in (5.10), we get

$$c'\log(\pi_B^2/c') + c'\log(A_2) + A_3\sqrt{A_2} - (C^* - c_0) = 0 \quad (5.12)$$

Case 2:

If variance is fixed, then substituting n' and v_h in (5.1), we get

$$(c'/A_2) + (A_3/\sqrt{A_2}) - V = 0 . \quad (5.13)$$

Conclusion

It is observed that by using a log-linear cost function, one can select more samples for a given fixed cost in Table 2 and in Table 3, the expected cost is much lower for a fixed variance. So generally it is preferable to use the log-linear cost function for selecting the samples in stratified randomized response survey.

Acknowledgement:

The first author wishes to thanks the facilities provided by the University of Southern Maine, USA during his post-doctoral research work in 2003.

References

1. Cochran W. G., 1977, *Sampling Techniques,* (3rd edt.), Wiley and Sons: New York.
2. Greenberg B. G; Abul-Ela; Abdel-Latif A; Simmons W. R., and Horvitz D. G., 1969, *The unrelated question RR model-theoretical frame-work*, J. Amer. Statist. Assoc., 64, 520-539.
3. Hong, K; Yum, J and Lee, H., 1994, *A stratified randomized response technique*, Korean J. App. Statist., 7, 141-147.
4. Khare B. B., 1987, *Allocation in stratified random sampling in presence of nonresponse*, Metrika, XLV, 1-2, 2132-221.
5. Kim J-M. and Warde W. D., 2003, *A stratified Warner's randomized response model*, *J. Statist. Plann. Infer.,* 1-12 (To be appear).
6. Mangat N. S., 1994, *An improved randomized response strategy*, J. Roy. Statist. Soc, B, 56(1), 93-95.
7. Mangat N. S and Singh R., 1990, *An alternative randomized response procedure*, Biometrika, 77, 439-442.
8. Moors, J. J. A., 1971, *Optimization of the unrelated question randomized response model*, J. Amer. Statist. Assoc. 66, 627-629.
9. Rao J. N. K., 1973, *On double sampling for stratification and analytical survey*, Biometrika, 60, 125-133.
10. Singh, S; Singh R and Mangat, N. S., 2000, *Some alternative strategies to Moor's model in randomized response mode*l, J. Statist. Plann. Infer., 83, 243-255.
11. Warner, S. L., 1965, *Randomized response: a survey technique for eliminating evasive bias*, J. Amer. Statist. Assoc. 60, 63-69.