The Korean Communications in Statistics **Vol. 14** No. 3, 2007, pp. 593–600

Nonresponse in Repeated Surveys

Hyeonah Park[1)] Seongryong Na[2)] and Jongwoo Jeon[3)]

Abstract

Under repeated surveys, missing values often appear for various reasons and are replaced by new samples. It is investigated that the existing estimator in repeated survey by Jessen (1942), which has been originally developed for the new samples of fixed size, can be used in such situation where the size of new samples is random. It is shown that the proposed estimator has smaller variance than the sample mean.

Keywords: Response probability; rotation sampling; sample of random size,

1. Introduction

Generally, the samples which are selected from a sampling design are continuously used for a given period. The losses of samples, however, frequently occur during repeated surveys because of moving, nonresponse, and so on. We often replace the missing values with new samples, which is similar to substitution in the imputation method. Various methods of rotation sampling under repeated surveys where the size of new samples at each occasion is fixed have been studied by many authors. Successive sampling with auxiliary variables was considered by Sen (1972). The methods using least square estimator under regression model were studied by Fuller (1990), Chhikara and Deng (1992) and Fuller and Breidt (1999). Scott and Smith (1974), Binder and Dick (1989) and Bell and Hillmer (1990) considered the repeated survey techniques in time series models.

1) Postdoctoral, Statistical Research Center for Complex Systems, Seoul National University, Seoul 151-747, Korea.
Correspondence : parkha03@yahoo.co.kr
2) Professor, Department of Information and Statistics, Yonsei University, Wonju 220-710, Korea.
E-mail : nasr@yonsei.ac.kr
3) Professor, Department of Statistics, Seoul National University, Seoul 151-747, Korea.
E-mail : jwjeon@plaza.snu.ac.kr

We consider the sampling method under repeated surveys where the size of new samples replaced is random, which seems more realistic since the nonresponses of the selected samples occur randomly. In dealing with the nonresponse in repeated surveys, the response probability at each observation is usually assumed to be constant and the size of new samples has the binomial distribution. Here, the existing estimator of repeated survey by Jessen (1942) is extended to the random size case. We show that the variance of this estimator is smaller than that of the usual sample mean.

This paper is organized as follows. In Section 2, we introduce the existing estimator with random size of new samples and discuss the properties of the estimator. In Section 3, results from a limited simulation study are presented.

2. Nonresponse in Two Occasions

Let the finite population be of size N at each occasion and be indexed from 1 to N, where N is assumed to be known. Let the parameter of interest be the population mean $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_{ji}$, where y_{ji} is the study variable of unit i on the j^{th} occasion. Let $\mathcal{F} = \{y_{j1}, \ldots, y_{jN}\}$ be the collection of the study variables in the finite population on occasion j. We assume that sampling on two occasions is performed and simple random sampling is used at each occasion. It is also assumed that missing values appear at the second occasion because of nonresponse, moving and so on, which are changed by new samples.

Let n, m and u be the sample size at each occasion, the size of samples with response at the second occasion, and the size of nonresponses or new samples at the second occasion, respectively. Note that $n = m + u$, where the sample size n at each occasion is fixed and the size of samples with response at the second occasion m is random. We define the estimator based on the new samples at the second occasion by

$$\bar{y}_{2u} = u^{-1}\sum_{i\in U_2} y_{2i},$$

where U_2 is the set of indices of the new samples at the second occasion. The estimator $\bar{y}_{2u}$ is defined to be zero when $U_2 = \emptyset$, which assignment is also applied to other estimators. Let $\bar{y}_{2m}$ be the estimator based on the responses at the second occasion defined by

$$\bar{y}_{2m} = m^{-1}\sum_{i\in M_2} y_{2i},$$

where M_2 is the set of indices of the responses. The estimator based on the

sample at the first occasion is defined by

$$\bar{y}_1 = n^{-1} \sum_{i \in A_1} y_{1i},$$

where A_1 is the set of indices for the sample at the first occasion. We also define

$$\bar{y}_{1m} = m^{-1} \sum_{i \in M_1} y_{1i}$$

to be the estimator based on the samples at the first occasion which respond at the second occasion, where M_1 is the set of corresponding indices and, in fact, $M_1 = M_2$.

The existing estimator which has been examined by Jessen (1942) has the form of

$$\bar{y}_2' = \phi_1 \bar{y}_{2u} + (1 - \phi_1)\bar{y}_{2m}',$$

where $\bar{y}_{2m}' = \bar{y}_{2m} + b(\bar{y}_1 - \bar{y}_{1m})$. Here, ϕ_1 is the constant that minimizes $\mathrm{Var}(\bar{y}_2'|\mathcal{F})$ and is given by $\phi_1 = (W_{2u} + W_{2m})^{-1} W_{2u}$, where $W_{2u}^{-1} = \mathrm{Var}(\bar{y}_{2u}|\mathcal{F})$ and $W_{2m}^{-1} = \mathrm{Var}(\bar{y}_{2m}'|\mathcal{F})$. The estimator b is defined by

$$b = b_N / b_D,$$

where $b_N = [\sum_{i \in M1}(y_{1i} - \bar{y}_{1m})(y_{2i} - \bar{y}_{2m})]$ and $b_D = [\sum_{i \in M1}(y_{1i} - \bar{y}_{1m})^2]$. Note that the existing estimator b is based on the samples of random size m.

The following theorem deals with the asymptotic properties of $\bar{y}_2'$.

Theorem 2.1 *Consider a sequence of finite populations for which y_{ji} have the finite second moments as in Isaki and Fuller (1982). Let $B = B_N / B_D$, where $B_N = E[\sum_{i \in M1}(y_{1i} - \bar{y}_{1m})(y_{2i} - \bar{y}_{2m})]$ and $B_D = E[\sum_{i \in M1}(y_{1i} - \bar{y}_{1m})^2]$.*

Then, we have

$$\bar{y}_{2m}' = \bar{y}_{2m} + B(\bar{y}_1 - \bar{y}_{1m}) + o_P(n^{-1/2}) \tag{2.1}$$

and

$$\bar{y}_2' = \phi_1 \bar{y}_{2u} + (1 - \phi_1)\{\bar{y}_{2m} + B(\bar{y}_1 - \bar{y}_{1m})\} + o_P(n^{-1/2}). \tag{2.2}$$

Let S_2^2 be the population variance at the second occasion. Then, for the correlation coefficient ρ between two occasions and the response probability p,

$$Var(\bar{y}_2'|\mathcal{F}) = n^{-1} S_2^2 [1 - \rho^2 p(1-p) + E(T|\mathcal{F})] + o(n^{-1}), \tag{2.3}$$

where $T = n^{-2}(n^2 - u^2\rho^2)^{-1}[u^3\rho^4(u-n)]$.

Proof: We know that

$$E[(b_N - B_N)^2|\mathcal{F}] = O(n^{-1}) \text{ and } E[(b_D - B_D)^2|\mathcal{F}] = O(n^{-1}).$$

Using Corollary 5.1.1.1 of Fuller (1996),

$$b_N - B_N = O_P(n^{-1/2}) \text{ and } b_D - B_D = O_P(n^{-1/2}).$$

By Taylor expansion,

$$\begin{aligned} b - B &= B_D^{-1}[(b_N - B_N) - B_D^{-1}B_N(b_D - B_D)] + o_P(n^{-1/2}) \\ &= O_P(n^{-1/2}). \end{aligned} \tag{2.4}$$

We express $\bar{y}'_{2m}$ as

$$\bar{y}'_{2m} = \bar{y}''_{2m} + (b - B)(\bar{y}_1 - \bar{y}_{1m}),$$

where $\bar{y}''_{2m} = \bar{y}_{2m} + B(\bar{y}_1 - \bar{y}_{1m})$. We also write that

$$\bar{y}'_2 = \bar{y}''_2 + (1 - \phi_1)(b - B)(\bar{y}_1 - \bar{y}_{1m}),$$

where $\bar{y}''_2 = \phi_1\bar{y}_{2u} + (1-\phi_1)\bar{y}''_{2m}$. Note that the distribution of m is binomial with the response probability p and the number of trials n. By the technique of two phase sampling and the distribution of m, we obtain that

$$E[(\bar{y}_1 - \bar{y}_{1m})^2|\mathcal{F}] = O(n^{-1}).$$

Then, using Corollary 5.1.1.1 of Fuller (1996), we have that

$$\bar{y}_1 - \bar{y}_{1m} = O_P(n^{-1/2}),$$

which, together with (2.4), implies (2.1) and (2.2).

We now deal with (2.3). From (2.1), we obtain that

$$\phi_1 = W_{2u}/(W_{2u} + W'_{2m}) + o(1),$$

where $W'_{2m} = \mathrm{Var}(\bar{y}''_{2m}|\mathcal{F})^{-1}$. Furthermore, since

$$\mathrm{Var}(\bar{y}_{2u}|\mathcal{F}) = E[\mathrm{Var}(\bar{y}_{2u}|\mathcal{F}, m)]$$

and

$$\mathrm{Var}(\bar{y}''_{2m}|\mathcal{F}) = E[\mathrm{Var}(\bar{y}''_{2m}|\mathcal{F}, m)],$$

we see that

$$\phi_1 - \phi_1' = o_P(1), \tag{2.5}$$

where $\phi_1' = W_{2u}''/(W_{2u}'' + W_{2m}'')$, $W_{2u}'' = \mathrm{Var}(\bar{y}_{2u}|\mathcal{F}, m)^{-1}$ and $W_{2m}'' = \mathrm{Var}(\bar{y}_{2m}''|\mathcal{F}, m)^{-1}$. Here, we have used the elementary fact that

$$Z_n = E(Z_n) + O_P(\mathrm{Var}(Z_n)^{1/2})$$

for any sequence of random variables Z_n with finite second moment.

Observe now that

$$\mathrm{Var}(\bar{y}_2'|\mathcal{F}) = E[\mathrm{Var}(\bar{y}_2'|\mathcal{F}, m)] + \mathrm{Var}[E(\bar{y}_2'|\mathcal{F}, m)].$$

From (2.1) and (2.5),

$$\bar{y}_2' = \phi_1'\bar{y}_{2u} + (1 - \phi_1')\bar{y}_{2m}'' + o_P(n^{-1/2}). \tag{2.6}$$

If the fpc is ignored, it follows from (2.6) and the technique of two phase sampling that

$$\mathrm{Var}(\bar{y}_2'|\mathcal{F}, m) = S_2^2[n^2 - (n-m)^2\rho^2]^{-1}[n - (n-m)\rho^2] + o_P(n^{-1}).$$

Then, using the Taylor expansion, we obtain that

$$\mathrm{Var}(\bar{y}_2'|\mathcal{F}, m) = n^{-1}S_2^2[1 + n^{-1}(n^{-1}m^2 - m)\rho^2 + T] + o_P(n^{-1}),$$

which leads to

$$\begin{aligned} E[\mathrm{Var}(\bar{y}_2'|\mathcal{F}, m)] = {} & n^{-1}S_2^2[1 - \rho^2 p(1-p) \\ & + E(T|\mathcal{F})] + o(n^{-1}). \end{aligned} \tag{2.7}$$

From (2.2) and with the aid of the technique of two phase sampling,

$$E(\bar{y}_2'|\mathcal{F}, m) = \bar{Y} + o_P(n^{-1/2}). \tag{2.8}$$

Finally, (2.3) follows from (2.7) and (2.8). □

Remark 2.1 It is easily seen that $\mathrm{Var}(\bar{y}_2'|\mathcal{F})$ decreases as ρ approaches to 1 and n increases. The asymptotic unbiasedness of $\bar{y}_2'$ is a direct consequence of (2.8).

Remark 2.2 Using new samples instead of the missing values under repeated surveys, we generally use the sample mean $\bar{y}_2$ of the form $\bar{y}_2 = n^{-1}\sum_{i\in A_2} y_{2i}$, where A_2 is the set of indices of the sample at the second occasion. It easily follows that

$$\mathrm{Var}(\bar{y}_2|\mathcal{F}) = n^{-1}S_2^2.$$

Then, we obtain that

$$\mathrm{Var}(\bar{y}_2|\mathcal{F}) - \mathrm{Var}(\bar{y}_2'|\mathcal{F}) \doteq n^{-1}S_2^2[\rho^2 p(1-p) - E(T|\mathcal{F})] > 0,$$

where $E(T|\mathcal{F}) < 0$ can be easily verified.

Remark 2.3 If it is assumed that each unit at the second occasion responds with the same probability under the repeated survey, then the number of observations with response has the binomial distribution with the uniform response probability. In this situation, it would be preferable to use the estimator based on the samples of randomly varying size. We have verified that $\bar{y}_2'$ is more effective than the usual sample mean $\bar{y}_2$ in the estimation of the population mean.

3. Simulation Results

In this section we provide the results of a limited simulation study performed to test our theory. In the simulation study, $B = 1,000$ samples of size $n = 100$ were generated by

$$\begin{pmatrix} y_{1i} \\ y_{2i} \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & (1-\rho^2)^{1/2}\sigma_2 \end{pmatrix} \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},$$

where $z_{ji} \sim N(0,1)$ for $i = 1,\ldots,n$ and $j = 1,2$, $\sigma_1 = \sigma_2 = 10$ and $\mu_1 = \mu_2 = 10$. We used the constant response probabilities p, where $p = 0.9, 0.7, 0.5$ and 0.3. We generate response variable r_{2i} using p, where r_{2i} have 0 or 1. Note that the nonresponses at the second occasion, that is, $r_{2i} = 0$ were replaced by new samples.

Using B samples of (y_{1i}, y_{2i}, r_{2i}) $i = 1,\ldots,n$, we computed the empirical values of relative efficiency $\mathrm{Var}(\bar{y}_2|\mathcal{F})/\mathrm{Var}(\bar{y}_2'|\mathcal{F})$ and of expectation $E(\bar{y}_2'|\mathcal{F})$. Table 3.1 contains the simulated values for relative efficiency and expectation. Each cell in Table 3.1 contains $\mathrm{Var}(\bar{y}_2|\mathcal{F})/\mathrm{Var}(\bar{y}_2'|\mathcal{F})$ and, in parenthesis, $E(\bar{y}_2'|\mathcal{F})$ for varying response probability p and correlation coefficient ρ.

As anticipated, it is observed in Table 1 that $\bar{y}_2'$ outperforms $\bar{y}_2$ and is unbiased for $\mu_2 = 10$.

Table 3.1: Relative efficiency (expectation) of $\bar{y}_2'$

p	0.9	0.7	0.5	0.3
$\rho = 1.0$	1.093 (10.050)	1.364 (10.001)	1.550 (9.989)	1.577 (9.998)
$\rho = 0.9$	1.079 (9.984)	1.247 (10.001)	1.348 (10.001)	1.345 (9.964)
$\rho = 0.8$	1.077 (9.966)	1.167 (9.972)	1.191 (10.009)	1.229 (9.959)
$\rho = 0.7$	1.014 (10.005)	1.118 (9.985)	1.157 (9.985)	1.182 (9.972)
$\rho = 0.6$	1.030 (10.012)	1.062 (10.009)	1.065 (9.999)	1.088 (9.992)
$\rho = 0.5$	1.005 (10.012)	1.022 (9.957)	1.055 (10.027)	1.048 (10.034)

We conclude that the above simulation results show that the proposed estimator reveals more efficiency than the sample mean especially when the correlation coefficient is not too small. In repeated surveys when the correlation coefficient is properly large, the utilization of $\bar{y}_2'$ can be recommended irrespective of response probability.

So far, we have assumed that ϕ_1 is all known. In many realistic cases, we may have to estimate in order to use the estimator proposed in this paper. Some consistent estimation methods must be considered so that the decent properties discussed here can remain valid, which will be an object of the future study.

Acknowledgements

This work was supported by the SRC/ERC program of MOST/KOSEF (R11-2000-073-00000)

References

Bell, W. R. and Hillmer, S. C. (1990). The time series approach to estimation for repeated surveys. *Survey Methodology*, **16**, 195–215.

Binder, D. A. and Dick, J. P. (1989). Modelling and estimation for repeated surveys. *Survey Methodology*, **15**, 29–45.

Chhikara, R. S. and Deng, L. Y. (1992). Estimation using multiyear rotation design sampling in agricultural surveys. *Journal of the American Statistical Association*, **87**, 924–932.

Fuller, W. A. (1990). Analysis of repeated surveys. *Survey Methodology*, **16**, 167–180.

Fuller, W. A. (1996). *Introduction to Statistical Time Series*, 2nd ed., John Wiley & Sons, New York.

Fuller, W. A. and Breidt, F. J. (1999). Estimation for supplemented panels, *Sankyā*, **61**, 58–70.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, **77**, 89–96.

Jessen, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experimental Station Research Bulletin*, **304**, 54–59

Scott, A. J. and Smith, T. M. F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association*, **69**, 674–678.

Sen, A. R. (1972). Successive sampling with p ($p \geq 1$) auxiliary variables. *The Annals of Mathematical Statistics*, **43**, 2031–2034.

[Received May 2007, Accepted October 2007]