Bayesian approach of weighting cell estimator

Sangeun Lee[1], Juyoung Lee[2], Jinhee Lee, Minwoong Shin[3]

ABSTRACT

A simple random sample is taken from a population and a particular survey item is subject to nonresponse that corresponds to random subsampling of the sampled values within adjustment cells. Our object is to estimate Bayesian probability interval of the population mean.

KEY WORDS : MAR, nonresponse, Quasi - Randomization, Weighting cell.

I. INTRODUCTION

Most of sample survey, some of the units contacted do not respond to at least some items being asked. Such nonresponse, which we will call survey nonresponse. The problem created by survey nonresponse is, of course, the data value intended by survey design to be observed are in fact missing. These missing data not only mean less efficient estimates because of the reduced the sample size but also the standard complete data method can not be immediately used to analyze the data. Moreover possible biases are caused by nonrespondents which are often systematically different from the respondents. Specially, these biases are difficult to eliminate because of the unknown reasons about nonrespondent.

Sometimes the aim of the nonresponse problem is that find the technique to analyze the data by collecting with first calls, which are then corrected with information about the probability of finding the respondent. In this study it will be proposed the method of estimate the probability of nonrespondent with prior information and also estimate the Bayesian probability interval .

2. QUASI - RANDOMIZATION INFERENCE

Suppose inferences are required for a population with N cases or units and let $Y=(y_{ij})$, where $y_i=(y_{i1}, y_{i2},\dots,y_{ik})$ represents a vector of k items for unit i , $i=1,2,\dots,N$.

[1] Department of Applied Information Statistics, Kyonggi University, Kyonggi-do, Korea, 442-760.

[2] Division of Epidemiology Dept. of infetious disease, National Institute of Health #5 Nokbun-dong, Eunpyung-gu, Seoul, Kore, 122-701.

[3] Department of Statistics, Hankuk University of Foreign Studies, Kyonggi-do, Korea, 449-791.

$$I = (I_1, I_2, \ldots, I_N)^t. \tag{2.1}$$

Let $inc = \{i \mid I_i = 1\}$. Sample selection processes can be characterized by a distribution for I given Y. Suppose that n units are selected by SRS(simple random sampling) and let

$$R_i = \begin{cases} 1 & y_i \text{ } responds \text{ if } sampled \\ 0 & otherwise \end{cases} \tag{2.2}$$

Values of Y are recorded *iff* $R_i = I_i = 1$. Cochran(1963) shows that the mean of the responding units.

$$\bar{y}_k = \sum_i^N I_i R_i y_i / \sum_i^N I_i R_i \tag{2.3}$$

is a biased estimate of $\bar{Y}$ with approximate bias

$$b(\bar{y}_k \mid Y, R) = \bar{Y}_k - \bar{Y} = (1 - \lambda_k)(\bar{Y}_k - \bar{Y}_{NR}) \tag{2.4}$$

where λ_k is the proportion of responding units and $\bar{Y}_{NR}$ is the mean of the nonresponding unit.

The elements of the quasi-randomization approach are

(1) A known distribution $f(I \mid Z)$ of sample selection, as for complete survey data.

(2) An assumed distribution for the response indicators R given I, Y and Z.

Let $Z = (z_{ij})$, where i th row, z_i, represents information about unit i known before the survey.

Consider a population of size N with $\sum_{i=1}^N R_i = M$ respondents. A simple random sample of size n is selected and $\sum_{i=1}^N R_i I_i = m$. Suppose that the distribution of R given I, Y

$$f(R \mid I, Y) = \begin{cases} \binom{N}{M}^{-1} & \sum R_i = M \\ 0 & \sum R_i \neq M \end{cases} \tag{2.5}$$

The probability of response is M/N and does not depend on the units sampled or values of the items. Let $D_i = R_i I_i$ and $D = (D_1, D_2, \ldots, D_N)^T$.

$$f(D \mid I, Y) = \begin{cases} \binom{N}{m}^{-1} & \sum D_i = m \\ 0 & otherwise \end{cases} \tag{2.6}$$

which is the distribution of a simple random sample of size m.

$$\bar{y}_R \pm 1.96 \sqrt{(m^{-1} - N^{-1}) S_{YR}^2}$$

where $\bar{y}_R$ and S_{YR}^2 are the mean and variance of the responding units.

The strong assumption is that R is independent of I and Y. The weighting cell estimators weaken this MCAR assumption by restricting it to hold only within subclasses of the population, so that MAR holds but MCAR does not.

3. WEIGHTING CELL ESTIMATORS

One way of viewing probability sampling is that a unit selected with probability π_i is "representing" π_i^{-1} units in the population, and hence should be given the weight π_i^{-1} in estimates of the population quantities $\pi_i = n_j / N_j$ for units i in stratum j. Horvitz-Thompson estimator for population total T.

$$t = \sum_i^N y_i I_i \pi_i^{-1}$$

The population mean $\overline{Y}$ may be estimated by

$$\overline{y}_w = \sum_i^N w_i y_i \tag{3.1}$$

where

$$w_i = I_i \pi_i^{-1} / \sum_k I_k \pi_i^{-1}$$

Note. since

$$E(I_i \mid Y) = \pi_i$$

$$E(t \mid Y) = \sum^N y_i \pi_i \pi_i^{-1} = T$$

t and $\overline{y}_w$ can only be calculated in the absence of nonresponse so that y_i is observed whenever $I_i = 1$.

Weighting cell estimators extend this approach to handle nonresponse by weighting responding units by the inverse of the probability of selection and response. Suppose we can divide the population into J adjustment cells, within which response is independent of (Y, I). Define an adjustment cell variable C that takes value j for all unit in cell j.

$$f(R \mid I, Y, C) = \begin{cases} \Pi_{j=1}^J \binom{N_j}{M_j} & \sum R_j = M_j \quad \text{for } all\ j \\ 0 & \sum R_j \neq M_j \quad \text{for } any\ j \end{cases} \tag{3.2}$$

where N_j is the number of units in cell j, M_j is the number of units that respond if sampled, and $\phi_j = M_j / N_j$ is the response rate in cell j. If values of ϕ_j were known, Horvitz-Thompson estimators of means and totals would be obtained by weighting responding unit i in cell j by $(\pi_j \phi_j)^{-1}$. $R(j)$ = the set of sampled units in adjustment cell j that respond

$$\hat{\phi}_j = m_j / n_j$$

4. BAYESIAN THEORY WITH COMPLETE RESPONSE.

We treated survey nonresponse from the quasi-randomization viewpoint, where the values of variables (Y, Z) were treated as fixed and inferences were based on the known sampling distribution $f(I \mid Z)$ and the modeled response distribution $f(R \mid I, Y, Z)$.

For inference about finite population quantities, a Bayesian approach where prior distribution are specified for unknown parameters in the model.

The survey design variables Z and the sample indicator variable I are known for all units in the population, and the item variables Y are recorded for the n sampled units with $I_i = 1$.

Write $Y = (Y_{inc}, Y_{exc})$, where Y_{inc} is the set of Y values included in the sample, and Y_{exc} the set excluded from the sample.

4.1 STRATIFIED RANDOM SAMPLING.

Let Z be a variable indicating J strata in the population, such that $Z_i = j$ if unit i belongs to stratum j, $j = 1, 2, \dots, J$.

Let Y be the survey item measured for units in the sample. The distribution $f(Y\,|Z)$ is specified as

$$f(Y\,|Z) = \int f(Y\,|Z\,,\,\theta) f(\theta\,|Z)\,d\theta \tag{4.1}$$

where $\theta = \{(\mu_i, \sigma_j^2), j = 1, 2, \dots, J\}$ are immediate parameters in the model, with reference prior

$$f(\theta) = \prod_{j=1}^{J} \sigma_j^{-2} \tag{4.2}$$

and $f(Y\,|Z\,,\,\theta) = \Pi_{I=1}^{N} f(y_i\,|Z_i\,,\,\theta)$, where for units in stratum j.

$$f(y_i\,|Z_i = j\,,\,\theta) = (2\,\pi\sigma_j^2)^{-1/2} \exp[-(y_i - \mu_k)^2 / 2\,\sigma_j^2] \tag{4.3}$$

the normal distribution with mean μ_j and variance σ_j^2.

The distribution $f(I\,|\,Y, Z)$ corresponds to stratified random sampling of n_j out of N_j units in stratum j, that is $f(I\,|\,Y, Z)$ is constant for all samples $I = (I_1, I_2, \dots, I_N)^T$ that have n_j units in stratum j for $j = 1, 2, \dots, J$ and is zero otherwise. Let $\bar{y}_j$ and s_j^2 denote the sample mean and variance in stratum j. In large samples, the posterior distribution of μ_i is normal with mean $\bar{y}_j$ and variance s_j^2 / n_j. Now the population mean $\bar{Y}_j$ in stratum j has the from

$$\bar{Y}_j = (\sum_{i \in inc} y_{ij} + \sum_{i \in exc} y_{ij}) / N_j$$

the posterior mean of $\bar{Y}_j$ is

$$E(\bar{Y}_j\,|\,Y_{inc}, Z) = (\sum_{i \in inc} y_{ij} + \sum_{i \in exc} E(y_{ij}\,|\,Y_{inc}, Z)) / N_j$$

$$\begin{aligned} E(y_{ij}\,|\,Y_{inc}, Z) &= E[E(y_{ij}\,|\mu_j, Y_{inc}, Z)\,|\,Y_{inc}, Z] \\ &= E(\mu_j\,|\,Y_{inc}, Z) \\ &= \bar{y}_j\,. \end{aligned}$$

It can also be shown that the posterior variance of $\overline{Y}_j$ is

$$Var(\ \overline{Y}_j \mid Y_{inc},\ Z) = (1 - n_i / N_j)\, s_j^2 / n_j .$$

Thus finite Population correction $(1 - n_i / N_j)$ appears in the estimate of precision for the $\overline{Y}_j$ in the same way it arises in stratified random sampling in randomization theory. Inferences about parameters (μ_j) differ from inferences about their finite population analogs $(\ \overline{Y}_j\ )$ by finite population corrections, which can be ignored if the proportions sampled (n_j / N_j) are small. In large samples, then, the posterior distribution of $\overline{Y}$ is normal with mean

$$E(\ \overline{Y} \mid Y_{inc},\ Z) = \sum_{j=1}^{J} P_j\ \overline{y}_j \tag{4.4}$$

and variance

$$Var(\ \overline{Y} \mid Y_{inc},\ Z) = \sum_{j=1}^{J} P_j^2\, (n_j^{-1} - N_j^{-1})\, s_j^2 \tag{4.5}$$

where $P_j = N_j / N$

4.2 BAYESIAN MODELS FOR SURVEY DATA

Divide the survey variables into two groups U and Y, where U are observed for all sampled items and Y are subject to nonresponse. The response pattern for Y is described by the response indicators matrix $R = (R_{ij})$ where $R_{ij} = 1$ if y_{ij} is recorded for unit i and $R_{ij} = 0$ otherwise.

Values of U, R, Y included in the sample are denoted by U_{inc}, R_{inc}, Y_{inc} and the excluded items by U_{exc}, R_{exc}, Y_{exc} respectively. The included items of Y_{inc} are divided into observed value Y_{obs} and missing values Y_{mis}..

5. ADJUSTMENT CELL MODEL.

Suppose a simple random sample of n units is taken from a population of N units, and a particular survey item Y is subject to nonresponse that corresponds to random subsampling of the sampled values within adjustment cells formed from a variable U, recorded for all units in the sample.

Let N_j and $\overline{Y}_j$ respectively, denote the population size and the population mean of Y in adjustment cell $U = j$. Our objective is to estimate the overall mean.

$$\overline{Y} = \sum P_j\, \overline{Y}_i$$

where $P_j = N_j / N$.

Suppose we specify that values of Y within adjustment cell j are *iid* normal with mean μ_j and variance σ_j^2 and that μ_j and $\ln \sigma_j^2$ have locally uniform priors.

Assuming large samples and known $\{N_j\}$, the posterior distribution of $\overline{Y}$ given the respondent data can be shown to be normal with mean

$$E(\overline{Y} \mid Y_{obs}, U_{inc}, \{N_j\}) = \sum P_j \overline{Y}_{jR} \quad (5.1)$$

and variance

$$Var(\overline{Y} \mid Y_{obs}, U_{inc}, \{N_j\}) = \sum P_j^2 (n_j^{-1} - N_j^{-1}) s_{jR}^2 \overline{Y}_{jR} \quad (5.2)$$

where m_j is the respondent sample size and s_{jR}^2 are the sample mean and variance of the respondent Y values in cell j.

Observe that (5.1) is the poststratified estimator, and (5.2) is its sampling variance. Hence these expression yield Bayesian probability intervals identical to confidence intervals from the frequentist approach.

REFERENCE

Cochran(1963) W. G. (1963). Sampling Techniques. New York, Wiley.

Little R. J. A and Rubin D. B. (1987).Statistical analysis with missing data. John Wiley & Sons.

Lohr S. L. (1997). Sampling : Design and analysis. Duxbury Press.

Rubin D. B. (1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons.