A Study on Estimates for the Proportion in the Sample Survey with the Nonresponse

Kay O. Lee*
Sung H. Park**

1. INTRODUCTION

When we estimate the population proportion of the individuals in the population for the attribute or the characteristic, we consider the sample survey. We can consider many methods of the sample survey, as mail questionnaire, visits, personal calls, etc. When we have the list of units in the population, we usually make use of the mail questionnaire. It is economical and free from the investigator's effect on the respondent, but it has some objections. The principal objection is that it involves a large nonresponse rate that might cause a significant bias in the result. The bias arises from the differences in the characteristics under investigation between those who respond and those who do not respond.

We must consider the design of sample survey and the estimators for the population proportion to decrease the bias. The sample design considered in this paper is an application of the double sampling. The different proposal of the estimators for the population proportion results in different precisions. The estimators proposed in this paper are minimum variance unbiased estimator, three kinds of Bayesian estimators which are minimax estimator, Bayesian estimator with a noninformative prior, and Bayesian estimator with a uniform prior.

* Department of Mathematics, Air Force Academy.
** Department of Computer Science & Statistics, Seoul National University.

The comparison of the estimators is needed to find the best estimator among them. In order to compare the estimators, we make use of some loss functions and the risk for the estimators.

2. DESIGN OF THE SAMPLE SURVEY

We generally consider the cost and precision in the sample survey. The mail questionnaire in the sample survey is employed to obtain a number of informations at the minimum cost. The principal objection to this method of collecting factual informations is that it generally involves a large nonresponse rate. Without consideration of the effect of non-respondents, we cannot obtain the result with the required precision. In order to increase the precision, we can take the personal calls or visits that may elicit a complete response, but the cost is very high. Therefore, the compromise method which combines the advantages of both methods is desirable. This idea was suggested by Hansen and Hurwitz [8] and the following general guidelines was proposed by them: (a) select a fairly large size of random sample and send a mail questionnaire to all of them, (b) after the deadline is over, identify the nonresponse class and select a relatively small size of subsample from the nonresponse class, (c) collect data in the subsample by personal calls or visits, and (d) combine data from the two parts of the survey to attain a better result at moderate cost.

We have to determine the initial sample size and the sampling fraction in the second survey so that we can conduct the sample survey.

In order to simplify the following investigations, let's define the notations:

n is the size of sample taken in the first survey.

n_1 is the number of units in the first sample survey that provides the response.

n_2 is the number of units in the first sample survey that does not provide the response.

r_2 is the size of sample in the second survey.

N_1 is the size of response class.

N_2 is the size of nonresponse class.

$N=N_1+N_2$ is the size of population.

K is the reciprocal of the sampling fraction in the second survey.

$$n_2=K\cdot r_2,\quad K>1 \tag{2.1}$$

$$E(n_1/N_1)=E(n_2/N_2)=K\cdot E(r_2/N_2) \tag{2.2}$$

Let us find the values of n and K under a specified precision at the lowest cost.

The cost in taking the first and the second sample is

$$c=c_0n+c_1n_1+c_2r_2, \tag{2.3}$$

where c_0 is the cost of making the first survey.

c_1 is the cost of processing the results from the first survey.

c_2 is the cost of getting and processing the data in the second survey.

The expected cost is given by

$$E(c)=c_0n+c_1w_1n+\frac{c_2w_2n}{K}, \tag{2.4}$$

where w_1n and w_2n are the expected values of n_1 and n_2, since n_1 and n_2 are not known until the first survey is made.

Theorem The values of n and K are determined to minimize the expected cost $E(c)$ for a preassigned expected variance $V(\bar{y})$.

$$V(\bar{y})=\frac{N-n}{N}\cdot\frac{S^2}{n}+\frac{(K-1)w_2}{n}S_2^2,$$

$$K=\sqrt{\frac{c_2(S^2-w_2S_2^2)}{S_2^2(c_0+c_1w_1)}},\qquad n=\frac{N[S^2+(K-1)w_2S_2^2]}{NV_0+S^2},$$

where V_0 is the value of a specified variance of the estimated population mean. [See [4] for references.]

3. ESTIMATORS OF THE POPULATION PROPORTION

We suppose that the population can be devided into two classes, those who will respond and who will not. The two classes will be called the response class and nonresponse class respectively.

We assume that the characteristics in the sample survey are classified into two categories: "favorable" category and "unfavorable" category.

The main subject of this paper will be the estimators for the proportion of the favorable characteristic in the nonresponse class.

In order to avoid confusion in the following investigations, we define some notations as follows:

X_2 is the number of units possessing the "favorable" attribute in the second survey.

m_1 is the number of units possessing the "favorable" attribute in the first survey.

M_1 is the number of units possessing the "favorable" attribute in the response class.

M_2 is the number of units possessing the "favorable" attribute in the nonresponse class.

$\hat{P}$ is the minimum variance unbiased estimator for the "favorable" category proportion.

$\check{P}$ is Bayesian estimator for the "favorable" category proportion.

$\hat{P}_i$ is the minimum variance unbiased estimator for the "favorable" category proportion in the i-th survey.

$\check{P}_i$ is Bayesian estimator for the "favorable" category proportion in the i-th survey.

(1) Minimum Variance Unbiased Estimator

As previously noted, if n_1 units in the sample of size n respond and n_2 do not respond, then we may regard n_1 as a random sample of the response class and n_2 as a random sample of the nonresponse class. Let r_2 denote the size of the random subsample from n_2 to be visited (or called), and let X_2 be the units among r_2 that possess the attribute under study. Clearly, m_1/n_1 and X_2/r_2 are the minimum variance unbiased estimates of P_1 and P_2, respectively. Therefore, the minimum variance unbiased estimator for the category proportion P is given by

$$\hat{P}=\frac{n_1}{n}\hat{P}_1+\frac{n_2}{n}\hat{P}_2, \tag{3.1}$$

where

$$\hat{P}_1=\frac{m_1}{n_1} \quad \text{and} \quad \hat{P}_2=\frac{X_2}{r_2}. \tag{3.1}$$

Thus

$$\hat{P}=\frac{n_1}{n}\cdot\frac{m_1}{n_1}+\frac{n_2}{n}\cdot\frac{X_2}{r_2} \tag{3.2}$$

Let us show that $\hat{P}$ is the minimum variance unbiased estimator.

$$E(\hat{P})=E[E(\hat{P}\,|\,n_1,\ n_2)]=E\left[\frac{n_1}{n}\cdot\frac{M_1}{N_1}+\frac{n_2}{n}\cdot\frac{M_2}{N_2}\right]$$

$$=\frac{M}{N}E(n_1/n)+\frac{M}{N}E(n_2/n)=\frac{M_1+M_2}{N}\equiv P$$

(2) The Minimax Estimator

From this section, we investigate three kinds of Bayesian estimators.

Since the size of n_1 is assumed to be large, the differences of the accuracy between the minimum variance unbiased estimator $\hat{P}_1$ and Bayesian estimators of P_1 will be small. However, the sample size r_2 is small because of the economics involved, and the differences of the accuracy between the $\hat{P}_2$ and Bayesian estimators of P_2 might be sizable depending on the prior distribution of P_2. Therefore, Bayesian approaches are suggested to estimate P_2 from the subsample, which will eventually lead to Bayesian estimates of P.

$$\tilde{P}=\frac{n_1}{n}\tilde{P}_1+\frac{n_2}{n}\tilde{P}_2. \tag{3.3}$$

Since $\hat{P}_1$ is approximately equal to $\tilde{P}_1$ in the accuracy, we may consider

$$\tilde{P}=\hat{P}_1\cdot n_1/n+\tilde{P}_2\cdot n_2/n$$

as a Bayesian estimator. Hence, let us investigate the Bayesian estimates only in the second survey.

We can find the minimax estimator by way of the following steps: (a) select a loss function, (b) determine the prior distribution which minimizes the risk of the selected loss function, and (c) obtain estimator which makes the risk function for the determined prior distribution a constant function. Therefore, we have only to obtain the Bayes estimator with the constant risk.

(3) The Bayesian Estimator with respect to a Uniform Prior Distribution of P_2

As noted in the previous subsection, the estimator will be studied only in the subsample.

We are able to obtain the Bayesian estimator with a uniform prior distribution by way of the following steps: (a) collect data in the subsample, and compute the posterior distribution. [See [5] for reference.]

The posterior distribution of P_2 for given X_2 is written as

$$f(P_2|X_2)=\frac{f(P_2)\cdot f(X_2|P_2)}{\int_0^1 f(P_2)\cdot f(X_2|P_2)dP_2}, \tag{3.4}$$

where $f(P_2)$ is a uniform prior density function of P_2, and $f(X_2|P_2)$ is the density function of X_2 for a given P_2.

We can be given a Bayesian estimate as follows:

$$\begin{aligned}\tilde{P}_2&=\int_0^1 P_2 f(P_2|X_2)dP_2\\&=\frac{\int_0^1 P_2\binom{r_1}{X_2}P_2^{X_2}(1-P_2)^{r_2-X_2}dP_2}{\int_0^1\binom{r_2}{X_2}P_2^{X_2}(1-P_2)^{r_2-X_2}dP_2}\\&=\frac{\Gamma(X_2+2)\cdot\Gamma(r_2-X_2+1)}{\Gamma(r_2+3)}\cdot\frac{\Gamma(r_2+2)}{\Gamma(X_2+1)\cdot\Gamma(r_2-X_2+1)}\\&=\frac{X_2+1}{r_2+2}.\end{aligned} \tag{3.5}$$

(4) The Bayesian Estimator with Noninformative Prior Distribution

If we meet with the situation where "little is known a priori", it will be worthwhile determining noniformative prior and estimating P_2. [See [3] for reference.]

We can obtain the estimate P_2 through the following steps: (a) determine the prior distribution, (b) compute the posterior distribution, and (c) estimate P_2 by the mean of the posterior distribution.

We can find that the noninformative prior for P_2 is proportional to $P_2^{-\frac{1}{2}}\times(1-P_2)^{-\frac{1}{2}}$. Substitution of the appropriate normalizing constant teaches us that

the corresponding posterior distribution for P_2 is given by

$$f(P_2|X_2)=\frac{\Gamma(r_2+1)}{\Gamma\left(X_2+\frac{1}{2}\right)\Gamma\left(r_2-X_2+\frac{1}{2}\right)}P_2^{X_2-\frac{1}{2}}(1-P_2)^{r_2-X_2-\frac{1}{2}}.$$

We can obtain a Bayesian estimate as follows:

$$\begin{aligned}\tilde{P}_2&=\frac{\Gamma(r_2+1)}{\Gamma\left(X_2+\frac{1}{2}\right)\Gamma\left(r_2-X_2+\frac{1}{2}\right)}\int_0^1 P_2 P_2^{X_2-\frac{1}{2}}(1-P_2)^{r_2-X_2-\frac{1}{2}}dP_2\\&=\frac{\Gamma(r_2+1)}{\Gamma\left(X_2+\frac{1}{2}\right)\Gamma\left(r_2-X_2+\frac{1}{2}\right)}\ \frac{\Gamma\left(X_2+\frac{3}{2}\right)\Gamma\left(r_2-X_2-\frac{1}{2}\right)}{\Gamma(r_2+2)}\\&=\frac{1}{}\ \frac{\Gamma\left(X_2+1+\frac{1}{2}\right)}{\Gamma\left(X_2+\frac{1}{2}\right)}\\&=\frac{X_2+\frac{1}{2}}{r_2+1}\end{aligned}\qquad(3.6)$$

4. COMPARISON

In the previous section, we have investigated the methods of estimation for the category proportion only within the nonresponse class. From now on, let us find which of them is the best estimator by comparing them. We employ the risk function for the given loss function and Bayes risk obtained by integrating the risk function for the criterion of comparison.

In order to simplify the following investigation, assign 16 to the size of the second survey and denote the minimum variance unbiased estimator by $\hat{P}_2$, the minimax estimator by $\tilde{P}_{21}$, a Bayesian estimator with respect to a uniform prior by $\tilde{P}_{22}$ and a Bayesian estimator with noninformative prior by $\tilde{P}_{23}$.

Frist, we consider the square error as the loss function.

(1) Loss Function: $L(\hat{P}_2, P_2)=(\hat{P}_2-P_2)^2$

Since the mean square error accounts for both variance and bias, the mean square error seems to be a good criterion for comparison.

Before comparing them, let us find the minimax estimator for this loss func-

tion. First we find the estimator which minimizes

$$\int_0^1 [\tilde{P}_{21}-P_2]^2 c P_2^{\alpha-1}(1-P_2)^{\beta-1} b(r_2,P_2)\,dP_2, \tag{4.1}$$

where $b(r_2, P_2)=\binom{r_2}{X_2}P_2^{X_2}(1-P_2)^{r_2-X_2}$, $c=\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$, $\alpha>1$, $\beta>1$.

since the family of beta distribution is a conjugate family for sample from a Bernoulli distribution.

Differentiating Eq. (4.1) with respect to $\tilde{P}_{21}$, we find the solution.

$$\tilde{P}_{21}=\frac{X_2}{r_2+\alpha+\beta}+\frac{1}{r_2+\alpha+\beta} \tag{4.2}$$

Determine the values of α and β that make the risk function with respect to $\tilde{P}_{21}$ a constant function.

$$\tilde{P}_{21}=\frac{X_2}{\sqrt{r_2}(1+\sqrt{r_2})}+\frac{1}{2(1+\sqrt{r_2})}. \tag{4.3}$$

Next, let us obtain the risk function for the proposed estimator and compare the risk functions. For notational convenience, denote the risk function for the minimum variance unbiased estimator by $R_1(P_2)$, the risk function for the minimax estimator by $R_2(P_2)$, the risk function for the Bayesian estimator with respect to a uniform prior by $R_3(P_2)$, and the risk function for the Bayesian estimator with noninformative prior by $R_4(P_2)$, respectively:

$$R_1(P_2)=\sum_{X_2=0}^{r_2}\left(\frac{X_2}{r_2}-P_2\right)^2 b(r_2,P_2)=\frac{P_2(1-P_2)}{r_2},$$

$$R_2(P_2)=\frac{1}{4(1+\sqrt{r_2})^2},$$

$$R_3(P_2)=\sum_{X_2=0}^{r_2}\left(\frac{X_2+1}{r_2+2}-P_2\right)^2 b(r_2,P_2)$$

$$=\frac{1}{(r_2+2)^2}[P_2^2(4-r_2)+P_2(r_2-4)+1],$$

$$R_4(P_2)=\sum_{X_2=0}^{r_2}\left(\frac{X_2+\frac{1}{2}}{r_2+1}-P_2\right)^2 b(r_2,P_2).$$

Figure 1 shows that the best estimator is $\hat{P}_2$ in (0, 0.09) and (0.91, 1). $\tilde{P}_{21}$ in (0.24, 0.76), $\tilde{P}_{22}$ in (0.18, 0.24) and (0.76, 0.82), $\tilde{P}_{23}$ in (0.09, 0.18) and (0.82, 0.91).

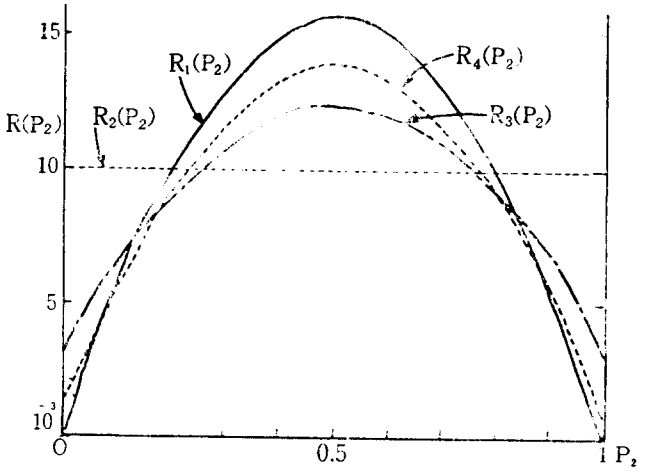

Figure 1. Comparison of the risk function for the square error.

When we have a lot of information for the P_2, we can choose the best estimator among them in the Figure 1, but when we meet with the situation where little information of P_2 is available, which should be chosen as the best estimator? We can regard the Bayes risk as a good criterion of comparison. We may obtain the Bayes risk for the risk function as follows:

$$R_i=\int_0^1 R_i(P_2)dP_2,\ i=1,2,3,4. \tag{4.4}$$

The results of the computation of Eq. (4.4) are given by

$$R_1=1/96,\ R_2=0.01,\ R_3=1/108,\ R_4=0.0095.$$

Therefore, when we have little information with respect to P_2, we may decide that the best estimator is $\tilde{P}_{22}$ obtained by comparing the Bayes risks.

(2) The Loss Function: $L(\hat{P}_2,\ P_2)=\dfrac{|\hat{P}_2-P_2|}{P_2(1-P_2)}$

When we think much of the aspects of economics, it is of great value to consider this formulation as the loss function. This loss function increases, as the population proportion approaches nearer to zero and unity.

Since it is a very complicated task to find the minimax estimator for this loss function, we regard Eq. (4.3) as the minimax estimator for this loss function.

As in the previous subsection, we can compute the risk function.

$$R_1(P_2)=\sum_{X_2=0}^{r_2}\frac{|X_2/r_2-P_2|}{P_2(1-P_2)}b(r_2,P_2)$$

$$=\frac{2}{1-P_2}[I_{P_2}(K_1-1,\ r_2-K_1+1)-I_{P_2}(K_1,\ r_2-K_1+1)],$$

where K_1 is the smallest integer such that $X_2\geqq r_2P_2$.

$$I_{P_2}(\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^{P_2}t^{\alpha-1}(1-t)^{\beta-1}dt$$

$$=\sum_{X_2=K_1}^{r_2}b(r_2,P_2).$$

$$R_2(P_2)=\sum_{X_2=0}^{r_2}\frac{\left|\frac{X_2}{r_2+\sqrt{r_2}}+\frac{1}{2+2\sqrt{r_2}}-P_2\right|}{P_2(1-P_2)}b(r_2,P_2)$$

$$=\frac{2r_2}{(r_2+\sqrt{r_2})(1-P_2)}I_{P_2}(K_2-1,r_2-K_2+1)-\frac{1}{2P_2(1-P_2)(1+\sqrt{r_2})}$$

$$+\left\{\frac{1}{(1+\sqrt{r_2})P_2(1-P_2)}-\frac{2}{1-P_2}\right\}I_{P_2}(K_2,r_2-K_2+1)$$

$$-\frac{r_2}{(r_2+\sqrt{r_2})P_2}+\frac{1}{1-P_2},$$

where K_2 is the smallest integer such that $\frac{X_2}{r_2+\sqrt{r_2}}+\frac{1}{2+2\sqrt{r_2}}\geqq P_2$.

$$R_3(P_2)=\sum_{X_2=0}^{r_2}\frac{\left|\frac{X_2+1}{r_2+1}-P_2\right|}{P_2(1-P_2)}b(r_2,P_2)$$

$$=\frac{r_2}{(r_2+2)(1-P_2)}[2I_{P_2}(K_3-1,r_2-K_3+1)-1]-\frac{1}{(r_2+2)P_2(1-P_2)}$$

$$+\frac{2-2(r_2+2)P_2}{(r_2+2)P_2(1-P_2)}I_{P_2}(K_3,r_2-K_3+1)+\frac{1}{1-P_2},$$

where K_3 is the smallest integer such that $X_2+1\geqq P_2(r_2+2)$.

$$R_4(P_2)=\sum_{X_2=0}^{r_2}\frac{\left|\frac{X_2+\frac{1}{2}}{r_2+1}-P_2\right|}{P_2(1-P_2)}b(r_2,P_2).$$

Figure 2 shows that the best estimator is $\hat{P}_2$ in (0, 0.05) and (0.95, 1), $\tilde{P}_{21}$ in (0.22, 0.78) and $\tilde{P}_{23}$ in (0.05, 0.22) and (0.78, 0.95), $\tilde{P}_{22}$ is never the best estimator.

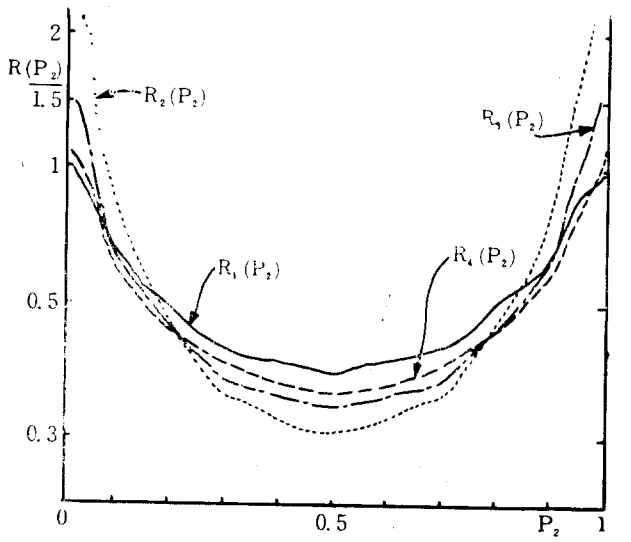

Figure 2. Comparison of the risk function for a new loss function.

As noted previously, the Bayes risks for the risk functions are given by

$$R_1=0.57815,\quad R_2=1.0434,\quad R_3=0.74039,\quad R_4=0.60402.$$

When we make use of Bayes risk as the criterion of comparison, it is easy to find the best estimator among them.

5. CONCLUSION

In the previous sections, the design of sample survey has been constructed and four kinds of estimation procedures have been studied for estimating the category proportion of a population under the existence of nonresponse. In particular, four estimators of P_2 have been discussed and compared. The conclusions based on the comparisons of the risk functions and Bayes risk in the previous section are such that:

(1) If any a priori information is available for the approximated value of P_2, the best estimator among them could be chosen by the value of P_2.

(2) If no information is available for the approximated value of P_2, $\tilde{P}_{22}$ has been chosen to the best estimator by comparing the Bayes risk for the loss

function which is square error.

(3) $\hat{P}_2$ has been chosen to the best estimator by comparing the Bayes risk for the loss function which is the absolute error divided by $P_2(1-P_2)$.

REFERENCES

[1] G.M. Kaufman and Benjamin King, "A Bayesian Analysis of Nonresponse in Dichotomous Process," *Journal of the American Statistical Association,* Vol. 68(1973), 676-678.

[2] Morris H. Degroot, *Optimal Statistical Decisions,* McGraw-Hill, 1970.

[3] G.E.P Box and G.C. Tiao, *Bayesian Inference in Statistical Analysis,* Addison-Wesley, 1970.

[4] William G. Cochran, *Sampling Techniques,* John Wiley & Sons, Inc., 1963.

[5] Sung H. Park, "A Study on Estimation Methods of Proportions Under the Existence of Nonresponse in Survey Sampling," 「統計」, 제 3 권 제 3 호, 1977.

[6] E.T. Whittaker and G.N. Watson, *A Course of Modern Analysis,* Cambridge Univ. Press, 1958.

[7] E.L. Lehmann, *Notes on the Theory of Estimation,* Univ. of California, 1962.

[8] Morris H. Hansen and William N. Hurwitz, "The Problem of Non-response in Sample Survey," *Journal of the American Statical Association,* Vol.41(1946), 517-529.