ESTIMATION THROUGH THE PROCEDURE OF SUBSAMPLING THE NONRESPONDENTS

Poduri S. R. S. Rao, University of Rochester
J. Edward Jackson, Eastman Kodak Company

ABSTRACT

When nonresponse occurs, Hansen and Hurwitz (1946) suggest the procedure of estimating the population mean by subsampling the nonrespondents of the initial sample. We consider the case where not all the subsampled units respond, and suggest estimators that are suitable for some practical situations. We evaluate the relative merits of our estimators by finding their biases and Mean Square Errors.

1. INTRODUCTION

Consider a finite population of size N with mean $\mu = (\Sigma y_i/N)$ of the characteristic of interest y. When a simple random sample of size n is drawn without replacement from the N units and only n_1 of them respond, the sample mean $\bar{y}_1 = (\Sigma y_i/n_1)$ is clearly a biased estimator of $\bar{Y}$. Suppose that the population consists of the "responding" stratum of size N_1 with mean $\bar{Y}_1$ and the "nonresponding" stratum of size $N_2 = N-N_1$ with mean $\bar{Y}_2$. The sample mean $\bar{y}_1$ is unbiased for $\bar{Y}_1$ but for estimating $\bar{Y}$ it has a bias equal to $W_2(\bar{Y}_1-\bar{Y}_2)$, where $W_1 = (N_1/N)$ and $W_2 = (1-W_1)$.

Hansen and Hurwitz (1946) suggest drawing a subsample of size $r = (n_2/k)$, where $k(\geq 1)$ is predetermined, from the $n_2 = (n-n_1)$ nonrespondents and eliciting responses from all of them. If $\bar{y}_{2r}$ is the mean of the r units, an unbiased estimator of μ is

$$\hat{\mu} = \frac{n_1\bar{y}_1 + n_1\bar{y}_2}{n} . \qquad (1)$$

The procedures for determining the optimum values of n and k are described and illustrated in Cochran (1977). J. N. K. Rao (1973) and Srinath (1971) suggest a modified procedure for determining the sample sizes at the initial and second phases. Review and discussion of the different procedures for determining the sample sizes at the two phases are given by the first author in Rao (1983a, 1983b).

In several practical situations, rarely all the r subsampled units respond, even in the case where the initial sample is conducted through the mail, and telephone or personal interviews are employed at the second phase. When only r_2 of the r subsampled units respond, El-Badry (1956) considers a subsample from the $(r-r_2)$ units, and suggests continuing this procedure until all the units in the subsample at the final phase respond. However, in practice, it is not usually convenient to consider subsamples beyond the second phase; even if it is convenient, the units may not respond at the subsequent phases since they are of the "hard core" type or they have no interest in the characteristic under study.

In this article, we consider the population to be composed of three strata of sizes (N_1, N_2, N_3), $(N = N_1 + N_2 + N_3)$, with means $(\bar{Y}_1, \bar{Y}_2, \bar{Y}_3)$ and variances (S_1^2, S_2^2, S_3^2). These strata correspond respectively to the units that respond in the initial sample, in the subsample at the second phase, and the units that do not respond even after these two attempts. The population mean is $\mu = W_1\bar{Y}_1 + W_2\bar{Y}_2 + W_3\bar{Y}_3$, where $W_1 = N_1/N$, $W_2 = N_2/N$ and $W_3 = 1 - W_1 - W_2$.

When only r_2 of the r subsampled units respond, in Jackson and Rao (1983) we have discussed the possibility of different estimators for μ based on different assumptions regarding the nonresponding units. In Section 2 of this article, we construct estimators with assumptions that may be valid in practice and derive their biases and variances. To examine the biases and Mean Square Errors (MSE's) of the estimators for departures from the assumptions, we have computed them for different values of (N_1,N_2,N_3), $(\bar{Y}_1,\bar{Y}_2,\bar{Y}_3)$ and (S_1^2, S_2^2). This investigation is described in Section 3. The results of our investigation are presented in Section 4.

2. ESTIMATORS AND THEIR BIASES & VARIANCES

If there were no nonrespondents, a random sample of size n from the N units would have resulted in n_1, n_2 and n_3 observations in the three strata with sample means $\bar{y}_1$, $\bar{y}_2$ and $\bar{y}_3$ respectively. With nonresponse, only the size of the respondents n_1 and their mean $\bar{y}_1$ are available. Neither (n_2, n_3) nor $(\bar{y}_2, \bar{y}_3)$ are observed. When a subsample of size $r = (n-n_1)/k$ units is drawn from the $(n-n_1)$ nonrespondents, only $r_2(\leq r)$ units respond. Let $\bar{y}_{2r}$ denote their mean. The estimators in the following sections are derived with different assumptions about $\bar{Y}_3$. Derivations of the expectation and variance are described only for the first estimator in some detail, and a similar approach is adopted for the rest of the estimators.

2.1 THE OVERALL SAMPLE MEAN

The mean of the responses from the two phases of sampling is

$$\hat{\mu}_m = \frac{n_1\bar{y}_1 + r_2\bar{y}_{2r}}{n_1 + r_2} . \qquad (2)$$

The approximate expectation and variance of this estimator are

$$E(\hat{\mu}_m) = \frac{W_1\bar{Y}_1 + (W_2/k)\bar{Y}_2}{W_1 + W_2/k} - \frac{W_1(1-W_1-W_2/k)}{n(W_1 + W_2/k)}(\bar{Y}_1-\bar{Y}_2).$$

and

$$V(\hat{\mu}_m) = \left(\frac{n_1}{n_1+r_2}\right)^2 \left[\frac{N_1 - n_1}{N_1-n_1}\right]S_1^2 + \left(\frac{r_2}{n_1 + r_2}\right)^2 \left[\frac{r-r_2}{rr_2} + \frac{N_2 - n_2}{N_2n_2}\right] S_2^2. \quad (4)$$

If we replace n_1, n_2 and r_2 by their means, after simplification the average variance is

$$V(\hat{\mu}_m) = \left(\frac{k}{kW_1 + W_2}\right) W_1T_1 + \frac{W_2}{(kW_1+W_2)^2} T_2. \quad (5)$$

where

$$T_1 = \left(\frac{N-n}{N}\right)\frac{S_1^2}{n}$$

and

$$T_2 = \left[\frac{N - n}{N} + \frac{kW_3}{1-W_1}\right]\frac{S_2^2}{n}.$$

2.2 The Means of the Second and Third Strata Are Equal

In some cases, inability of the interviewer or inadequate time may be responsible for obtaining only r_2 responses from the r sampled units at the second phase. If the questionnaire is of a sensitive nature, the $(r-r_2)$ units may not repond to telephone calls or personal interviews due to reasons of confidentiatlity. In such situations, it is possible that $\bar{Y}_3$ is almost equal to $\bar{Y}_2$, and with this optimistic assumption the population mean becomes $\mu_o = W_1\bar{Y}_1+(1-W_1)\bar{Y}_2$. An unbiased estimator of this mean is

$$\hat{\mu}_o = \frac{n_1\bar{y}_1+(n-n_1)\bar{y}_{2r}}{n}. \quad (6)$$

Thus,

$$E(\hat{\mu}_o) = W_1\bar{Y}_1 + (W_2 + W_3)\bar{Y}_2 \quad (7)$$

and $\hat{\mu}_o$ in (6) is a biased estimator for μ_o if $\bar{Y}_3 \neq \bar{Y}_2$.

The expected variance is

$$V(\hat{\mu}_o) = W_1T_1 + \frac{(1-W_1)^2}{W_2} T_2. \quad (8)$$

2.3 The Mean of the Third Stratum is a Linear Combination of the Means of the First Two Strata

The assumption in Section 2.3 that $\bar{Y}_3=\bar{Y}_2$ may be thought to be an extreme one. A conservative assumption is that $\bar{Y}_3 = W_1\bar{Y}_1 + W_2\bar{Y}_2 / (W_1 + W_2)$. In that case, the expression for the population mean μ becomes the same as the above one for $\bar{Y}_3$ and it may be denoted by μ_c. Since kr_2 is unbiased for n_2, an estimator for this mean is

$$\hat{\mu}_o = \frac{n_1\bar{y}_1 + kr_2\bar{y}_{2r}}{n_1 + kr_2}. \quad (9)$$

The expectation of $\hat{\mu}_c$ is

$$E(\hat{\mu}_c) = \frac{W_1\bar{Y}_1 + W_2\bar{Y}_2}{W_1 + W_2} - \frac{W_1(1-W_1)-W_1W_2}{n(W_1 + W_2)^2}(\bar{Y}_1-\bar{Y}_2) \quad (10)$$

and its average variance is

$$V(\hat{\mu}_c) = \frac{W_1}{(W_1+W_2)^2} T_1 + \frac{W_2}{(W_1+W_2)^2} T_2 \quad (11)$$

2.4 Neglible Value for the Mean of the Third Stratum

If the value of the survey item is of an insignificant amount for all the units of the third stratum, they may not respond to the initial survey nor to the second attempt. For this situation, the population mean becomes $\mu_\ell = W_1\bar{Y}_1 + W_2\bar{Y}_2$, which is the lower bound of μ, and its estimator is

$$\hat{\mu}_\ell = \frac{n_1\bar{y}_1 + kr_2\bar{y}_{2r}}{n}. \quad (12)$$

This estimator is clearly unbiased for the above mean but is biased for μ unless $\bar{Y}_3 = 0$ and has average variance

$$V(\hat{\mu}_\ell) = W_1T_1 + W_2T_2 \quad (13)$$

2.5 A Method of Extrapolation

When responses to a survey arrive in successive waves, Hendricks (1949) suggests a method of extrapolation. The approach in this and the following Sections are analogous to this method. At the end of the initial survey, the response rate is $x_1 = (n_1/n)$ and the sample mean is $z_1 = \bar{y}_1$. At the end of the second attempt, the response rate is $x_2 = (n_1 + r_2)/n$ and the sample mean is $z_2 = (n_1\bar{y}_1 + r_2\bar{y}_{2r})/(n_1 + r_2)$. Assume that the regression of z on x is linear, and let α and β denote the estimates of the intercept and slope coefficients of the regression line. If the response is 100 percent, that is x = 1, the estimate of E(z) is $(\hat{\alpha}+\hat{\beta})$. Estimating these parameters from the above observations we find the projected estimator for μ to be

$$\hat{\mu}_{h1} = \frac{2n_1+r_2-n}{n_1 + r_2}\bar{y}_1 + \frac{n - n_1}{n_1+r_2}\bar{y}_{2r}. \quad (14)$$

The approximate expectation of this estimator is

$$E(\hat{\mu}_{h1}) = \frac{2W_1 + (W_2/k)-1}{W_1+(W_2/k)}\bar{Y}_1 + \frac{1 - W_1}{W_1+(W_2/k)}\bar{Y}_2 - \frac{W_1(1-W_1)-W_1W_2/k}{n(W_1 + W_2/k)^2}(\bar{Y}_1 - \bar{Y}_2) \quad (15)$$

and its average variance is

$$V(\hat{\mu}_{h1}) = \left(\frac{2W_1 + W_2/k - 1}{W_1 + W_2/k}\right)^2 \frac{T_1}{W_1} + \left(\frac{1 - W_1}{W_1+W_2/k}\right)^2 \frac{T_2}{W_2} . \quad (16)$$

2.6 A Second Type of Extrapolation

The values of x_1 and z_1 are the same as in Section 2.5. However, $x_2 = (n_1+kr_2)/n$ and $z_2 = (n_1\bar{y}_1 + kr_2\,\bar{y}_{2r})/(n_1+kr_2)$. We note that (kr_2) is unbiased for n_2 and $\bar{y}_{2r}$ is unbiased for $\bar{y}_2$. Now, with the approach described in the above Section, we find the estimator for the population mean to be

$$\hat{\mu}_{h2} = \frac{(kr_2+2n_1-n)\,\bar{y}_1 + (n-n_1)\,\bar{y}2r}{n_1 + kr_2} . \quad (17)$$

The expectation of this estimator is approximately equal to

$$E(\hat{\mu}_{h2}) = \frac{2W_1 + W_2 - 1}{W_1 + W_2}\bar{Y}_1 + \frac{1-W_1}{W_1 + W_2}\bar{Y}_2 - \frac{W_1(1-W_1) - W_1W_2}{n(W_1 + W_2)^2}(\bar{Y}_1 - \bar{Y}_2) \quad (18)$$

and its average variance is

$$V(\hat{\mu}_{h2}) = \left(\frac{2W_1 + W_2 - 1}{W_1 + W_2}\right)^2 \frac{T_1}{W_1} + \left(\frac{1 - W_1}{W_1 + W_2}\right)^2 \frac{T_2}{W_2} . \quad (19)$$

3. Numerical Investigation

The biases in the six estimators arise due to two reasons. First, the actual population mean may not be the anticipated one. For instance, with the optimistic assumption that $\bar{Y}_2 = \bar{Y}_3$, the population mean is μ_o, and $\hat{\mu}$ may have a large bias if the mean is other than μ_o. The second reason for the bias is that except for $\hat{\mu}_o$ and $\hat{\mu}_\ell$ the sample size in the denominator is a random variable, and the estimators are of the ratio type. The variances of the estimators arise of course due to the sampling at both the phases.

To investigate the differences among the biases, variances and MSE's of the six estimators, we have computed them for serveral combinations of the population sizes N_i, means $\bar{Y}_i$, and variances (S_1^2, S_1^2). These population values are chosen to represent practical situations. We have also computed them for different values of the sample size n and the subsampling fraction (1/k). Conclusions of our investigation are presented in the following Section.

4. Comparisons and Conclusions

As expected, the Optimistic estimator $\hat{\mu}_o$ has smaller bias and MSE relative to the remaining estimators when the means of the second and third strata are equal. Similarly, $\hat{\mu}_c$ has relatively smaller bias and MSE when the mean of the third stratum is a weighted average of the means of the first and second strata. We have also found that for either of the above cases $\hat{\mu}_o$, $\hat{\mu}_c$ and $\hat{\mu}_{h2}$ have smaller biases and MSE's than the remaining three estimators. We note that $\hat{\mu}_o$ is based on the total sample size n, and μ_c as well as μ_{h2} are based on the unbiased estimator $(n_1 + kr_2)$ of n. Further, for all these three estimators appropriate weights are given to the observations from the 2 phases of sampling.

If the mean of the third stratum is negligible, as expected $\hat{\mu}_\ell$ has smaller bias and MSE than the above three estimators. For this situation, $\hat{\mu}_{h1}$ may also be preferred to them.

The overall mean $\hat{\mu}_m$ has relatively larger bias and MSE than the rest of the estimators. We note that this estimator as well as $\hat{\mu}_{h1}$ are based on the actual observed number of responses and not on an estimate of the intial sample size n.

All the estimators are almost unbiased when the anticipated situations are met, even when the sampling fraction (n/N) is small, say 5%. If the expected conditions are not met, the estimators of course are biased. For instance, $\hat{\mu}_c$ will have negligible amount of bias when the means of the second and third strata are equal; otherwise it will have some amount of bias.

The variances of all these estimators are very small relative to their biases, even when the sampling fraction (n/N) is small. As a result, the MSE's of these estimators are not much larger than their squared biases.

As indicated earlier, increasing the sample size n has a negligible effect on the biases of the estimators. The variances of the estimators decrease as n increases, as expected. However, the relative merits of the estimators remain almost the same for large or small values of n.

Increasing or decreasing the subsampling fraction (1/k) has an effect on the biases of only $\hat{\mu}_m$ and $\hat{\mu}_{h1}$ but not of the remaining estimators. The biases of these two estimators decrease as (1/k) increases. As mentioned earlier the remaining four estimators are based on n or its unbiased estimator. The variances of all the estimators of course decrease with increasing subsampling fraction. However, the relative biases and MSE's remain nearly the same whether k is large or small.

REFERENCES

Cochran, W. G. (1977). Sampling Techniques, pp. 370-374. New York: Wiley.

El-Badry, M. A. (1956). A sampling procedure for mailed questionnaires. Journal of the American Statistical Association 51: 209-227.

Hansen, M. H., and Hurwitz, W. N. (1946). The problem of nonresponse in sample surveys. Journal of the American Statistical Association 41: 517-529.

Hendricks, W. A., (1949). Adjustment for bias by non-response in mailed surveys. Agr. Econ. Res., 1, 52-56.

Jackson, J. E., and Rao, P. S. R. S., (1983). Estimation procedures in the presence of non-response. American Statistical Association. 1983 Proceedings of the Section on Survey Research Methods.

Rao, J. N. K. (1973). On double sampling for stratification and analytical surveys. Biometrika 60: 125-133.

Rao, P. S. R. S., (1983a). Randomization Approach. Incomplete Data in Sample Surveys Vol. 2, Part III, 97-105. Madow, W. G., Olkin, I., and Rubin, D. (Eds.), Academic Press.

Rao, P. S. R. S., (1983b). Hansen-Hurwitz Method for Subsampling Nonrespondents. Encyclopedia of Statistical Sciences, Vol. 3, 573-574. Kotz, S., Johnson, N. L. and Read, C. B., (Eds.), Wiley Interscience.

Srinath, K. P. (1971). Multiphase sampling in non-response problems. Journal of the American Statistical Association 66: 583-586.