A Theory of Minimum Mean Square Estimation in Surveys with Nonresponse

Crop
Reporting
Board

ıtistical Reporting
rvice

U.S. Department
of Agriculture

Washington, D.C.

June 1977

A THEORY OF MINIMUM MEAN SQUARE ESTIMATION IN SURVEYS WITH NONRESPONSE

By

Harold F. Huddleston
Statistical Reporting Service
U.S. Department of Agriculture

June 1977

PREFACE

This paper presents a theory for handling a problem for which survey practitioners wish statisticians would find a solution. That is, the theory places greater weight on large selective samples of respondents and less weight on small samples of nonrespondents than the unbiased estimator which is commonly used. The theory presented herein suggests a way of taking a small step in this direction.

A procedure is described for making estimates in situations where nonresponse arises because of difficulties of accessibility of a portion of the population. However, it is assumed this portion of the population is still accessible for obtaining the desired survey information but at a substantial increase in cost, or delay in time. Consequently, the unbiased estimator first described by Hansen and Hurwitz can be employed. The biased estimator described in this paper permits the mean square error of the estimator to be determined. In addition, it indicates under what conditions the mean square error is less than the sampling error of the Hansen-Hurwitz estimator. These results also indicate under what condition the sample of nonrespondents may be reduced and the mean square error remain less than the error of the Hansen-Hurwitz estimator. This is a direct result of the ratio of the mean of the respondents and the expected standard error of the nonrespondents. Alternatively, this ratio may be thought of as the coefficient of variation of the nonrespondent mean when it is equal to the respondent mean. Consequently, the proposed estimator for the population mean may have a smaller sampling error or permit a smaller sample of nonrespondents than the more conventional estimator proposed by Hansen and Hurwitz. For agencies making repeated surveys of the same or similar populations, such information may be readily available and the proposed estimator can be used with confidence.

The author wishes to acknowledge the valuable guidance and encouragement given by H. O. Hartley, Institute of Statistics, Texas A&M University, in pursuing this approach.

CONTENTS

Page

CHAPTER 1 - INTRODUCTION

1.1 Objective.. 1

1.2 Review of Literature.................................... 1

CHAPTER 2 - SIMPLE RANDOM SAMPLE FROM LIST FRAME

2.1 One Hundred Percent Sampling of Frame.................. 3

2.2 The Classical Unbiased Estimator....................... 3

2.3 Simple Minimum Mean Square Estimator of Mean............ 4

2.4 The Optimum Weights for Strata......................... 5

2.5 Sample Estimator for Mean.............................. 6

2.6 Variance and Mean Square Error......................... 7

2.7 A Minimum Mean Square Ratio Estimator.................. 11

2.8 Optimum Weight Using Ratio Estimator................... 12

2.9 Sample Estimators for Ratio Method..................... 13

CHAPTER 3 - TWO-PHASE SELECTION OF SAMPLE UNITS

3.1 Simple Random Sample of Frame Units.................... 15

3.2 The Classical Unbiased Estimator and Variance.......... 15

3.3 Simple Minimum Mean Square Estimator................... 16

3.4 The Optimum Weights for Domain Means................... 17

3.5 Sample Estimators...................................... 18

3.6 Variance and Mean Square............................... 19

3.7 Ratio Estimator for Random Subsample of Units.......... 25

3.8 Optimum Weight Using Ratio Estimator................... 27

3.9 Sample Estimators for Combined Ratio................... 28

Page

CHAPTER 4 - STRATIFIED SAMPLE FROM LIST FRAME

4.1 Simple Random Sampling from All Strata.................. 30

4.2 Estimators for 100 Percent Sampling of List in Each Stratum.. 30

4.3 Ratio Estimators for 100 Percent Sampling of List from Strata.. 37

4.4 Random Subsample of List Available for Stratum........... 42

4.5 Ratio Estimator for Random Subsample of Stratified List.. 46

4.6 Sample Estimators for Separate Ratio Estimators.......... 48

CHAPTER 5 - AN EXAMPLE FOR A STRATIFIED LIVESTOCK SURVEY

5.1 Nature of Survey.. 49

5.2 Survey Means and Error Estimates........................ 49

5.3 Comments on Comparison with Classical Estimator.......... 50

REFERENCES... 53

CHAPTER 1 - INTRODUCTION

1.1 Objective

A problem frequently encountered in sample surveys is how to deal with nonresponse. That is, the desired information is not secured for a significant part of the sample on the initial attempt. Considerable effort has gone into seeking meaningful procedures to handle such problems that arise in single frame surveys. There are several types of nonresponse which might be encountered: (1) A survey is conducted by mail but only a fraction of questionnaires are returned; (2) A large-scale area interview survey of households is conducted but many persons are not at home; (3) A study of a select group of people over time results in many persons moving or otherwise not being available. In most of the above cases, a follow-up procedure is advisable, but the second phase or follow-up sample of nonrespondents is considerably more expensive than the initial method of sampling.

Several techniques have been suggested which attempt to avoid the more costly follow-up phase. Certain difficulties arise with such procedures, but they deserve mention since they do provide a degree of adjustment for nonresponse. Their greatest difficulty lies in the fact that they do not provide a measure of accuracy for the estimator.

It is the purpose of this paper to deal with estimation for the total population in such a way that a measure of accuracy is available and to determine under what conditions the expected mean square error of the estimator will be less than the error of the classical method of Hansen and Hurwitz.

1.2 Review of Literature

Early workers utilizing mail surveys attempted to adjust for the nonresponse by use of regression methods. That is, a concomitant variable was available for the response and nonresponse groups or strata. Normally, the covariate was available for an earlier point in time. Where the correlation between the covariate and the characteristics being estimated was high, the adjustment for nonresponse without the follow-up phase of sampling was reasonably satisfactory. For cases where the timeliness of the survey was not affected by repeated application of the initial sampling method, a trend in the means related to time segments was frequently found to exist. There is evidence, for instance, to support the assumption that the magnitude of the characteristic may be related to the availability of the person or the person's willingness to supply the information requested. For situations where the nonresponse is due to the "resistance" of individuals to responding, a technique set forth by Hendricks (1949, 1956) based on a series of follow-up phases has been verified for several agricultural populations.

While these techniques may be successful in reducing the bias due to nonresponse, the variability for the nonresponse strata is unknown. A method suggested by Hartley (1946) and applied by Politz and Simmons (1949, 1950) makes use of the availability of persons during the survey period or previous week to provide probability weights for the not-at-homes in a survey. This method does provide a measure of the survey precision.

The double sampling technique, which provides unbiased estimation and sampling errors, appears to have been first given in a paper by Hansen and Hurwitz (1946). Their procedure provided for a random sample of nonrespondents to be selected for follow-up interviews. An extension of this result for two-stage sample has been given by Faradori (1962).

CHAPTER 2 - SIMPLE RANDOM SAMPLE FROM LIST FRAME

2.1 One Hundred Percent Sampling of Frame

The methods developed in this chapter are concerned with the situation in which the frame units are classified into two strata called respondents and nonrespondents. The strata means are to be combined linearly based on weights derived from sample data to estimate the mean for all N units.

The classical procedure considers a frame of N units which corresponds exactly to the target population to be surveyed. A survey of all N of the units yields information for N_1 units, leaving N_2 units for which no information is obtained. For all surveys which require measurement of survey error, a second stage of sampling is completed by selecting a random sample of n_2 units from N_2 units for which information is then obtained.

Unbiased estimation requires that the strata means be combined using $\frac{N_1}{N}$ and $\frac{N_2}{N}$ as weights for the respondents and nonrespondents strata where $N_1 + N_2 = N$.

2.2 The Classical Unbiased Estimator

A procedure due to Hansen and Hurwitz (1946) was first developed for surveys in which the initial attempt was made to secure information by mail. A subsample of persons who did not return a complete questionnaire by mail was visited to secure information by personal interview. We assume that the more expensive method of personal interview is successful in securing information for all units. The estimator used for the mean of the population is:

$$\bar{Y} = \frac{N_1}{N}\bar{Y}_1 + \frac{N_2}{N}\bar{Y}_2 \tag{2.2.1}$$

and the sample estimator is:

$$\bar{y} = \frac{N_1}{N}\bar{Y}_1 + \frac{N_2}{N}\bar{y}_2 \tag{2.2.2}$$

where $\bar{Y}_1$ = population mean for mail respondents

$\bar{Y}_2$ = population mean for nonrespondents

$\bar{y}_2$ = sample mean for nonrespondents

$\bar{N}_1$ = number of respondents

N_2 = number of nonrespondents = $N - N_1$

The population variance of $\bar{Y}$ is:

$$V(\bar{Y}) = (\frac{N_2}{N})^2 (1 - \frac{n_2}{N_2}) \frac{\sigma_2^2}{n_2} \qquad (2.2.3)$$

where σ_2^2 = variance of nonrespondent strata

n_2 = size of nonrespondent sample selected for personal interview (a fixed size for each survey)

The sample estimate of the variance of (2.2.2) is:

$$v(\bar{y}) = (\frac{N_2}{N})^2 (1 - \frac{n_2}{N_2}) \frac{s_2^2}{n_2} \qquad (2.2.4)$$

where $$s_2^2 = \frac{\sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2}{n_2 - 1}$$

y_{2i} = the i^{th} nonrespondent interviewed

2.3 Simple Minimum Mean Square Estimator of Mean

It is proposed that a class of biased estimators based on a linear combination of the strata means be considered such that the nonrespondent weight W_2 be less than $\frac{N_2}{N}$ and the respondent weight is $W_1 = 1 - W_2$.

The estimator proposed is:

$$\bar{Y}_M = W_1 \bar{Y}_1 + W_2 \bar{Y}_2 \qquad (2.3.1)$$

and its sample estimator is:

$$(2.3.2) \qquad \bar{y}_m = W_1 \bar{Y}_1 + W_2 \bar{y}_2$$

since $E(\bar{y}_2 | N_2) = \bar{Y}_2$ for a simple random sample of units from N_2. The bias of (2.3.1) is:

$$(2.3.3) \qquad \text{Bias} = \bar{Y}_M - \bar{Y} = (W_2 - \frac{N_2}{N}) \; (\bar{Y}_2 - \bar{Y}_1)$$

and variance

$$(2.3.4) \qquad V \; (\bar{Y}_M) = W_2^2 \; (1 - \frac{n_2}{N_2}) \frac{\sigma_2^2}{n_2}$$

Therefore, the mean square error of (2.3.1) is:

$$(2.3.5) \qquad \text{M.S.E.} \; (\bar{Y}_M) = (W_2 - \frac{N_2}{N})^2 \; (\bar{Y}_2 - \bar{Y}_1)^2 + W_2^2 \; (1 - \frac{n_2}{N_2}) \frac{\sigma_2^2}{n_2}$$

2.4 The Optimum Weights for Strata

The value of W_2 which will minimize the mean square error is desired. The derivative of (2.3.5) with respect to W_2 is set equal to zero and solved for W_2 . This optimum value is designated W* .

$$(2.4.1) \qquad f' \; (\text{M.S.E.}) = 2W_2 \; (1 - \frac{n_2}{N_2}) \frac{\sigma_2^2}{n_2} + 2(W_2 - \frac{N_2}{N}) \quad (\bar{Y}_2 - \bar{Y}_1)^2 = 0$$

$$(2.4.2) \qquad W_2^* = \frac{\frac{N_2}{N} \; (\bar{Y}_2 - \bar{Y}_1)^2}{(1 - \frac{n_2}{N_2}) \frac{\sigma_2^2}{n_2} + (\bar{Y}_2 - \bar{Y}_1)^2}$$

If the following change of variable is made,

$$\text{let } T = \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{(1 - \frac{n_2}{N_2}) \frac{\sigma_2^2}{n_2}}}$$

Then (2.4.2) may be rewritten as

$$(2.4.3) \qquad W_2^* = \frac{\frac{N_2}{N} T^2}{1 + T^2} = \frac{\frac{N_2}{N}}{\frac{1}{T^2} + 1}$$

When $(\bar{Y}_2 - \bar{Y}_1)$ is different than zero, $\frac{1}{T^2} > o$ and W_2^* is less than $\frac{N_2}{N}$.

When T is quite large, $W_2^* \doteq \frac{N_2}{N}$.

The optimum value of W_2 results in giving greater weight to the respondents than the classical unbiased estimator as long as the difference in the means $(\bar{Y}_2 - \bar{Y}_1)$ is not too large relative to the variance.

2.5 Sample Estimator for Mean

It is proposed that the estimators for the mean, bias, and weights (W_2^*) be constructed from heuristic considerations. The justification will be sought by a study of its mean square error which is given in the next section. The proposed sample estimator of the mean is:

$$(2.5.1) \qquad \bar{y}_m = (1 - \hat{W}_2^*) \bar{Y}_1 + W_2^* \bar{y}_2 = \bar{Y}_1 + \hat{W}_2^* (\bar{y}_2 - \bar{Y}_1)$$

where $\bar{Y}_1$ = the mean of the N_1 respondents

$\bar{y}_2$ = the sample mean from a simple random sample of n_2 units selected from the N_2 nonrespondent

W_2^* = the sample weight which is to be estimated by

$$(2.5.2) \qquad W_2^* = \frac{\frac{N_2}{N} t^2}{1 + t^2}$$

where $$t = \frac{(\bar{y}_2 - \bar{Y}_1)}{\sqrt{(1 - \frac{n_2}{N_2}) \frac{S_2^2}{n_2}}}$$

$$s_2^2 = \frac{\sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2}{n_2 - 1}$$

y_{2i} = the i^{th} nonrespondent interviewed

The square of the bias is given approximately by

$$2.5.3) \qquad \hat{B}^2 \doteq (\hat{W}_2^* - \frac{N_2}{N})^2 \; (\bar{y}_2 - \bar{Y}_1)^2$$

In practice, the sample variance and mean square error will be derived from the tables given in the next section, and the variance of the classical estimator.

2.6 Variance and Mean Square Error

The mean of the nonrespondents, $\bar{y}_2$, is approximately normally distributed if the nonrespondent sample size, n_2 , is moderately large; hence, $\bar{y}_2 - \bar{Y}_1$ will be approximately normal, and the error in S_2^2 negligible. Substituting for W_2^* , (2.3.1) is

$$(2.6.1) \qquad \bar{Y}_M = \bar{Y}_1 + \frac{\frac{N_2}{N}(\bar{Y}_2 - \bar{Y}_1)^3}{(1 - \frac{n_2}{N_2}) \frac{\sigma_2^2}{n_2} (\bar{Y}_2 - \bar{Y}_1)^2} = \bar{Y}_1 + F\;(\bar{Y}_2, \sigma_2^2)$$

since $\bar{Y}_1$ is known and n_2 is fixed for repeated sampling of nonrespondents.

Therefore, the variance of $\bar{Y}_M$ is:

$$(2.6.2) \qquad V\;(\bar{Y}_M) = (\frac{N_2}{N})^2 \quad V\,[\frac{(\bar{Y}_2 - \bar{Y}_1)^3}{(1 - \frac{n_2}{N_2}) \frac{\sigma_2^2}{n_2} + (\bar{Y}_2 - \bar{Y}_1)^2}]$$

Before proceeding to a study of the variance of $\bar{Y}_M$ we generalize this estimator which arose from one attempt to construct an estimate with minimum mean square error. Since it is necessary to estimate the unknown parameters $(\bar{Y}_2 - \bar{Y}_1)$ and σ_2^2 from the data, $\bar{y}_m$ has strictly speaking lost the property of minimum mean square error and there is no reason why modification should not be considered to reduce the M.S.E.

The generalization of (2.6.1) and (2.6.2) is as follows:

$$(2.6.3) \qquad \bar{Y}_M = \bar{Y} + \frac{\frac{N_2}{N}\,(\bar{Y}_2 - \bar{Y}_1)^{2\beta + 1}}{\{(1 - \frac{n_2}{N_2})\,\frac{\sigma_2^2}{n_2} + (\bar{Y}_2 - \bar{Y}_1)^2\,\}^{\beta}}$$

and

$$(2.6.4) \qquad V\,(\bar{Y}_M) = (\frac{N_2}{N})^2 \; V\,[\,\frac{(\bar{Y}_2 - \bar{Y}_1)^{2\beta + 1}}{\{(1 - \frac{n_2}{N_2})\,\frac{\sigma_2^2}{n_2} + (\bar{Y}_2 - \bar{Y}_1)^2\,\}^{\beta}}\,]$$

where $\beta \geq o$.

To study (2.6.3), let $u = z - \Delta$

where $z \sim N\,(0,1)$ and $\Delta =$ constant.

Explicitly, $u = \dfrac{\bar{Y}_2 - \bar{Y}_1}{\sigma_2 \div \sqrt{n_2}}$ and $\Delta = \dfrac{\bar{Y}_1}{\sigma_2 \div \sqrt{n_2}}$

We now examine the variance of (2.6.3) and the mean square error by numerical integration in terms of z for values of β and Δ using as our variable

$$(2.6.5) \qquad y_i = \frac{(z_i - \Delta)^{2\beta + 1}}{[(z_i - \Delta)^2 + 1]^{\beta}}$$

where without loss of generality we have let $\bar{Y}_2 = 0$, and $\frac{\sigma_2^2}{n_2} = 1$.

The sampling fraction is assumed small so the finite population factor may be disregarded and the variance of $\frac{s_2^2}{n_2}$ is assumed negligible.

$$E\,(y_i) = E\,[\Phi(z_i)] = \sum_{i=1}^{14} f_i\,[\Phi(z_i)] \quad \text{where } z_i$$

is the midpoint and f_i the class frequencies for the normal distribution. The midpoint and class frequencies are given in Appendix 1, page 14.

a b.

The variance and mean square error are shown in Tables 1 and 2 apart from the factor $(\frac{N_2}{N})^2$.

The corresponding variance of the classical unbiased estimator $\bar{Y}$ is:

$$V\ (\bar{Y}) = (\frac{N_2}{N})^2 \ \frac{\sigma_2^2}{n_2} = (\frac{N_2}{N})^2 \quad \text{for} \quad \frac{\sigma_2^2}{n_2} = 1 \ .$$

The variance of $\bar{Y}$ which is comparable with the values in Table 1 is 1.0. The mean square error, apart from the factor $\frac{N_2}{N}$ is given by the following expression:

$$(2.6.6) \qquad V\ [\Phi(z_i)] + [E\ \Phi\ (z_i)]^2$$

Table 1--Ratio of Variances: $V\ (\bar{Y}_M) \div V\ (\bar{Y})$

β	Δ							
	0	.25	.50	.75	1.00	1.50	2.00	3.00
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
.25	.8148	.8237	.8459	.8798	.9201	.9971	1.0476	1.0675
.50	.6714	.6842	.7206	.7748	.8408	.9713	1.0628	1.1080
.75	.5638	.5791	.6237	.6908	.7732	.9421	1.0677	1.1415
1.00	.4780	.4969	.5460	.6211	.7139	.9107	1.0650	1.1691
1.50	.3585	.3765	.4287	.5105	.6137	.8450	1.0435	1.2092
2.00	.2757	.2932	.3445	.4260	.5316	.7700	1.0082	1.2332
2.50	.2166	.2329	.2814	.3595	.4629	.7153	.9651	1.2444
3.00	.1729	.1879	.2323	.3060	.4049	.6554	.9180	1.2455

Table 2--Ratio of Mean Square Error to Variance: MSE ($\bar{Y}_M$) ÷ V ($\bar{Y}$)

β	Δ							
	0	.25	.50	.75	1.00	1.50	2.00	3.00
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
.25	.8148	.8244	.8486	.8851	.9280	1.0083	1.0587	1.0745
.50	.6714	.6866	.7297	.7927	.8679	1.0103	1.1025	1.1346
.75	.5638	.5837	.6410	.7254	.8258	1.0197	1.1491	1.1986
1.00	.4780	.5040	.5725	.6746	.7958	1.0340	1.1975	1.2661
1.50	.3585	.3887	.4745	.6040	.7589	1.0719	1.2979	1.4099
2.00	.2757	.3103	.4092	.5593	.7410	1.1166	1.4003	1.5626
2.50	.2166	.2546	.3638	.5303	.7342	1.1643	1.5026	1.7218
3.00	.1729	.2137	.3312	.5114	.7342	1.2129	1.6037	1.8855

Table 1 indicates the variance for any value of β greater than zero and Δ less than 1.5 that the proposed estimator will have a smaller variance. In Table 2 the mean square error is seen to be larger if Δ is as large as 1.5. The mean square error becomes much larger than the classical estimator if β > 1 and Δ > 1.5. If no prior knowledge is available on the value of Δ for the population of interest and all values appear equally likely over the range 0 to 3, a value β = 1 should be used. For repetitive surveys where prior information on the magnitude of Δ will be available, the choice of β will depend primarily on whether Δ is greater than 1.5 or less than 1.5. A value for $\beta \geq 1$ is desirable for $\Delta \leq 1.0$. If the value of Δ expected is quite small, say ½ or less, a value of β = 3 would result in considerable reduction in the error.

While Tables 1 and 2 provide a basis for judging the usefulness of the proposed estimator (2.3.1), the tables also provide a means of calculating the variance and mean square error of sample estimator (2.5.1) based on the variance of the classical estimator. If Table 1 values are referred to as $V(\beta,\hat{\Delta})$ and the sample estimators of Δ,

$(\bar{y}_2 - \bar{Y}_1) \div \frac{S_2}{\sqrt{n_2}}$, is used to enter the table along with β, the variance of (2.5.1) is:

$$(2.6.6) \qquad v\ (\bar{y}_m) = (\frac{N_2}{N})^2\ V\ (\beta,\hat{\Delta})\ \frac{s_2^2}{n_2}$$

and the values in Table 2 referred to as M $(\beta,\hat{\Delta})$ then the sample estimator of the mean square error is:

$$(2.6.7) \qquad \text{M.S.E.}\ (\bar{y}_m) = (\frac{N_2}{N})^2\ M\ (\beta,\hat{\Delta})\ \frac{S_2}{n_2} \quad \text{ignoring f.p.c.}$$

However, an alternative estimator of (2.6.7) based on the sample estimator (2.5.3) and (2.6.6) is:

$$(2.6.7') \qquad \text{M.S.E.}\ (\bar{y}_m) = (\frac{N_2}{N})^2\ V\ (\beta,\hat{\Delta})\ \frac{s_2^2}{n_2} + (\hat{W}_2^* - \frac{N_2}{N})^2\ (\bar{y}_2 - \bar{Y}_1)^2$$

2.7 A Minimum Mean Square Ratio Estimator

A ratio estimator for the nonrespondent stratum mean is proposed which combined with the respondent mean to form a linear combination of the two means. The mean of the concomitant variable $\bar{X}_2$ and the total X_2 are assumed known for the nonrespondents. The proposed estimator is:

$$(2.7.1) \qquad \bar{Y}_R = W_1\ \bar{Y}_1 + W_2\ \bar{Y}_{2R} = W_1\ \bar{Y}_1 + W_2\ R_2\ \bar{X}_2$$

where W_1 and W_2 are fixed weights different than $\frac{N_1}{N}$ and $\frac{N_2}{N}$, but such that $W_1 + W_2 = 1$, and $N_2 = N - N_1$.

The bias of the proposed estimator is:

$$(2.7.2) \qquad E\ (\bar{Y}_R - \bar{Y}) = (W_2 - \frac{N_2}{N})\ (\bar{Y}_{2R} - \bar{Y}_1)$$

which will be zero if $\frac{N_2}{N} = W_2$ or $\bar{Y}_{2R} = \bar{Y}_1$. The variance of the estimator, using the usual approximation for a ratio, is:

$$(2.7.3) \qquad V\ (\bar{Y}_R) = W_2^2\ (1 - \frac{n_2}{N_2})\ \frac{1}{n_2}\ [V\ (Y_2) + V\ (X_2) - 2R\ \text{Cov}(Y_2,X_2)]$$

The mean square error based on (2.7.2) and (2.7.3) is:

$$(2.7.4) \qquad \text{M.S.E.}\ (\bar{Y}_R) = (2.7.3) + (2.7.2)^2\ .$$

2.8 Optimum Weight Using Ratio Estimator

The value of W_2 which will minimize the mean square error (2.7.4) is desired. Setting the derivative of (2.7.4) with respect to W_2 equal to zero, we obtain the optimum value.

(2.8.1) $$f' \text{ (M.S.E.)} = 2W_2 \left(1 - \frac{n_2}{N_2}\right) \frac{1}{n_2} [V(Y_2) + V(X_2) - 2R \text{ Cov}(Y_2,X_2)]$$
$$+ 2 \left(W - \frac{N_2}{N}\right) (\bar{Y}_{2R} - \bar{Y}_1)^2 = 0$$

(2.8.2) $$W_2^* = \frac{\frac{N_2}{N}(\bar{Y}_{2R} - \bar{Y}_1)^2}{\left(1 - \frac{n_2}{N_2}\right) \frac{1}{n_2} [V(Y_2) + V(X_2) - 2R \text{ Cov}(Y_2,X_2)] + (\bar{Y}_{2R} - \bar{Y}_1)^2}$$

letting $V\ (R) = \left(1 - \frac{n_2}{N_2}\right) \frac{1}{n_2} [V(Y_2) + V(X_2) - 2R \text{ Cov}(Y_2,X_2)]$

(2.8.3) $$W_2^* = \frac{\frac{N_2}{N}\ (\bar{Y}_{2R} - \bar{Y}_1)^2}{V\ (R) + (\bar{Y}_{2R} - \bar{Y}_1)^2}$$

If $T = \frac{(\bar{Y}_{2R} - \bar{Y}_1)}{\sqrt{V\ (R)}}$

$$W_2^* = \frac{\frac{N_2}{N}\ T^2}{1 + T_2} = \frac{\frac{N_2}{N}}{1 + \frac{1}{T^2}}$$

which is similar to the results in Section 2.5 except the quality T is based on a ratio estimate for the nonrespondent mean and the usual approximation for the variance of a ratio.

2.9 Sample Estimators for Ratio Method

The sample estimators of the mean, bias, W_2, variance and mean square error are obtained in a manner analogous to Sections (2.5) and (2.6).

The sample estimators are:

Mean

$$(2.9.1) \qquad \bar{y}_R = \bar{Y}_1 + \hat{W}_2^* \left(\frac{\bar{y}_2}{\bar{x}_2} \bar{X}_2 - \bar{Y}_1\right) = \bar{Y}_1 + \hat{W}_2^* (\bar{y}_{2R} - \bar{Y}_1)$$

Bias

$$(2.9.2) \qquad b = \left(\hat{W}_2^* - \frac{N_2}{N}\right)(\bar{y}_{2R} - \bar{Y}_1)$$

Weight

$$(2.9.3) \qquad \hat{W}_2^* = \frac{\frac{N_2}{N}(\bar{y}_{2R} - \bar{Y}_1)^2}{v(\hat{R}) + (\bar{y}_{2R} - \bar{Y}_1)^2}$$

$$\text{where } v(\hat{R}) = \left(1 - \frac{n_2}{N}\right) \frac{1}{n_2} \left[s_{Y2}^2 + s_{X2}^2 - 2 \frac{\bar{y}_2}{\bar{x}_2} \text{Cov}(y_2, x_2)\right]$$

The sample estimates of the variance and mean square error can be derived using Table 1, in Section 2.6 with formulas (2.6.6) and (2.6.7')

where $\frac{s_2^2}{n_2}$ and $\bar{y}_m$ are replaced by $v(\hat{R})$ and $\bar{y}_R$ respectively. These

expressions for the variance and mean square error are adjusted by

$$+ \left(\frac{N_2}{N}\right)^2 \frac{1}{n_2} \left[v(x_2) - 2\hat{R}\, \text{Cov}(y_2, x_2)\right]$$

to allow for the difference in the variance of the unbiased estimator and the ratio estimator.

Appendix 1

Class Marks And Cell Frequencies Used For N(0,1)

Class Marks	Cell Frequencies
-3.25	.00135
-2.75	.00486
-2.25	.01654
-1.75	.04406
-1.25	.09189
- .75	.14980
- .25	.19150
.25	.19150
.75	.14980
1.25	.09189
1.75	.04406
2.25	.01654
2.75	.00486
3.25	.00135

CHAPTER 3 - TWO-PHASE SELECTION OF SAMPLE UNITS

3.1 Simple Random Sample of Frame Units

The methods developed in this chapter are for respondent and nonrespondent "domains." The number of respondents and nonrespondents are not known for the population, but only for the particular sample of n units selected in the first phase. Since totals are not to be estimated by "domains," but only for the population, post-stratification theory is appropriate for the mean and total of the population. The population size, N, is known for the frame.

3.2 The Classical Unbiased Estimator and Variance

The classical double sampling procedure considers a complete list of size N from which a sample of size n is selected using simple random sampling. A survey (possibly conducted by mail) of n units yields n_1 responses leaving n_2 units which are labeled nonrespondents. The n_1 and n_2 are random samples from populations with unknown sizes N_1 and N_2 . A random sample of $k = fn_2$ nonrespondents is selected and information is obtained by a more expensive data collection method (usually personal interview). The value of f (or k), the fraction of nonrespondents sampled, is determined in advance of the survey and assumed constant in the subsequent development.

The estimator of the mean of the population is:

(3.2.1) $$\bar{Y} = \frac{N_1}{N}\bar{Y}_1 + \frac{N_2}{N}\bar{Y}_2$$

The sample estimate of the mean is:

(3.2.2) $$\bar{y} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}_2 \qquad \text{where } n_1 + n_2 = n \quad .$$

The sample estimate of the population total is:

(3.2.3) $$\hat{Y} = \frac{N}{n}\left(\sum_{i=1}^{n_1} y_{1i} + \frac{n_2}{k}\sum_{j=1}^{k} y_{2j}\right)$$

The variance of the estimator for (3.2.1)

(3.2.4) $$V\,(\bar{Y}) = \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n} + \left(\frac{n_2}{k} - 1\right)\frac{N_2}{N}\frac{\sigma_2^2}{n}$$

and the sample estimate of the variance of (3.2.3)

$$(3.2.5) \qquad v\,(\hat{y}) = N^2(\frac{N-n}{N})\,\frac{S^2}{n} + N^2(\frac{n_2}{k} - 1)\,\frac{n_2}{n}\,S_2^2$$

where S^2 is the variance corresponding to the population from which the n units were selected, and S_2^2 is the variance of the nonrespondents.

3.3 Simple Minimum Mean Square Estimator

It is proposed that a biased estimator identical to that developed in Section 2.3 be used for the simple random sample of n units selected at the first stage from the N units in the frame. The following diagram shows the strata for the frame.

	Respondents	Nonrespondents	Total
Population Sizes	N_1	N_2	N
Population Means	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}$
Population Variances	σ_1^2	σ_2^2	σ^2
1st Phase Sample Size	n_1	n_2	n
1st Phase Means	$\bar{y}_1$	$\bar{y}_2$	$\bar{y}$
2nd Phase Sample Size		$k = fn_2$	k
2nd Phase Mean		$\bar{y}_{2k}$	
Sample Variances	s_1^2	s_2^2	s^2

It shall be assumed that Prob $(n_1 < 2, k < 2)$ is small so it is reasonable to consider $n_1 \geq 2$, $n_2 \geq 2$, and $k = \max\,(2,\ fn_2)$.

The proposed estimator for the population mean is:

$$(3.3.1) \qquad \bar{Y}_M = W_1\,\bar{Y}_1 + W_2\,\bar{Y}_2 \quad \text{where } W_1 + W_2 = 1\ .$$

The bias of the estimator is: $\bar{Y} - \bar{Y}_M$

$$(3.3.2) \qquad B = (W_2 - \frac{N_2}{N})\,(\bar{Y}_2 - \bar{Y}_1)$$

The estimator of the population total is:

$$(3.3.3) \qquad Y_M = N \, (W_1 \, \bar{Y}_1 + W_2 \, \bar{Y}_2) = N \, \{\bar{Y}_1 + W_2 \, (\bar{Y}_2 - \bar{Y}_1)\}$$

The variance of estimator of the mean

$$(3.3.4) \qquad V \, (\bar{Y}_M | \, n_1 \geq 2, \, k \geq 2) = W_1^2 \, \frac{\sigma_1^2}{n_1} \, (1 - \frac{n_1}{N_1}) + W_2^2 \, \frac{\sigma_2^2}{k} \, (1 - \frac{k}{N_2})$$

The expectation over all values of n_1 and k

$$E \, (\frac{1}{n_1}) \doteq \frac{1}{n \, \frac{N_1}{N}} \quad \text{and} \quad E \, (\frac{1}{k}) \doteq \frac{1}{f \, n \, \frac{N_2}{N}} \; .$$

Therefore the variance of $\bar{Y}_M$ becomes

$$(3.3.5) \qquad V \, (\bar{Y}_M) \doteq W_1^2 \, \sigma_1^2 \, (\frac{N}{n \, N_1} - \frac{1}{N_1}) + W_2^2 \, \sigma_2^2 \, (\frac{N}{f \, n \, N_2} - \frac{1}{N_2})$$

And the mean square error of (3.3.1), neglecting the finite population correction factors is:

$$(3.3.6) \qquad M.S.E. = \frac{N}{n} \, [(1 - W_2)^2 \, \frac{\sigma_1^2}{N_1} + W_2^2 \, \frac{\sigma_2^2}{f \, N_2}] + (W_2 - \frac{N_2}{N})^2 \, (\bar{Y}_2 - \bar{Y}_1)^2$$

3.4 The Optimum Weights for Domain Means

The value of W_2 which will minimize the mean square error of (3.3.6) is desired. The derivative of (3.3.6) with respect to W_2 set equal to zero is:

$$(3.4.1) \qquad f'(M.S.E.) = -2 \, (1 - W_2) \, \sigma_1^2 \, \frac{N}{n \, N_1} + 2 \, W_2 \, \sigma_2^2 \, \frac{N}{f \, n \, N_2}$$
$$+ (W_2 - \frac{N_2}{N}) \, (\bar{Y}_2 - \bar{Y}_1)^2 = 0$$

The optimum value of W_2 found by solving (3.4.1) is:

$$(3.4.2) \qquad W_2^* = \frac{\frac{\sigma_1^2}{N_1} + \frac{n \, N_2}{N^2} \, (\bar{Y}_2 - \bar{Y}_1)^2}{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{f \, N_2} + \frac{n}{N}(\bar{Y}_2 - \bar{Y}_1)^2} = \frac{\frac{N}{n} \, \frac{\sigma_1^2}{N_1} + \frac{N_2}{N} \, (\bar{Y}_2 - \bar{Y}_1)^2}{\frac{N}{n} \, (\frac{\sigma_1^2}{N_1} + \frac{\sigma_1^2}{f \, N_2}) + (\bar{Y}_2 - \bar{Y}_1)^2}$$

If we let

$$T = \frac{(\bar{Y}_2 - \bar{Y}_1)}{\sqrt{\frac{N}{n} (\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{f\ N_2})}}$$

and

$$C = \frac{\frac{N}{n\ N_1} \sigma_1^2}{\frac{N}{n} (\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{f\ N_2})}$$

then (3.4.2) becomes

(3.4.3) $$W_2^* = \frac{C + \frac{N_2}{N} T^2}{1 + T^2} \qquad \text{where } o \leq c \leq 1$$

and W_2^* will be approximately equal to $\frac{N_2}{N}$ for large values of T.

3.5 Sample Estimators

It is proposed that the population parameters in the estimators for the mean, bias, and weights be replaced by sample estimates. The justification of the resulting estimators will be sought by a study of the mean square errors. The estimators proposed are:

The mean

(3.5.1) $$\bar{y}_m = (1 - \hat{W}_2^*) \bar{y}_1 + \hat{W}_2^* \bar{y}_{2k} = \bar{y}_1 + \hat{W}_2^* (\bar{y}_{2k} - \bar{y}_1)$$

The bias squared

(3.5.2) $$b^2 = (\hat{W}_2^* - \frac{n_2}{n})^2 (\bar{y}_{2k} - \bar{y}_1)^2$$

The weight

$$(3.5.3) \qquad \hat{W}_2^* = \frac{c + \frac{n_2}{n} t^2}{1 + t^2}$$

where $$c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{k}}$$

and $$t^2 = \frac{(\bar{y}_{2k} - \bar{y}_1)^2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{k}}$$

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2}{n_1 - 1} \quad , \qquad s_2^2 = \frac{\sum_{i=1}^{k} (y_{2i} - \bar{y}_{2k})^2}{k - 1}$$

y_{1i} = the i^{th} respondent

y_{2i} = the i^{th} nonrespondent interviewed

The sample variance and mean square error are to be obtained using the tables given in the next section and the sample statistics.

3.6 Variance and Mean Square

The means of the respondents and nonrespondents, $\bar{y}_1$ and $\bar{y}_{2k}$, will be approximately normally distributed if the sample sizes are moderately large. Since n_1 and n_2 represent random samples from their respective populations, the two means are independent with a bivariate normal distribution. It is proposed to study the characteristics of the estimator (3.3.1) through the bivariate normal distribution of $\bar{y}_{2k}$ and $\bar{y}_1$. It will be assumed without loss of generality that $\bar{y}_{2k} = 0$. To simplify the study and insure that the error of the estimated population variance is negligible, equal variances within "domains" will be assumed with the sample variances being pooled in estimating $\hat{\sigma}^2$.

Hence, the means have the following marginal distributions

(3.6.1) $\bar{y}_{2k} \sim N\ (0, \frac{\sigma^2}{fnP})$

(3.6.2) $\bar{y}_1 \sim N\ (\Delta(1-P\ \frac{\sigma^2}{n(1-P)})$

where $P = \frac{N_2}{N}$, $\Delta = \bar{Y}_1 \div \frac{\sigma^2}{n} (\frac{1}{fP} + \frac{1}{(1-P)})$, and f is the fraction of nonrespondents interviewed.

The quality T will be distributed as:

(3.6.3) $T \sim N\ [-\ \Delta\ (1-P), \frac{\sigma^2}{n} (\frac{1}{fP} + \frac{1}{1-P})]$

and when $\frac{\sigma^2}{n} = 1$

(3.6.3') $T \sim N\ [-\ \Delta\ (1-P), (\frac{1}{fP} + \frac{1}{1-P})]$

and the value for C in (3.5.3) is the constant $C_o = \frac{fp}{fP + (1-P)}$ for fixed f and P.

If the difference between $\bar{Y}_2$ and $\bar{Y}_1$ is fixed, the $\hat{W}_2^*$ can be calculated with the variance and mean square error will be functions of the sample values for $\bar{y}_{2k}$ and $\bar{y}_1$.

(3.6.4) $V\ (\bar{Y}_M) = V\ [\bar{y}_1 + (\frac{C + T^2}{1 + T^2})\ (\bar{y}_{2k} - \bar{y}_1)] = V\ [\Phi\ (\bar{y}_1, \bar{y}_{2k})]$

(3.6.5) $M.S.E.\ (\bar{Y}_M) = V\ [\Phi\ (\bar{y}_1, \bar{y}_{2k})] + [E\ \Phi\ (\bar{y}_1, \bar{y}_{2k}) - \bar{Y}]^2$

where $\bar{Y} = (1-P)\ \bar{Y} + P\ (\bar{Y}_2)$.

The variance of the classical unbiased estimator is:

(3.6.6) $V\ (\bar{Y}) = (1-P) + \frac{P}{f}$

It is proposed to evaluate (3.6.4) and (3.6.5) for comparison to (3.6.6) by numerical integration over the bivariate normal distribution of $\bar{y}_1$ and $\bar{y}_{2k}$.

The variance and mean square error of $\bar{y}_m$ were determined by numerical integration for selected values of P, f and Δ. A bivariate normal distribution with a total of 196 cells was used to determine 160 distributions for which values of V $(\bar{y}_m)$ and M.S.E. $(\bar{y}_m)$ are shown in Tables 3 and 4. Table 4 indicates that gains in efficiency are to be expected for small values of f and Δ. The largest gain is to be realized for situation in which the response rate is near .50. When the sampling rate of the nonrespondents is .50 or greater, the classical estimator should be used. The tables also bear out the fact that the variance plays a dominate role in the mean square error. However, the bias term contributes relatively less to the mean square errors in Table 4 than for the corresponding value of β = 1 in Table 2. No general evaluation of alternate forms of (3.5.1) based on raising the terms T^2 anu $(1 + T^2)$ to various powers of β was undertaken as was done in Tables 1 and 2. The presence of the term C in (3.5.3) which begins to dominate as P and f become large relative to the role of T for small values which makes the outcome of such an evaluation dependent upon a second condition. The prior knowledge of both variances by individual domains and Δ seem unrealistic for general application.

For the most favorable situation, f = .05 and p = .5, the gains in efficiency are somewhat less than those in Table 2, but the loss in efficiency for Δ = 3 are less. The values in Tables 3 and 4 provide a basis for calculating the sample variance and mean square error, that is:

$$(3.6.7) \qquad v\ (\bar{y}_m) = [(\frac{N-n}{N})\ \frac{S^2}{n} + (\frac{n_2}{k} - 1)\ \frac{n_2}{n}\ s_2^2]\ V\ (p,f,\Delta)$$

$$(3.6.8) \qquad \text{M.S.E.}\ (\bar{y}_m) = [(\frac{N-n}{N})\ \frac{S^2}{n} + (\frac{n_2}{k} - 1)\ \frac{n_2}{n}\ s_2^2]\ M\ (p,f,\Delta)$$

The values for the 196 cells for the bivariate normal from which the variances and mean square errors in Table 4 were derived from the normal (0,1) distribution with same marginal distribution class marks used in Tables 1 and 2 of Chapter 2. The 196 cell frequencies (P_{ij}) are the product of the marginal cell frequencies.

The marginal values corresponding to the distribution of the two means are:

for $\bar{Y}_2 : X_{i}\cdot = z_i\sqrt{\frac{1.0}{f\ P}}$;

for $\bar{Y}_1 : X\cdot_{j} = z_i\sqrt{\frac{1.0}{(1-P)}} + \Delta\sqrt{\frac{(1-P)\ +\ fP}{fP\ (1-P)}}$

The 196 cell values for the bivariate normal distributions were

$$X_{ij} = X_{i}\cdot - X\cdot_{j}$$

The T values were calculated for each cell

$$T_{ij} = X_{ij} \div \sqrt{\frac{(1-P)\ +\ fP}{fP\ (1-P)}}$$

The variable to be studied y_{ij} corresponding the estimator was

$$y_{ij} = X\cdot_{j} + (X_{ij})\ [\ \frac{\frac{fP}{fP\ +\ (1-P)} + P\ T_{ij}^2}{1 + T_{ij}^2}\ ]$$

where

$$V\ (\bar{y}_m) = \sum^{196} P_{ij}\ y_{ij}^2 - (\ \sum^{196} P_{ij}\ y_{ij})^2$$

$$B^2(y_m) = [\ \sum^{196} P_{ij}\ y_{ij} - \Delta(1-P)\sqrt{\frac{(1-P)\ +\ fP}{fP\ +\ (1-P)}}\ ]^2$$

$$\text{M.S.E.}\ (\bar{y}_m) = V\ (\bar{y}_m) + B^2(\bar{y}_m)$$

The values given by the relationships lead to symmetry about P = .5; observe the tabled values for P = .30 and .70.

Table 3--Ratio of Variances, $V(\bar{Y}_M) \div V(\bar{Y})$, for Double Sampling ($\beta = 1$)

Non-response Rate	Sampling Fraction of N.R.	Δ							
		0	.25	.50	.75	1.00	1.50	2.00	3.00
P = .1	f = .05	.6853	.6958	.7261	.7727	.8301	.9520	1.0475	1.1119
	f = .10	.7916	.7988	.8195	.8512	.8903	.9783	1.0384	1.0823
	f = .30	.9501	.9523	.9585	.9682	.9801	1.0054	1.0252	1.0386
	f = .50	.9957	.9965	.9986	1.0018	1.0058	1.0143	1.0210	1.0255
	f = .70	1.0126	1.0128	1.0134	1.0143	1.0153	1.0176	1.0194	1.0206
P = .3	f = .05	.5924	.6058	.6446	.7041	.7775	.9333	1.0554	1.1378
	f = .10	.6794	.6901	.7210	.7684	.8268	.9508	1.0480	1.1136
	f = .30	.8812	.8855	.8981	.9173	.9410	.9913	1.0308	1.0575
	f = .50	.9677	.9693	.9740	.9811	.9899	1.0087	1.0234	1.0333
	f = .70	1.0044	1.0048	1.0061	1.0081	1.0105	1.0157	1.0198	1.0226
P = .5	f = .05	.5776	.5916	.6317	.6933	.7692	.9303	1.0567	1.1420
	f = .10	.6580	.6694	.7022	.7526	.8147	.9465	1.0498	1.1196
	f = .30	.8625	.8674	.8816	.9034	.9303	.9874	1.0322	1.0625
	f = .50	.9591	.9609	.9664	.9747	.9850	1.0069	1.0241	1.0357
	f = .70	1.0024	1.0029	1.0045	1.0068	1.0097	1.0158	1.0207	1.0239
P = .7	f = .05	.5924	.6058	.6446	.7041	.7775	.9333	1.0554	1.1379
	f = .10	.6794	.6901	.7210	.7684	.8268	.9508	1.0480	1.1137
	f = .30	.8814	.8858	.8983	.9175	.9412	.9915	1.0309	1.0575
	f = .50	.9678	.9694	.9740	.9812	.9900	1.0087	1.0234	1.0332
	f = .70	1.0048	1.0053	1.0066	1.0085	1.0110	1.0162	1.0202	1.0229

Table 4--Ratio of Mean Square Error to Variance, MSE ($\bar{Y}_M$) ÷ V(Y), for Double Sampling (β = 1)

Non-response Rate	Sampling Fraction of N.R.	Δ							
		0	.25	.50	.75	1.00	1.50	2.00	3.00
P = .1	f = .05	.6853	.7002	.7425	.8058	.8808	1.0283	1.1295	1.1720
	f = .10	.7916	.8018	.8306	.8737	.9248	1.0253	1.0943	1.1232
	f = .30	.9501	.9532	.9619	.9751	.9906	1.0212	1.0421	1.0510
	f = .50	.9957	.9968	.9997	1.0041	1.0093	1.1014	1.0267	1.0297
	f = .70	1.0126	1.0129	1.0137	1.0149	1.0163	1.0190	1.0209	1.0217
P = .3	f = .05	.5924	.6114	.6656	.7464	.8423	1.0309	1.1603	1.2146
	f = .10	.6794	.6946	.7377	.8020	.8783	1.0284	1.1314	1.1747
	f = .30	.8812	.8874	.9049	.9309	.9619	1.0228	1.0646	1.0822
	f = .50	.9677	.9700	.9765	.9862	.9977	1.0204	1.0360	1.0426
	f = .70	1.0044	1.0050	1.0068	1.0095	1.0127	1.0190	1.0233	1.0251
P = .5	f = .05	.5776	.5974	.6534	.7370	.8362	1.0313	1.1652	1.2214
	f = .10	.6580	.6742	.7200	.7884	.8695	1.0290	1.1386	1.1846
	f = .30	.8625	.8695	.8893	.9189	.9540	1.0232	1.0706	1.0906
	f = .50	.9591	.9617	.9693	.9807	.9941	1.0206	1.0389	1.0465
	f = .70	1.0024	1.0032	1.0053	1.0085	1.0123	1.0197	1.0248	1.0269
P = .7	f = .05	.5924	.6114	.6656	.7464	.8423	1.0309	1.1603	1.2147
	f = .10	.6794	.6946	.7377	.8020	.8783	1.0284	1.1315	1.1748
	f = .30	.8814	.8876	.9050	.9311	.9621	1.0230	1.0648	1.0823
	f = .50	.9678	.9701	.9766	.9863	.9978	1.0204	1.0360	1.0424
	f = .70	1.0048	1.0055	1.0073	1.0099	1.0131	1.0194	1.0237	1.0254

3.7 Ratio Estimator for Random Subsample of Units

Since the original listing of the N frame units was subsampled yielding n units for the initial survey contact, the total number of respondents, N_1, and nonrespondents, N_2, are not known.

Since separate ratio estimates are not needed by "domains" but only for the population, a ratio estimator is considered for which a concomitant variable X is known only for the total population. Therefore, the "combined" ratio estimator is proposed. The estimators for the population mean and total are:

$$(3.7.1) \qquad \bar{Y}_R = \frac{W_1 \bar{Y}_1 + W_2 \bar{Y}_2}{W_1 \bar{X}_1 + W_2 \bar{X}_2} \bar{X} = \frac{\bar{Y}_M}{\bar{X}_M} \bar{X} = R_M \bar{X}$$

$$(3.7.2) \qquad Y_R = R_M X$$

where $\bar{Y}_1$ = population mean for the N_1 respondents

$\bar{Y}_2$ = population mean for the N_2 nonrespondents

$\bar{X}_1$ = population mean for the N_1 respondents

$\bar{X}_2$ = population mean for the N_2 nonrespondents

$\bar{X}$ = population mean for the N units

X = population total for the N units

$\bar{Y}$ = population mean for the N units .

The bias of the proposed estimator of the mean

$$(3.7.3) \qquad \bar{Y}_R - \bar{Y} = \frac{\bar{Y}_M}{\bar{X}_M} \bar{X} - \bar{Y} = \frac{\bar{Y}_M \bar{X} - \bar{X}_M \bar{Y}}{\bar{X}_M}$$

Now $\bar{Y}_M = \bar{Y}_1 + (W_2 - \frac{N_2}{N}) (\bar{Y}_2 - \bar{Y}_1)$

$$\bar{X}_M = \bar{X}_1 + (W_2 - \frac{N_2}{N}) (\bar{X}_2 - \bar{X}_1)$$

Making these substitutions in (3.7.3) we obtain

$$(3.7.3') \qquad \bar{Y}_R - \bar{Y} = \frac{(W_2 - \frac{N_2}{N})[\bar{X}\ (\bar{Y}_2 - \bar{Y}_1) - \bar{Y}\ (\bar{X}_2 - \bar{X}_1)]}{\bar{X}_1 + (W_2 - \frac{N_2}{N})\ (\bar{X}_2 - \bar{X}_1)}$$

The bias will be zero if either

(1) $\quad W_2 = \frac{N_2}{N}$

or (2) $\quad \bar{X}\ (\bar{Y}_2 - \bar{Y}_1) = \bar{Y}\ (\bar{X}_2 - \bar{X}_1)\ ; \quad \frac{\bar{Y}}{\bar{X}} = \frac{(\bar{Y}_2 - \bar{Y}_1)}{(\bar{X}_2 - \bar{X}_1)}$

Since the variables Y and X are positively correlated (usual assumption for ratio estimate to be used efficiently) hence, the means of the nonrespondents $(\bar{Y}_2, \bar{X}_2)$ must both be greater (or less) than the means of the respondents $(\bar{Y}_1, \bar{X}_1)$ which implies $\bar{Y} : \bar{X}$ is positive if this condition is to be satisfied.

The variance of the estimator (3.7.1)

$$(3.7.4) \qquad V\ (\bar{Y}_R) = (1 - \frac{n}{N})[V\ (\bar{Y}_M) + V\ (\bar{X}_M) - 2R\ \text{Cov}(\bar{Y}_M, \bar{X}_M)]$$

where

$$V\ (\bar{Y}_M) = W_1^2\ \sigma_{y_1}^2\ (\frac{N}{n\ N_1} - \frac{1}{N_1}) + W_2^2\ \sigma_{y_2}^2\ (\frac{N}{f\ n\ N_2} - \frac{1}{N_2})$$

$$V\ (\bar{X}_M) = W_1^2\ \sigma_{x_1}^2\ (\frac{N}{n\ N_1} - \frac{1}{N_1}) + W_2^2\ \sigma_{x_2}^2\ (\frac{N}{f\ n\ N_2} - \frac{1}{N_2})$$

and

$$\text{Cov}(\bar{Y}_M, \bar{X}_M) = W_1^2\ \text{Cov}(\bar{Y}_1, \bar{X}_1) + W_2^2\ \text{Cov}(\bar{Y}_2, \bar{X}_2)$$

since the means of the respondents are independent of the nonrespondents means.

The mean square of (3.7.1) is:

$$(3.7.5) \qquad \text{M.S.E.}\ (\bar{Y}_R) = (3.7.4) + (3.7.3')^2$$

3.8 Optimum Weight Using Ratio Estimator

The value of W_2 which will minimize (3.7.5) is desired. Since the numerator and denominator of (3.7.1) may be written as the mean of the population plus the bias, we have

$$(3.8.1) \qquad \bar{Y}_R = \frac{\bar{Y} + (W_2 - \frac{N_2}{N})(\bar{Y}_2 - \bar{Y}_1)}{\bar{X} + (W_2 - \frac{N_2}{N})(\bar{X}_2 - \bar{X}_1)} \bar{X}$$

If the numerator and the denominator of the R.H.S. (3.8.1) are divided by $\bar{X}$, and the approximate bias of $\bar{Y}_R$ derived by taking the expectation of the Taylor Expansion in terms of the denominator of the R.H.S. If only the first two terms in the series are retained, the bias is:

$$(3.8.2) \qquad \text{Approx. Bias} = (W_2 - \frac{N_2}{N})[(\bar{Y}_2 - \bar{Y}_1) - (\bar{X}_2 - \bar{X}_1)\frac{\bar{Y}}{\bar{X}}]$$

rather than (3.7.3') which is valid if $(W_2 - \frac{N_2}{N})(\frac{\bar{X}_2 - \bar{X}_1}{\bar{X}}) < 1.$

It is proposed to use this estimate in place of (3.7.3'). Setting the derivative of (3.7.5) with respect to W_2 equal to zero

$$\begin{aligned}(3.8.3) \qquad f'\,(\text{M.S.E.}) = &- 2\,(1-W_2)\,\sigma^2_{y_1}(\frac{N}{n\,N_1} - \frac{1}{N_1}) + 2W_2\,\sigma^2_{y_2}(\frac{N}{f\,n\,N_2} - \frac{1}{N_2}) \\ &- 2\,(1-W_2)\,\sigma^2_{x_1}(\frac{N}{n\,N_1} - \frac{1}{N_1}) + 2W_2\,\sigma^2_{x_2}(\frac{N}{f\,n\,N_2} - \frac{1}{N_2}) \\ &+ 4\,(1-W_2)\,R\,\mathrm{Cov}(\bar{Y}_1,\bar{X}_1) - 4W_2\,R\,\mathrm{Cov}(\bar{Y}_2,\bar{X}_2) \\ &+ 2\,(W_2 - \frac{N_2}{N})[(\bar{Y}_2 - \bar{Y}_1) - (\bar{X}_2 - \bar{X}_1)\frac{\bar{Y}}{\bar{X}}]^2\end{aligned}$$

where $R = \frac{\bar{Y}}{\bar{X}}$, and $\sigma^2_{x_1}$, $\sigma^2_{y_1}$, $\sigma^2_{x_2}$, $\sigma^2_{y_2}$ are the variances of the variables x,y for the two domains.

Collecting terms involving W_2, the following expression is obtained for the optimum value of W_2:

$$(3.8.4) \qquad W_2^* = \frac{\sigma^2_{R_1} + \frac{N_2}{N}\,[(\bar{Y}_2 - \bar{Y}_1) - (\bar{X}_2 - \bar{X}_1)\,\frac{\bar{Y}}{\bar{X}}\,]^2}{\sigma^2_{R_1} + \sigma^2_{R_2}[(\bar{Y}_2 - \bar{Y}_1) - (\bar{X}_2 - \bar{X}_1)\,\frac{\bar{Y}}{\bar{X}}\,]^2}$$

where

$$\sigma^2_{R_1} = \sigma^2_{y_1}\,(\frac{N}{n\,N_1} - \frac{1}{N_1}) + \sigma^2_{x_1}\,(\frac{N}{n\,N_1} - \frac{1}{N_1}) - 2R_1\ \mathrm{Cov}(\bar{Y}_1, \bar{X}_1)$$

$$\sigma^2_{R_2} = \sigma^2_{y_2}\,(\frac{N}{f\,n\,N_2} - \frac{1}{N_2}) + \sigma^2_{x_2}\,(\frac{N}{f\,n\,N_2} - \frac{1}{N_2}) - 2R_2\ \mathrm{Cov}(\bar{Y}_2, \bar{X}_2)$$

in order to simplify the notation in (3.8.4).

3.9 Sample Estimators for Combined Ratio

The sample estimators for the mean, bias, W_2, variance and mean square error are obtained in a manner analogous to Sections (2.5) and (2.6). The sample estimators are:

Mean

$$(3.9.1) \qquad \bar{y}_r = \frac{\bar{y}_m}{\bar{x}_m}\ \bar{X} = r_m\ \bar{X}$$

Bias

$$(3.9.2) \qquad b = \frac{(\hat{W}_2^* - \frac{n_2}{n})[\bar{x}\ (\bar{y}_2 - \bar{y}_1) - \bar{y}\ (\bar{x}_2 - \bar{x}_1)]}{\bar{x}_1 + (\hat{W}_2^* - \frac{n_2}{n})\,(\bar{x}_2 - \bar{x}_1)}$$

The sample estimate of W_2^* is:

$$(3.9.3) \qquad \hat{W}_2^* = \frac{s_{R_1}^2 + \frac{n_2}{n}[(\bar{y}_2 - \bar{y}_1) - (\bar{x}_2 - \bar{x}_1)\frac{\bar{y}}{\bar{x}}]^2}{s_{R_1}^2 + s_{R_2}^2 + [(\bar{y}_2 - \bar{y}_1) - (\bar{x}_2 - \bar{x}_1)\frac{\bar{y}}{\bar{x}}]^2}$$

where $s_{R_1}^2$ and $s_{R_2}^2$ are based on domain variances and covariances.

The sample estimator of the variance is obtained by using the values given in Table 3 based on sample estimates of $\hat{p}$, $\hat{\Delta}$ and the classical variance given by (3.2.5) which is then adjusted by:

$$+ (\frac{N-n}{N}) [v (\bar{X}_M) - 2\hat{R}_M \text{Cov}(\bar{Y}_M, \bar{X}_M)]$$

The value of $\dot{f}$ is assumed fixed in advance for the survey. The mean square error is obtained by adding the bias squared, given by (3.9.2), to the variance.

CHAPTER 4 - STRATIFIED SAMPLE FROM LIST FRAME

4.1 Simple Random Sampling from All Strata

In the previous chapters, a frame of addresses was considered where either (1) all N addresses were initially contacted resulting in a group of nonrespondents and a simple random sample of n_2 of the N_2 nonrespondents were interviewed, or (2) a simple random sample of n of the N addresses was selected and contacted but for some units no response was obtained; hence, k of the n_2 nonrespondents were interviewed. In the present chapter a frame of addresses will be considered in which the N addresses may be stratified into L strata such that $N_1 + N_2 + \ldots. + N_L = N$.

4.2 Estimators for 100 Percent Sampling of List in Each Stratum

In Chapter 2 a simple minimum mean square estimator was proposed based on weighted means of two response strata.

(4.2.1) $$\bar{Y}_M = W_1 \bar{Y}_1 + W_2 \bar{Y}_2$$

where $W_1 + W_2 = 1$

and $$W_2 = \frac{\frac{N_2}{N} (\bar{Y}_2 - \bar{Y}_1)^2}{(1 - \frac{n_2}{N_2}) \frac{\sigma_2^2}{n_2} + (\bar{Y}_2 - \bar{Y}_1)^2} = \frac{\frac{N_2}{N}}{1 + \frac{1}{T^2}}$$

For a stratified list with nonrespondents sampled in each stratum, the following minimum mean square estimator is proposed:

(4.2.2) $$\bar{Y}_{MS} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{Y}_{Mh}$$

where $N_1 + N_2 + \ldots. + N_h = N$

and

(4.2.3) $$\bar{Y}_{Mh} = W_{1h} \bar{Y}_{1h} + W_{2h} \bar{Y}_{2h}$$

Two cases are considered:

I. W_{2h} is determined independently for each stratum to minimize the mean square error of (4.2.3).

II. W_{2h} is determined simultaneously across all strata to minimize the mean square error of (4.2.4).

For Case I we may appeal directly to the results of Chapter 2 and for W_{2h} while for Case II we seek a joint solution involving all h strata.

A. Case I:

The estimator in (4.2.3) is not equal to the usual unbiased estimator $\bar{Y}_h$. The bias of Y_{Mh} is:

$$(4.2.4) \qquad \bar{Y}_{Mh} - \bar{Y}_h = (W_{2h} - \frac{N_{2h}}{N_h}) (\bar{Y}_{2h} - \bar{Y}_{1h})$$

writing (4.2.4) as the squared bias and summing over all strata we obtain the bias squared of (4.2.2).

$$(4.2.4') \qquad (\text{Bias})^2 = \sum_{h=1}^{L} [\frac{N_h}{N} (W_{2h} - \frac{N_{2h}}{N_h}) (\bar{Y}_{2h} - \bar{Y}_{1h})]^2$$

The sample estimators of (4.2.2) and (4.2.4') would be

$$(4.2.5) \qquad \bar{y}_{Ms} = \sum_{h=1}^{L} \frac{N_h}{N} [(1 - W_{2h}) \bar{Y}_{1h} + W_{2h} \bar{y}_{2h}]$$

$$= \sum_{h=1}^{L} \frac{N_h}{N} [\bar{Y}_{1h} + W_{2h} - Y_{1h})]$$

$$(4.2.6) \qquad b^2 = \sum_{h=1}^{L} [\frac{N_h}{N} (W_{2h} - \frac{N_{2h}}{N_h}) (\bar{y}_{2h} - \bar{Y}_{1h})]^2$$

The population variance based on the results of Chapter 2 would be:

$$(4.2.7) \qquad V\,(\bar{Y}_{MS}) = \sum_{h=1}^{L} W_{2h}^{2}\,\left(1 - \frac{n_{2h}}{N_{2h}}\right) \frac{\sigma_{2h}^{2}}{n_{2h}}$$

The mean square error of (4.2.2) is obtained from (4.2.4) and (4.2.7)

$$(4.2.8) \qquad MSE\,(\bar{Y}_{MS}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 W_{2h}^{2} \left(1 - \frac{n_{2h}}{N_{2h}}\right) \frac{\sigma_{2h}^{2}}{n_{2h}} + \sum_{h=1}^{L} \left[\frac{N_h}{N} \left(W_{2h} - \frac{N_{2h}}{N_h}\right) (\bar{Y}_{2h} - \bar{Y}_{1h})\right]^2$$

The sample estimators of (4.2.7) and (4.2.8) are:

$$(4.2.9) \qquad v\,(\bar{y}_{MS}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(\frac{N_{2h}}{N_h}\right)^2 \left(1 - \frac{n_{2h}}{N_{2h}}\right) \frac{S_{2h}^{2}}{n_{2h}} \cdot V(\beta_h, \hat{\Delta}_h)$$

from (2.6.6)

and,

$$(4.2.10) \qquad M.S.E.\,(\bar{y}_{MS}) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(\frac{N_{2h}}{N_h}\right)^2 \left(1 - \frac{n_{2h}}{N_{2h}}\right) \frac{S_{2h}^{2}}{n_{2h}} \cdot M(\beta_h, \hat{\Delta}_h)$$

from (2.6.7)

We derive the value of W_{2h} which will yield the minimum mean square error for each stratum from (2.4.2) of Chapter 2.

$$(4.2.11) \qquad W_{2h} = \frac{\frac{N_{2h}}{N_h} (\bar{Y}_{2h} - \bar{Y}_{1h})^2}{\left(1 - \frac{n_{2h}}{N_{2h}}\right) \frac{\sigma_{2h}^{2}}{n_{2h}} + (\bar{Y}_{2h} - \bar{Y}_{1h})^2}$$

or

$$(4.2.12) \qquad W_{2h} = \frac{\frac{N_{2h}}{N_h}}{1 + \frac{1}{T_h^2}}$$

For the sample estimator we use

$$(4.2.13) \qquad W_{2h} = \frac{\frac{N_{2h}}{N_h} (\bar{y}_{2h} - \bar{Y}_{2h})^2}{(1 - \frac{n_{2h}}{N_h}) \frac{s_{2h}^2}{n_{2h}} + (\bar{y}_{2h} - \bar{Y}_{1h})^2}$$

The other results of Chapter 2 will likewise apply for individual strata.

B. <u>Case II</u>:

The estimator for $\bar{Y}_{MS}$ is the same as Case I except the values of W_h are determined to minimize the mean square error for the population mean rather than individual strata means. Writing the bias of (4.2.2) and adding (4.2.4) over strata

$$(4.2.14) \qquad \bar{Y}_{MS} - \bar{Y}_{ST} = \sum_{h=1}^{L} (W_{2h} - \frac{N_{2h}}{N_h}) (Y_{2h} - Y_{1h})$$

The sample estimators of (4.2.2) and 4.2.4) are:

$$(4.2.15) \qquad \bar{y}_{MS} = \sum_{h=1}^{L} \frac{N_h}{N} [Y_{1h} + \hat{W}_{2h} (\bar{y}_{2h} - \bar{Y}_{1h})]$$

$$(4.2.16) \qquad b = \sum_{h=1}^{L} \frac{N_h}{N} (\hat{W}_{2h} - \frac{N_{2h}}{N_h}) (\bar{y}_{2h} - \bar{Y}_{1h})$$

The population variance of the estimator $\bar{Y}_{MS}$ is the same as in Case I, namely

$$(4.2.17) \qquad V\,(\bar{Y}_{MS}) = \sum_{h=1}^{L} (\frac{N_h}{N})^2 \; W_{2h}^2 \; (1 - \frac{n_{2h}}{N_{2h}}) \; \frac{\sigma_{2h}^2}{n_{2h}}$$

However, the sample estimate of the variance from Table I of Chapter 2 is not available when the alternate method is used for deriving W_{2h} on pages 36 and 37. A multivariate normal distribution would need to be specified and numerically integrated to derive a table similar to Table 2. Until a satisfactory method of specifying variances and covariances, other than L univariate distribution which are all identical, Case II is of little practical interest unless the quantities $(\bar{B} - \bar{B}_i)$ are known and a new Table 2 is derived based on (4.2.25) below.

A comparison of the biased squared terms in (4.2.8) and (4.2.14) plus (4.2.17) is made to show the expected difference in the minimum mean square error of $\bar{Y}_{MS}$ when the criterions of Cases I and II are used to determine W_h. If we make the following substitutions in the bias terms

$$\text{Let} \qquad P_1 = \frac{N_1}{N}\,,\; P_2 = \frac{N_2}{N}\,,\; \ldots\; P_h = \frac{N_h}{N}\; \ldots.,\; P_L = \frac{N_L}{N}$$

$$\text{and} \qquad B_h = (W_{2h} - \frac{N_{2h}}{N_h})\;(\bar{Y}_{2h} - \bar{Y}_{1h})$$

Then comparing the second terms of the R.H.S. of (4.2.8) and (4.2.19) we have

$$(4.2.21) \qquad \sum_{h=1}^{L} P_h^2\, B_h^2 \text{ and } [\sum_{h=1}^{L} P_h\, B_h]^2 = \sum_{h=1}^{L} P_h^2\, B_h^2 + \sum_{h=1}^{L} P_j\, B_j\, P_k\, B_k$$

where $\sum_{h=1}^{L} P_h\, B_h$ is the average bias $\bar{B}_{ST}$

If the B_h's are all positive, that is, $\bar{Y}_{2h} \geq \bar{Y}_{1h}$ for all h strata

$$(4.2.22) \qquad \sum_{h=1}^{L} P_h^2\, B_h^2 < [\sum_{h=1}^{L} P_h\, B_h]^2 = \sum_{h=1}^{L} P_h^2\, B_h^2 + \sum_{j=k}^{L} P_j\, B_j\, P_k\, B_k$$

since all $P_h > o$. Hence the bias contribution in Case II is greater than in Case I. However, if some $\bar{Y}_{2h} \leq \bar{Y}_{1h}$ and other $\bar{Y}_{2h} > \bar{Y}_{1h}$ then the bias contribution in Case II will be less. The relationship of the values of W_h in Case II to those in Case I for individual stratum can be best seen below in equation (4.2.25) below.

The values of W_{2h} were h = 1, ... L. which will yield the minimum mean square error for $\bar{Y}_{MS}$ will be obtained by taking the derivatives in (4.2.19) with respect to W_{21}, W_{22}, ... W_{2L}. The L equations we obtain are of the form

$$(4.2.23) \qquad \frac{\partial(M.S.E.)}{\partial W_i} = 2(\frac{N_i}{N})^2 W_{2i} (1 - \frac{n_{2i}}{N_{2i}}) \frac{\sigma^2_{2i}}{n_{2i}} - \frac{2\ N_i^2}{N^2}$$

$$\cdot (W_{2i} - \frac{N_{2i}}{N_i})(Y_{2i} - Y_{1i}) - 2 \sum_{j \neq 1}^{L} \frac{N_i}{N} \frac{N_j}{N} (Y_{2i} - Y_{1i})$$

$$\cdot (W_{2j} - \frac{N_{2j}}{N_j}) (Y_{2j} - Y_{1j})$$

By setting (4.2.23) equal to zero and solving for W_{2i} we get the following expression:

$$(4.2.24) \qquad W_{2i} = \frac{(\frac{N_i}{N})^2 (\frac{N_{2i}}{N_i}) (\bar{Y}_{2i} - \bar{Y}_{1i})^2 + \frac{N_i}{N} \sum_{j \neq 1}^{L} \frac{N_j}{N} (W_{2j} - \frac{N_{2j}}{N_j}) (\bar{Y}_{2i} - \bar{Y}_{1i}) (\bar{Y}_{2j} - \bar{Y}_{1j})}{(\frac{N_i}{N})^2 (1 - \frac{n_{2i}}{N_{2i}}) \frac{\sigma^2_{2i}}{n_{2i}} + (\frac{N_i}{N})^2 (\bar{Y}_{2i} - \bar{Y}_{1i})^2}$$

By adding and subtracting $(\frac{N_i}{N})(W_{2i} - \frac{N_{2i}}{N_i}) (\bar{Y}_{2i} - \bar{Y}_{1i}) (\bar{Y}_{2i} - \bar{Y}_{1i})$

from the numerator in (4.2.24) we obtain W_{2i} as an expression involving only the parameters for the i^{th} stratum and the average bias, $\bar{B}$, given by (4.2.21), or

(4.2.25) $$W_{2i} = \frac{(\frac{N_i}{N})^2(\frac{N_{2i}}{N_i})(\bar{Y}_{2i} - \bar{Y}_{1i})^2 + \frac{N_i}{N}(\bar{Y}_{2i} - \bar{Y}_{1i})(\bar{B} - B_i)}{(\frac{N_i}{N})^2(1 - \frac{n_{2i}}{N_{2i}})\frac{\sigma^2_{2i}}{n_{2i}} + (\frac{N_i}{N})^2(Y_{2i} - Y_{1i})^2}$$

In this form a direct comparison with (4.2.11) is possible, and nature of the difference is determined by $(\bar{B} - B_i)$. If the bias in the h^{th} stratum equal the average bias, the same value for W_{2h} is obtained as in Case I.

The simultaneous solution to the L equations given by (4.2.23) being set equal to zero is:

For simplification of the notation

let $$P_{2h} = \frac{N_{2h}}{N_h} ,$$

$$P_h = \frac{N_h}{N} ,$$

$$\Delta_h = \bar{Y}_{2h} - \bar{Y}_{1h} ,$$

and $$V_{2h} = (1 - \frac{n_{2h}}{N_{2h}}) \frac{\sigma^2_{2h}}{n_{2h}} = (\frac{1}{n_{2h}} - \frac{1}{N_{2h}}) \sigma^2_{2h}$$

The system of equation will be of the form given below for h = 1.

$$W_{21} (P_1^2 V_{21} + P_1^2 \Delta_1^2) + W_{22} (P_1 P_2 \Delta_1 \Delta_2) + W_{23} (P_1 P_3 \Delta_1 \Delta_3)$$
$$+ \ldots\ldots + W_{2L} (P_1 P_L \Delta_1 \Delta_L)$$
$$= P_1^2 P_{21} \Delta_1^2 + P_1 P_2 P_{22} \Delta_1 \Delta_2 + P_1 P_2 P_{23} \Delta_1 \Delta_3$$
$$+ \ldots\ldots + P_1 P_L P_{2L} \Delta_1 \Delta_L$$

and for h = 1

$$W_{21}\ (P_1\ P_L\ \Delta_1\ \Delta_L) + W_{22}\ (P_2\ P_L\ \Delta_2\ \Delta_L) + W_{23}\ (P_3\ P_L\ \Delta_2\ \Delta_L)$$

$$+ \ldots\ldots + W_{2L}\ (P_L^2\ V_{2L} + P_L^2\ \Delta_L^2)$$

$$= P_1\ P_L\ P_{2L}\ \Delta_1\ \Delta_L + P_2\ P_L\ P_{2L}\ \Delta_2\ \Delta_L$$

$$+ P_3\ P_L\ P_{2L}\ \Delta_3\ \Delta_L + \ldots\ldots + P_L^2\ P_{2L}\ \Delta_L^2$$

In matrix notation the system of equation can be represented as

(4.2.26) $(A_{ij})_{LXL}\ (W_{2i})_{LX1} = (Y_i)_{LX1}$

and the solution for W_2 will be unique if the inverse to the systematical matrix A exists, that is:

(4.2.27) $W_2 = A^{-1}\ Y$.

4.3 Ratio Estimators for 100 Percent Sampling of List from Strata

A ratio estimator for the nonrespondent stratum mean is proposed which is to be combined with the respondent mean to form a linear combination of the two means. The mean of the concomitant variable $\bar{X}_{2h}$ and the total X_{2h} are assumed known for the nonrespondents.

(4.3.1) $\bar{Y}_{Rh} = W_{1h}\ \bar{Y}_{1h} + W_{2h}\ R_{2h}\ \bar{X}_{2h} = W_{1h}\ \bar{Y}_{1h} + W_{2h}\ \bar{Y}_{2Rh}$

is the estimator for the hth strata, and the minimum mean square ratio estimator for the population is:

(4.3.2) $$\bar{Y}_{RS} = \frac{\sum_{h=1}^{L} N_h\ \bar{Y}_{Rh}}{N}$$

where $W_{1h} + W_{2h} = 1$

and $N_1 + N_2 + \ldots. + N_L = N$.

A. Case I

The bias of the estimator in (4.3.1) is:

$$(4.3.3) \qquad \bar{Y}_h - \bar{Y}_{Rh} = \left(\frac{N_{2h}}{N_h} - W_{2h}\right)\left(\bar{Y}_{2Rh} - \bar{Y}_{1h}\right)$$

The squared bias summing over all strata is:

$$(4.3.4) \qquad (\text{Bias})^2 = \sum_{h=1}^{L} \left[\frac{N_h}{N} \left(W_{2h} - \frac{N_{2h}}{N_h}\right)\left(\bar{Y}_{2Rh} - \bar{Y}_{1h}\right)\right]^2$$

The sample estimators are:

$$(4.3.5) \qquad \bar{y}_{RS} = \sum_{h=1}^{L} \frac{N_h}{N} \left[\left(1 - \hat{W}_{2h}\right) \bar{Y}_{1h} + \hat{W}_{2h} \bar{y}_{2Rh}\right]$$

$$(4.3.6) \qquad b^2 = \sum_{h=1}^{L} \left[\frac{N_h}{N} \left(\hat{W}_{2h} - \frac{N_{2h}}{N_h}\right)\left(\bar{y}_{2Rh} - \bar{Y}_{1h}\right)^2\right.$$

$$(4.3.7) \qquad v\left(\bar{y}_{RS}\right) = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(\frac{N_{2h}}{N_h}\right)^2 \frac{s_{2h}^2}{n_{2h}} \; v\left(\Delta_h\right)$$

from Section 2.9 and Table 1. The adjustment of this expression by term

$$\sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(\frac{N_{2h}}{N_h}\right)^2 \left(1 - \frac{K_h}{N_{2h}}\right) \left[v\left(\bar{X}_{Mh}\right) - 2\,\hat{R}_{Mh}\,\text{Cov}\left(\bar{Y}_{Mh}, \bar{X}_{Mh}\right)\right] .$$

The population variance based on the results of previous results, (2.7.3) and standard ratio variances, is:

$$(4.3.7') \qquad V\left(\bar{Y}_{RS}\right) = \sum_{h=1}^{L} \left\{W_{2h}^2 \left(1 - \frac{n_{2h}}{N_{2h}}\right) \frac{1}{n_{2h}} \left[V\left(Y_{2h}\right) + V\left(X_{2h}\right) - 2R_{2h}\text{Cov}\left(Y_{2h}, X_{2h}\right)\right]\right\}$$

The mean square is:

$$(4.3.8)\quad \text{M.S.E.}(\bar{Y}_{RS}) = \sum_{h=1}^{L} W_{2h}^{2}\left(1 - \frac{n_{2h}}{N_{2h}}\right) \frac{1}{n_{2h}} [V(Y_{2h}) + V(X_{2h}) - 2R_{2h}\text{Cov}(Y_{2h},X_{2h})]$$

$$+ \sum_{h=1}^{L} \left[\frac{N_h}{N} \left(W_{2h} - \frac{N_{2h}}{N_h}\right) (\bar{Y}_{2Rh} - \bar{Y}_{1h})\right]^2$$

Using the results of (2.8.2) and (2.9.2)

$$(4.3.9)\quad W_{2h} = \frac{\frac{N_{2h}}{N_h} (\bar{Y}_{2Rh} - \bar{Y}_{1h})^2 \left(\frac{N_h}{N}\right)^2}{\left(1 - \frac{n_{2h}}{N_{2h}}\right) \frac{1}{n_{2h}} V(R_{2h}) + (\bar{Y}_{2Rh} - \bar{Y}_{1h})^2 \left(\frac{N_h}{N}\right)^2}$$

or

$$(4.3.10)\quad W_{2h} = \frac{\frac{N_{2h}}{N_h} \left(\frac{N_h}{N}\right)^2}{1 + \left(\frac{N_h}{N}\right)^2 \frac{1}{T_h^2}}$$

For the sample estimator we use

$$(4.3.11)\quad W_{2h} = \frac{\frac{N_{2h}}{N_h} \left(\frac{N_h}{N}\right)^2 (\bar{y}_{2Rh} - \bar{Y}_{1h})^2}{\left(1 - \frac{n_{2h}}{N_{2h}}\right) \frac{1}{n_{2h}} v(R_{2h}) + (\bar{y}_{2Rh} - \bar{Y}_{1h})^2 \left(\frac{N_h}{N}\right)^2}$$

It is clear that if the usual relationships between y_{2h} and x_{2h} hold as when the classical ratio is more efficient than the simple mean per unit, the (4.3.2) will be more efficient than (4.2.2).

B. Case II

The estimator is the same as Case I, but the bias term is similar to (4.2.16).

$$(4.3.12) \qquad \bar{Y}_{RS} - \bar{Y}_{ST} = \sum_{h=1}^{L} \frac{N_h}{N} (W_{2h} - \frac{N_{2h}}{N_h})(\bar{Y}_{2Rh} - \bar{Y}_{1h})$$

The variance is the same as (4.3.7).

The mean square error of Y_{RS} would be

$$(4.3.13) \qquad M.S.E.(\bar{Y}_{RS}) = \sum_{h=1}^{L} \hat{W}_{2h}^{2}(1 - \frac{n_{2h}}{N_{2h}}) \frac{1}{n_{2h}} [V(Y_{2h}) + V(\bar{X}_{2h} - 2\hat{R}_h Cov(Y_{2h}, X_{2h})]$$

$$+ \sum_{h=1}^{L} [\frac{N_h}{N} (W_{2h} - \frac{N_{2h}}{N_h})(\bar{Y}_{2Rh} - \bar{Y}_{1h})]^2$$

Use the results of (4.2.23), (4.2.24), and (4.2.25)

$$(4.3.14) \qquad W_{2i} = \frac{(\frac{N_i}{N_h})^2 (\frac{N_{2i}}{N_i})(Y_{2Ri} - Y_{1i})^2 + \frac{N_i}{N} \sum_{j \neq 1}^{L} \frac{N_j}{N} (W_{2j} - \frac{N_{2j}}{N_j})(\bar{Y}_{2Ri} - \bar{Y}_{1i})(\bar{Y}_{2Rj} - \bar{Y}_{1j})}{(1 - \frac{n_{2i}}{N_{2i}}) \frac{1}{n_{2i}} V(R_h) + (\frac{N_i}{N})^2 (\bar{Y}_{2Ri} - \bar{Y}_{1i})^2}$$

$$(4.3.15) \qquad W_{2i} = \frac{(\frac{N_i}{N})^2 (\frac{N_{2i}}{N_i}) (\bar{Y}_{2Ri} - \bar{Y}_{1i})^2 + \frac{N_i}{N} (\bar{Y}_{2Ri} - \bar{Y}_{1i})(\bar{B} - B_i)}{(1 - \frac{n_{2i}}{N_{2i}}) \frac{1}{n_{2i}} V(R_h) + (\frac{N_i}{N})^2 (\bar{Y}_{2Ri} - \bar{Y}_{1i})^2}$$

The simultaneous solution of the L equations given by (4.3.14) lead to results similar to that obtained in (4.2.26) and (4.2.27).

Namely, the vector $(W_{2i})_{LX1}$ will be unique if the A matrix has an universe:

$$W_2 = A^{-1} Y .$$

The sample estimators of the weight, bias, and variance are:

$$(4.3.16)\quad \hat{W}_{2h} = \frac{(\frac{N_h}{N})^2(\frac{N_{2h}}{N_h})(\bar{y}_{2h}-\bar{Y}_{1h})^2 + \frac{N_h}{N}\sum_{j=h}^{L}\frac{N_j}{N}(W_{2j}-\frac{N_{2j}}{N_j})(Y_{2Rh}-\bar{Y}_{1h})(\bar{y}_{2Rj}-\bar{Y}_{1j})}{(1-\frac{n_{2h}}{N_{2h}})\frac{1}{n_{2h}}\,v(\hat{R}_h) + (\frac{N_h}{N})^2(\bar{y}_{2Rh}-\bar{Y}_{1h})^2}$$

$$(4.3.17)\quad b^2 = [\sum_{h=1}^{L}(W_{2h}-\frac{N_{2h}}{N_h})(\bar{y}_{2Rh}-\bar{Y}_{1h})]^2$$

$$(4.3.18)\quad v(\bar{y}_{RS}) = \sum_{h=1}^{L}\hat{W}_{2h}^2(1-\frac{n_{2h}}{N_{2h}})\frac{1}{n_{2h}}[v(y_{2h}) + v(x_{2h}) - 2\hat{R}_h\mathrm{Cov}(y_{2h},x_{2h})]$$

The difficulty in evaluating this estimator as pointed out on page 36 still remains, and Case II is of little practical value unless tables corresponding to Tables 1 and 2 can be derived.

4.4 Random Subsample of List Available for Stratum

The results of Chapter 3 may be applied in the formulation of the estimator, variance, and mean square error. The bias term will depend on whether Case I or Case II criterion are used. The results of Sections 4.2 and 4.3 will be used for Case III.

The simple minimum mean square error estimator is:

(4.4.1) $\bar{Y}_{Mh} = W_{1h} \bar{Y}_{1h} + W_{2h} \bar{Y}_{2h}$ for each stratum, and

(4.4.2) $\bar{Y}_{MS} = \sum_{h=1}^{L} \frac{N_h}{N} (W_{1h} \bar{Y}_{1h} + W_{2h} \bar{Y}_{2h})$ for the population mean,

where $W_{1h} + W_{2h} = 1$ and $N_1 + N_2 + \ldots. + N_L = N$.

The sample estimates of (4.4.1) and 4.4.2) are:

(4.4.3) $\bar{y}_{Mh} = \hat{W}_{1h} \bar{y}_{1h} + \hat{W}_{2h} \bar{y}_{2h}$

(4.4.4) $\bar{y}_{MS} = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_{Mh}$

where $\hat{W}_{1h}$ and $\hat{W}_{2h}$ will depend on Case I or Case II criterion.

The variance of the estimator $\bar{Y}_{MS}$, using (3.3.5), is approximately:

(4.4.5) $V(\bar{Y}_{MS}) = \sum_{h=1}^{L} (\frac{N_h}{N})^2 \{W_{1h}^2 \sigma_{1h}^2 (\frac{N_h}{n_h N_{1h}} - \frac{1}{N_{1h}}) + W_{2h}^2 \sigma_{2h}^2 (\frac{N_h}{f_h n_h N_{2h}} - \frac{1}{N_{2h}})\}$

A. Case I

The bias term for the population mean $\bar{Y}_{MS}$ from (3.3.2)

(4.4.6) $B^2 = \sum_{h=1}^{L} (\frac{N_h}{N})^2 (W_{2h} - \frac{N_{2h}}{N_h})^2 (\bar{Y}_{2h} - \bar{Y}_{1h})^2$

The sample estimator of (4.4.6)

$$(4.4.7)\quad b^2 = \sum_{h=1}^{L} (\frac{N_h}{N})^2 (\hat{W}_{2h} - \frac{n_{2h}}{n_h})^2 \; (\bar{y}_{2h} - \bar{y}_{1h})^2$$

where $\hat{W}_{2h}$ is the optimum value for the h^{th} stratum.

The mean square of $\bar{Y}_{MS}$ is:

$$(4.4.8)\quad \text{M.S.E.}(\bar{Y}_{MS}) = (4.4.5) + (4.4.6)$$

The optimum value of W_{2h} is determined from (4.4.8) by setting the derivative with respect to W_{2h} equal to zero, and solving

$$(4.4.9)\quad W^*_{2h} = \frac{\sigma^2_{2h}(\frac{N_h}{n_h N_{1h}} - \frac{1}{N_{1h}}) + (\frac{N_{2h}}{N_h})(\bar{Y}_{2h} - \bar{Y}_{1h})^2}{\sigma^2_{2h}(\frac{N_h}{n_h N_{1h}} - \frac{1}{N_{1h}}) + \sigma^2_{2h}(\frac{N_h}{f_h n_h N_{2h}} - \frac{1}{N_{2h}}) + (\bar{Y}_{2h} - \bar{Y}_{1h})^2}$$

For the sample estimate of W_{2h} we obtain

$$(4.4.10)\quad W^*_{2h} = \frac{(\frac{N_h - n_h}{N_h}) \; \frac{s^2_{1h}}{n_{1h}} + (\frac{n_{2h}}{n_h}) \; (\bar{y}_{2h} - \bar{y}_{2h})^2}{(\frac{N_h - n_h}{N_h}) \; \{ \frac{s^2_{1h}}{n_{1h}} + s^2_{2h}(\frac{1}{k_h} - \frac{1}{n_{2h}}) + (\bar{y}_{2h} - \bar{y}_{1h})^2 \}}$$

The comparisons of the variance and mean square error with the classical estimator for the individual stratum are the same as given in Tables 3 and 4 of Chapter 3. That is:

$$(4.4.11)\quad v(\bar{Y}_{MS}) = \sum_{h=1}^{L} (\frac{N_h}{N})^2 [(\frac{N_h - n_h}{N_h}) \; \frac{s^2_h}{n_h} + (\frac{n_{2h}}{k_h} - 1) \; \frac{n_{2h}}{n_h} \, s^2_{2h}] \; V(P_h, f_h, \Delta_h)$$

$$(4.4.12)\quad \text{M.S.E.}(Y_{MS}) = \sum_{h=1}^{L} (\frac{N_h}{N})^2 [(\frac{N_h - n_h}{n_h}) \frac{S_h^2}{n_h} + (\frac{n_{2h}}{k_h} - 1) \frac{n_{2h}}{n_h} S_{2h}^2] M(P_h, f_h, \Delta_h)$$

B. Case II

The estimator for $\bar{Y}_{MS}$ is the same as in Case I, but the bias term differs since it is determined with respect to $\bar{Y}_{ST}$ rather than the strata means $\bar{Y}_h$.

$$(4.4.13)\quad B^2 = [\sum_{h=1}^{L} (\frac{N_h}{N}) (W_{2h} - \frac{N_{2h}}{N_h}) (\bar{Y}_{2h} - \bar{Y}_{1h})]^2$$

The sample estimator of the bias is:

$$(4.4.14)\quad b^2 = [\sum_{h=1}^{L} (\frac{N_h}{N}) (\hat{W}_{2h} - \frac{n_{2h}}{n_h}) (\bar{y}_{2h} - \bar{y}_{1h})]^2$$

The mean square error of $\bar{Y}_{MS}$ is:

$$(4.4.15)\quad \text{M.S.E.}(\bar{Y}_{MS}) = (4.4.5) + 4.4.12)$$

We wish to minimize (4.4.14) with respect to the L parameters of W_{2h}. Taking L partial derivatives and setting each equal to zero, we get for the h^{th} equation:

$$(4.4.16)\quad \frac{\partial\ \text{M.S.E.}(\bar{Y}_{MS})}{\partial\ W_{2h}} = (\frac{N_h}{N})^2 \sigma_{1h}^2 (\frac{N_h}{n_h N_{1h}} - \frac{1}{N_{1h}}) + \sigma_{2h}^2 (\frac{N_h}{f_h n_h N_h} - \frac{1}{N_{2h}})$$

$$+ -2 (\frac{N_h}{N})^2 (W_{2h} - \frac{N_{2h}}{N_h}) (\bar{Y}_{2h} - \bar{Y}_{1h})^2$$

$$+ -2 \sum_{j=h}^{L} \frac{N_h}{N} \frac{N_j}{N} (W_{2j} - \frac{N_{2j}}{N_j}) (\bar{Y}_{2h} - \bar{Y}_{1h}) (\bar{Y}_{2j} - \bar{Y}_{1j})$$

The set of L equations are solved simultaneously for W_{2h} yielding a unique solution if the inverse of the A matrix exists. Each of the L equations to be solved are of the following form.

To simplify the notation

let $$P_{2h} = \frac{N_{2h}}{N_h}$$

$$P_h = \frac{N_h}{N}$$

$$\Delta_h = \bar{Y}_{2h} - \bar{Y}_{1h}$$

$$V_h = (\frac{N_h - n_h}{N_h}) \{ (\frac{N_{1h} - 1}{N_h}) \sigma^2_{1h} + (\frac{N_{2h} - 1}{N_h}) \sigma^2_{2h} + \frac{N_{1h} N_{2h}}{N_h} (Y_{2h} - Y_{1h}) \}^2$$

$$V_{2h} = \sigma^2_{2h} (\frac{N_h}{f_h n_h N_{2h}} - \frac{1}{N_{2h}})$$

$$V_{1h} = \sigma^2_{1h} (\frac{N_h}{n_h N_{1h}} - \frac{1}{N_{1h}})$$

For h=1, we would have

$$(4.4.17)\quad W_{21} (P_1^2 V_{1h} + P_1^2 V_{2h} + P_1^2 \Delta_1^2) + W_{22} (P_1 P_2 \Delta_1 \Delta_2) + \ldots + W_{2L}$$

$$= (P_1 P_L \Delta_1 \Delta_L)(P_1^2 V_{1h} + P_1^2 P_{21} \Delta_1^2)$$

$$+ P_{22} (P_1 P_2 \Delta_1 \Delta_2) + \ldots + P_{2L} (P_1 P_L \Delta_1 \Delta_L)$$

In matrix notation the system of equations would be

$$(4.4.18)\quad [A_{ij}]_{LXL} [W_{2i}]_{LX1} = [Y_i]_{LX1}$$

Sample estimates are available for all the elements of the A matrix and the vectors W and Y.

Where the elements in the first rows are:

$$[A_{ij}] = [(P_1^2\ V_{1h} + P_1^2\ V_{21} + P_1^2\ \Delta_1^2),\ (P_1\ P_2\ \Delta_1\ \Delta_2),\ (P_1\ P_3\ \Delta_1\ \Delta_3),$$

$$\ldots\ldots,\ (P_1\ P_L\ \Delta_1\ \Delta_L)]$$

$$[W_{21}] = W_{21}$$

$$[Y_1\ ] = [P_{21}\ (P_1^2\ V_{11} + P_1^2\ \Delta_1^2),\ P_{22}\ (P_1\ P_2\ \Delta_1\ \Delta_2),\ P_{23}\ (P_1\ P_3\ \Delta_1\ \Delta_3),$$

$$\ldots\ldots,\ P_{2L}\ (P_1\ P_L\ \Delta_1\ \Delta_L)]$$

While these values of W_{2h} may result in the population mean square error being less than in Case I, the difficulty in obtaining a sample estimate of the variance makes the criterion in Case I more practical. In addition, strata means are generally of considerable interest in a stratified design.

4.5 Ratio Estimator for Random Subsample of Stratified List

The results for Case I are stated from Sections 3.7, 3.8, 3.9, and 4.4.

The estimators for the population mean and total are obtained independently for each strata.

$$(4.5.1) \quad \bar{Y}_{Rh} = \frac{W_{1h}\ \bar{Y}_{1h} + W_{2h}\ \bar{Y}_{2h}}{W_{1h}\ \bar{X}_{1h} + W_{2h}\ \bar{X}_{2h}}\ \bar{X}_h = \frac{\bar{Y}_{Mh}}{\bar{X}_{Mh}}\ \bar{X}_h = R_{Mh}\ \bar{X}_h$$

$$(4.5.2) \quad \bar{Y}_R = \sum_{h=1}^{L} \frac{N_h}{N}\ \bar{Y}_{Rh}$$

$$(4.5.3) \quad Y_R = \sum_{h=1}^{L} N_h\ \bar{Y}_{Rh}$$

The variance and bias squared are:

$$(4.5.4) \qquad V(\bar{Y}_R) = \sum_{h=1}^{L} (\frac{N_h}{N})^2 \; V(\bar{Y}_{Rh})$$

where $V(\bar{Y}_{Rh})$ is from (3.7.4)

$$(4.5.5) \qquad V(\bar{Y}_{Rh}) = (1 - \frac{n_h}{N_h}) \; [V(\bar{Y}_{Mh}) + V(\bar{X}_{Mh}) - 2R_{Mh} \; Cov(Y_{Mh}, X_{Mh})]$$

$$(4.5.6) \qquad B^2 = \sum_{h=1}^{L} (\frac{N_h}{N})^2 \; B_h^2$$

where B_h^2 is from (3.7.3')

$$(4.5.7) \qquad B_h^2 = [\frac{(W_{2h} - \frac{N_{2h}}{N_h}) \; \{\bar{X}_h \; (\bar{Y}_{2h} - \bar{Y}_{1h}) - \bar{Y}_h \; (\bar{X}_{2h} - \bar{X}_{1h})\}}{\bar{X}_{1h} + (W_2 - \frac{N_{2h}}{N_h}) \; (\bar{X}_{2h} - \bar{X}_{1h})}]^2$$

The mean square error is:

$$(4.5.8) \qquad M.S.E.(\bar{Y}_R) = (4.5.4) + (4.5.6)$$

The value of W_{2h} which will minimize the mean square error using (3.8.4) is:

$$(4.5.9) \qquad W_{2h}^* = \frac{\sigma_{R1h}^2 + \frac{N_{2h}}{N_h} \, [(\bar{Y}_{2h} - \bar{Y}_{1h}) - (\bar{X}_{2h} - \bar{X}_{1h}) \; \frac{\bar{Y}_h}{\bar{X}_h}]^2}{\sigma_{R1h}^2 + \sigma_{R2h}^2 + [(\bar{Y}_{2h} - \bar{Y}_{1h}) - (\bar{X}_{2h} - \bar{X}_{1h}) \frac{\bar{Y}_h}{\bar{X}_h}]^2}$$

The variance is calculated by (4.4.11) and adjusted by:

$$\sum_{h=1}^{L} (\frac{N_h}{N})^2 (\frac{N_h - n_h}{N_h}) [v(\bar{X}_{Mh}) - 2\hat{R}_{Mh} \text{Cov}(\bar{Y}_{Mh},\bar{X}_{Mh})]$$

The mean square error is derived by adding (4.5.6) to the variance.

4.6 Sample Estimators for Separate Ratio Estimators

The sample mean is:

$$(4.6.1) \qquad \bar{y}_R = \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_{Rh} = \sum_{h=1}^{L} \frac{N_h}{N} (\frac{\bar{y}_{Mh}}{\bar{x}_{Mh}} \bar{X}_h)$$

The bias is:

$$(4.6.2) \qquad b^2 = [\frac{(\hat{W}^*_{2h} - \frac{n_{2h}}{n_h}) \{ \bar{X}_h (\bar{y}_{2h} - \bar{y}_{1h}) - \bar{y}_h (\bar{x}_{2h} - \bar{x}_{1h})\}}{\bar{x}_{1h} + (\hat{W}^*_{2h} - \frac{n_{2h}}{n_h})(\bar{x}_{2h} - \bar{x}_{1h})}]^2$$

The sample variance from (3.7.4) and Table 3:

$$(4.6.3) \qquad v(\hat{R}) = \sum_{h=1}^{L} (\frac{N_h}{N})^2 \; v(\hat{R}_h) \cdot V(P_h,f_h,\Delta_h)$$

The mean square error based on (3.7.4) and Table 4:

$$(4.6.4) \qquad \text{M.S.E.}(\hat{R}) = \sum_{h=1}^{L} (\frac{N_h}{N})^2 \; v(\hat{R}_h) \cdot M(P_h,f_h,\Delta_h)$$

where it has been assumed that for the variable Y_h that the variance is the same for both respondent and nonrespondent domains, and for the variable X_h the domain variances are likewise equal. In practice, pooled variances within each strata are used for each variable.

The corresponding expressions for Case II are not set forth due to the inability to develop an appropriate sample estimate for the variance.

CHAPTER 5 - AN EXAMPLE FOR A STRATIFIED LIVESTOCK SURVEY

5.1 Nature of Survey

A livestock survey conducted in March of 1968 is used to illustrate the estimates derived from the theory developed in this paper. The survey used was a multiple purpose survey to estimate the inventories of cattle, hogs, and sheep with several subclasses for each. The sample was designed as a routine operational survey by the Statistical Reporting Service. The survey was conducted by mail with a fixed number of nonrespondent follow-up interviews. The estimation by species was to be based on the classical double sampling theory within each strata as set forth by Hansen and Hurwitz (1946). The strata were constructed based on the 1967 Illinois State Farm Census; that is, the data used as a basis for stratification related to several livestock characteristics as of January 1, 1967. The basis and justification for the strata are not of concern in this study.

5.2 Survey Means and Error Estimates

The results in Table 5 relate to only one of the survey items - total number of cattle on farms March 31, 1968. The estimated mean number of cattle per farm and the variance of the mean derived using the classical unbiased estimator are given in columns 1 and 2 of Table 5. These statistics were calculated using the following sample estimates for the mean and its variance:

$$\bar{y}_h = \frac{n_{h1}}{n_h} \bar{y}_{n1} + \frac{n_{h2}}{n_h} \bar{y}_{h2}$$

$$s^2_{\bar{Y}_h} = \left(\frac{N_{2h}}{N_h}\right)^2 \left(1 - \frac{k_h}{N_{2h}}\right) \frac{s^2_{2h}}{k_h} \qquad \text{(Strata 1 to 7)}$$

$$s^2_{\bar{y}h} = \left(\frac{N_h - n_h}{N_h}\right) \frac{s^2_h}{n_h} + \left(\frac{n_{h2}}{k_h} - 1\right) \frac{n_{h2}}{n_h} s^2_{h2} \qquad \text{(Strata 8 to 19)}$$

The corresponding estimates based on application of the techniques given in Chapters 2, 3, and 4 are shown in columns 3, 4, and 5 of Table 5. The results for strata 1 through 7 are based on Chapter 2 and Section 4.2 of Chapter 4 while the results for strata 8 through 19 are based on Chapter 3 and Section 4.4 of Chapter 4. In particular, the estimator used for the mean was either 4.2.5 or 4.4.3; the variance based on either (4.2.9) or (4.4.11); the mean square error of the mean on either 4.2.10 or 4.4.12.

The weights for the nonrespondent substrata were based on either (4.2.13) or (4.4.10). The weights for the classical estimator were $N_{h2} + N_h$ for strata 1 through 7 and $n_{h2} + n_h$ for strata 8 through 19. For all strata a value of 1.0 was used for β in Tables 1 and 2. The population and list sizes by strata along with the sample sizes and substrata weights are given in Table 6. The quantity delta (Δ) used to calculate the variances and mean square errors in columns 4 and 5 of Table 5 are given in the last column on the right of Table 6. The variances and mean square error were derived by multiplying the variance of the classical estimator by a factor derived by linear interpolation using Tables 1 and 2 of Chapter 2 and Tables 3 and 4 of Chapter 3.

Where the sample value of delta (Δ) exceeded 2.0 for strata 1 through 7 and 1.5 for strata 8 through 19, the classical unbiased estimator of the mean should be used. These values correspond approximately to the point in the tables where the minimum mean square estimator becomes less efficient than the classical estimates.

5.3 Comments on Comparison with Classical Estimator

Due to the extremely small population and substrata sizes for strata 2, 6, and 7 and no inferences appear warranted except to note that the sampling errors remain fairly large even after enumerating most of the nonrespondents. A 100 percent sample is probably required if it is necessary to control the error for such small populations.

The mean square error at the state level is about 11.0 percent less than the variance of the classical estimator while the mean is increased by about one percent. At this level of aggregation, the possibility of bias appears to be quite small. This is evident in the state mean and also when the pairs of means are plotted for individual strata. For strata 2, 9, 12, 14, 15, and 16 where the minimum mean square estimator should be rejected based on delta, the mean and the bias at the state level would both have been reduced if the classical estimator, which is shown in column 1 of Table 5, had been used in deriving the state average. That is, the state average would have been 23.8 rather than the 24.2. For the 13 strata in which the minimum mean square estimator was considered appropriate the unbiased estimator of the mean was 21.2 as compared to 21.0 for the mean square estimator.

For the 6 strata (2, 9, 12, 14, 15, and 16) the weights computed for minimum mean square estimator are not too different from the weights for the unbiased estimator. While the mean square estimator was not rejected based on delta, the derived weights were such that no serious bias would have been introduced by using the minimum mean square estimator even though it was inefficient. Hence, the procedure tends to have the characteristics of substituting a weight not too different from $N_{h2} + N_h$ (or $n_{h2} + n_h$) when the estimator becomes inefficient.

Table 5--Comparison of Estimates of Means by Strata (Cattle Per Farm)

Strata	Classical Mean	Variance of Classical Mean	Minimum Mean Square Mean	Variance of Minimum Mean Square Mean	M.S.E. of Minimum Mean Square Mean	Relative Efficiency Col. 5 + 2
1	212.2	585.1	210.5	283.9	285.5	.49
2	422.8	2038.1	431.1	2169.4	107570.7	52.78
3	3.7	80.9	2.3	40.0	40.5	.50
4	1357.5	183506.7	1332.5	88853.9	89294.4	.49
5	138.1	1352.5	124.2	813.8	876.8	.65
6	113.5	205.7	106.4	161.9	181.9	.88
7	1319.1	509432.1	1258.0	249316.1	251506.6	.49
8	26.2	15.3	25.0	11.5	11.8	.77
9	37.7	4.7	38.1	5.1	5.4	1.15
10	111.9	175.5	108.4	140.0	143.8	.82
11	25.8	5.6	26.4	5.1	5.6	1.00
12	104.5	115.7	122.3	120.3	129.9	1.12
13	276.4	3356.4	291.1	3090.8	3343.2	1.00
14	3.9	1.0	4.0	1.1	1.1	1.10
15	6.4	7.9	6.1	8.0	8.7	1.10
16	4.1	3.1	5.4	3.1	3.4	1.10
17	3.7	3.7	3.5	3.1	3.2	.86
18	10.1	5.0	11.3	4.1	4.3	.86
19	25.5	91.2	22.9	77.0	81.8	.90
State	1/ 23.9	2/ 1.27	1/ 24.2	2/ 1.08	2/ 1.13	.89

1/ Derived from individual strata means with $\frac{N_h}{N}$ being used as weights.

2/ Derived from individual strata squared errors with $(\frac{N_h}{N})^2$ being used as weights.

Table 6--Population and Substrata Sizes with Nonrespondent Weights

Strata	Population Size N_h	Total List Sample n_h	Total Sample Nonrespondents n_{h2}	Nonrespondents Interviewed K_h	Classical Weight $\frac{n_{h2}}{n_h}$	Minimum MSE Weight W_{h2}	Δ
1	319	319	243	34	.762	.004	.10
2	14	14	10	8	.714	.689	7.33
3	54	54	40	14	.741	.018	.21
4	52	52	37	14	.712	.002	.08
5	317	317	210	28	.662	.113	.69
6	39	39	31	24	.795	.374	1.19
7	7	7	4	2	.571	.004	.15
8	9451	235	167	22	.711	.290	.50
9	16127	1152	857	124	.743	.712	3.40
10	5516	785	581	83	.739	.403	.64
11	25383	632	434	62	.685	.562	1.37
12	3060	219	147	21	.667	.606	2.27
13	229	115	83	16	.712	.571	1.33
14	4906	187	155	20	.732	.727	2.80
15	4175	297	214	28	.717	.670	1.67
16	1312	187	142	20	.754	.558	1.91
17	40995	297	227	35	.761	.612	.84
18	18107	722	507	76	.701	.372	.88
19	3581	293	216	32	.734	.521	.96
State	133644	5923	4294	663			

REFERENCES

1. Cochran, W. G., Sampling Techniques, New York, John Wiley and Sons, 1965, 2nd Ed.

2. Faradori, G. T., "Some Nonresponse Sampling Theory for Two-Stage Designs," Unpublished dissertation, Consolidated University of North Carolina, 1962.

3. Finkner, A. L., "Adjustment for Nonresponse Bias in a Rural Survey," Agricultural Economics Research, Vol. 4, (1952), pp. 77-82.

4. Hansen, M. H. and Hurwitz, W. N., "The Problem of Nonresponse in Sample Surveys," Journal of the American Statistical Association, Vol. 41, (1946), pp. 517-529.

5. Hartley, H. O., "Analytic Studies of Survey Data," Institute of Statistics, University of Rome, 1959.

6. Hartley, H. O., "Discussion of Paper by F. Yates," Journal of the Royal Statistical Association, 109, 37.

7. Hendricks, W. A., "Adjustment for Bias by Nonresponse in Mailed Surveys," Agricultural Economics Research, Vol. 1, (1949), pp. 52-56.

8. Hendricks, W. A., The Mathematical Theory of Sampling, New Brunswick, N. J., Scarecrow Press, 1956.

9. Huddleston, H. F., "Methods Used in a Survey of Orchards," Agricultural Economics Research, Vol. 2, (1950), pp. 126-130.

10. Politz, A. N. and Simmons, W. R., "An Attempt to Get the 'Not at Homes' in the Sample Without Callbacks," Journal of the American Statistical Association, Vol. 44, (1949), pp. 9-31 and Vol. 45, (1950), pp. 136-137.

11. Simmons, W. R., "A Plan to Account for 'Not at Homes' by Combining Weighting and Callback," Journal of Marketing, Vol. 11, (1954), pp. 42-53.