Survey Methodology, June 1993
Vol. 19, No. 1, pp. 95–101
Statistics Canada

Double Sampling for Stratification

R.P. TREDER and J. SEDRANSK[1]

ABSTRACT

Double sampling is a common alternative to simple random sampling when there are expected to be gains from using stratified sampling, but the units cannot be assigned to strata prior to sampling. It is assumed throughout that the survey objective is estimation of the finite population mean. We compare simple random sampling and three allocation methods for double sampling: (a) proportional, (b) Rao's (Rao 1973a,b) and (c) optimal. There is also an investigation of the effect on sample size selection of misspecification of an important design parameter.

KEY WORDS: Optimal sample sizes; Two phase sampling.

1. INTRODUCTION

Suppose we wish to estimate the finite population mean in a stratified population, but the units cannot be assigned to strata prior to sampling. Typically, the number of units in each stratum is unknown. Then, double sampling is commonly considered as an alternative to simple random sampling. With double sampling, a simple random sample of size n' is selected from a finite population of N units with n_i' units identified as members of stratum i, $i = 1, \ldots, L$. The second phase sample is a set of L independent simple random subsamples where, in stratum i, n_i units are selected from the n_i' identified in the first phase. Letting y_{ij} denote the value of Y for the j-th unit in the second phase sample in stratum i, the finite population mean, $\bar{Y}$, is estimated by

$$\hat{\bar{Y}} = \sum_{i=1}^{L} w_i \bar{y}_i,$$

where $w_i = n_i'/n'$ and $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$.

Let $\sigma(n_i')$ and $\sigma(n_i)$ denote, respectively, the set of values for first phase and second phase sample units in stratum i, $\boldsymbol{n}' = (n_1', \ldots, n_L')$ and $\sigma(\boldsymbol{n}')$ the set of values for all first phase sample units. Also, let $\bar{y}_{n'}$ be the mean of the values in $\sigma(\boldsymbol{n}')$, $\bar{y}_i'$ the sample mean of $\sigma(n_i')$, $s_i'^2 = \sum_{j=1}^{n_i'} (y_{ij} - \bar{y}_i')^2/(n_i' - 1)$ the sample variance of $\sigma(n_i')$, $S_i^2 = \sum_{j=1}^{N_i} (Y_{ij} - \bar{Y}_i)^2/(N_i - 1)$ the population variance in stratum i and S^2 the analogous finite population variance. It is assumed throughout that n' is sufficiently large that $Pr(n_i' = 0)$ is negligible. Noting that $1 \le n_i \le n_i'$,

$$E(\hat{\bar{Y}}) = E_{\sigma(n')}\{E(\hat{\bar{Y}} \mid \sigma(\boldsymbol{n}'))\} = \bar{Y}$$

and

$$\begin{aligned} V(\hat{\bar{Y}}) &= V_{\sigma(n')}E\{\hat{\bar{Y}} \mid \sigma(\boldsymbol{n}')\} \\ &\quad + E_{\sigma(n')}\{V(\hat{\bar{Y}} \mid \sigma(\boldsymbol{n}'))\} \\ &= V_{\sigma(n')}(\bar{y}_{n'}) \\ &\quad + E_{\sigma(n')}\left\{\sum_{i=1}^{L} w_i^2 s_i'^2 \left(\frac{1}{n_i} - \frac{1}{n_i'}\right)\right\} \qquad (1.1) \\ &= S^2\left(\frac{1}{n'} - \frac{1}{N}\right) \\ &\quad + E_{n'}\left\{\sum_{i=1}^{L} w_i^2 S_i^2 \left(\frac{1}{n_i} - \frac{1}{n_i'}\right)\right\}. \qquad (1.2) \end{aligned}$$

We assume the linear cost function

$$C = c'n' + \sum_{i=1}^{L} c_i n_i, \qquad (1.3)$$

where c' is the per unit cost associated with sampling a first phase unit, and c_i is the per unit cost of measuring Y in stratum i. The sample sizes, n' and the n_i, are selected subject to fixed total cost or to fixed total expected cost.

In this paper we compare three double sampling designs, differentiated by the way that the sample sizes, n' and the n_i, are chosen. We also compare these methods with a simple random sample having the same fixed total cost.

The alternative designs are presented in Section 2 and compared in Section 3. Section 4 presents the results of an investigation of the effect on sample size selection of misspecification of an important design parameter.

[1] R.P. Treder, Statistical Sciences, Inc. Seattle, Washington; J. Sedransk, State University of New York at Albany, Albany, New York.

2. ALTERNATIVE METHODS

2.1 Proportional Allocation

For proportional allocation, $n_i = nw_i$ where $n = \sum_{i=1}^{L} n_i$. Then, using (1.2), the variance of $\hat{\bar{Y}}$ under proportional allocation, V_P, can be shown to be

$$V_P = \left(\frac{1}{n'} - \frac{1}{N}\right) S^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) \sum_{i=1}^{L} W_i S_i^2, \quad (2.1)$$

where $W_i = N_i/N$ is the population proportion of units in stratum i. Substituting $n_i = nw_i$ in (1.3), the expected total cost is

$$\bar{C}_P = c'n' + cn, \quad (2.2)$$

where $c = \sum_{i=1}^{L} W_i c_i$. Choosing n' and n to minimize (2.1) subject to fixed total expected cost, $\bar{C}_P = C^*$, yields

$$n' = \frac{C^*}{c' + \sqrt{c'cG}}, \quad (2.3a)$$

$$n = \frac{C^*}{c + \sqrt{c'c/G}}, \quad (2.3b)$$

where $G = S_W^2/S_B^2$, $S_W^2 = \sum_{i=1}^{L} W_i S_i^2$ and $S_B^2 = S^2 - S_W^2$.

Using (2.3),

$$V_P = \frac{1}{C^*}\left\{\left(c' + \sqrt{c'cG}\right) S_B^2 + \left(c + \sqrt{c'c/G}\right) S_W^2\right\} - \frac{S^2}{N}. \quad (2.4)$$

2.2 Rao's Allocation

Rao (1973a,b) proposes selecting $n_i = \nu_i n_i'$ where the ν_i $(0 < \nu_i \leq 1)$ are constants fixed in advance of sampling. Using this allocation in (1.2), the variance of $\hat{\bar{Y}}$ under Rao's allocation, V_R, can be shown to be

$$V_R = \left(\frac{1}{n'} - \frac{1}{N}\right) S^2 + \frac{1}{n'} \sum_{i=1}^{L} W_i S_i^2 \left(\frac{1}{\nu_i} - 1\right). \quad (2.5)$$

The corresponding expected cost, $\bar{C}_R$, is

$$\bar{C}_R = c'n' + n' \sum_{i=1}^{L} c_i \nu_i W_i. \quad (2.6)$$

The ν_i which minimize (2.5) subject to $\bar{C}_R = C^*$ satisfy

$$\nu_i^0 = \frac{S_i\sqrt{c'}}{S_B\sqrt{c_i}}, \quad (2.7)$$

provided that the right side of (2.7) does not exceed 1 for any i. Otherwise, an algorithm is required to determine the optimal ν_i (see Rao 1973a,b). Since Rao minimizes the *unconditional* variance, the optimal ν_i do not depend on the observed n_i'. After determining the ν_i, n' is obtained from (2.6). Assuming that $\nu_i^0 \leq 1$ for each i,

$$V_R = \frac{1}{C^*}\left(\sum_{i=1}^{L} W_i S_i \sqrt{c_i} + S_B\sqrt{c'}\right)^2 - \frac{S^2}{N}. \quad (2.8)$$

2.3 Optimal Allocation

The optimal allocation of the sample sizes can be obtained by minimizing (1.2) directly. For *fixed* n' and $\boldsymbol{n}'$, select the n_i to minimize

$$\sum_{i=1}^{L} w_i^2 S_i^2 \left(\frac{1}{n_i} - \frac{1}{n_i'}\right), \quad (2.9)$$

subject to fixed remaining cost, $C^* - c'n' = \sum_{i=1}^{L} c_i n_i$ and $n_i \leq n_i'$. An algorithm is required to determine the optimal n_i given the n_i'; see Hughes and Rao (1979) and Treder (1989). One may find the optimal value of n' by evaluating (1.2) for a sequence of "trial" values of n'. For each such n', one estimates the expected value of (2.9) using Monte Carlo sampling of $\boldsymbol{n}'$ (see Booth and Sedransk (1969) and Treder (1989)). Note that the algorithm needed to find the optimal n_i is straightforward, and the Monte Carlo sampling of $\boldsymbol{n}'$ given n' is simple. There are several differences between the optimal allocation and Rao's allocation. In the former, total costs will not exceed C^* while in the latter the allocation only guarantees that the budget will be satisfied on the average. In the latter, the ν_i are fixed in repeated sampling while in the former, allocation of the n_i depends on the observed $\boldsymbol{n}'$. Of course, additional effort (*i.e.*, the Monte Carlo sampling) is needed to find the optimal allocation. In contrast to the optimal allocation, Rao's method permits selection of the second phase sampling fractions *prior* to observing the n_i' (see (2.7)). See Sections 3 and 4 for additional discussion.

3. COMPARISONS

3.1 Proportional *vs* Rao's Allocation

Assuming that $\nu_i^0 \leq 1$, $i = 1, \ldots, L$, and using (2.4) and (2.8), it can be shown that

$$V_P - V_R = \frac{1}{C^*}\left(S_W - \frac{\bar{S}_c}{\sqrt{c}}\right) \times \left\{2S_B\sqrt{c'c} + c\left(S_W + \frac{\bar{S}_c}{\sqrt{c}}\right)\right\}, \quad (3.1)$$

Table 1

Percent decrease in variance, R, for Rao's allocation compared to proportional allocation for a selection of textbook examples

Reference	L	S^2	S_W^2	G	C^*	R for $c' = 1$ and $c =$ 1	2	5	25
Cochran (1977), p. 93	2	52,448	17,646	0.51	30	15.1	16.6	18.6	21.6
Hansen *et al.* (1953), p. 205	3	2,835,856	1,467,632	1.07	1,000	48.7	55.1	62.3	70.9
Sukhatme *et al.* (1984), p. 118	4	72,238	23,509	0.48	100	11.8	13.5	15.7	18.9
Cochran (1977), p. 111	7	619	343	1.25	1,000	11.2	11.7	12.4	13.7
Hansen *et al.* (1953), p. 202	8	47,393	45,595	25.36	1,000	10.5	11.0	11.5	12.0
Hansen *et al.* (1953), p. 202	11	47,393	44,974	18.59	1,000	22.9	24.1	25.4	26.7
Hansen *et al.* (1953), p. 235	11	2,039,184	820,722	0.67	1,000	21.3	24.8	29.1	35.1
Hansen *et al.* (1953), p. 202	12	47,393	40,252	5.64	1,000	16.7	18.3	19.8	21.6

Note: $R = 100(V_P - V_R)/V_P$ with V_P and V_R defined in (2.1) and (2.5) and C^* is the total budget. The cost function is defined in (1.3), and the variances (S^2, S_W^2, G) in (2.3).

where $\bar{S}_c = \sum_{i=1}^{L} W_i S_i \sqrt{c_i}$. Recalling that $c = \sum_{i=1}^{L} W_i c_i$ and using the Cauchy-Schwarz inequality, $S_w - \bar{S}_c/\sqrt{c} \geq 0$. Thus, as expected, $V_P - V_R \geq 0$. Defining $\bar{S} = \sum_{i=1}^{L} W_i S_i$ and $\bar{S}_\gamma = \sum_{i=1}^{L} W_i S_i \sqrt{\gamma_i}$ with $\gamma_i = c_i / \sum_{j=1}^{L} W_j c_j$, and using (3.1), it can be shown that

$$V_P - V_R = \frac{1}{C^*}\left\{2\sqrt{c'c}\left(\frac{S_B}{S_W + \bar{S}}\right) + c\right\} \times \left(S_W^2 - \bar{S}^2\right)$$

$$+ \frac{1}{C^*}\left\{2\sqrt{c'c}S_B + c(\bar{S} + \bar{S}_\gamma)\right\} \times \left(\bar{S} - \bar{S}_\gamma\right). \tag{3.2}$$

The first term in (3.2) is the reduction in variance if all sampling costs are equal while the second term in (3.2) is the reduction if all strata variances are equal. As expected, if $c_i = c$ and $S_i = S$, $V_P = V_R$.

We present in Table 1, the values of $R = 100(V_P - V_R)/V_P$ corresponding to a set of textbook examples with $c_i = c$. In parallel columns we give characteristics of the associated populations (L, S^2, S_W^2, $G = S_W^2/S_B^2$) and C^* together with the values of R corresponding to $c/c' = 1$, 2, 5 and 25. This set of examples represents a broad range of conditions where stratified sampling may be used. For a given value of c, the range of values of R indicates the wide range of gains that may be attained. It is clear from Table 1 that there may be substantial reductions in variance if one uses Rao's allocation, even when second phase strata sampling costs are equal and in situations when the stratification is not especially effective (note the large values of G for three examples). As c increases, R increases at a rate that is approximately constant (see Table 1).

3.2 Comparisons with Simple Random Sampling

For comparability with Rao and proportional allocations, assume a simple random sample of size n^* with expected cost $n^* \sum_{i=1}^{L} W_i c_i = n^* c$ (see (1.3)). Thus, for a fixed expected cost, C^*, $n^* = C^*/c$ and

$$\text{Var}(\bar{y}_{n^*}) = S^2\left(\frac{c}{C^*} - \frac{1}{N}\right) \equiv V_S, \tag{3.3}$$

where $\bar{y}_{n^*}$ is the sample mean. Using (2.4) and (3.3),

$$V_S - V_P = \frac{1}{C^*}\left\{(c - c')S_B^2 - 2S_B S_W\sqrt{c'c}\right\}. \tag{3.4}$$

It can be shown that $V_S - V_P \geq 0$ if, and only if,

$$\frac{c}{c'} \geq \left(\sqrt{G} + \sqrt{1 + G}\right)^2 = LB_P, \tag{3.5}$$

where $G = S_W^2/S_B^2$. Using (2.8) and (3.3),

Table 2

Percent decrease in variance for proportional (R_P) and Rao's (R_R) allocation compared to simple random sampling for a selection of textbook examples

Reference	L	LB_P	LB_R	R_P			R_R		
				$c = 1$	5	25	$c = 1$	5	25
Cochran (1977), p. 93	2	3.8	2.6	−177.9	11.9	45.7	−136.0	28.3	57.4
Hansen *et al.* (1953), p. 205	3	6.1	1.1	−102.8	−6.1	26.4	−4.1	59.9	78.6
Sukhatme *et al.* (1984), p. 118	4	3.7	2.7	−132.8	12.8	46.6	−105.3	26.5	56.7
Hansen *et al.* (1953), p. 210	4	17.4	0.7	−127.7	−21.3	3.6	23.0	58.9	69.4
Cochran (1977), p. 111	7	6.8	4.5	−197.8	−9.8	23.3	−164.5	3.9	33.8
Hansen *et al.* (1953), p. 202	8	103.4	5.6	−38.2	−14.1	−4.0	−23.7	−0.9	8.5
Hansen *et al.* (1953), p. 202	11	76.4	1.7	−44.0	−15.6	−3.9	−11.0	13.7	23.8
Hansen *et al.* (1953), p. 235	11	4.5	2.2	−105.8	4.0	37.9	−62.0	32.0	59.7
Hansen *et al.* (1953), p. 202	12	24.5	4.0	−71.6	−19.9	0.2	−42.8	3.9	21.8

Note: Using (2.4), (2.8) and (3.3), $R_P = 100(V_S - V_P)/V_S$, $R_R = 100(V_S - V_R)/V_S$, and (LB_P, LB_R) are defined in (3.5) and (3.7). For these examples, $c' = 1$ and C^*, the total budget for each of the methods, is as in Table 1.

$$V_S - V_R = \frac{c}{C^*}\left\{S^2 - \left(\bar{S}_\gamma + S_B\sqrt{c'/c}\right)^2\right\}, \qquad (3.6)$$

where it is again assumed that $v_i^0 \leq 1$ for all i (see (2.7)). It is easily seen that $V_S - V_R \geq 0$ if, and only if,

$$\frac{c}{c'} \geq \frac{S_B^2}{(S - \bar{S}_\gamma)^2} = LB_R. \qquad (3.7)$$

In practice, one will estimate LB_P and LB_R in (3.5) and (3.7) and compare them with the cost ratio, c/c′, to decide if it will be beneficial to use double sampling with proportional or Rao's allocation rather than simple random sampling. In Table 2 we present the values of LB_P and LB_R for each of the examples in Table 1. We also include for $c = 1$, 5, and 25 the values of R, the per cent reduction in variance accruing from using a double sampling method rather than simple random sampling. As noted above, this set of examples represents a broad range of conditions where stratified sampling may be used. For a given value of c, the range of values of R_P and R_R indicates the wide range of gains (over simple random sampling) that may be obtained.

While $LB_P \geq LB_R$ is true in general, $LB_P \gg LB_R$ for many of the examples. The results point to potentially large gains for double sampling, especially using Rao's allocation, when c/c' is large. Conversely, if c/c' is relatively small, gains are modest and, in some cases, simple random sampling is preferred. This argues for careful estimation of LB_P, LB_R and c/c'.

3.3 Optimal *vs* Rao's Allocation

To compare the optimal allocation with that proposed by Rao, we have considered a wide range of values of the design parameters c', S^2 and $\{(c_i, S_i^2, W_i) : i = 1, \ldots, L\}$. We took $C^* = 1{,}000$ and considered $L = 2$ and 3. The values of the design parameters for $L = 2$ are listed in Table 3. Note that for these examples $G = S_W^2/S_B^2$ ranges from 0.01 to 10.00. We assume throughout that N is sufficiently large that S^2/N in (1.2) is negligible.

Table 3

Values of design parameters for the case of $L = 2$ strata

Parameter	Values
c'	0.125, 0.250, 0.500, 1.000
c_1	1, 4, 16
c_2	16
W_1	0.5, 0.6, 0.7, 0.8, 0.9
S^2	70.4, 128, 704
S_1^2	1, 4, 16, 64
S_2^2	64

Note: All 720 combinations of the above parameters were used. In addition, we also studied all arrangements of c', S^2, and S_1^2 as above together with

(a) $c_1 = 16$; $c_2 = 1, 4, 16$ and $W_1 = 0.5, 0.6, 0.7, 0.8, 0.9$,

(b) $W_1 = 0.1, 0.2, 0.3, 0.4$; $c_1 = 1, 4, 16$; $c_2 = 16$, and

(c) $W_1 = 0.1, 0.2, 0.3, 0.4$; $c_1 = 16$; $c_2 = 1, 4, 16$.

To ensure comparability of the two allocations we proceeded as indicated below for each specification of the design parameters.

1. Fix a single value of n'. We used both the value of n' identified as best using (a) Rao's method and (b) the optimal allocation.
2. From each of K Monte Carlo replications ($K = 200$ or 500) we obtain $\boldsymbol{n}' = (n_1', \ldots, n_L')$ and then $\boldsymbol{n} = (n_1, \ldots, n_L)$ using the optimal allocation and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_L)$ from Rao's method. For the latter we use the algorithm which makes appropriate adjustments when the right side of (2.7) exceeds 1 for one or more strata.

   Since neither $\boldsymbol{n}$ from the optimal method nor $\boldsymbol{n}$ from Rao's method ($n_i = \nu_i n_i'$) are necessarily integers we round the n_i and adjust them so that for each sample the budget is satisfied (up to the approximation necessitated by having integer values of n' and $\boldsymbol{n}$). We found that if these adjustments were not made there were anomalous results where the variance of $\hat{\bar{Y}}$ using Rao's allocation was less than the corresponding variance using the optimal allocation. This occurred when the total cost associated with Rao's procedure was larger than that for the optimal procedure.
3. To obtain estimates, $\bar{V}_{(c)O}$ and $\bar{V}_{(c)R}$, of the conditional variances, $E_{n'}\{\sum_{i=1}^{L} w_i^2 S_i^2 (1/n_i - 1/n_i')\}$, corresponding to the optimal and Rao's allocation, we used the average of $\sum_{i=1}^{L} w_i^2 S_i^2 (1/n_i - 1/n_i')$ over the K replications. The estimates of the unconditional variance, $\mathrm{Var}(\hat{\bar{Y}})$, in (1.2) are denoted by $\bar{V}_{(u)O}$ and $\bar{V}_{(u)R}$ where $\bar{V}_{(u)R} = \bar{V}_{(c)R} + (S^2/n')$.

The precision of these estimates was assessed by estimating the standard errors and coefficients of variation of $\bar{V}_{(u)R}$ and $\bar{V}_{(c)R}$. All of the standard errors were less than 0.0022. The coefficients of variation for $\bar{V}_{(u)R}$ and $\bar{V}_{(c)R}$ were below 0.0074 and 0.023, respectively. Thus, $\bar{V}_u$ and $\bar{V}_c$ provide precise estimates of the unconditional and conditional variances.

We present in Table 4 estimates of the per cent increase in the average unconditional variance for Rao's allocation, $I_u = 100\,(\bar{V}_{(u)R} - \bar{V}_{(u)O})/\bar{V}_{(u)O}$, for some of the design parameters listed in Table 3. We include results only for the value of n' identified as optimal by the optimal procedure. These results are typical of those seen for the other specifications in Table 3, those that we considered for the case $L = 3$, and those which use the value of n' identified as optimal by Rao's method. It is clear from Table 4 that improvements in precision are small, ranging from none to about 4%.

We obtained somewhat similar results for the per cent increase in the *conditional* variance for Rao's allocation, $I_c = 100(\bar{V}_{(c)R} - \bar{V}_{(c)O})/\bar{V}_{(c)O}$, where $\bar{V}_{(c)R}$ and $\bar{V}_{(c)O}$ are obtained by estimating $E\{\sum_{i=1}^{L} w_i^2 S_i^2 (1/n_i - 1/n_i')\}$ using, respectively, Rao's allocation and the optimal allocation. The results, based on 200 Monte Carlo replications and presented using boxplots in Treder (1989, Figures 2.8.2 and C.1 - C.3), can be summarized as follows. For *all* parameter specifications, the medians of the distributions of I_c are near 0. Most of the values of I_c are small: about 95% of the parameter specifications have distributions of I_c with third quartiles less than 10%. However, occasionally, there are large values of I_c: about 15% of the parameter specifications have the maximal value of I_c larger than 20%.

Table 4

Percent increase, I_u, in the average unconditional variance $\bar{V}_{(u)}$ for Rao's allocation compared to optimal allocation for a selection of design parameters with $S^2 = 70.4$, $S_2^2 = 64$, $c_2 = 16$ and $c' = 1$

S_1^2	G	c_1 = 16	c_1 = 4	c_1 = 1
	a. $(W_1, W_2) = (.9, .1)$			
64	10.000	0.0	0.4	1.4
16	0.419	0.1	0.1	0.1
4	0.166	0.1	0.1	0.4
1	0.116	0.1	0.3	0.8
	b. $(W_1, W_2) = (.7, .3)$			
64	10.000	0.0	0.7	3.6
16	0.760	0.0	0.2	0.7
4	0.455	0.1	0.3	1.4
1	0.394	0.0	0.7	0.9
	c. $(W_1, W_2) = (.5, .5)$			
64	10.000	0.0	1.0	4.1
16	1.316	0.0	0.4	0.9
4	0.934	0.0	0.6	1.8
1	0.858	0.0	0.2	0.0

Note: $I_u = 100(\bar{V}_{(u)R} - \bar{V}_{(u)O})/\bar{V}_{(u)O}$. See the note to Table 1 for definitions of the costs and variances.

These results can be explained, in part, by defining the *optimal* second phase sample size in stratum i by $n_i = \xi_i(\boldsymbol{n}') \cdot n_i'$ where the dependence of n_i on the observed $\boldsymbol{n}'$ is emphasized by writing $\xi_i(\boldsymbol{n}')$ and $0 < \xi_i(\boldsymbol{n}') \leq 1$. Then, one may find the optimal allocation by choosing the $\xi_i(\boldsymbol{n}')$ to minimize (for fixed $\boldsymbol{n}'$)

$$\frac{1}{n'} \sum_{i=1}^{L} \frac{w_i S_i^2}{\xi_i(\boldsymbol{n}')}, \tag{3.8}$$

subject to $\sum_{i=1}^{L} c_i n_i' \cdot \xi_i(\boldsymbol{n}') = C^* - c'n'$ (see 2.9).

By contrast, for the Rao allocation, for fixed n', one selects the ν_i to minimize

$$\frac{1}{n'}\sum_{i=1}^{L}\frac{W_iS_i^2}{\nu_i}, \tag{3.9}$$

subject to $n'\sum_{i=1}^{L} c_iW_i\nu_i = C^* - c'n'$, *i.e.* fixed expected cost.

Minimizing (3.8) rather than (3.9) will yield a smaller conditional and, thus, unconditional variance. However, when n' is large, the difference between (3.8) and (3.9) will be small.

3.4 Recommendations

Given reasonable estimates of the design parameters, one should first compare the cost ratio, c/c', with lower bounds, LB_P and LB_R, in (3.5) and (3.7) to see whether it is preferable to use double sampling rather than simple random sampling. These assessments must be done carefully because inappropriate use of double sampling may result in a *reduction* in precision. If there are good estimates of the design parameters, using Rao's allocation is preferable to proportional allocation.

Given the importance of adhering to a fixed budget we recommend the use of a modification of Rao's procedure:

> Use Rao's procedure to find the "optimal" value of n'. Then, given the n_i', use the optimal allocation procedure (*i.e.* minimize (2.9)) to find the n_i. This method guarantees that the budget will be satisfied for each sample, preserves most of the (small) gain in precision from using the optimal allocation and is easy to implement.

An alternative is to use Rao's procedure to find the "optimal" values of n' and the ν_i. Then implement an algorithm to round and modify the n_i ($n_i = \nu_i n_i'$) to ensure that the budget is satisfied for each sample. Unfortunately, it is difficult to develop the part of the algorithm needed to insure against cost overruns.

However, to avoid the large values of the proportional error in the *conditional* variance (*i.e.* I_c) that occur occasionally, one must use the *optimal* values of n' and the n_i.

Each of these methods requires knowledge of some design parameters. For Rao's allocation, the optimal ν_i require that the W_i and S_i^2 be specified. One can see from (2.9) that for the optimal allocation, the optimal n_i depend on the S_i^2 but not on the W_i. However, the optimal choice of n' requires that the W_i be specified. Alternatively, Srinath (1971) and Rao (1973a) have suggested a procedure which requires knowledge of the S_i^2 but not the W_i. Clearly, Rao's allocation requires the greatest knowledge of the design parameters and Srinath's procedure the least. Since the choice of n' is, typically, robust to misspecification of design parameters (see, *e.g.*, Sedransk 1965, Section 4.2.3), the optimal method may work well in the circumstances for which Srinath's method was designed.

4. SENSITIVITY OF ALLOCATIONS TO ESTIMATION OF DESIGN PARAMETERS

The preceding analysis assumes that the sample allocations are minimally affected by errors in the specification of the design parameters. In this section we investigate, in a simple case, the effect on $\mathrm{Var}(\hat{\bar{Y}})$ of the misspecification of an important design parameter. With proportional allocation, the choice of n' and n depends only on $G = S_W^2/S_B^2$, c' and c (see (2.3)). Estimating G by $\hat{G}$ and substituting the resulting values of n' and n from (2.3) in (2.1),

$$\frac{V_P(\hat{\bar{Y}})_{\hat{G}}}{S_W^2} = \frac{1}{C^*}\left(\frac{c' + \sqrt{c'c\hat{G}}}{G} + c + \sqrt{c'c/\hat{G}}\right) - \frac{1}{N}\left(1 + \frac{1}{G}\right), \tag{4.1}$$

where G is the correct value of S_W^2/S_B^2 and $\hat{G}$ is used only to determine n' and n.

Table 5

Per cent increase in unconditional variance, I, for proportional allocation when G is estimated by $\hat{G}$. $C^* = 1{,}000$, $c' = 1$ and $c_1 = c_2 = 16$

G	$\hat{G}$								
	1/100	1/36	1/16	1/4	1	4	16	36	100
1/100	0.0	6.0	19.8	69.3	174.9	389.8	817.3	817.3	817.3
1/36	6.2	0.0	4.4	33.1	103.8	251.4	547.7	547.7	547.7
1/16	21.9	3.9	0.0	12.1	57.1	156.2	357.9	357.9	357.9
1/4	71.7	30.7	12.6	0.0	11.8	51.2	138.5	138.5	138.5
1	128.9	67.2	37.3	7.3	0.0	7.5	35.9	35.9	35.9
4	179.1	101.6	63.3	22.3	5.7	0.0	5.4	5.4	5.4
16	210.2	123.4	80.3	33.5	12.9	2.3	0.0	0.0	0.0
36	220.4	130.7	86.0	37.4	15.7	4.0	0.0	0.0	0.0
100	225.9	134.6	89.1	39.5	17.2	4.9	0.0	0.0	0.0

Note: I is defined in (4.3), $G = S_W^2/S_B^2$ and the cost function is given by (1.3).

The optimal value of Var $(\hat{\bar{Y}})$ (*i.e.* when using G) in (2.4) can be expressed as

$$\frac{V_P(\hat{\bar{Y}})_G}{S_W^2} = \frac{1}{C^*}\left(\frac{c'}{G} + c + 2\sqrt{c'c/G}\right) - \frac{1}{N}\left(1 + \frac{1}{G}\right). \qquad (4.2)$$

If $(1/N)(1 + 1/G)$ is negligible, the per cent increase in variance due to estimating G, $I = 100\{V_P(\hat{\bar{Y}})_G - V_P(\hat{\bar{Y}})_G\} / V_P(\hat{\bar{Y}})_G$, is, from (4.1) and (4.2),

$$I = \frac{(1 - G) + \sqrt{c/c'}\{\sqrt{\hat{G}} - 2\sqrt{G} + (G/\hat{G})\}}{(1 + \sqrt{cG/c'})^2} \times 100. \qquad (4.3)$$

Note that (4.3) depends only on G, $\hat{G}$ and c/c'.

We present in Table 5 the values of I for $C^* = 1{,}000$, $c' = 1$, $c_1 = c_2 = 16$ and nine values of G and $\hat{G}$. The following conclusions are based on the results in Table 2.10.1 of Treder (1989) which includes additional values of G and $\hat{G}$. As long as $\hat{G}$ is within the interval $[G/4, 4G]$, using $\hat{G}$ to find (n', n) increases the variance by no more than 15%, typically less. If $\hat{G}$ is in the interval $[G/2, 2G]$, the increase in variance due to misspecification is about 4% or less. As G increases, the increase in variance associated with such intervals (*e.g.*, $[G/4, 4G]$) decreases. This happens because for large G, one has $n' = n$ and both $\hat{G}$ and G yield the same allocation. One manifestation of this result is the array of zeros in the lower right corner of Table 5. When G is small, that is when stratification is good, the sample allocation is more sensitive to incorrect specification of G than when G is large. These findings are little influenced by the values assigned to $c_1 = c_2$. In summary, for proportional allocation, fairly large misspecifications of the design parameter (G) lead to relatively small increases in variance.

REFERENCES

BOOTH, G., and SEDRANSK, J. (1969). Planning some two-factor comparative surveys. *Journal of the American Statistical Association*, 64, 560-573.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd Ed.). New York: John Wiley.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*, (Vol. 1). New York: John Wiley.

HUGHES, E., and RAO, J.N.K. (1979). Some problems of optimal allocation in sample surveys involving inequality constraints. *Communications in Statistics - Theory and Methods A*, 8(15), 1551-1574.

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

RAO, J.N.K. (1973b). On double sampling for stratification and analytical surveys. *Biometrika*, 60, 669.

SEDRANSK, J. (1965). A double sampling scheme for analytical surveys. *Journal of the American Statistical Association*, 60, 985-1004.

SRINATH, K.P. (1971). Multiphase sampling in nonresponse problems. *Journal of the American Statistical Association*, 66, 583-586.

SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S., and ASOK, C. (1984). *Sampling Theory of Surveys with Applications*, (3rd Ed.). Ames, IA: Iowa State University Press.

TREDER, R.P. (1989). Some problems in double sampling for stratification. Unpublished Ph.D. dissertation, University of Washington.