DESIGNING SEVERAL FACTOR ANALYTICAL SURVEYS

Gordon Booth and J. Sedransk

Iowa State University

1. Introduction

In many sample surveys the principal objective is to compare several sectors of a finite population. Specifically, there may be several factors (criteria of classification) of interest and each of these factors may have been divided into several categories. Then, for each factor, one may wish to compare these categories. For instance, a proposed study of hospitals in Iowa includes the factors "size of hospital" and "type of ownership." To take a simple example, one may wish to contrast large and small hospitals as well as public and private ones. (Here, for the factor "size of hospital," there are two categories - large and small.) Assuming this general specification of the problem, one must proceed to suggest comparisons of interest, and appropriate estimators. Then it is possible to select the sample so that specified precision for the estimates of the comparisons is attained at minimum cost, or maximum precision is achieved subject to a given budget.

For simplicity of presentation, attention is restricted to two-factor studies with two categories for each criterion of classification. For many of the topics discussed, extensions to more complex surveys are immediately clear.

In Section 2, it is assumed that one may select an independent random sample in each of the population sectors to be studied. (Thus, using the illustrative example, one may choose a random sample of Iowa's small, public hospitals independently of the large, private ones.) Then, utilizing a general type of comparison, a procedure to obtain the optimal sample size allocation is given. Since electronic computer algorithms may be needed to calculate the optimal allocation, approximate solutions are suggested. In a numerical study these approximate solutions are contrasted with the corresponding exact solutions.

If one cannot select independent samples in the subpopulations under study, a double sampling procedure may be feasible. Utilizing the results obtained in Section 2, a double sampling scheme is given in Section 3. This entails the formulation of a complex sampling rule for selecting the main sample from the large preliminary sample. A procedure for choosing the optimal sizes for the two samples (assuming a given budget) is also presented.

2. Single-phase Sampling

2.1. Comparisons

The two factors under investigation, each having two categories, may be represented by a 2 x 2 table with the (i,j)-th cell denoting the i-th category of the first factor, and the j-th category of the second factor. There are N_{ij} elements in the (i,j)-th cell (i,j = 1, 2), from which a random sample of size n_{ij} is selected. The sample mean is denoted by $\bar{y}_{ij}$, and the (population) within-cell variance by σ^2_{ij}. Let $N_{i.} = \sum_j N_{ij}$, $N_{.j} = \sum_i N_{ij}$, $N = \sum_i \sum_j N_{ij}$, with analogous definitions for $n_{i.}$, $n_{.j}$, and n.

The type of comparison to be employed depends on the assessment of the presence of interaction between the factors. If the interaction is deemed to be sufficiently small so that it may be neglected, and if a linear model is assumed to provide an adequate representation of the population, the contrasts of interest are readily apparent from the assumed model. Making such assumptions, Sedransk (1967) has obtained optimal sample size allocations which are applicable to studies where equal precision is required for all comparisons. He considers the two, three and four-factor cases where each factor has two categories of interest to the investigator.

However, if the interaction effect is large, and if one wishes to make an overall comparison, the choice of type of comparison is not so evident. Denoting the first factor by α and the second by τ, the two categories for each factor may be compared by considering

$$D_\alpha = \{N_{.1}(\mu_{11} - \mu_{21})/N\} + \{N_{.2}(\mu_{12} - \mu_{22})/N\}$$

and

$$D_\tau = \{N_{1.}(\mu_{11} - \mu_{12})/N\} + \{N_{2.}(\mu_{21} - \mu_{22})/N\} \qquad (2.1.1)$$

where μ_{ij} is the population mean for the (i,j)-th cell.

Unbiased estimators for D_α and D_τ are given by

$$\hat{D}_\alpha = \{N_{.1}(\bar{y}_{11} - \bar{y}_{21})/N\} + \{N_{.2}(\bar{y}_{12} - \bar{y}_{22})/N\}$$

and

$$\hat{D}_\tau = \{N_{1.}(\bar{y}_{11} - \bar{y}_{12})/N\} + \{N_{2.}(\bar{y}_{21} - \bar{y}_{22})/N\}. \qquad (2.1.2)$$

Rewriting $\hat{D}_\alpha$ as

$$\left(\frac{N_{.1}}{N}\bar{y}_{11} + \frac{N_{.2}}{N}\bar{y}_{12}\right) - \left(\frac{N_{.1}}{N}\bar{y}_{21} + \frac{N_{.2}}{N}\bar{y}_{22}\right), \qquad (2.1.3)$$

it may be noted that each of the two terms in (2.1.3) is an estimator similar to one suggested by Yates [1960, p. 134, (2)], except that, in Yates' example, some of the weights $\{N_{i.}, N_{.j}\}$ are to be estimated from the sample.

The other estimator cited by Yates [1960, p. 134, (1)] as appropriate if interaction is present, can be obtained from (2.1.2) by setting $N_{.j} = N_{i.} = N/2$ for $(i,j = 1, 2)$. Finally, all results obtained by using (2.1.2) are applicable to <u>any</u> pre-specified choice of weights $\{W_1, 1-W_1, W_2, 1-W_2\}$ with W_1 replacing $N_{.1}/N$, $(1-W_1)$ replacing $N_{.2}/N$, etc.

The "proportionately weighted estimators" given by (2.1.2) are considered extensively in the sequel. They often provide a reasonable mode of comparison when overall contrasts are required. Also, the difficulties in obtaining optimal sample size allocations are well illustrated by assuming comparisons of this form. As noted above, the choice of particular (pre-specified) weights in (2.1.2) does not affect the ensuing analysis.

In some situations, it may be preferable to consider "simple effects" rather than composite comparisons such as (2.1.1). (Here, the "simple effects" are $\mu_{11} - \mu_{12}$, $\mu_{21} - \mu_{22}$, $\mu_{11} - \mu_{21}$ and $\mu_{12} - \mu_{22}$.) Such an approach is certainly more reasonable if the main objective is to select (separately) for each category of the α-factor the "better" category of the τ-factor. However, it may be unsatisfactory to look only at the simple effects if one wishes to obtain overall appraisals.

2.2. Procedures to obtain optimal allocations

For the proportionately weighted estimators, it is desired to find those values of the $n_{ij} \geq 0$ that

I. minimize $\sum_i \sum_j c_{ij} n_{ij}$

subject to

$$\mathrm{Var}(\hat{D}_\alpha) = \sum_i \sum_j N_{.j}^2 \sigma_{ij}^2/(N^2 n_{ij}) = V_1$$

and

$$\mathrm{Var}(\hat{D}_\tau) = \sum_i \sum_j N_{i.}^2 \sigma_{ij}^2/(N^2 n_{ij}) = V_2 \qquad (2.2.1)$$

where c_{ij} is the cost of sampling an element in the (i,j)-th cell and V_1, V_2 are constants specified in advance.

This problem is equivalent to the one considered by Cochran (1963, pp. 123-124). Using standard calculus methods, there is no simple explicit solution for the optimal values of the n_{ij}, and Cochran presents a complex iterative method that may be used to obtain the solution.

Alternatively, one may reformulate problem I:

II. minimize $\sum_i \sum_j c_{ij} n_{ij}$

subject to

$$\mathrm{Var}(\hat{D}_\alpha) \leq V_1, \quad \mathrm{Var}(\hat{D}_\tau) \leq V_2,$$

$$0 \leq n_{ij} \leq N_{ij} \quad \text{for} \quad i,j = 1, 2. \qquad (2.2.2)$$

This is a convex programming problem, and numerical solutions may be obtained by using an appropriate electronic computer algorithm. For example, Hartley and Hocking (1963) describe a method of convex programming by tangential approximation.

Since the convex programming approach is the more general one, it is preferable to the specification denoted by I, but neither is completely adequate. The convex programming method depends on the availability of appropriate computer facilities, and the procedure described by Cochran is somewhat cumbersome to carry out. Moreover, neither approach yields explicit algebraic expressions for the optimal values of the n_{ij}. (Such expressions are very useful in planning a double sampling procedure of the type described in Section 3.) Hence, it appears profitable to explore an alternative formulation (III) of the problem, and compare the resultant optimal allocation with the corresponding one obtained by using the convex programming method. Thus, the $n_{ij} \geq 0$ are to be selected to

III. minimize $\sum_i \sum_j c_{ij} n_{ij}$

subject to

$$W_1 \mathrm{Var}(\hat{D}_\alpha) + (1-W_1)\mathrm{Var}(\hat{D}_\tau) = V^* \qquad (2.2.3)$$

where, to approximate problem formulation I or II, one might set $W_1 = V_2/(V_1 + V_2)$ and $V^* = 2V_1V_2/(V_1 + V_2)$.

Then, it is easily shown that the optimal value of n_{ij}, n_{ij}^*, is given by

$$n_{ij}^* = \left[\frac{\sigma_{ij}^2}{c_{ij}}\right]^{1/2} \left[W_1 N_{.j}^2 + (1-W_1)N_{i.}^2\right]^{1/2} \cdot \frac{\left\{\sum\sum \sigma_{ij}\left[c_{ij}\left\{W_1 N_{.j}^2 + (1-W_1)N_{i.}^2\right\}\right]^{1/2}\right\}}{N^2 V^*} \qquad (2.2.4)$$

For each of thirty numerical examples, a comparison of the allocation given by (2.2.4) with the corresponding optimal allocation from the convex programming approach (II) is given in Section 2.3.

For the "simple effects" estimators, one might choose those values of the n_{ij} that minimize $\Sigma\Sigma c_{ij} n_{ij}$ subject to $\mathrm{Var}(\bar{y}_{11} - \bar{y}_{12}) = \mathrm{Var}(\bar{y}_{21} - \bar{y}_{22}) = \mathrm{Var}(\bar{y}_{11} - \bar{y}_{21}) = \mathrm{Var}(\bar{y}_{12} - \bar{y}_{22}) = \bar{V}$. The optimal allocation can be obtained by standard methods.

2.3. Evaluation of an approximation to the optimal sample size allocation

In this section, it is desired to ascertain the efficacy of using the allocation given by (2.2.4) as an approximation to the optimal allocation of the n_{ij} obtained by using the "convex programming" approach (II). There is no loss of generality by taking $c_{ij} = 1$ for (i,j = 1, 2), and for simplicity, it is assumed that $V_1 = V_2 = \bar{V}'$. Then, the optimal allocation is obtained by using a convex programming algorithm to

$$\text{minimize} \quad \Sigma\Sigma n_{ij}$$

subject to

$$\mathrm{Var}(\hat{D}_\alpha) = \Sigma\Sigma N_{.j}^2 \sigma_{ij}^2/(N^2 n_{ij}) \leq \bar{V}'$$

$$\mathrm{Var}(\hat{D}_\tau) = \Sigma\Sigma N_{i.}^2 \sigma_{ij}^2/(N^2 n_{ij}) \leq \bar{V}'$$

$$0 \leq n_{ij} \leq N_{ij} \quad \text{for} \quad i,j = 1, 2. \qquad (2.3.1)$$

The approximation to this optimal allocation is obtained by

$$\text{minimizing} \quad \Sigma\Sigma n_{ij}$$

subject to

$$\{\mathrm{Var}(\hat{D}_\alpha) + \mathrm{Var}(\hat{D}_\tau)\}/2 = \bar{V}'. \qquad (2.3.2)$$

From (2.2.4), the approximate optimal allocation is given by

$$n_{ij}^* = \sigma_{ij}(N_{.j}^2 + N_{i.}^2)^{\frac{1}{2}} \{\Sigma\Sigma\sigma_{ij}(N_{i.}^2 + N_{.j}^2)^{\frac{1}{2}}\}/2N^2\bar{V}'. \qquad (2.3.3)$$

For each of thirty numerical examples, the following general procedure was used: (1) Applying the Hartley-Hocking technique to (2.3.1), the optimal allocation $\{n_{ij}\}$ was obtained; (2) The approximate optimal allocation $\{n_{ij}^*\}$ was calculated from (2.3.3); (3) $\mathrm{Var}(\hat{D}_\alpha)$ and $\mathrm{Var}(\hat{D}_\tau)$ were computed using each of the allocations given by (1) and (2) above; (4) The total sample size ($n = \Sigma\Sigma n_{ij}$) was computed for each of the allocations (1) and (2). Now define as follows:

$C_{(i)}$ = total cost (sample size) using the allocation given under (i) for i = 1, 2,

$V_{(2)} = \mathrm{Max}\,[\mathrm{Var}(\hat{D}_\alpha), \mathrm{Var}(\hat{D}_\tau)]$, computed using the allocation given under (2),

$$V_{(1)} = \begin{cases} \mathrm{Var}(\hat{D}_\alpha), \text{ computed using the allocation under (1), if } V_{(2)} = \mathrm{Var}(\hat{D}_\alpha) \\ \mathrm{Var}(\hat{D}_\tau), \text{ computed using the allocation under (1), if } V_{(2)} = \mathrm{Var}(\hat{D}_\tau), \end{cases}$$

$$S_{.j} = \Sigma_i \sigma_{ij}^2 \quad \text{and} \quad S_{i.} = \Sigma_j \sigma_{ij}^2,$$

$$R_1 = \mathrm{Max}\,[N_{.1}/N_{.2}, N_{.2}/N_{.1}] \text{ and}$$

$$R_2 = \mathrm{Max}\,[N_{1.}/N_{2.}, N_{2.}/N_{1.}],$$

$$R_3 = \mathrm{Max}\,[S_{.1}/S_{.2}, S_{.2}/S_{.1}] \text{ and}$$

$$R_4 = \mathrm{Max}\,[S_{1.}/S_{2.}, S_{2.}/S_{1.}];$$

(5) The following values were computed for each of the numerical examples:

(a) $P_c = (C_{(1)} - C_{(2)})(100)/C_{(1)}$ = percent decrease in cost by using (2.3.3) rather than the solution to (2.3.1).

(b) $P_v = (V_{(2)} - V_{(1)})(100)/V_{(1)}$ = percent increase in variance by using (2.3.3) rather than the solution to (2.3.1).

(c) $R_N = \mathrm{Max}\,[R_1/R_2, R_2/R_1]$

(d) $R_S = \mathrm{Max}\,[R_3/R_4, R_4/R_3]$.

It is easily shown that P_v and P_c are invariant under changes in scale. That is, if each of the N_{ij} is multiplied by a constant k_1, and each of the σ_{ij}^2 by a constant k_2, the values of P_v and P_c are not altered. Thus, the numerical examples presented in Table 1 cover a very wide range of values of the N_{ij} and σ_{ij}^2.

The results of the thirty numerical examples, for the most part, encourage the use of the approximate solution (2.3.3). For many of the examples, the allocations from (2.3.3) and (2.3.1) were essentially the same, and most of the values of P_v and P_c in Table 1 are small. (In choosing the examples, the objective was to identify those cases where P_v is large.) Note that large positive values of P_c are accompanied by large positive values of P_v. In such situations, $\mathrm{Max}\,[\mathrm{Var}(\hat{D}_\alpha), \mathrm{Var}(\hat{D}_\tau)]$ is much larger

Table 1

Values of P_v and P_c for thirty specifications* of the N_{ij} and σ^2_{ij}.

Example	N_{11}	N_{12}	N_{21}	N_{22}	σ^2_{11}	σ^2_{12}	σ^2_{21}	σ^2_{22}	P_c	P_v
1	1	1	1	1	1	1	1	1	0.00	0.00
2	1	1	1	1	1	1	1	50	0.00	0.00
3	1	1	1	1	1	1	1	1000	0.00	0.00
4	1	1	1	1	1	10	1	10	-0.06	-0.06
5	1	1	1	1	1	100	1	100	0.02	0.02
6	1	1	1	1	1	1000	1	1000	0.00	0.00
7	1	1	1	1	100	1	100	1	0.02	0.02
8	1	1	1	1	8	4	2	1	-0.06	-0.06
9	1	1	1	90	1	1	1	90	0.01	0.01
10	1	1	1	90	1	1	1	900	0.34	0.23
11	1	1	1	100	100	100	100	1	0.00	0.00
12	1	1	1	1000	1	1	1	1	0.17	0.02
13	1	10	1	10	1	1	1	1	0.62	5.24
14	1	100	1	100	1	1	1	1	1.89	10.05
15	1	100	1	100	1	100	1	100	31.72	52.49
16	1	100	1	100	100	1	100	1	38.29	71.02
17	1	1000	1	1000	1	1	1	1	1.87	10.50
18	1	1000	1	1000	1000	1	1000	1	46.13	89.45
19	1	100	100	1	100	1	1	100	0.17	0.16
20	1	100	100	1	1	100	100	1	0.08	0.08
21	1	10	100	1000	1	10	100	1000	10.89	18.83
22	100	100	100	1	1	1	1	50	0.00	0.00
23	100	100	100	1	1	1	1	100	0.00	0.00
24	100	1	1	100	100	1	1	100	0.17	0.16
25	1000	100	10	1	1000	100	10	1	10.95	18.84
26	1000	100	10	1	1	10	100	1000	23.68	52.85
27	2	2	1	1	4	3	2	1	4.08	6.86
28	21	7	9	7	1	1	1	1	0.00	0.00
29	245	213	119	117	99036	99036	91023	91023	0.00	0.00
30	26	18	46	51	5208	833	3333	1875	1.75	2.68

*Most of the sets of N_{ij} and σ^2_{ij} have been scaled for convenience of presentation.

than Min $[\mathrm{Var}(\hat{D}_\alpha), \mathrm{Var}(\hat{D}_\tau)]$ under allocation (2). Thus, the reduction in total cost obtained by using the approximate solution is achieved at the cost of having one estimate with variance exceeding $\bar{V}'$ by a substantial amount. (Note that rounding errors in two examples cause P_c and P_v to be slightly negative.)

To identify the situations where P_v is large, the indices R_N and R_S are used. In Table 2, note that when R_N is significantly larger than one, P_v is generally quite large. Also, when <u>both</u> R_N and R_S are simultaneously much greater than one, both P_c and P_v are larger than they would be if only R_N were greater than one. (Compare example 14 with 15, and 17 with 18.) Note that when R_N exceeds one, even if R_S is equal to one, P_v may be moderately large (examples 13, 14 and 17). Finally, R_S alone is not a good indicator of the conditions under which P_v is large (examples 4, 5, 6 and 7). From the above, the following procedure is suggested: (1) Compute R_N. If R_N is near one, use (2.3.3). If R_N is much larger than one (perhaps, conservatively, larger than two), compute R_S. (2) If R_S is near one, then the allocation given by (2.3.3) may still be satisfactory. If R_S is also much larger than one, (2.3.3) is likely to be unsatisfactory.

Note that (except for rounding errors) the reduction in cost (P_c) is never as large as the increase in variance (P_v). Nevertheless, if P_v is only moderately large (examples 14, 17 and 27), then the resultant reduction in cost might make the use of (2.3.3) attractive. (This may be true for those examples with R_N large and $R_S = 1$.) Finally, it should be observed that it will always be possible to calculate R_N <u>prior</u> to sampling.

Table 2

Indices of the effectiveness of (2.3.3).

Example	R_N	R_S	P_c	P_v
1	1.00	1.00	0.00	0.00
2	1.00	1.00	0.00	0.00
3	1.00	1.00	0.00	0.00
4	1.00	10.00	-0.06	-0.06
5	1.00	100.00	0.02	0.02
6	1.00	1000.00	0.00	0.00
7	1.00	100.00	0.02	0.02
8	1.00	2.00	-0.06	-0.06
9	1.00	1.00	0.01	0.01
10	1.00	1.00	0.34	0.23
11	1.00	1.00	0.00	0.00
12	1.00	1.00	0.17	0.02
13	10.00	1.00	0.62	5.24
14	100.00	1.00	1.89	10.05
15	100.00	100.00	31.72	52.49
16	100.00	100.00	38.29	71.02
17	1000.00	1.00	1.87	10.50
18	1000.00	1000.00	46.13	89.45
19	1.00	1.00	0.17	0.16
20	1.00	1.00	0.08	0.08
21	10.00	10.00	10.89	18.83
22	1.00	1.00	0.00	0.00
23	1.00	1.00	0.00	0.00
24	1.00	1.00	0.17	0.16
25	10.00	10.00	10.95	18.84
26	10.00	10.00	23.68	52.85
27	2.00	1.55	4.08	6.86
28	1.22	1.00	0.00	0.00
29	1.76	1.09	0.00	0.00
30	2.12	2.72	1.75	2.68

3. Two-phase Sampling

If the elements in the four sub-populations (represented by the cells in a 2 x 2 table) are not identifiable in advance, one cannot sample independently in each of them. Thus, for the example cited earlier, it is assumed that there is no comprehensive list of the small, public hospitals in Iowa. However, one may select a large preliminary sample, and identify the sub-population to which each sampled element belongs. Then, for each sub-population, a sub-sample is selected for further study. Such a "two-phase" or "double" sampling procedure will be useful if the cost of identifying an element is small relative to the cost of securing the necessary information in the main survey.

For simplicity of presentation, only the proportionately weighted estimators $(\hat{D}_\alpha, \hat{D}_\tau)$ are considered, and the single "composite" precision statement given by (3.1) is utilized[1]. The numerical analysis in Section 2.3 suggests that if equal precision is desired for $\hat{D}_\alpha$ and $\hat{D}_\tau$,

$$V = \{\mathrm{Var}(\hat{D}_\alpha) + \mathrm{Var}(\hat{D}_\tau)\}/2 \qquad (3.1)$$

is a reasonable composite precision statement.

It is assumed, at first, that the weights in D_α and D_τ are selected prior to sampling. It may be noted that the ensuing analysis is not affected by selecting weights other than those given in (2.1.1). If the weights given in (2.1.1) are to be estimated from the sample results, modifications in the theory are necessary. These are given in the last part of this section.

Further, it is assumed that a given budget (C*) is available to the investigator, and that the sizes of the preliminary sample (n') and main sample (n) are related by the cost equation

$$C^* = c'n' + cn \qquad (3.2)$$

where c' is the cost of identifying an element, and c is the cost of securing the necessary information from an element in the main sample.

The overall objective is to find that value of n' which satisfies (3.2), and minimizes V. This is most easily accomplished by evaluating V for a series of values of n' satisfying (3.2), and selecting that value of n' giving the minimum value of V.

To evaluate V one must first derive the sampling rule (S.R.) to be used. For fixed, but arbitrary, values of n' and n, the S.R. specifies, for any preliminary sample, how the sub-sampling is to be carried out. Let the random variable n'_{ij} denote the number of elements in the preliminary sample that are members of the (i,j)-th cell ($\Sigma\Sigma n'_{ij} = n'$). Then, conditional on the observed n'_{ij}, the S.R. gives the sample sizes (denoted by n_{ij}) for the main sample.

For a given S.R.,

$$V = \frac{1}{2}\left[E\{\mathrm{Var}(\hat{D}_\alpha|\{n'_{ij}\})\} + \mathrm{Var}\{E(\hat{D}_\alpha|\{n'_{ij}\})\}\right.$$

$$\left. + E\{\mathrm{Var}(\hat{D}_\tau|\{n'_{ij}\})\} + \mathrm{Var}\{E(\hat{D}_\tau|\{n'_{ij}\})\}\right]$$

$$= \frac{1}{2} E\left[\Sigma\Sigma_{ij} \left(\frac{N_{\cdot j}^2 + N_{i\cdot}^2}{N^2}\right) \frac{\sigma_{ij}^2}{n_{ij}}\right] = E\left[\Sigma\Sigma\, g_{ij}^2/n_{ij}\right] \qquad (3.3)$$

$$= \{E\,[V(n_{ij})]\}/2 \qquad (3.4)$$

where $\{n'_{ij}\}$ represents a fixed set of the n'_{ij} and the expectation in (3.3) and (3.4) is taken over all possible sets of n'_{ij} with the values of the n_{ij} determined by the S.R.

From (3.4) it is clear that, for a given

[1] The general procedure outlined in Section 3 is similar to the one given by Sedransk (1965) who investigated comparisons among two and three sub-populations.

set of n'_{ij}, one should select the n_{ij} to minimize $V(n_{ij})$ - that is, to minimize the sum of the two conditional variances. The "complete" sampling rule (described below) achieves this objective by utilizing the standard optimal allocation, $n_{ij} \propto g_{ij}$, when it is possible to do so, and by making appropriate modifications when the observed n'_{ij} are too small. (The convexity of $V(n_{ij})$ suggests the adjustments to be made.) Define

$$g_{ij}^2 = (N_{.j}^2 + N_{i.}^2)\sigma_{ij}^2/2N^2 \quad \text{for} \quad i,j = 1, 2,$$

$$a_{ij} = ng_{ij}/\Sigma\,\Sigma\, g_{ij} \quad \text{for} \quad i,j = 1, 2,$$

$$b_{ij} = (n - n'_{11})g_{ij}/(g_{12} + g_{21} + g_{22}) \quad \text{for} \quad (i,j) \neq (1,1),$$

$$d_{ij} = (n - n'_{11} - n'_{12})g_{ij}/(g_{21} + g_{22}) \quad \text{for} \quad (i,j) = (2,1),(2,2),$$

$$\bar{n} = n - n'_{11} - n'_{12} \quad \text{and} \quad \bar{n}' = n'_{21} + n'_{22}.$$

The complete S.R. is defined as follows. (Note that when using the S.R., a re-labeling of the n'_{ij} may be necessary.)

(1) If $n'_{ij} \geq a_{ij}$ for all (i,j), let

$$n_{ij} = a_{ij} \text{ for all } (i,j).$$

(2) If only one $n'_{ij} < a_{ij}$, say $n'_{11} < a_{11}$, but if $n'_{ij} \geq a_{ij}$ for all $(i,j) \neq (1,1)$, let

$$n_{11} = n'_{11}.$$

For the determination of the remaining n_{ij}, there are three possible cases to be considered.

Case one:

If $n'_{ij} \geq b_{ij}$ for all $(i,j) \neq (1,1)$, let

$$n_{ij} = b_{ij} \quad \text{for all } (i,j) \neq (1,1).$$

Case two:

If only one of the $n'_{ij} < b_{ij}$, for $(i,j) \neq (1,1)$, say $n'_{12} < b_{12}$, but if $n'_{21} \geq b_{21}$ and $n'_{22} \geq b_{22}$, let

$$n_{12} = n'_{12}$$

$$n_{21} = d_{21}, \quad \text{if } d_{21} < n'_{21} \leq \bar{n}' - (\bar{n} - d_{21})$$

$$= n'_{21}, \quad \text{if } n'_{21} \leq d_{21}$$

$$= \bar{n} - n'_{22}, \quad \text{if } \bar{n}' - (\bar{n} - d_{21}) < n'_{21}$$

$$n_{22} = \bar{n} - n_{21}.$$

Case three:

If only one of the $n'_{ij} > b_{ij}$, for $(i,j) \neq (1,1)$, say $n'_{22} > b_{22}$, but if $n'_{12} < b_{12}$ and $n'_{21} < b_{21}$, let

$$n_{12} = n'_{12}, \qquad n_{21} = n'_{21}$$

$$n_{22} = n - n'_{11} - n'_{12} - n'_{21}.$$

(3) If two of the $n'_{ij} < a_{ij}$, say $n'_{11} < a_{11}$ and $n'_{12} < a_{12}$, then utilize the allocation given by Case two of (2).

(4) If only one of the $n'_{ij} > a_{ij}$, say $n'_{22} > a_{22}$, then let

$$n_{11} = n'_{11}, \quad n_{12} = n'_{12}, \quad n_{21} = n'_{21},$$

$$n_{22} = n - n_{11} - n_{12} - n_{21}.$$

Thus, the complete S.R. specifies, for any preliminary sample, the exact subsampling procedure. Having selected a random sample of n' elements, the investigator would identify each of the elements and ascertain the values of the n'_{ij}. From the S.R., the number of elements (n_{ij}) to be selected from the (i,j)-th cell for the main sample would be determined. The S.R. is somewhat cumbersome to present, but simple to use for a single problem.

To find the optimal value of n', one must be able to evaluate (assuming the S.R. given above) $2V = E\,[V(n_{ij})]$ for specified values of n' and n. Noting that the expectation in (3.4) refers to all possible sets of the n'_{ij} (with the values of the n_{ij} determined by the S.R.), it is clear that an analytical evaluation is not feasible. Moreover, even with specified values of the necessary parameters and the aid of a computer, a complete <u>numerical</u> evaluation of $E\,[V(n_{ij})]$ is impractical because of the extremely large number of possible sets of the n'_{ij}.

In lieu of a complete evaluation of $E[V(n_{ij})]$, one may use Monte Carlo sampling to estimate $E[V(n_{ij})]$. Thus, to determine the n'_{ij}, a sample of size n' is drawn from the multinomial distribution with inclusion probabilities $\{\pi_{ij} = N_{ij}/N\}$. Using the S.R., the n_{ij} are found, and $V(n_{ij})$ calculated. This procedure is replicated K-1 times, and the sample mean $(\bar{V})$ and variance of $V(n_{ij})$ calculated. The entire procedure is repeated for different values of n', and the "optimal value" of n' is chosen on the basis of minimum $\bar{V}$. If one selects a large value for K (perhaps, K = 500) it appears, from some numerical examples, that the Monte Carlo procedure provides an unequivocal choice for the optimal value of n'. The sample standard error of $\bar{V}$ is so small, and the pattern of the $\bar{V}$ (as n increases) is so regular that an error in selecting the optimal n' value is very unlikely. Even if K = 100, one is unlikely to make a costly error in selecting the optimal value of n'. To illustrate this, consider the following specification: C* = 220, c' = 1, c = 10, π_{11} = .15, π_{12} = .20, π_{21} = .25, π_{22} = .40, g_{11} = 1, g_{12} = 1, g_{21} = 1, g_{22} = 4, K = 100. Then, for each trial value of n', the following quantities are given in Table 3: the corresponding value of n, $\bar{V}$ and $s(\bar{V})$ = (Monte Carlo) sample estimate of the standard error of $\bar{V}$. The best choice for n' is seen to be 30.

Table 3

Monte Carlo estimates of V for various values of n'.

n'	n	$\bar{V}$	$s(\bar{V})$
20	20	3.27	0.082
30	19	2.68	0.015
40	18	2.79	0.006
50	17	2.95	0.001

To estimate $E[V(n_{ij})]$, one must specify values for the π_{ij}. Although some of the marginal totals of the N_{ij} might be known, the N_{ij} will usually be unknown. However, in many situations, reasonable estimates of the π_{ij} can be made. It must be emphasized that once n' is determined, the sampling rule depends only on the weights in D_α and D_τ. (As noted earlier, such weights do not have to be functions of the N_{ij}.)

If a computer is not available, one may roughly approximate $\{E[V(n_{ij})]\}/2$ using a procedure given by Sedransk[2] (1965). Thus, assuming that $\pi_{11} \le \pi_{12} \le \pi_{21} \le \pi_{22}$, consider the following approximate determination of the values of the $E(n_{ij})$:

$$E(n_{11}) = ng_{11}/g^*, \quad \text{if} \quad E(n'_{11}) > ng_{11}/g^*$$
$$= E(n'_{11}), \quad \text{otherwise;}$$

$$E(n_{12}) = (n-En_{11})g_{12}/(g^*-g_{11}),$$
$$\text{if} \quad E(n'_{12}) > (n-En_{11})g_{12}/(g^*-g_{11})$$
$$= E(n'_{12}), \quad \text{otherwise;}$$

$$E(n_{21}) = (n-En_{11}-En_{12})g_{21}/(g_{21}+g_{22}),$$
$$\text{if} \quad E(n'_{21}) > (n-En_{11}-En_{12})g_{21}/(g_{21}+g_{22})$$
$$= E(n'_{21}), \quad \text{otherwise;}$$

$$E(n_{22}) = n - En_{11} - En_{12} - En_{21} \qquad (3.5)$$

where $g^* = \Sigma\Sigma\, g_{ij}$, $E(n'_{ij}) = n'\pi_{ij}$, and for the special case of n' = n, take $E(n_{ij}) = E(n'_{ij})$.

Finally, approximate $\{E[V(n_{ij})]\}/2$ by

$$V' = \frac{1}{2}\Sigma\Sigma\left(\frac{N_{\cdot j}^2 + N_{i\cdot}^2}{N^2}\right)\frac{\sigma_{ij}^2}{E(n_{ij})}. \qquad (3.6)$$

Then, one may evaluate V' [using (3.5)] for different choices of n', and choose the "optimal" value of n' as before. For the cases considered by Sedransk (1965), such a procedure proved to be very satisfactory. To further investigate the utility of this approximate method, 22 examples have been worked using both the approximate method, and the Monte Carlo sampling procedure. At least 100 Monte Carlo replications were used for each trial value of n' considered. The examples were chosen to represent the wide variety of possible relationships among the $\pi_{ij} = N_{ij}/N$ and g_{ij}.

In Table 4, the following quantities are presented for each example: (1) The values of the g_{ij} and π_{ij}. Using the definition of g_{ij}, it is easily verified that the examples include a number of patterns of the σ_{ij}^2; (2) The "optimal" values of n' as determined by the approximate and Monte Carlo methods (n'_a and n'_0, respectively); (3) The estimate, A, from Monte Carlo sampling, of the per cent increase in the variance, $V = E[V(n_{ij})]/2$, by using the "optimal" n' value from the approximate method rather than the optimal n' obtained from the Monte Carlo calculations; (4) The ratio $B = \bar{V}/V'$ evaluated at the "optimal" value of n' as determined by using the approximate method. (Recall that $\bar{V}$ is the Monte Carlo estimate of V.) B indicates the utility of using V' as an estimator for V, the value of n' chosen being the most appropriate

[2] For further details, see Section 4.2.2 of that paper.

Table 4

Efficiency of a procedure to estimate the optimal sample sizes: $C^* = 220 = n' + 10n$.

Ex.	g_{11}	g_{12}	g_{21}	g_{22}	π_{11}	π_{12}	π_{21}	π_{22}	n'_a	n'_0	A*	B**	C***
1	1	1	1	4	.15	.20	.25	.40	30	30	0.0	1.04	1.22
2	4	1	1	1	.15	.20	.25	.40	60	60	0.0	1.05	2.24
3	4	4	2	2	.15	.20	.25	.40	40	40	0.0	1.09	1.49
4	4	3	2	1	.15	.20	.25	.40	40	50	5.4	1.14	1.70
5	1	2	3	4	.15	.20	.25	.40	20	30	15.7	1.22	1.16
6	2	2	4	4	.15	.20	.25	.40	20	30	16.5	1.22	1.17
7	4	1	4	1	.15	.20	.25	.40	40	50	2.1	1.12	1.85
8	1	1	1	4	.10	.40	.10	.40	30	30	0.0	1.07	
9	4	1	1	1	.10	.40	.10	.40	70	80	3.7	1.13	
10	4	3	2	1	.10	.40	.10	.40	50	60	0.2	1.16	
11	2	2	4	4	.10	.40	.10	.40	50	50	0.0	1.09	
12	4	1	4	1	.10	.40	.10	.40	60	70	4.8	1.22	
13	1	1	1	4	.20	.20	.30	.30	30	40	4.4	1.14	
14	4	1	1	1	.20	.20	.30	.30	50	50	0.0	1.06	
15	4	3	2	1	.20	.20	.30	.30	30	40	3.2	1.12	
16	2	2	4	4	.20	.20	.30	.30	20	30	12.8	1.22	
17	4	1	4	1	.20	.20	.30	.30	30	40	5.7	1.11	
18	1	1	1	4	.10	.10	.20	.60	30	30	0.0	1.08	
19	4	1	1	1	.10	.10	.20	.60	70	70	0.0	1.12	
20	4	3	2	1	.10	.10	.20	.60	60	70	1.0	1.18	
21	2	2	4	4	.10	.10	.20	.60	30	40	4.1	1.17	
22	4	1	4	1	.10	.10	.20	.60	50	60	9.1	1.18	

*A is the (Monte Carlo) estimate of the per cent increase in variance, V, by using the "optimal" n' value from the approximate method rather than the optimal n' obtained from the Monte Carlo calculations.

**B is the ratio $\bar{V}/V'$ calculated at the "optimal" value of n' as determined by using the approximate method.

***C is the ratio $\bar{V}_{20}/\bar{V}_{opt}$ where $\bar{V}_{20}$ is the (Monte Carlo) estimate of V for $n' = n = 20$, and $\bar{V}_{opt}$ is the corresponding estimate of V at the optimal value of n'.

one if the approximate method is employed; (5) The ratio $C = \bar{V}_{20}/\bar{V}_{opt}$ where $\bar{V}_{20}$ is the (Monte Carlo) estimate of V for $n' = n = 20$, and $\bar{V}_{opt}$ is the corresponding estimate of V for $n' = n'_0$.

The results (presented in Table 4) are very similar to those found by Sedransk (1965, pp. 996-999), and indicate that the loss of precision by using the "optimal" n' value from the approximate method is generally small. The values of A range from 0.0% to 16.5% with only three examples having values exceeding 10.0%. The cases where A is large coincide with values of $n'_a = 20$. This may be explained by the observation that increasing n' a little from the single-phase sampling position ($n' = n$) usually reduces the variance considerably, whereas further increases in n' produce smaller reductions in the variance. Thus, when the true optimum value of n' is only slightly larger than $n' = n = 20$, taking $n'_a = 20$ may result in a large increase in variance. To be conservative, if the value of n'_a is near n (i.e., single-phase sampling is indicated to be optimal), a further investigation is indicated. However, in most of the examples the optimum is flat; that is, $\bar{V}$ varies insubstantially with n' as n' moves in either direction from its optimal value. From this discussion, it appears that if n'_a is not (nearly) equal to n, one may be confident about using the approximate method.

For some specifications of the π_{ij} and choices of n', the probability of obtaining a zero value for at least one of the n'_{ij} may not be negligible. In such cases, one may wish to take a larger value of n' than that given by n'_0 or n'_a. For the examples presented in Table 4, and including all trial values of n' considered, all estimates are based on samples having $n'_{ij} > 0$ for $i, j = 1, 2$.

From B, it is clear that V' underestimates $\bar{V}$ (and, therefore, should underestimate V) for all examples. This agrees with the findings of Sedransk (1965, p. 998). The ratio C, presented for the first seven examples, indicates the efficacy of double sampling. If one used single-phase sampling (i.e., selected n' = n = 20), rather than double sampling with n' = n_0', the per cent increase in variance ranged from 16 to 124. Greater increases could be expected for examples where n_0' is very large.

Finally, it should be noted that Sedransk (1965, pp. 998-999) has presented some numerical evidence that the double sampling procedure is robust with regard to specifying the values of the g_{ij}. That is, the loss in unconditional precision because of choosing a non-optimal value of n' will generally be moderate, even if there are fairly large errors in specifying the g_{ij}. Also, he indicates that all of these conclusions seem to persist if the cost relationship c/c' is altered.

If the approximate method is considered to be unsatisfactory for some situations, one might continue to use V' as given in (3.6), but approximate the $E(n_{ij})$ more closely. (The first order Taylor series approximation is likely to be satisfactory since, in most applications, n' and n will be large.) Using an "approximate" S.R. akin to that given by Sedransk (1965, p. 994), one may obtain $E(n_{11})$, $E(n_{12}) = E\{E(n_{12}|n_{11})\}$, and $E(n_{21}) = E\{E(n_{21}|n_{11},n_{12})\}$. This is most easily accomplished by using normal approximations to the distributions of the n_{ij}', but one will have to use some numerical integration to evaluate a few of the terms in $E(n_{21})$.

If the proportionately weighted estimators are employed with the weights estimated from the preliminary sample, the estimators of D_α and D_τ are given by

$$\hat{D}_\alpha' = \{n_{.1}'(\bar{y}_{11} - \bar{y}_{21})/n'\} + \{n_{.2}'(\bar{y}_{12} - \bar{y}_{22})/n'\}$$

$$\hat{D}_\tau' = \{n_{1.}'(\bar{y}_{11} - \bar{y}_{12})/n'\} + \{n_{2.}'(\bar{y}_{21} - \bar{y}_{22})/n'\}. \tag{3.7}$$

Then, following the procedure used to derive (3.3),

$$[\mathrm{Var}(\hat{D}_\alpha') + \mathrm{Var}(\hat{D}_\tau')]/2 =$$

$$\frac{1}{2}\left[E\left\{\Sigma\Sigma\left(\frac{(n_{.j}')^2 + (n_{i.}')^2}{(n')^2}\right)\frac{\sigma_{ij}^2}{n_{ij}}\right\}\right.$$

$$+ V\left[\frac{n_{.1}'}{n'}(\mu_{11} - \mu_{21}) + \frac{n_{.2}'}{n'}(\mu_{12} - \mu_{22})\right]$$

$$\left. + V\left[\frac{n_{1.}'}{n'}(\mu_{11} - \mu_{12}) + \frac{n_{2.}'}{n'}(\mu_{21} - \mu_{22})\right]\right] \tag{3.8}$$

where E and V in (3.8) refer to all possible sets of the n_{ij}' with the corresponding n_{ij} determined by the S.R.

It is clear from (3.8) that the choice of a S.R. depends only on the term in curly brackets. However, this term is equivalent to (3.3) with $(n_{.j}'/n')$ and $(n_{i.}'/n')$ replacing $(N_{.j}/N)$ and $(N_{i.}/N)$. Hence, the complete S.R. given below (3.4) may be utilized for the current problem with the appropriate substitutions for $(N_{i.}/N)$ and $(N_{.j}/N)$.

If the estimators given by (3.7) are used, it is difficult to determine the optimal value of n'. For a fixed value of n', if one uses rough approximations for the π_{ij}, Monte Carlo sampling may be employed to estimate the first term in (3.8). As suggested above, this procedure may be repeated for several different values of n', and the "optimal" value thus located. However, the variance terms in (3.8) depend on n', the π_{ij} and the μ_{ij}. Since the values of the μ_{ij} (and the π_{ij}) may determine the true optimal value of n', it appears that rough estimates of the two variance terms in (3.8) should be combined with the Monte Carlo estimates of the first term to approximate $\frac{1}{2}[\mathrm{Var}(\hat{D}_\alpha') + \mathrm{Var}(\hat{D}_\tau')]$ more closely.

4. An Additional Application of the Two-phase Sampling Procedure

In classical double sampling for stratification, one selects a simple random sample of size n' with n_h' elements subsequently identified as being members of stratum h (h = 1, 2, ..., L). Then, one may select (by simple random sampling) a subsample of n_h elements from the n_h' elements found to belong to stratum h. Assuming that n' is sufficiently large so that the probability of obtaining any $n_h' > 0$ is negligible, $\bar{y}_{st}^* = \sum_h n_h' \bar{y}_h/n'$ is an unbiased estimator of the population mean with

$$\mathrm{Var}(\bar{y}_{st}^*) = V\left[\sum_h n_h' \bar{Y}_h/n'\right] + E\left\{\sum_h (n_h')^2 s_h^2/(n')^2 n_h\right\}$$

$$- E\left[\sum_h (n_h')^2 s_h^2/(n')^2 N_h\right] \tag{4.1}$$

where E and V in (4.1) refer to all sets of the n_h' ($n_h' \neq 0$) with the corresponding n_h determined by some sampling rule.

From (4.1), it is clear that for given values of n' and n, one should choose the sampling rule to minimize the term in curly brackets.

(None of the other terms are functions of the n_h.) This sampling rule (for L = 4 strata) can be obtained immediately from the one presented in Section 3. Finally, note the similarity between (4.1) and (3.8). (The two expressions are essentially identical if the f.p.c. terms are omitted from (4.1) or included in (3.8).) Thus, the problem of finding the optimal value of n' (assuming a linear cost function with prespecified budget) for double sampling with stratification is identical with that presented earlier.

Acknowledgment

This research has been supported, in part, by the U. S. Office of Education under Contract OEC-3-6-002041-2041.

References

Cochran, W. G. (1963). Sampling Techniques, Second Edition, New York: John Wiley.

Hartley, H. O. and Hocking, R. (1963). "Convex programming by tangential approximation," Management Science, 9, 600-612.

Sedransk, J. (1965). "A double sampling scheme for analytical surveys," Jour. Amer. Stat. Assoc., 60, 985-1004.

Sedransk, J. (1967). "Designing some multifactor analytical studies," Jour. Amer. Stat. Assoc., 62, (to appear).

Yates, F. (1960). Sampling Methods for Censuses and Surveys, Third Edition, London: Griffin.