Survey Methodology, June 1990
Vol. 16, No. 1, pp. 105–116
Statistics Canada

Estimation Using Double Sampling and Dual Stratification

DONALD B. WHITE[1]

ABSTRACT

The problem considered is that of estimation of the total of a finite population which is stratified at two levels: a deeper level which has low intrastratum variability but is not known until the first phase of sampling, and a known pre-stratification which is relatively effective, unit by unit, in predicting the deeper post-stratification. As an important example, the post-stratification may define two groups corresponding to responders and non-responders in the situation of two-phase sampling for non-response. The estimators of Vardeman and Meeden (1984) are employed in a variety of situations where different types of prior information are assumed. In a general case, the standard error relative to that of the usual methods is studied via simulation. In the situation where no prior information is available and where proportional sampling is employed, the estimator is unbiased and its variance is approximated. Here, the variance is always lower than that of the usual double sampling for stratification. Also, without prior information, but with non-proportional sampling, using a slight modification of the second phase sampling plan, an unbiased estimator is found along with its variance, an unbiased estimator of its variance, and an optimal allocation scheme for the two phases of sampling. Finally, applications of these methods are discussed.

KEY WORDS: Two-phase sampling; Prior information; Variance estimation; Optimal allocation; Non-response.

1. INTRODUCTION

Various stratified sampling designs employ various types of prior information. For example, the usual stratification model assumes full prior knowledge of individual stratum memberships. Post-stratification is useful when there is global information on stratum sizes but no information on individuals. Double sampling for stratification, on the other hand, assumes no prior information on strata. Further, some knowledge of the population values is necessary, for example, for the allocation of sampling resources among strata (see, for example, Cochran 1977, pp. 96-99 and 331-332).

The rigid assumptions inherent to these sampling designs and population models often are not satisfied due to the discrepancy between the population under study and the (possibly dated) prior information. Seeking to appropriately handle this discrepancy, Vardeman and Meeden (1984) have introduced a pair of estimators which combine information on stratum memberships, stratum sizes, and stratum averages with analogous information gained from the current sample. Their two estimators apply to two essentially different situations. The first is where the prior information is global only, *i.e.,* only on stratum sizes and averages. The second estimator applies where there is also partial information on individual stratum memberships. Here, the population is stratified according to various factors, some of which are known and some of which, though not known, may be inexpensive to determine on a first phase of sampling.

[1] Donald B. White, Department of Statistics, State University of New York at Buffalo, 249 Farber Hall, Buffalo, New York 14214.

As an example, consider the use of sampling to determine the spread of an infectious disease. If detection of infection is expensive, then stratification, according to risk categories, is desirable to reduce the second phase sample size. Factors determining risk categories may include gender, age, place of residence, ethnicity, health habits, and contact with potential carriers. As some of these factors are not known prior to sampling, the model of Vardeman and Meeden can be employed since the true risk categories can be predicted by the known factors.

Another example is two-phase sampling for non-response. Extending the method of Hansen and Hurwitz (1946), we have a population which is divided into two post-strata, *i.e.*, responders and non-responders. The methods discussed here apply when there is some prior information which classifies units into pre-strata which are then used to predict whether or not the unit will be in the group of responders.

The notion of employment of prior information in two-phase designs is not without precedent in the sampling literature. As an example, Han (1973) has used prior information on an auxiliary regression variable (to be measured in a first phase sample) to construct a simple hypothesis (say H_0) regarding the mean of that variable. The first phase sample measurements are then used to test H_0. If H_0 is accepted, the value specified by H_0 is used in the estimator; if it is rejected, the sample average is used.

A discussion of the use of the first estimator of Vardeman and Meeden (global information only) can be found in White (1987). There, optimal choices of the weighting constants for prior information relative to the information contained in the current sample were determined. Here, the situation considered is where prior information is also available on individual stratum memberships. After introducing the necessary notation in Section 2, we explore a simulated example in Section 3. In Section 4, in two different sampling situations, unbiased estimators are analyzed in terms of variance, unbiased estimation of the variance, and optimal allocation of sampling resources. In Section 5, applications of these techniques are discussed.

2. THE POPULATION MODEL AND SAMPLING SCHEME

We now present the population model and the proposed sampling design. We begin with a finite population P of units labelled 1, 2, ..., N with associated unknown values y_1, y_2, ..., y_N. Denote the population total by $\tau = \sum_{i=1}^{N} y_i$. For $1 \leq i \leq N$, unit i also possesses an unknown post-stratum membership j_i, $1 \leq j_i \leq J$, and a known pre-stratum membership k_i, $1 \leq k_i \leq K$.

A variety of population quantities require a specialized notation. Such quantities include sizes of groups, group averages and group variances. Subscripts will identify the group involved: no subscript implies reference to the entire population, "$k\cdot$" refers to pre-stratum k, $1 \leq k \leq K$, "$\cdot j$" refers to post-stratum j, $1 \leq j \leq J$, and the subscript "kj" refers to the intersection of pre-stratum k with post-stratum j. The base symbols N, $\bar{Y}$ and S^2 refer to number of elements, y-average, and finite population variance, respectively. Also, we let P, $P_{k\cdot}$, $P_{\cdot j}$ and P_{kj} denote the subsets of P corresponding to the four categories given above. For example, we have

$$S_{k\cdot}^2 = \frac{1}{N_{k\cdot} - 1} \sum_{i \in P_{k\cdot}} (y_i - \bar{Y}_{k\cdot})^2.$$

Also, we can write

$$\tau = \sum_j N_{.j} \bar{Y}_{.j} . \qquad (1)$$

We finally let $W_{kj} = N_{kj}/N_{k.}$, *i.e.*, W_{kj} is the proportion of units in pre-stratum k which fall into true stratum j.

We now discuss the sampling technique. In the first phase of sampling, a stratified simple random sample without replacement s' is selected, with $n'_{k.}$ units (first phase sampling fraction denoted by $f'_{k.} = n'_{k.}/N_{k.}$) selected from pre-stratum k. Samples from different pre-strata are independent. For these $n' = \sum_k n'_{k.}$ units, post-strata, j_i, are observed. Following the notational pattern given above, we let n'_{kj} denote the number of units in s' sampled from pre-stratum k which happen to fall in post-stratum j. Also, $n'_{.j} = \sum_k n'_{kj}$ is the total number of units in s' which fall in post-stratum j. This set of units is denoted by $s'_{.j}$. These quantities are observed, while quantities involving y-values, such as $\bar{y}'$ and $s^{2'}$ (with all four types of subscripts), remain unobserved. Here, and in the following, the average of any empty collection is taken as zero, and, if the size of a group is one or zero, we take its variance s^2 to be zero. We note that for $1 \leq k \leq K$, the random vectors $(n'_{k1}, \ldots, n'_{kJ})$ are independent with each possessing a multivariate hypergeometric distribution.

For the second phase of sampling, we partition s' into $\cup_{j=1}^{J} s'_{.j}$, *i.e.*, by post-stratification. For each j, let $v_j(\cdot)$ denote a known function on and into the non-negative integers with $v_j(0) = 0$ and $1 \leq v_j(x) \leq x$ if $x \geq 1$. The second phase sample s is also stratified, but now is a subsample of s' and stratified according to the post-stratification. The sample from $s'_{.j}$ is denoted $s_{.j}$ and is of size $n_{.j} \equiv v_j(n'_{.j})$. Here, y-values are observed, yielding quantities such as $\bar{y}_{.j}$ and $s^2_{.j}$, the y-average and finite population variance of the units in the phase two sample and stratum j.

The estimates of τ given by Vardeman and Meeden include the option of inclusion of prior guesses for the relative stratum sizes within each pre-stratum and for the stratum averages. Thus, we have prior guesses for the values W_{kj} and $\bar{Y}_{.j}$ which are given by Π_{kj} and $\mu_{.j}$, respectively. In the estimator introduced below, these guesses are given weighting constants which reflect the confidence in the guess relative to the confidence in the corresponding information yielded by the current sample. For each k, the confidence value allotted to the collection $(\Pi_{k1}, \ldots, \Pi_{kJ})$ is denoted $\tilde{M}_{k.} \in [0,\infty]$ and for each j, the confidence value given to $\mu_{.j}$ is denoted $M_{.j} \in [0,\infty]$. In the current sample, the collection $(W_{k1}, \ldots, W_{kJ})$ is estimated by $(n'_{k1}/n'_{k.}, \ldots, n'_{kJ}/n'_{k.})$ and is based on a simple random sample of size $n'_{k.}$. Thus, the confidence in Π_{kj}, say, as opposed to $n'_{kj}/n'_{k.}$, is reflected by the size of $\tilde{M}_{k.}$ versus that of $n'_{k.}$. Similarly, in the current sample, $\bar{Y}_{.j}$ is estimated by $\bar{y}_{.j}$ and is based on a sample of size $n_{.j}$; thus, the relative confidence in the prior guess and the current estimate is reflected by the relative sizes of $M_{.j}$ and $n_{.j}$. Any confidence weight for prior information equal to zero corresponds to no use of the prior information, and, as in the use of stratum sizes in the usual post stratification model, a value of infinity implies no use of the corresponding information in the current sample.

Using the prior guesses, current estimates and confidence weights, we estimate W_{kj} and $\bar{Y}_{.j}$ by $\hat{\Pi}_{kj} = (\tilde{M}_{k.}\Pi_{kj} + n'_{kj})/(\tilde{M}_{k.} + n'_{k.})$ and $\hat{\mu}_{.j} = (M_{.j}\,\mu_{.j} + n_{.j}\bar{y}_{.j})/(M_{.j} + n_{.j})$, respectively. Finally, an estimate $\hat{\tau}$ of the population total τ is constructed by replacing in the formula (1) for τ any unobserved quantity by its estimate given above. Thus, we employ

$$\hat{\tau} = \sum_{j=1}^{J} \left\{ n_{.j}\bar{y}_{.j} + (n'_{.j} - n_{.j})\hat{\mu}_{.j} + \sum_{k=1}^{K} (N_{k.} - n'_{k.})\hat{\Pi}_{kj}\hat{\mu}_{.j} \right\} . \qquad (2)$$

Computation of the bias and variance of $\hat{\tau}$ in the general case is left open by Vardeman and Meeden. The case $K = 1$ and $M_{.j} = 0, 1 \leq j \leq J$, has been studied in White (1987). Before proceeding to a result in a more complex situation, we first explore the results of a simulation on a hypothetical population.

3. A MONTE CARLO STUDY

Here we present a specific population and sampling scheme which is modelled after the introductory example regarding estimation of the spread of an infectious disease. For a population of 10,000 individuals who are susceptible, the disease is assumed to be more prevalent among the 5,000 who live in the western section of the area considered. Since this is a known characteristic, the population is partitioned according to the east-west boundary into $K = 2$ pre-strata. Next, we assume that certain easily obtained additional information enables the sampler to categorize the individual as low, medium, or high risk for becoming infected. See Table 1 for the details of the construction of the population.

For estimation of the total number infected ($\tau = 2302$), we assume no prior knowledge of the stratum proportions $\bar{Y}_{.1}$, $\bar{Y}_{.2}$, and $\bar{Y}_{.3}$ and thus take $M_{.1} = M_{.2} = M_{.3} = 0$. There remain four major ingredients to the estimation process: 1) the prior guesses $\{\Pi_{kj}$: $k = 1, 2$, $j = 1, 2, 3\}$ for the distribution of individuals from pre-strata to post-strata, 2) the weighting constants $\tilde{M}_1$ and $\tilde{M}_2$ given to these prior guesses, 3) the first phase sample design and outcome, and 4) the second phase sample design and outcome. These are detailed in the following.

First, in White (1987) it was found for the $K = 1$ case that an effective choice of weighting constants was to select M equal to the sample size on which the previous information was based. Following that notion, we allowed, for each simulation, the collection $\{\Pi_{kj}\}$ to select itself through a preliminary sample of size m (either 500 or 2500) from each pre-stratum. That is, Π_{kj} is taken to be the proportion of the m individuals from pre-stratum k falling in post-stratum j.

Second, for each run, the weighting constants were taken as $\tilde{M}_{1.} = \tilde{M}_{2.} = M$ for all $M \in \{0, 100, 200, 300, \ldots, 10{,}000, \infty\}$. Recall that $M = \infty$ corresponds to the situation of the usual post-stratification where no use is made of the current sample to estimate group sizes.

Third, the first phase sample is stratified according to pre-strata with sampling fractions $f'_{k.}$ taken to be $f'_{1.} = f'_{2.} = f$, $f \in \{.10, .20, .30, .40, .50\}$. Recall that in this phase of sampling, only post-stratification is observed. This information is, presumably, inexpensive to obtain.

Table 1

Number Infected/Group Size for the Pre-strata and Post-strata Combinations

Location of Residence / Risk Group j	Low 1	Medium 2	High 3	Total
East ($k = 1$)	40/4000	80/800	100/200	220/5000
West ($k = 2$)	2/200	80/800	2000/4000	2082/5000
Total	42/4200	160/1600	2100/4200	2302/10000

On the other hand, sampling a unit in phase 2, where the presence of infection is determined, is assumed to be rather expensive. The individuals selected are a subsample of the phase one sample, stratified according to post-strata. The sampling fractions in various strata are again taken as equal ($v_j(n'_{\cdot j}) = [c_j n'_{\cdot j}]$ for $n'_{\cdot j}$ large enough, and $c_1 = c_2 = c_3 = c$) and so that different simulations can be compared, c is selected so that the fraction of the entire population which appears in the phase 2 sample remains constant at .10.

Now, the following process is repeated $R = 50{,}000$ times: obtain a preliminary sample of size m from which prior guesses Π_{kj} for W_{kj} are constructed. Next, a sample, stratified according to pre-strata with sampling fractions f, is obtained. Only post-stratification is observed. Then, a subsample, stratified according to post-strata with sampling fractions c, is obtained and units in this sample are classified as infected or not infected. Finally, on each run, $\hat{r}$ is obtained for each value of M considered. The standard error of $\hat{r}$ is estimated using the R simulated values of $\hat{r}$. Recall, however, that in a real-life application, the standard error of an estimate will depend on the particular values of Π_{kj} used; here, these values are different on each run and thus the estimated standard error should be viewed as a long run average for a mixture of distributions of $\hat{r}$, mixed according to the distribution of the Π_{kj} based on the preliminary sample.

The simulations were performed on an IBM3031 computer. For this example, where $y_i \in \{0,1\}$ for all i, all random quantities are functions of independent hypergeometric or multivariate hypergeometric variables. Using the fact that the conditional distribution of a univariate marginal of a multivariate hypergeometric distribution given any subcollection of the other coordinates is itself hypergeometric, all random quantities were simulated using the IMSL 92DP hypergeometric simulation subroutine GGHPR. For the first combination of m and f (500 and .10), the simulation process was repeated five times to check internal consistency.

Tables 2 and 3 summarize pertinent characteristics of the variation of the simulated SE($\hat{r}$) as a function of M for the five repeated simulations (Table 2), and the simulations for various values of f and m (Table 3). Table 2 gives only highlights which demonstrate internal consistency and confirm that the number of repetitions is chosen large enough. Note that M_0 denotes the value of M for which SE ($\hat{r}$) is minimized. In Table 3, also given is a comparison with the better of the possible usual techniques (regular two-phase or stratified according to pre-strata) relative to the ideal where the true strata are regarded as known. The standard error of an estimator based on stratified sampling using pre-strata only is 113.27, and for stratified according to true strata, it is 105.47. Thus, letting the estimator in regular two-phase sampling be denoted by $\hat{r}_2$ and realizing that SE($\hat{r}_2$) depends upon f and c, the values appearing in the columns headed Percent Relative Reduction in SE($\hat{r}$) are 100 $[\min(\mathrm{SE}(\hat{r}_2), 113.27)] - \mathrm{SE}(\hat{r})/[\min(\mathrm{SE}(\hat{r}_2), 113.27) - 105.47]$.

Table 2

Key Features of the Repeated Runs with $m = 500, f = .10$ and $c = 1.0$

Run #	M_0	SE($\hat{r}$)			
		$M = 0$	$M = m$	$M = M_0$	$M = \infty$
1	600	113.55	109.67	109.62	112.00
2	700	113.42	109.50	109.45	111.80
3	700	113.92	109.86	109.78	112.00
4	600	113.61	109.71	109.66	112.07
5	600	113.56	109.74	109.70	112.17

Table 3

Key Features of SE($\hat{\tau}$) as a Function of M

					SE($\hat{\tau}$)				Percent Relative Reduction in SE($\hat{\tau}$)		
m	f'	c	SE($\hat{\tau}_2$)	M_0	$M = M_0$	$M = 0$	$M = m$	$M = \infty$	$M = 0$	$M = m$	$M = \infty$
500	.10	1.00	126.29	600	109.62	113.55	109.67	112.00	−3.6	46.2	16.3
500	.20	.50	115.19	600	107.95	109.02	107.97	110.72	54.5	67.9	32.7
500	.30	.33	111.80	600	107.87	108.25	107.87	110.38	56.1	62.1	22.4
500	.40	.25	109.22	750	106.51	106.76	106.52	108.29	65.6	72.0	24.8
500	.50	.20	107.98	700	106.17	106.28	106.18	107.55	67.7	71.7	17.1
2500	.10	1.00	126.29	*	≤ 106.20	113.33	106.42	106.20	−0.8	87.8	90.6
2500	.20	.50	115.19	*	≤ 105.76	108.67	106.02	105.76	59.0	92.9	96.3
2500	.30	.33	111.80	*	≤ 106.63	108.18	106.87	106.63	57.2	77.9	81.7
2500	.40	.25	109.22	*	≤ 105.77	106.59	105.94	105.77	70.1	87.5	92.0
2500	.50	.20	107.98	*	≤ 105.81	106.34	105.96	105.81	65.3	80.5	86.5

* -- > 10,000

A variety of important results can be discerned from Table 3. First is that for $m = 500$, M_0 is very close to, although always slightly larger than, m. This is the result predicted by the $K = 1$ situation from White (1987). For $m = 2500$, though in every case $M_0 > 10{,}000$, one discovers that SE($\hat{\tau}$) at $M = m$ is very close to the minimum at $M = M_0$.

Second is that at $M = m$, the percent relative reduction in SE($\hat{\tau}$) ranges from a minimum of 46% to over 90%. Also, at $M = 0$, corresponding to the situation of dual stratification with no prior information on any population characteristic, the percent relative reduction in SE($\hat{\tau}$) is always over 50% except in the case of the smallest first phase sampling fraction, $f = .10$. In that case, when prior information is not available and the first phase sample size is small, one is better off to use the pre-strata and ignore the true stratification. On the other hand, if one does have a set of prior guesses available for the collection of W_{kj}, but is uncertain of what weights to attach to these values, one could use the usual post-stratification notion of using weight $M = \infty$. If the prior information is good, as in our case $m = 2500$, then the percent relative reduction in SE($\hat{\tau}$) is always over 80%. Even if the prior information is only moderately accurate, as in the case $m = 500$, the reduction in standard error is between 16% and 33%.

In summary, if one is able to identify a weighting constant applicable to prior information on the distributions of units among strata, then a substantial reduction in standard error can be obtained using these methods. Even if one cannot identify such a constant or does not have applicable prior information, one can still decrease standard error using dual stratification by taking $M = 0$ if the prior information on W_{kj} is either poor or non-existent, or $M = \infty$ with accurate prior information. In particular, it thus turns out that the case $M = 0$ is important. This case is examined in detail in the next section.

4. BIAS, STANDARD ERROR, AND OPTIMAL ALLOCATION WITH NO PRIOR INFORMATION

When no prior information is available, we set $M_j = 0$ and $\tilde{M}_k = 0$ for each $1 \leq j \leq J$ and $1 \leq k \leq K$. In this section, we at first also assume that sampling in both phases is proportional to the size of the group from which the sample is drawn, that is, for each

k, $n'_{k\cdot} = fN_{k\cdot}$ (*i.e.*, $f'_{k\cdot} = f$, all k) and for each j, $n_{\cdot j} = cn'_{\cdot j}$ (*i.e.*, $v_j(x) = cx$, all j). This, of course, immediately introduces an approximation (referred to in what follows as approximation A1), since the resulting sample sizes are not necessarily integers. However, in reasonably large populations, and for reasonably large sampling fractions f and c, this approximation has little impact on the derivations that follow.

In this situation, $\hat{\mu}_{\cdot j}$ reduces to $\bar{y}_{\cdot j}$ and $\hat{\Pi}_{kj}$ reduces to $n'_{kj}/n'_{k\cdot}$ and, thus, we have $\hat{\tau} = 1/f \sum_{j=1}^{J} n'_{\cdot j}\bar{y}_{\cdot j}$. The derivations of the expectation and variance of $\hat{\tau}$ are summarized in the appendix. The key features are two conditioning arguments: first, we condition on s' since the second phase sample is a function of s' and, second, because of the multivariate hypergeometric nature of the phase one sample, we condition on the values n'_{kj}, the sizes of the various pre-stratum and post-stratum combinations in the first phase sample.

In the appendix, we show first that $\hat{\tau}$ in this case is unbiased (aside from approximation A1) and that an approximation of its variance is given by

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{f}\sum_{k} N_{k\cdot}S^2_{k\cdot} + \frac{1-c}{fc}\sum_{j} N_{\cdot j}S^2_{\cdot j}. \qquad (3)$$

As discussed in the appendix in more detail, formula (3) 1) gives answers close to the simulated values, 2) is based on approximations whose error is small for large populations and reasonably large samples, and 3) reduces to the exact formula in all three of the standard situations. In addition, it is easy to show that the variance given by (3) is always smaller than that of the situation of regular two phase sampling.

Now as in any stratification model, there is a question of optimal design. The problem addressed here is that of minimum variance given a fixed cost. To this end, we let $T_1 = \sum_k N_{k\cdot}S^2_{k\cdot}$ and $T_2 = \sum_j N_{\cdot j}S^2_{\cdot j}$. We assume, for the design question at hand, that these are known. In reality, of course, only guesses are available. Next, we let D denote the total budget, d_0, the start-up cost, d_1, the cost per unit in the phase one sample, and d_2, the cost per unit in the phase two sample. Letting D_a denote the number of dollars available for sampling per population unit, we have

$$D_a = \frac{D - d_0}{N} = f(d_1 + cd_2). \qquad (4)$$

With f and c subject to constraint (4), we seek to minimize (3), var($\hat{\tau}$), now given by

$$\text{var}(\hat{\tau}) \approx \frac{1-f}{f}T_1 + \frac{1-c}{fc}T_2. \qquad (5)$$

The solution is easily found to be given by

$$c = \left[\frac{d_1T_2}{d_2(T_1 - T_2)}\right]^{1/2} \qquad (6)$$

with f found using (4). If $T_1 \leq T_2$, we automatically take $c = 1$ since then the pre-stratification is more effective than the post-stratification.

In the case of non-proportional sampling, the estimator given is biased and calculations of the bias and standard error in this more general situation are prohibitive. However, a slight modification of the second phase sampling design along with the associated change in the estimator $\hat{\tau}$ yields an estimator which is unbiased. Following a description of the required modification, we compute the variance and an unbiased estimator of the variance and we find an optimal method of allocating sampling resources to the various pre- and post-strata.

The modification to the sampling plan is to leave the second phase sample within pre-strata rather than pooling within post-strata across pre-strata. Thus, given n'_{kj} units appearing in $s' \cap P_{kj}$, we have a function $v_{kj}(\cdot)$ (like $v_{\cdot j}(\cdot)$ in Section 2) which defines a sample size $n_{kj} = v_{kj}(n'_{kj}) = c_{kj} n'_{kj}$ to be taken by simple random sampling from $s' \cap P_{kj}$. Based upon this sample, we obtain the quantities $\bar{y}_{kj}$ and s^2_{kj} which were defined in Section 2. The estimator is now $\hat{\tau} = \sum_k 1/f'_{k\cdot} \sum_j n'_{kj} \bar{y}_{kj}$.

Now, since samples (and thus estimators) are independent between pre-strata, $\hat{\tau}$ is the sum of independent estimators of the K pre-stratum totals, where each estimator is based on a regular double sampling scheme. Thus, the results of Rao (1973) apply to each pre-stratum and we first observe that $\hat{\tau}$ is unbiased because its summands are unbiased estimators of their respective pre-stratum totals. Second, using Rao's results, we have

$$\mathrm{var}(\hat{\tau}) = \sum_k \frac{1}{f_{k\cdot}} \left[(N_{k\cdot} - n'_{k\cdot}) S^2_{kj} + \sum_j N_{kj} S^2_{kj} (1/c_{kj} - 1) \right]. \tag{7}$$

Also, an unbiased estimator of var($\hat{\tau}$) is given by

$$\widehat{\mathrm{var}}(\hat{\tau}) = \sum_k N_{k\cdot} \left[(N_{k\cdot} - 1) \sum_j \left(\frac{n'_{kj} - 1}{n'_{k\cdot} - 1} - \frac{n'_{kj} - 1}{n'_{k\cdot} - 1} \right) \frac{n'_{kj} s^2_{kj}}{n'_{k\cdot} n_{kj}} \right.$$

$$\left. + \frac{N_{k\cdot} - n'_{k\cdot}}{N_{k\cdot}(n'_{\kappa\cdot} - 1)} \sum_j \frac{n'_{kj}}{n'_k} \left(\bar{y}_{kj} - \sum_{j'} \frac{n'_{kj'}}{n'_k} \bar{y}_{kj'} \right)^2 \right]. \tag{8}$$

We note at this point that in the case of proportional sampling considered earlier in this section, we have proposed two different estimators for τ, one based on a pooled second phase sample, the other unpooled. In both cases, the estimator was found to be unbiased, and, also, reduction of formula (7) to the case where $f'_{k\cdot} = f$ for all k and where $c_{kj} = c$ for all k and all j yields formula (3), *i.e.*, the approximate variance for the pooled second phase sampling estimator.

Finally, again following the results in Rao, we derive an optimal allocation of sampling resources. Say that D dollars are available for the two phases of sampling, where sampling a unit in phase 1 from $P_{k\cdot}$ costs $d'_{k\cdot}$ dollars and sampling a unit in phase 2 from $P_{\cdot j}$ costs $d_{\cdot j}$ dollars. Given these costs, we wish to find the values of $f'_{k\cdot}$ and c_{kj} which minimize the variance of $\hat{\tau}$. Using the Cauchy inequality for the phase 2 sample in each pre-stratum, we observe that no matter what the value of $f'_{k\cdot}$, the sampling fraction from post-stratum j is given by

$$c_{kj} = S_{kj} \left(\frac{d'_{k\cdot}}{d_{\cdot j} (S^2_{k\cdot} - \sum_j W_{kj} S^2_{kj})} \right)^{1/2}. \tag{9}$$

Now, the effective expected cost (over both phases of sampling) for each unit sampled in phase 1 and in pre-stratum k is given by

$$d_{k\cdot}^{(e)} = d'_{k\cdot} + \sum_{j} W_{kj}\, c_{kj}\, d_j\,. \tag{10}$$

When viewed in this way, for cost considerations, the first phase of sampling can be seen as a regular stratified sample with (effective) cost of a unit sampled in $P_{k\cdot}$ given by (10). Thus, Cochran (1977,p.97) provides the required formulation of the first phase allocation:

$$\frac{n'_{k\cdot}}{n'} = \frac{N_{k\cdot}\; S_{k\cdot}/\sqrt{d_{k\cdot}^{(e)}}}{\sum_{k'} N_{k'\cdot}\; S_{k'\cdot}/\sqrt{d_{k'\cdot}^{(e)}}} \tag{11}$$

where

$$n' = \sum_{k} n'_{k\cdot} = D \sum_{k} \frac{N_{k\cdot}\; S_{k\cdot}/\sqrt{d_{k\cdot}^{(e)}}}{\sum_{k'} N_{k'\cdot}\; S_{k'\cdot}/\sqrt{d_{k'\cdot}^{(e)}}}\,. \tag{12}$$

Following the modifications suggested by Rao, one can handle the situation where one or more of the c_{kj} turn out to be greater than one. One can also modify the results in the usual way to minimize sampling cost in the case of pre-determined variance.

5. APPLICATIONS

One can employ the method of dual stratification presented here at two levels. At one level, double sampling with pre-strata can be employed with no use of prior information on stratum sizes or stratum averages. At a more complex level, if one has in hand prior information on the number of units in each stratum coming from each pre-stratum, and if the sampler has a level of confidence for this information, then a further reduction in standard error can be obtained by employing this prior information.

This two phase sampling and estimation technique could be used in the proposed nationwide survey to determine the extent of spread of the HTLV-III (Acquired Immune Deficiency Syndrome) virus. The extended incubation period, estimated to be on the average 4.5 years (Lui *et al.* 1986), makes the survey approach imperative, yet there are psychosocial and financial factors which make such a survey extremely difficult to carry out. Thus, methods which assist in reducing sample size while maintaining accuracy must be pursued.

Allen (1984) provides data which suggests a partition of the American population according to a variety of factors which can be used to define risk categories. Known factors, which could be used to define pre-strata, include age, gender, presence of certain diseases, nationality, immigration status, and geographical location. Unknown factors, which could be determined via interview, include sexual preference and drug use. Data on the prevalence of HTLV-III within various subgroups can be both 1) incorporated into the overall estimate of prevalence and 2) used to determine sampling allocations. Such data is available, for example, for blood donors (Kuritsky *et al.* 1986), military results (Redfield and Burke 1987), intravenous drug abusers in Queens, New York (Robert-Guroff *et al.* 1986) and male homosexuals in Greenwich Village (Casareale *et al.* 1984/5). Though this prior information can be used to reduce cost and increase accuracy, confidentiality and sensitivity/specificity of the HTLV-III test remain as significant obstacles which must be addressed carefully before such a study will provide meaningful results.

ACKNOWLEDGEMENT

The author would like to express his appreciation to the reviewer for helpful comments in the area of non-proportional sampling.

APPENDIX

Derivation of Expectation and Variance With No Prior Information and Proportional Sampling

Using the notation given in Section 2, we proceed first with the derivation of $E(\hat{\tau})$. The conditional expectation given s' is $E(\hat{\tau} \mid s') = 1/f \sum_j n'_{\cdot j} \bar{y}'_{\cdot j}$. Then, writing $n'_{\cdot j} \bar{y}'_{\cdot j}$ as $\sum_k n'_{kj} \bar{y}'_{kj}$, we find $E(\hat{\tau}) = E(E(\hat{\tau} \mid s')) = 1/f \sum_j \sum_k E\big(n'_{kj}\, E(\bar{y}'_{kj} \mid n'_{kj})\big) = 1/f \sum_j \sum_k E(n'_{kj})$ $\bar{Y}_{kj} = \tau$ since n'_{kj} is hypergeometric with sampling fraction f and N_{kj} units in pre-stratum k and post-stratum j. Thus, $\hat{\tau}$ is, in this case, unbiased (ignoring approximation A1).

Computation of the variance is along the same lines, yet much more technically detailed. Only certain elements of the computation will be presented and particular emphasis will be placed on the points in the derivation where approximations are made. First, some computation using the two phases of conditioning discussed above, yields

$$\operatorname{var}(E(\hat{\tau} \mid s')) = \frac{1-f}{f} \sum_k N_{k\cdot} S^2_{k\cdot}. \tag{13}$$

We next obtain

$$\operatorname{var}(\hat{\tau} \mid s') = \frac{1-c}{f^2 c} \sum_j \frac{n'_{\cdot j}}{n'_{\cdot j} - 1} \cdot \left[\sum_k (n'_{kj} - 1) s'^2_{kj} + \sum_k n'_{kj} (\bar{y}'_{kj} - \bar{y}'_{\cdot j})^2 \right]. \tag{14}$$

Our second and third approximations are to approximate $n'_{\cdot j} / (n'_{\cdot j} - 1)$ by one (A2) and $(n'_{kj} - 1)$ by n'_{kj} (A3) in equation (14). We now require the expectation of the first term in (14) and find

$$E\left[\frac{1-c}{f^2 c} \sum_j \sum_k n'_{kj} s'^2_{kj} \right] \approx \frac{1-c}{fc} \sum_j \sum_k N_{kj} S^2_{kj}. \tag{15}$$

In (15), one further approximation (A4) is necessary; we ignore the possibility of $n'_{kj} \leq 1$ for any k,j. We also require the expectation of the second term in (14). The exact formula turns out to be

$$\frac{1-c}{fc}\sum_j\left\{\sum_k N_{kj}(\bar{Y}_{kj}-\bar{Y}_{\cdot j})^2+a_1\sum_k S_{kj}^2-a_2\right\} \tag{16}$$

where $a_1 = 1 - f - E[n'_{kj}(1 - n'_{kj}/N_{kj})/n'_{\cdot j}]$ and $a_2 = E[(\sum_k n'_{kj}(\bar{Y}_{kj} - \bar{Y}_{\cdot j}))^2/n'_{\cdot j}]$. We note first that $|a_1| \leq 1$ and thus when combined with N_{kj} in (15), it can be ignored (approximation A5). Also, if in a_2 $n'_{\cdot j}$ is approximated (A6) by its expectation, $fN_{\cdot j}$, since $E[\sum_k n'_{kj}(\bar{Y}_{kj} - \bar{Y}_{\cdot j})] = 0$, we have

$$a_2 \approx \frac{1}{fN_{\cdot j}}\operatorname{var}\left(\sum_k n'_{kj}(\bar{Y}_{kj}-\bar{Y}_{\cdot j})\right) \approx (1-f)\sum_k\frac{N_{kj}}{N_{\cdot j}}(1-W_{kj})(\bar{Y}_{kj}-\bar{Y}_{\cdot j})^2$$

where we have finally approximated $(N_{k\cdot} - 1)$ by $N_{k\cdot}$ (A7) in computing the variance of the hypergeometric variable n'_{kj}. When compared to the similar term with coefficient N_{kj} in (16), we discover that a_2 itself is approximately negligible. Finally, once again ignoring differences between N_{kj} and $(N_{kj} - 1)$ or between $N_{\cdot j}$ and $(N_{\cdot j} - 1)$ (approximation A8), (15) and (16) can be combined to yield

$$E(\operatorname{var}(\hat{\tau}\mid s')) \approx \frac{1-c}{fc}\sum_j\frac{N_{\cdot j}}{N_{\cdot j}-1}\sum_k\left[(N_{kj}-1)S_{kj}^2+N_{kj}(\bar{Y}_{kj}-\bar{Y}_{\cdot j})^2\right]$$

$$=\frac{1-c}{fc}\sum_j N_{\cdot j}S_{\cdot j}^2. \tag{17}$$

Combining (13) and (17), we finally obtain

$$\operatorname{var}(\hat{\tau}) \approx \frac{1-f}{f}\sum_k N_{k\cdot}S_{k\cdot}^2+\frac{1-c}{fc}\sum_j N_{\cdot j}S_{\cdot j}^2. \tag{18}$$

The validity of this approximation rests on three facts. First, when (18) is evaluated in the five examples for which simulated data exists, the results compare very favorably. The approximated standard error given by (12) is 113.25, 108.97, 108.09, 106.77, and 106.32 for $f' = .10$, .20, .30, .40, and .50, respectively. These values are nearly equal to those in Table 3 and the column giving $SE(\hat{\tau})$ and $M = 0$ with m equal to 500 or 2500. Second, the error introduced by each approximation made was analyzed and found, with the possible exception of approximation A6, to be negligible in the case of relatively large population and sample sizes. Even in the case of A6, the law of large numbers indicates that $n'_{\cdot j}$ will be well approximated by its expectation if the sample sizes are reasonably large. Finally, as described in the following, this approximation formula reduces to the exact formula in all three standard situations. First, this situation reduces to the usual stratified sampling according to pre-strata when we take $J = K$, $P_{\cdot j} = P_{k\cdot}$ for $j = k$, and $c = 1$. Here, formula (18) reduces to $\operatorname{var}(\hat{\tau}) \approx (1 - f)/f \sum_k N_{k\cdot} S_{k\cdot}^2$ which is well known to be the exact formula. Also, the estimation scheme described reduces to the usual two phase sampling for stratification when we take $K = 1$ and (18) again reduces to the exact formula (see Cochran 1977, p. 329). Similarly, we obtain the situation of regular stratified sampling by post-strata if we take $f = 1$ (here, K and the pre-stratification become irrelevant), and formula (18) again reduces to the exact value.

REFERENCES

ALLEN, J.R. (1984). Epidemiology of the Acquired Immunodeficiency Syndrome (AIDS) in the United States. *Seminars in Oncology*, 11, 4-11.

CASAREALE, D. *et al.* (1984/5). Prevalence of AIDS-associated retrovirus and antibodies among male homosexuals at risk for AIDS in Greenwich Village. *AIDS Research*, 1, 407-421.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons, Inc.

HAN, C. (1973). Double sampling with partial information on auxiliary variables. *Journal of the American Statistical Association*, 68, 914-918.

HANSEN, M.H., and HURWITZ, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

KURITSKY, J.N. *et al.* (1986). Results of nationwide screening of blood and plasma for antibodies to HTLV-III. *Transfusion*, 26, 205-207.

LUI, K. *et al.* (1986). A model based approach for estimating the mean incubation period of transfusion-associated acquired immuno-deficiency symdrome. *Proceedings of the National Academy of Sciences*, 83, 3051-3055.

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

REDFIELD, R.R., and BURKE, D.S. (1987). Shadow on the land: the epidemiology of HIV infection. *Viral Immunology*, 1, 69-81.

ROBERT-GUROFF, M. (1986). Prevalence of antibodies to HTLV-I, -II, and -III in intravenous drug abusers from an AIDS endemic region. *Journal of the American Medical Association*, 255, 3133-3137.

VARDEMAN, S., and MEEDEN, G. (1984). Admissible estimators for the total of a stratified population that employ prior information. *Annals of Statistics*, 12, 675-684.

WHITE, D. (1987). Mean squared error of estimators using two stage sampling for stratification and prior information. *Proceedings of the Section on Survey Research Methods, American Statistical Association*.