The problem considered is that of estimating the total of a stratified finite population. Vardeman and Meeden (Ann. Stat. 12, 1984, 675-684) have introduced estimators for this situation which employ prior information regarding stratum sizes, averages, and memberships. In a two-stage sampling model for the situation of prior information on stratum sizes, we analyze their estimator for bias and for mean squared error relative to the usual estimators. A natural question then considered is the following: for given levels of confidence in the values specified by our prior information, what weighting constants for these values do we use in the estimators so as to minimize mean squared error? The answer to this question is found to depend upon some knowledge regarding the error in the prior information.

I. INTRODUCTION

Various stratified sampling designs employ various types of prior information. For example, the usual stratification model assumes full prior knowledge regarding stratum memberships. Poststratification is useful when there is global information on stratum sizes but no information on individuals. Two stage sampling for stratification, on the other hand, assumes no prior information on strata.

In each of these methods, the knowledge regarding stratum memberships or stratum sizes is assumed to be complete knowledge or complete lack of knowledge. However, the information required in these models is most commonly available through previous censuses or surveys or from parallel studies on other populations possessing similar global characteristics. It is apparent that usually the information is either dated or at least has applicability to the current study which is open to question. When forming an estimate, one therefore might wish to combine this prior information with analogous information gleaned from the current sample. Vardeman and Meeden (1984) have introduced a pair of estimators for this situation described by Vardeman and Meeden. Using their notation, we have a finite population \( \mathcal{P} \) of \( N \) units labelled \( 1, 2, \ldots, N \) with associated values \( y_1, \ldots, y_N \) which are unknown. Denote the population total by \( \tau = \sum y_i \), where here, and in the following, summations over \( i \) run from \( 1 \) to \( N \). Unit \( i \) possesses stratum membership \( j_i \in \{1, 2, \ldots, J\} \). Unknown constants of interest are the stratum averages \( \bar{y}_j \), stratum sizes \( N_j \) and relative stratum sizes \( p_j = N_j/N \), for \( 1 \leq j \leq J \). Prior guesses for relative stratum sizes are \( \pi_j \). In this section, we regard \( \pi_j \) as constant for each \( j \) since it does not depend upon the current samples. The constant \( M \in [0, 1] \) reflects the confidence in the prior guesses and will be described below.

The sample quantities of interest are the following. First, the stage one sample \( s^* \) is a simple random sample without replacement (WOR) of size \( n^* \) from \( \mathcal{P} \). Here, we observe only stratum memberships \( j_i \) for \( i \in s^* \) and we let \( n_j^* \) denote the number of units in the stage one sample which fall in stratum \( j \). We note that \( (n_1^*, n_2^*, \ldots, n_J^*) \) is multivariate hypergeometric with parameters \( N \) (population size), \( n^* \) (sample size) and \( N_1, \ldots, N_J \) (stratum or group sizes).

The second stage sample is a stratified sample with stratum sample sizes \( n_j = v_j(n_j^*) \), where for each \( j \), \( v_j \) is a function on the non-negative integers such that \( v_j(0) = 0 \), \( v_j(1) = 1 \) and \( 2 \leq v_j(1) \leq 1 \) if \( i \geq 2 \). The set of \( n = \sum n_j \) (summations over \( j \) run from \( 1 \) to \( J \)) units in this second stage sample is denoted by \( s \) and \( y \)-values are determined for each of these units. Thus, we let \( \bar{y}_j \) denote the average of the units in \( s \) and stratum \( j \). Technically,

\[
y_j = \begin{cases} \frac{1}{n_j} \sum_{i \in s, j_i = j} y_i & \text{if } n_j^* > 1 \\ n_j & \text{if } n_j^* = 0. \end{cases}
\]
We are now in a position to consider the meaning of the confidence coefficient \( M \). \( M \) represents the confidence in the collection of guesses of relative stratum sizes, \( \pi_1, \ldots, \pi_J \), and should be considered relative to the confidence in the sample estimates, \( \frac{n_j}{n} \), which is given by the sample size \( n \). Thus, if the prior guesses are given weight equal to that of the current estimates, we take \( M = n \). Extreme cases of no confidence in (and thus no use of) the prior guesses and total confidence in the prior guesses (and thus no use of the current estimates) correspond to \( M = 0 \) and \( M = \infty \), respectively.

We can now construct an estimate \( \hat{\pi}_j \) for \( \pi_j \) which is a weighted average of the prior guess and the sample estimate. This is

\[
\hat{\pi}_j = \frac{\pi_j + n_j}{M + n},
\]

Finally, an estimate \( \hat{\gamma} \) of the population total \( \gamma \) is constructed by replacing, in the formula for \( \gamma \), any unobserved quantity by its estimate. We thus note that, letting \( \mathbf{p}_j \) denote the units in stratum \( j \), the total can be written

\[
\hat{\gamma} = \sum_{j=1}^{J} \hat{\gamma}_j = \sum_{j=1}^{J} \left( \sum_{i \in s \cap \mathbf{p}_j} \gamma_i + \sum_{i \notin s \cap \mathbf{p}_j} \gamma_i \right)
\]

For \( i \in s \cap \mathbf{p}_j \), \( \gamma_i \) are observed and \( \sum_{i \in s \cap \mathbf{p}_j} \gamma_i = n_j \hat{\gamma}_j \). For \( i \in (s \cap \mathbf{p}_j) \setminus (s \cap \mathbf{p}_j) \), \( \gamma_i \) are not observed, but we know that there are \( n_j - n_j^* \) such units. Values \( \hat{\gamma}_i \) for \( i \in \mathbf{p}_j \setminus (s \cap \mathbf{p}_j) \) are likewise not observed; here, there are \( n_j - n_j^* \) such units but this must be estimated by \( (N - n) \hat{\pi}_j \).

Estimating all unobserved quantities, we obtain our estimate of the total,

\[
\hat{\gamma} = \sum_{j=1}^{J} n_j^* \hat{\gamma}_j + (N - n^*) \hat{\pi}_j \sum_{j=1}^{J} \hat{\gamma}_j.
\]

We now seek the bias, variance and mean squared error of \( \hat{\gamma} \). We begin with the expectation, and thus bias, of \( \hat{\gamma} \). All moments are computed using the following two-step conditioning argument. Since the second stage sample depends upon the first stage sample, we condition first on \( s \). The expectation or variance of the resulting function of \( s^* \) is then required. Since the first stage sample is a simple random sample without replacement, the units within a given stratum all have the same probability of selection. It is the number of units selected from each stratum \( n_j^* \) which is random. Thus, conditionally, given \( n_j^* \), the first stage sample from \( \mathbf{p}_j \) is a simple random sample without replacement of size \( n_j^* \), and the second conditioning step is to condition on \( n_j^* \).

Thus, proceeding piecewise, we have \( \mathbb{E}(n_j^* \hat{\gamma}_j) = \mathbb{E}(n_j^* \mathbb{E}(\hat{\gamma}_j | n_j^*)) = \mathbb{E}(n_j^* \hat{\gamma}_j) = n_j^* \hat{\gamma}_j \).

In a similar fashion, we have \( \mathbb{E}(\hat{\gamma} | s^*) = \mathbb{E}(\hat{\gamma} | n_j^*) = \mathbb{E}(\hat{\gamma} | n_j^*) = \mathbb{E}(\hat{\gamma} | n_j^*) \). Conditioned on the first stage sample, \( \hat{\gamma}_i \) is a sample average based on a simple random sample without replacement from units in \( s \cap \mathbf{p}_j \). Any average of a vacuous set of values will be taken as zero. Finally, we have

\[
\mathbb{E}(n_j^* \hat{\gamma}_j) = \mathbb{E}(n_j^* \mathbb{E}(\hat{\gamma}_j | n_j^*)) = \mathbb{E}(n_j^* \hat{\gamma}_j) = n_j^* \hat{\gamma}_j.
\]

Combining these results, we discover \( \mathbb{E}(\hat{\gamma}) = \sum \mathbb{E}(n_j^* \hat{\gamma}_j + (N - n^*) \hat{\pi}_j) \hat{\gamma}_j \). But, since \( N_j = N_j^* \), we note that \( \tau = \sum \mathbb{E}(n_j^* \hat{\gamma}_j + (N - n^*) \hat{\pi}_j \hat{\gamma}_j) \) and the bias, \( B(\hat{\gamma}) = \mathbb{E}(\hat{\gamma}) - \tau \), is therefore given by \( (N - n^*) \mathbb{E}(\mathbb{E}(\hat{\gamma} | n_j^*)) \). But \( \mathbb{E}(\hat{\gamma} | n_j^*) \)

\[
B(\hat{\gamma}) = \frac{M(N - n^*) \mathbb{E}(\hat{\gamma} | n_j^*)}{(M + n^*)}.
\]

We now turn to the derivation of the variance of \( \hat{\gamma} \). As before, we condition on \( s^* \). Letting \( g_j(n_j^*) = n_j^* + (N - n^*) \hat{\pi}_j \), we first obtain the random variables

\[
\mathbb{E}(\hat{\gamma} | s^*) = n_j^* \hat{\gamma}_j,
\]

and

\[
\mathbb{V}ar(\hat{\gamma} | s^*) = n_j^* \mathbb{V}ar(\hat{\gamma}_j | s^*).
\]

The latter formula follows from the fact that if \( j \neq j \), \( \hat{\gamma}_j \) and \( \hat{\gamma}_j \) are conditionally independent given \( s^* \). Also, conditioned on the first stage sample, \( \hat{\gamma}_j \) is a sample average based on a simple random sample without replacement from units in \( s^* \cap \mathbf{p}_j \). If we denote the finite population variance of these \( n_j^* \) units by \( \mathbb{V}ar(\hat{\gamma}_j | s^*) = s_j^2 \), then finally we obtain

\[
\mathbb{V}ar(\hat{\gamma} | s^*) = n_j^* \mathbb{V}ar(\hat{\gamma}_j | s^*) = n_j^* s_j^2.
\]
...n)j. We proceed by deriving $\text{var}(E(T|s^*)) = E(\text{var}(g(n_j)\hat{y}_j^*|n_j) + \text{var}(g(n_j)\hat{y}_j^*|n_j)\hat{y}_j^*)$.

These terms shall be evaluated separately.

First, we have $\text{var}(g(n_j)\hat{y}_j^*) = \text{var}(g(n_j)E(\hat{y}_j^*|n_j)) + E(\text{var}(g(n_j)\hat{y}_j^*)) = \sum_j \text{var}(g(n_j)\hat{y}_j^*)$, since the two variables are conditionally independent given $n^*$.

Second, we have $\text{cov}(g(n_j)\hat{y}_j^*, g(n_j)\hat{y}_j^*) = \sum_j \text{cov}(g(n_j)\hat{y}_j^*, g(n_j)\hat{y}_j^*)$, since the two variables are conditionally independent given $n^*$.

The finite population variance for the $N_j$ units in stratum $j$. We derive the covariances in an exactly analogous fashion. Here, for $j \neq j'$, $\text{cov}(g(n_j)\hat{y}_j^*, g(n_{j'})\hat{y}_{j'}^*) = \sum_j \text{cov}(g(n_j)\hat{y}_j^*, g(n_{j'})\hat{y}_{j'}^*)$, since the two variables are conditionally independent given $n^*$.

We finally note that since, conditioned on $n^*$, the first stage sample from stratum $j$ has the same distribution as a simple random without replacement, $E(\text{var}(T|s^*)) = E(g(n_j)^2(1/n_j - 1/n_j)) = \sum_j E(g(n_j)^2(1/n_j - 1/n_j))$. Combining these results, we obtain

$$\text{var}(\hat{\tau}) = \sum_j \text{var}(g(n_j)\hat{y}_j^*) = \sum_j \text{var}(g(n_j)\hat{y}_j^*)$$

$$+ \sum_j \text{cov}(g(n_j)\hat{y}_j^*, g(n_j)\hat{y}_j^*)$$

$$= \sum_j \text{var}(g(n_j)\hat{y}_j^*) + \sum_j \text{cov}(g(n_j)\hat{y}_j^*, g(n_j)\hat{y}_j^*)$$

where $g(n_j)^2 = g(n_j)^2$. (2.8)

In the usual two stage sampling for stratification situation, where there is no prior information and we take $M = 0$, this formula can be reduced to

$$\text{var}(\hat{\tau}) = \frac{N^2}{N} \text{var}(g(n_j)\hat{y}_j^*) + \frac{N^2}{N} \sum_j \text{var}(g(n_j)\hat{y}_j^*)$$

where $S_j$ is the finite population variance for all of $\hat{\tau}$. This formula is the exact version of Cochran's formula (12.8) (Cochran (1977), p. 329) and can be found in Tucker (1981), p. 144.

3. CHOICE OF WEIGHTING CONSTANTS

In this section, we shall examine the mean squared error of $\hat{\tau}$ and determine optimal values of the weighting constant $M$ according to the following two rules. First, we wish to minimize mean squared error and second, we wish to insure a small likelihood of having mean squared error larger than would be obtained by ignoring the prior information. In the following, we shall explore the general principles of how to obtain a gain in precision and how much gain to expect.

To carry out such an exploration, we take $J = 2$. The only other assumptions we shall make are that the stratum sizes are much larger than one, that $n_j = cn_j$, $j = 1, 2$, $c$ is large enough so that $P[n_j = 0] = 0$, and that $S_{12} = S_{21} = S^2$. We shall also use standardized and scaled $y$-values so that we can take $\hat{y}_1 = 1$ and $\hat{y}_2 = 0$. From the above, 2.4 and 2.8 we have

$$\text{MSE}(\hat{\tau}) = \left[ \frac{n^* M(1 - f^*)}{M + n^*} \right]^2 + \left[ \frac{M + n^*}{M + n^*} \right]^2$$

$$\times \left[ \frac{2}{c \rho_j} \sum_j \left( \frac{n_j + a^*}{n_j} \right)^2 \right]$$

$$+ n^* \rho_j(1 - \rho_j)(1 - f^*)$$

where $f^* = n^*/N$, the first stage sampling fraction and $a = \rho_j^2(N - n^*)/(M + N)$. The remaining moment can be expanded to $E(n_j^*) + 2a + a^2E(n_j^*)$. Letting $g(x) = 1/x$, using the linear approximation to $g(x)$ expanded about $E(n_j^*) = n^* \rho_j$ yields

$$E(n_j^*) = E(g(n_j^*)) = (n^* \rho_j)^{-1}.$$

The error to this approximation is bounded above by $(n^* \rho_j)^{-2} \text{var}(n_j^*)^{1/2} \sum_j \text{var}(g(n_j^*))$. The relative error, which is bounded by $(1 - \rho_j^2(1 - f^*)/n^* \rho_j)^{1/2}$, is small if $n^* \rho_j$ is large enough. With this approximation, $E(n_j^*) = (n_j^* + a^2/n_j^*)^{1/2}$ takes the particularly simple form $n_j^* \rho_j(1 + a^2/n_j^*)^{1/2}$. Using this approximation, and substituting $M = M/n^*$ for $M$, we have

$$\text{MSE}(\hat{\tau}) = \left[ \frac{n^* M(1 - f^*)}{M^* + 1} \right]^2 \left[ (1 - \rho_1^2)^2 \right]$$

$$+ n^* \left[ \frac{M^* + 1 - f^*}{M^* + 1} \right]^2 \sum_j \rho_j^2(1 - \rho_j)(1 - f^*)$$

$$+ \rho_j(1 - \rho_j)(1 - f^*)$$

(3.2)

If we regard the above as a function of $M^*$, say $h(M^*)$, then $h(0)$ is the mean squared error (actually variance) of $\hat{\tau}$ in the usual case where there is no consideration of prior information.
Since we wish to perform an analysis in terms of standard error, \( SE(\hat{\tau}) = \text{MSE}(\hat{\tau})^{1/2} \), then our current goal is analyze \( r(M') = (h(M')/h(0))^{1/2} \) for minimum and \( M' : r(M') \leq 1 \). The function \( r \) is given by

\[
r(x) = \frac{(Ax^2 + Bx + 1)^{1/2}}{x + 1}
\]

where

\[
A = \left[ n \left( 1 - f^* \right) \left( \Pi_1 - \rho_1 \right)^2 
+ \sum_{j=1}^{S_2} \sum_{j=1}^{P_2} \left( \frac{\Pi_j}{\rho_j} \right)^2 \right] 
+ \frac{s_2}{c^2} \frac{\rho_1^2 (1 - \rho_1)^2 (1 - f^*)^2}{c_1^2 + \rho_1^2 (1 - \rho_1) (1 - f^*)}
\]

and

\[
B = \frac{2s_2}{c^2} \frac{\rho_1^2 (1 - \rho_1) (1 - f^*)}{c_1^2 + \rho_1^2 (1 - \rho_1) (1 - f^*)}
\]

Key features of \( r(x) \) are the following:
1) \( r(0) = 1 \) and \( r'(0) < 0 \), implying that for at least some small values of \( M' \), we will obtain a decrease in \( SE(\hat{\tau}) \) from that of the usual estimate.
2) \( r \) is minimized at \( x_0 = (2 - B)/(2A - B) \); if \( 2A \leq B \), then \( r \) is strictly decreasing,
3) \( r(x) = 1 \) at \( x_1 = (2 - B)/(A - 1) \) as well as at \( x = 0 \), and
4) \( r(x) = \lim_{x \to \infty} r(x) = A^{1/2} \).

Except in a few extreme cases, \( r(x) \) is not very sensitive to \( f^* \) or the actual values of \( \rho_1 \) and \( \rho_2 \) (only relative to the guesses \( \Pi_1 \) and \( \Pi_2 \)). For our future analysis, we therefore fix an example: \( f^* = .10 \), \( \rho_1 = .25 \), and \( \rho_2 = .75 \). Since \( \Pi_2 / \rho_2 \) should be near unity for \( j = 1, 2 \), the sum in the middle term of the numerator of \( A \) should also be near unity. Thus, the key variables which dominate \( r(x) \) are \( n^{1/2} |\Pi_1 - \rho_1| \) and \( c^{-1/2} S_2 \) which shall be denoted \( e \) and \( v \), respectively. Figures

![Figure 1. Standard error ratios for various e and v.](image1)

![Figure 2.](image2)

\( M = M_0 \) yields the minimum value of \( r \).
\( M > r^{-1}(1) \) yields

\[ [\text{S.E.}(\hat{\tau})](M) > [\text{S.E.}(\hat{\tau})](0). \]

![Figure 3. Minimum and limiting standard error ratios.](image3)
value of M decreases, as does the value of M be-
known quantities e and v. We thus set n =

r(x) only depends on these four through the two
unknown quantities e and v. We thus set n =
approximated l/P by one, if e < /P1P2 = .137
for r-l(1), carry the following interpretations.
Approximating l/P by one, if e < /P1P2 = .137
then M0 is infinity and the usual poststratifica-
tion estimate is best. Further, if e < /P1P2(1 +
r2) = .454, then r-1(1) = and the standard er-
error using the methods given here is always lower
than when the prior information is ignored. Fi-
gures 3a, b, c, and d show for various v how
r(M0) and r(M) depend on e.

The reader should observe at this point that,
although the sampler may have an idea of the
value of S and thus v, by the definition of the
method being presented, he has little knowledge
of the value of |Pi - P1| except that he hopes it
is near zero. This problem can, however, be
studied from the following point of view. If the
Sampler has a notion of Pi as an estimate based on
a previous sample of size, say, m, then one might
ask what value of M is appropriate, relative to m
and to the known behavior of r(x) as a function
of |Pi - P1|. Since r(x) actually depends upon
|Pi - P1|2, we note the following: If the true
population fraction in stratum 1 at the time of
the previous sample was also P1, if m/N is not
too large and if Pi is the usual estimate of P1,
then mPi is approximately binomial,

E(Pi - P1)2 =

P1(1 - P1)/m and sd((Pi - P1)2) = \sqrt{2E(Pi - P1)2}

(see Johnson and Kotz (1969) p. 51). A reasona-
ble value to use for \hat{e}2 then is its expectation
under this model, \hat{e}2 = (n*/m)cP1(1 - P1). Appro-
ximating Pi/P1 by one and \hat{e}2 by zero we thus
select x = X0 = P1(1 - P1)/\hat{e}2 = m/n*
so the ratio is minimized by taking M = m. This result
reflects our original intuition in regarding M
as a value reflecting the confidence in Pi rela-
tive to the sample estimate, n*/n*. If this
guess at \hat{e}2 is correct, then SE(t) is as small

as we can make it. If this guess is an overesti-
mate of \hat{e}2, then M is smaller than the optimal
value and x < X0. This is relatively painless;
at least here, we are guaranteed that r(x) < 1.
The danger comes in underestimating \hat{e}2 by so much
that x > X1 and the estimate t has higher
standard error than does the usual estimate. Say
that \hat{e}2 = \alpha \hat{e}2, \alpha > 1. Then we select

x = X0 = \frac{\hat{e}2}{\hat{e}2} = \alpha \frac{\hat{e}2}{\hat{e}2} = \alpha x_0.

Noting that X1 \geq 2X0 since \alpha < 2, we are sure
that x < X1 if \alpha < 2, i.e., if (Pi - P1)2 <
2E(Pi - P1)2 = E(Pi - P1)2 + (1/\sqrt{2})sd(Pi - P1)2.

We are thus safe if the true squared error of Pi
is less than .71 standard deviations higher than
its expected value. This fact, when combined
with possible problems in the applicability of
the previous sample, may lead the sampler to sel-
lect M somewhat smaller than m in order to insure
against a higher standard error than when the
prior information is ignored.

4. CONCLUSIONS AND EXTENSIONS

The results in Section 3 strongly indicate
that there are situations where the use of this
estimate is a viable option. In any poststrati-
filed model where there is some information on
stratum sizes, one should put some effort into
either establishing bounds on the error in this
prior information or in approximating its distri-
bution. With such knowledge, even if it is
only a gross approximation, one can use the meth-
ods and results given here to make an educated
choice between 1) total trust in the prior infor-
mation (standard poststratification) if the error
is small enough, 2) no use of the prior informa-
tion (standard two stage sampling for stratifica-
tion) if the error is likely to be large, and 3)
use of the weighted average estimates discussed
here if the error in the prior guess for the dis-
tribution of units among strata is thought to be
moderate. If the latter route is chosen, one can also use these methods to determine a weighting
constant M which is close to optimal.

There are many areas for further study. Some
extensions of these results to the situations
when there is prior information available on in-
dividual stratum memberships and stratum averages
as well as relative stratum sizes is in prepara-
tion for publication. Also, the model, as it
stands, defies variance estimation. Adaptation
of the model to allow for variance estimation is
an important extension. Further work on how to
determine optimal weighting constants is also
needed. A primary need is to apply these methods
to current sampling situations. Finally, in the
theoretical side, further work in establishing
the Bayesian foundations to these methods is re-
quired.
REFERENCES


