POST-STRATIFYING WITH SAMPLE DATA ONLY: CAN USING THE OBVIOUS MODEL HELP?

Phillip S. Kott, National Agricultural Statistics Service
NASS, 3251 Old Lee Highway, Room 300, Fairfax, VA 22030

Key Words: Design-based; Model-dependent; Combined; Unbiased; Variance.

1. INTRODUCTION

Suppose the observations in a stratified simple random sample can be post-stratified into more homogenous groups, but one has no auxiliary information about the population sizes of those groups. It is difficult to effectively incorporate the post-stratification into the conventional design-based framework for estimating a population total. Chapman and Biemer (1984) provides an example of an unsatisfying attempt. If one is willing to assume the obvious model, however, it is easy to develop an estimator that is combined unbiased (its model expected design bias is zero) and has less combined variance (model expected design variance) than the conventional estimator.

The combined unbiased estimator proposed here is by no means new. A variant of it is used, for example, by the U.S. Census Bureau for its Quarterly Financial Report (QFR). The original impetus for the research described in this paper was an internal memorandum (Vogel 1990) at the National Agricultural Statistics Service (NASS). There are doubtless other users and potential users as well. The goal here is to introduce a theoretical framework under which this estimator is unbiased and has an estimable variance. Moreover, since this framework requires that a model be assumed, a test for the model is proposed and briefly discussed.

2. NOTATION

We begin with a stratified simple random sample that has been post-stratified into more homogenous groups. Let $h = 1, ..., H$ index the original sampling strata, and $k = 1, ..., K$ index the groups into which the sample has been post-stratified (and the population could, in principle, have been post-stratified).

Furthermore, let u denote a particular sampling unit, C_k denote the set of all population units in group k, N_k denote the size of C_k, c_k denote the set of those units from C_k in the sample, n_k denote the size of c_k, T_h denote the set of all population units in sampling stratum h, M_h denote the size of T_h, t_h denote the set of those units from T_h in the sample, m_h denote the size of t_h (which we assume to exceed 1), and y_u denote the value of interest for unit u.

We will assume that n_k is greater than unity for all k. Since the n_k are random variables, this means that $\text{Prob}(n_k > 1) \approx 1$. For simplicity, we will treat this near equality as an exact equality from now on.

The total we wish to estimate is

$$Y = \sum_{k=1}^{K} \sum_{u\epsilon C_k} y_u = \sum_{k=1}^{K} N_k\overline{Y}_k. \quad (1)$$

Note: Y_k is implicitly defined by equation (1)).

3. THE COMBINED UNBIASED ESTIMATOR

Consider the following estimator for Y:

$$\hat{Y}^C = \sum_{k=1}^{K} \hat{N}_k \, \overline{y}_k, \quad (2)$$

where $\overline{y}_k = \sum_{u\epsilon c_k} y_u/n_k$, $\hat{N}_k = \sum_{u\epsilon c_k} W_u$,

and $W_u = M_h/m_h$ for $u \epsilon T_h$ is the sampling weight of unit u. For simplicity, we are ignoring the possibility of nonresponse.

Under traditional design-based sampling theory $\hat{N}_k$ is an unbiased estimator for N_k, but y_k is not, in general, an unbiased estimator for Y_k. This is because not all of the sampling units in c_k necessarily have the same selection probability.

There is an obvious model under which y_k is an unbiased estimator for Y_k. Suppose the y_u can be treated as if they satisfy the stochastic equation:

$$y_u = \mu_k + e_u \text{ for } u \epsilon C_k, \quad (3)$$

where the e_u are independent random variables with $E_M(e_u) = 0$ and $Var_M(e_u) = \sigma_k^2$ for $u \epsilon C_k$, and the subscript M on the expectation (variance) operator denotes expectation (variance) with respect to the model. It is easy to see that y_k is an model unbiased estimator for Y_k in the sense that $E_M(y_k - Y_k) = 0$.

We will be combining design and model-based sampling theory in this analysis. Särndal (1978) provides an excellent introduction to the twin topics of design and model-based inference.

As already noted, $\hat{N}_k$ is a design unbiased estimator for N_k. We can write this formally as $E_D(\hat{N}_k) = N_k$. If the model in (3) holds, then the combined model and design bias of $\hat{Y}^C$ is zero in the sense that

$$E(\hat{Y}^C - Y) = E_M[E_D(\hat{Y}^C - Y)]$$

$$= E_D[E_M(\hat{Y}^C - Y)]$$

$$= E_D[E_M(\Sigma_k \hat{N}_k\overline{y}_k - \Sigma_k N_k\overline{Y}_k]$$

$$= E_D(\Sigma_k \hat{N}_k \mu_k - \Sigma_k N_k\mu_k)$$

$$= 0.$$

The proof is trivial once you realize that $\hat{N}_k$ and N_k are not random variables with respect to the model in (3). Consequently, we will say that $\hat{Y}^C$ is "combined unbiased." (Technical note: since we have assumed Prob$(n_k > 1) \approx 1$ for all k, there is no need to condition the E_D operator on $n_k > 1$.)

4. THE COMBINED VARIANCE OF $\hat{Y}^C$

Let us call the combined model and design expectation of $(\hat{Y}^C - Y)^2$ the "combined variance" of $\hat{Y}^C$. The combined variance of $\hat{Y}^C$ is then

$$\begin{aligned}\mathrm{Var}(\hat{Y}^C) &= E[(\hat{Y}^C - Y)^2] \\ &= E[(\Sigma_k \hat{N}_k \bar{y}_k - \Sigma_k N_k \bar{Y}_k)^2] \\ &= E[\{\Sigma_k (\hat{N}_k - N_k)\bar{Y}_k + \\ &\qquad \Sigma_k \hat{N}_k (\bar{y}_k - \bar{Y}_k)\}^2] \qquad (4) \\ &= E[(\Sigma_k (\hat{N}_k - N_k)\bar{Y}_k)^2] + \\ &\qquad \Sigma_k E_D[\hat{N}_k^2(1/n_k - 1/N_k)]\sigma_k^2 \\ &= \quad V_1 \quad + \quad V_2 .\end{aligned}$$

(We have made repeated use of the equality $E_M[Y_k(y_k - Y_k)] = 0$.) The component V_1 essentially measures the contribution to the combined variance of estimating the N_k with the $\hat{N}_k$, while V_2 measures the contribution to the combined variance of estimating the Y_k with the y_k. In most design-based and model-based analyses of stratified samples, the N_k are fixed and known, hence there is no variance component like V_1. It is important to realize that this is not the case here.

An estimator for V_2 is

$$v_2 = \Sigma_k(\hat{N}_k{}^2/n_k - \hat{N}_k)s_k{}^2,$$

where

$$s_k{}^2 = \sum_{u\epsilon c_k} y_u{}^2 - (\sum_{u\epsilon c_k} y_u)^2/n_k]/(n_k - 1).$$

Observe that $E_M(s_k{}^2) = E_M(\sigma_k{}^2)$, while $\hat{N}_k{}^2/n_k - \hat{N}_k$ is a nearly design biased estimator for $E_D[\hat{N}_k{}^2(1/n_k - 1/N_k)]$ (technically, it is asymptotically unbiased; see Särnal, Swensson, and Wretman (1991). Thus, if anything, v_2 may have a slight combined bias because $E_D(\hat{N}_k{}^2/N_k) \geq E_D(\hat{N}_k) = N_k$.

Now

$$\begin{aligned}V_1 &= \mathrm{Var}_D(\Sigma_k \hat{N}_k \bar{Y}_k) \\ &= \mathrm{Var}_D(\Sigma_h \sum_{u\epsilon t_h} W_u \bar{Y}_{(u)}),\end{aligned}$$

where $Y_{(u)} = Y_k$ when $u\epsilon c_k$. when $u\epsilon c_k$. This suggests the following estimator for V_1:

$$v_1 = \sum_h (M_h^2/m_h)(1 - m_h/M_h) \qquad (5)$$

$$[\sum_{u\epsilon t_h} \bar{y}_{(u)}^2 - (\sum_{u\epsilon t_h} \bar{y}_{(u)})^2/m_h]/(m_h - 1),$$

where $y_{(u)} = y_k$ when $u\epsilon c_k$. Since $E_M(y_k) \geq E_M(Y_k)$ implies

$$E_M[\sum_{u\epsilon t_h} \bar{y}_{(u)}^2 - (\sum_{u\epsilon t_h} \bar{y}_{(u)})^2/m_h] \geq$$

$$E_M[\sum_{u\epsilon t_h} \bar{Y}_{(u)}^2 - (\sum_{u\epsilon t_h} \bar{Y}_{(u)})^2/m_h],$$

v_1 may have a slight positive combined bias, but -- like v_2 -- it is nearly combined unbiased.

5. THE DESIGN-BASED ESTIMATOR

The conventional design-based estimator for Y is

$$\hat{Y}^D = \sum_k \sum_{u\epsilon c_k} W_u y_u = \sum_k \hat{N}_k \bar{y}_k{}^D, \qquad (6)$$

where

$$\bar{y}_k{}^D = \sum_{u\epsilon c_k} W_u y_u / \sum_{u\epsilon c_k} W_u = \sum_{u\epsilon c_k} w_u y_u,$$

and $w_u = W_u / \sum_{v \epsilon c_k} W_v$.

Since $\hat{Y}^D$ is design unbiased, it must also be combined unbiased.

It is fairly easy to show that the combined variance of $\hat{Y}^D$ has the form $V_1^D + V_2^D$, where $V_1^D = V_1$ from equation (4), and

$$V_2^D = \sum_k E_D[\hat{N}_k^2(\sum_{u \epsilon c_k} w_u^2 - 1/N_k)]\sigma_k^2. \tag{7}$$

The value of V_2^D in (7) is greater than or equal to the value of V_2 in (4) (since this is true for any sample, it must be true when averaged over all possible samples). Strict equality holds only when all the w_u within each of the C_k are equal, in which case $\hat{Y}^D = \hat{Y}^C$.

As long as the model is correct, $\hat{Y}^C$ is a better estimator than $\hat{Y}^D$ in terms of combined variance. The advantage of $\hat{Y}^D$ over $\hat{Y}^C$, of course, is that it need not rely on a model to assure its unbiasedness in some meaningful sense (i.e., under the design).

6. A STATISTICAL TEST

As noted in the previous section, $\hat{Y}^C$ is better than $\hat{Y}^D$ if the model in equation (3) is correct; but is it? If the model were correct, then the difference $\hat{Y}^C - \hat{Y}^D = \Sigma_k N_k(y_k - y_k^D)$ would have a model expectation of zero and a model variance of

$$\mathrm{Var}_M(\hat{Y}^C - \hat{Y}^D) =$$

$$= \sum_k \hat{N}_k^2 \sigma_k^2 (\sum_{u \epsilon c_k} w_u^2 - 1/n_k).$$

This means that the statistic

$$t = \frac{\hat{Y}^C - \hat{Y}^D}{[\sum_k \hat{N}_k^2 \, s_k^2 (\sum_{u \epsilon c_k} w_u^2 - 1/n_k)]^{1/2}}. \tag{8}$$

would have asymptotically a standard normal distribution.

By calculating t from the sample, we can test the model. If $|t| \leq 1$, then the model appears to be quite reasonable. Conversely, if $|t| \geq 2$, then belief in the model is very hard to sustain. If $|t|$ is between 1 and 2, we are in a twilight region in which it is impossible to determine with confidence whether the size of the difference between $\hat{Y}^C$ and $\hat{Y}^D$ can be wholly attributable to sampling error of not. Since we are equally concerned with type 2 error (accepting the model when it is false) as type 1 (rejecting it when it is true), it may be imprudent to adopt the conventional "reject the model only when $|t| \geq 2$" rule.

7. SOME REMARKS

The temptation to post-stratify a stratified simple random sample into groups and construct an estimator like $\hat{Y}^C$ in equation (2) is often great. As Chapman and Biemer (1984) demonstrates, however, a purely design-based analysis of $\hat{Y}^C$ is difficult and not very fruitful. More common, I suspect, are ad hoc and incomplete model-based analyses of $\hat{Y}^C$ that ignore the randomness of the estimators of the population group sizes (the N_k).

This paper used a combination of design and model-based principles to explore the use of $\hat{Y}^C$. We saw that if one accepts the obvious model, then $\hat{Y}^C$ is combined unbiased and has less combined variance than the design-based estimator, $\hat{Y}^D$ in equation (6). A test for the model was also proposed. When the data (through the test) or common sense tells us that the model is unreasonable, there seems little justification for preferring $\hat{Y}^C$ over $\hat{Y}^D$.

An estimator for the combined variance of $\hat{Y}^C$ was proposed that has a slight combined bias. The construction of a combined unbiased estimator for the combined variance of $\hat{Y}^C$ is possible, but it is very messy and sheds little light onto the analysis. It is left for the interested reader.

There are some surface similarities between $\hat{Y}^C$ and the synthetic estimator (see Särnal, Swensson, and Wretman 1991, p. 410). Both employ models to estimate post-stratification group means (the Y_k in equation (1)). In synthetic estimation, however, the group sizes (the N_k) are known, while $\hat{Y}^C$ incorporates design-based estimators for these values.

8. A POSTSCRIPT FROM REAL LIFE

This paper proposes a method for testing whether the model inherent in $\hat{Y}^C$ holds. Sadly, this method has never been applied to real data. The author proposed it first at NASS, where analysts determined that the variant of $\hat{Y}^C$ under investigation could not pass the "laugh test" (it produced estimates so far from $\hat{Y}^D$ as to be judged biased by appearances alone).

The version used for the Census Bureau's QFR survey appears to be more reasonable than the one contemplated at NASS. Nevertheless, the statisticians there are committed to using a fully design-based estimator in the future and don't feel the need for any elaborate testing. To be fair, the present program, which they want to replace, estimates variances using v_2 alone. Moreover, complicated sample rotation and nonresponse adjustment schemes make appropriate variance estimation a lot more difficult than presented here.

REFERENCES

CHAPMAN, D. W. and BIEMER, P. (1984). "A comparison of two estimators for the Quarterly Financial Report Survey." Statistical Research Division internal report, U.S. Bureau of the Census.

SÄRNDAL, C.E (1978). Design-based and model-based inference in survey sampling. Scandinavian Journal of Statistics: Theory and Applications, 5, 27-52.

SÄRNDAL, C.E, SWENSSON, B., and WRETMAN, J (1991). Model Assisted Survey Sampling. New York: Springer-Verlag.

VOGEL, F. (1990). "A Strawman Proposal." Internal NASS Memorandum.