United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research and
Applications
Division

NASS Staff Report
Number SRB-88-05

May 1988

Estimating the Non-Overlap Variance Component for Multiple Frame Agricultural Surveys

Phillip S. Kott
Read Johnston

ESTIMATING THE NON-OVERLAP VARIANCE COMPONENT FOR MULTIPLE FRAME AGRICULTURAL SURVEYS, by Phillip S. Kott and Read Johnston, Research and Applications Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC 20250, May 1988. Staff Report No. SRB-NERS-88-05.

ABSTRACT

This paper proposes a new procedure for estimating the variances of those parts of a Multiple Frame Agricultural Survey that rely on the enumeration of a subsample of JES non-overlap (NOL) tracts. This new variance estimator, which is a generalization of Cochran and Huddleston (1), has several advantages over the one currenty in use by the NASS staff. It is better from a theoretical perspective and also yields more realistic results than the current NOL variance estimator.

KEY WORDS

Bias, model, sample design, subsample, two stage (phase) sampling,

ACKNOWLEDMENTS

The authors would like to thank Bill Iwig and Barry Ford for the work that brought the issue of variance estimation in Agricultural Surveys to our attention. We would also like to thank Gretchen McClung for providing the data used in our analysis and for helpful comments on early drafts of this manuscript. We would further like to thank Ron Fecso, George Hanuschak, and Ron Bosecker for their helpful comments.

May, 1988

Washington, DC

TABLE OF CONTENTS

SUMMARY

The National Agricultural Statistics Service (NASS) currently employs a nonstandard two stage sampling technique for the non-overlap portion of its Multiple Frame Agricultural Surveys. This sampling design leads to unbiased and relatively efficient estimators. The variances of these estimators, however, are difficult to estimate. In fact, the variance estimation formula currently in use, while seemingly reasonable, has a major flaw - a tendency to significantly underestimate true variances.

An alternative, nearly unbiased, formula is suggested for agency use that does not have this flaw. An empirical analysis reveals that the current variance formula, on average, yields significantly lower state variance estimates than the alternative. This means that the current formula has a strong tendency to underestimate variances.

In addition to estimating the variance from the non-overlap portion of one of its Agricultural Surveys, NASS is also interested in separating the contributions to variance from the two stages of sample selection. A slight modification of the present estimator for the variance of the second stage of sampling is proposed here (the first stage variance estimator is revamped completely). The modified estimator is based on a reasonable model, while the one currently in use is ad hoc. An empirical analysis, however, fails to reveal statistically significant differences between the two approaches.

ESTIMATING THE NON-OVERLAP VARIANCE COMPONENT FOR MULTIPLE FRAME AGRICULTURAL SURVEYS

By Phillip S. Kott and Read Johnston

INTRODUCTION

The National Agricultural Statistics Service (NASS) draws an area frame sample of land segments for its June Enumerative Survey (JES). Subunits (tracts) from the June area segments are then subsampled and combined with samples drawn from the NASS list sampling frames for the Multiple Frame Agricultural Surveys. We are concerned here only with that part of an Agricultural Survey estimate originating from the area subsample. This is also known as the NOL (non-overlap/not on the list) estimate, since only farm tracts sampled in June that are not on a relevant list sampling frame are enumerated as part of the Agricultural Survey area subsample.

NASS statisticians have for some time been aware that the variances of the NOL estimates are difficult to estimate. This is because of the rather unusual two stage (or phase) design of an Agricultural Survey area sample. First a stratified sample of area segments is selected for the JES. During the JES, the segments are divided into tracts which are parts of separate farm operations or non-agricultural areas within the segment. Then the tracts within those segments are restratified based on their JES questionnaire responses and list status with no regard to segment or original stratum.

Restratification gives NASS the ability to use the information collected in June in a highly cost effective manner. Those NOL tracts likely to contain large farm values of interest to the Agricultural Survey are often placed into certainty strata; conversely, tracts likely to have no values of interest are placed into strata with very low subsampling fractions.

The Agricultural Survey NOL subsample of JES tracts is more efficient (i.e., produces estimates with less variance for the same cost) than a conventionally drawn subsample would be. This is because using information obtained from the JES questionnaire allows NASS to focus the subsample on tracts that

are likely to have positive quantities of the farm values being enumerated by the Agricultural Survey. On the other hand, while NASS statisticians are confident in having a procedure that produces estimates with relatively low variances, they are less certain about how those variances should be estimated.

Presently NASS uses a variation of a variance estimating method originally proposed by Hartley (3). Recent corrections to the operational formula suggested by McClung (4), which make it more consistent with Hartley's original proposal, have exposed some substantial flaws in the current variance estimating methodology. Bill Iwig of the Survey Sampling Branch has observed cases where, had all JES tracts been subsampled with certainty and their June values used in place of corresponding Agricultural Survey values, the estimated variances would be less than the matching JES variance estimates. This clearly should not be.

We will suggest an alternative estimation formula for Agricultural Survey NOL variances. Our approach is a generalization of a suggestion by Cochran and Huddleston (1), which like Hartley assumes that the JES sample is the result of unstratified simple random sampling (something that was very close to true at the time both papers were written).

We begin with a "simple" statement of our variance estimation formula. Due to the complex sampling design, unfortunately, this requires the development of some complicated notation. We then show the theoretical reasoning behind its development. Using December 1986 Agricultural Survey data for harvested acres of winter wheat in 41 states, we compare our formula with the McClung-corrected version (4) of the present operational variance estimation method. We find the latter to produce significantly lower estimates on average, as Iwig's observation suggests would be the case. A proposed innovation concerning the estimation of the variance contribution of the second stage of sampling is also studied using this data set. The results, unfortunately, are inconclusive.

It should be noted that throughout this paper we assume that every tract subsampled by the Agricultural Survey provides complete and accurate information on the value of interest. Unfortunately,

this assumption is not generally true. The problem of how to estimate variances in the presence of nonresponse and/or measurement errors is a very thorny one and will have to wait for another time.

THE NEW VARIANCE ESTIMATION FORMULA

Since there are two levels of stratification, we will follow the NASS summary system naming conventions and refer to the JES substrata as districts. (Note: the JES substrata were formerly known as paper strata; they are original land use strata subdivided geographically. The Agricultural Survey strata from which equal probability subsamples are actually drawn are called select strata. Because a single tract is often subsampled from a given select strata, NASS combines a number of select strata into summary strata for variance estimation (and other) purposes.

Strictly speaking, this practice makes design unbiased estimation of variances impossible. Nevertheless, as we will argue in the next section, one can construct a variance estimator with good properties under a reasonable model.

Let

$h=1,2, \ldots, H$ denote the summary strata (for a Agricultural Survey in some state),

$D=1,2, \ldots, L$ denote the districts,

$j=1,2, \ldots, n_D$ denote the JES sampled segments in district D,

n_D/N_D denote the JES sampling fraction of tracts in D,

x_i denote the farm value of interest for tract i,

T_h denote the number of JES tracts in h,

v_h denote the number of Agricultural Survey tracts in h,

w_i denote the second stage sampling weight for tract i (the inverse of its second stage selection probability),

$y_i=(N_D/n_D)x_i$ denote the first stage expanded farm value of tract i,

$z_i=w_iy_i$ denote the fully expanded farm value of tract i,

S^1 denote the set of all JES tracts whether enumerated by the Agricultural Survey or not,

S_j denote the set of all Agricultural Survey tracts in segment j,

S_D denote the set of all Agricultural Survey tracts in district D,

R_h denote the set of all Agricultural Survey tracts in stratum h,

$y_{jh}=\sum_{i\in S_j\cap R_h} y_i$ denote the total first stage expanded farm value of all Agricultural Survey tracts in stratum h and segment j,

$y_{Dh}=\sum_{i\in S_D\cap R_h} y_i$ denote the total first stage expanded farm value of all Agricultural Survey tracts in stratum h and district D,

$y_{\cdot h}=\sum_{i\in R_h} y_i$ denote the total first stage expanded farm value of all Agricultural Survey tracts in stratum h,

$z_{j\cdot}=\sum_{i\in S_j} z_i$ denote the total <u>fully</u> expanded farm value of all Agricultural Survey tracts in segment j,

$z_{D\cdot}=\sum_{i\in S_D} z_i$ denote the total fully expanded farm value of all Agricultural Survey tracts in district D, and

$Y=\sum_{i\in S^1} y_i$ denote the total first stage expanded farm value of <u>JES</u> tracts.

The Agricultural Survey estimator is

$$\hat{X}=\sum_{h=1}^{H}\sum_{i\in R_h} z_i=\sum_{h=1}^{H}\sum_{i\in R_h} w_iy_i.$$

The variance estimator for $\hat{X}$ we propose is

$$\hat{var} = \sum_{D=1}^{L} n_D/(n_D-1)\ [(\sum_{j=1}^{n_D} z_j.^2) - z_D.^2/n_D] +$$
$$\sum_{h=1}^{H} (([\sum_{i\epsilon R_h} w_i^2] - T_h)\ [1/(v_h(v_h-1))]\ \cdot \qquad (1)$$
$$[\sum_{D=1}^{L} n_D/(n_D-1)\{[\sum_{j=1}^{n_D} y_{jh}^2] - y_{Dh}^2/n_D\} - y.{}_h^2]\}.$$

NASS statisticians are also interested in breaking the variance down into two useful components: the between segment variance and the within (summary) stratum variance. The former is due to JES sampling and the latter to Agricultural Survey subsampling. Our estimator for the within stratum or second stage variance of $\hat{X}$ is

$$\hat{var}_2 = \sum_{h=1}^{H} ((\sum_{i\epsilon R_h} w_i^2) - T_h)\ [1/(v_h-1)][(\sum_{i\epsilon R_h} y_i^2) - y.{}_h^2/v_h].$$

This is exactly equal to the second stage variance estimator currently in use when all the w_i within each respective summary stratum are equal.

An estimator for the between segment or first stage variance is simply

$$\hat{var}_1 = \hat{var} - \hat{var}_2.$$

This is very much different than the first stage

variance estimator currently in use in both its original and McClung-corrected forms.

THE THEORETICAL DEVELOPMENT

The Agricultural Survey estimator $\hat{X}$ described in the previous section is a design unbiased estimate of X, which is the sum of the x_i across all tracts in the population whether enumerated by the Agricultural Survey or not. To see this, observe that $Y=\sum_{i\varepsilon S^1} y_i=\sum_{i\varepsilon S^1}(N_D/n_D)x_i$ is a design unbiased estimator of X with respect to the first stage of sampling, while $\hat{X}$ is a design unbiased estimator of Y with respect to the second sampling stage. Mathematically, $E_1(Y)=X$ and $E_2(\hat{X})=Y$, which implies $E(\hat{X})=E_1E_2(\hat{X})=X$.

From any textbook on design-based sampling theory (e.g., Cochran, 2, p. 276), we know that the variance of a two stage estimator like $\hat{X}$ is

$$\text{var}(\hat{X})= \text{var}_1[E_2(\hat{X})] + E_1[\text{var}_2(\hat{X})], \qquad (2)$$

where E_k and var_k denote, respectively, expectation and variance with respect to the k^{th} stage of sampling.

The first term in equation (2) is called the first

stage variance because it equals the variance that would be obtained if every JES tract were part of the Agricultural Survey subsample.

The second term in (2) is called the second stage variance, but that is not strictly speaking true. The second stage variance (really $var_2(\hat{X})$) can only be defined with respect to a given JES sample. The second term in (2) is actually the average of second stage variances taken over all possible JES samples (and weighted by the probability of drawing each sample).

Despite this slight confusion about the second stage variance, it is easier to estimate than the first stage variance and we will attack it first. The only difficulty results from the practice of collapsing select strata into summary strata.

The problem with first stage variance estimation (so easy in June when there is only one stage) is that total Agricultural Survey values for the segments in the JES sample can only be estimated using the Agricultural Survey subsample. As we shall see, putting estimated segment totals in place of real totals in the usual one stage variance formula biases the resulting estimator.

Second Stage Variance Estimation

First note that a formula for an unbiased estimator of $\text{var}_2(\hat{X})$ given any JES sample is automatically an unbiased estimator of $E_1[\text{var}_2(\hat{X})]$. Mathematically, if $E_2[v_2-\text{var}_2(\hat{X})]=0$ for some estimator v_2 given any JES sample, then the first stage expectation of $E_2[v_2-\text{var}_2(\hat{X})]$ must also be zero. Consequently, $E(v_2)=E_1E_2(v_2)=E_1[\text{var}_2(\hat{X})]$.

Were there no distinction between summary and select strata and were the subsample drawn using simple random sampling within each summary stratum, then

$$\hat{\text{var}}_2^D=\sum_{h=1}^{H}(T_h^2/v_h - T_h)\ [1/(v_h-1)][(\sum_{i\varepsilon R_h} y_i^2) - y_{\cdot h}^2/v_h]$$

would be the conventional design unbiased estimator for $\text{var}_2(\hat{X})$.

Up until this point we have followed standard practice and suppressed the prefix "design" from the terms "unbiased" and "variance." In this and the sections that follow we will be introducing models to aid in the analysis. As a result it is important to draw a distinction between design and model-based characteristics.

Since select strata are not identical to summary strata, we must rely on a model linking together all tracts within the same summary stratum. In particular, if all the y_i in each summary stratum h are assumed to be uncorrelated random variables with a common mean (m_h) and variance (s_h^2), then a model based analogue of $\hat{var}_2^D$ is

$$\hat{var}_2^M = \sum_{h=1}^{H} \left(\left\{ \sum_{i \in R_h} w_i^2 \right\} - T_h \right) [1/(v_h-1)] \left[\left\{ \sum_{i \in R_h} y_i^2 \right\} - y_{\cdot h}^2 / v_h \right]. \qquad (3)$$

This estimator is clearly a model unbiased estimator of the model variance of $\hat{X}$ as an estimator of Y:

$$E^M[(\hat{X}-Y)^2] = \sum_{h=1}^{H} \left(\left\{ \sum_{i \in R_h} w_i^2 \right\} - T_h \right) [1/(v_h-1)] s_h^2,$$

where E^M denotes expectation with respect to the model.

The estimator $\hat{var}_2^M$ is also an estimator of the second stage design variance of $\hat{X}$ (that is, $var_2(\hat{X})$). Although not necessarily design unbiased, it is not difficult to show that the model expectation of the design bias of $\hat{var}_2^M$ is zero. Consequently, $\hat{var}_2^M$ can be said to be an <u>almost</u> design unbiased estimator of the second stage design variance of $\hat{X}$ under a reasonable model.

Observe that if all the w_i are equal within each respective summary strata (as would be the case had tracts been selected via simple random sampling within summary strata) then $\hat{var}_2^M$ collapses into $\hat{var}_2^D$. As a result, $\hat{var}_2^M$ can be viewed, at the very least, as an ad hoc generalization of the standard design unbiased variance estimator.

NASS is forced to an ad hoc technique in this situation because some probability select strata contain only one tract, which renders conventional design unbiased variance estimation impossible. It seems prudent under such circumstances to choose an estimator of the design variance of $\hat{X}$ with good properties under a reasonable model.

The McClung-corrected version of the ad hoc estimator currently in use is

$$\hat{var}_2^M = \sum_{h=1}^{H} (1 - [v_h/T_h])\ v_h/(v_h-1)\left(\left(\sum_{i\in R_h} z_i^2\right) - z_{\cdot h}^2/v_h\right). \qquad (4)$$

(Note: In the operational estimator, the number of JES tracts in stratum h (T_h) has mistakenly been replaced by the first stage estimate of the total number of tracts in h.)

Unlike $\hat{var}_2^M$, $\hat{var}_2^N$ is not a model unbiased estimator of $var_2(\hat{X})$ when the w_i vary within some summary stratum. This can most easily be seen by considering the following special case of the model.

Suppose all the y_i within each summary stratum were equal (in model notation, all $s_h^2=0$). Then $var_2(\hat{X})$ would be zero since every possible second stage sample would yield the same result. From (3) we see that $\hat{var}_2^M$ would also be zero; $\hat{var}_2^N$, on the other hand, would only be zero if all the z_i in each summary strata were equal. That, in turn, requires that w_i be constant within each summary stratum.

First Stage Variance Estimation

Consider a segment j within district D. The value $z_j.$ is an unbiased estimator of (N_D/n_D) times the total farm value among all tracts in segment j whether in the Agricultural Survey subsample or not. Consequently, $E_2(z_j.)$ is exactly (N_D/n_D) times the total farm value among all tracts in segment j. With this in mind, the following would be a (nearly) unbiased estimator of the first stage variance of $\hat{X}$:

$$\hat{\text{var}}_1[E_2(\hat{X})] = \sum_{D=1}^{L} [n_D/(n_D-1)][\sum_{j=1}^{n_D} (E_2(z_j.))^2 - (E_2(z_D.))^2/n_D]. \quad (5)$$

(This assumes that the JES sample was drawn using a with replacement sample of segments. This assumption, while strictly speaking false, is close enough to the truth and simplifies matters considerably.)

Taken as is, equation (5) is useless since it supposes we know what the $(E_2(z_j.))^2$ and $(E_2(z_D.))^2$ are. Nevertheless, it does suggest that $\text{var}_1[E_2(\hat{X})]$ would be estimated in a design unbiased manner if one could find design unbiased estimators for the $(E_2(z_j.))^2$ and $(E_2(z_D.))^2$ to plug appropriately into (5).

Observe first that $z_j.^2$ and $z_D.^2$ are <u>not</u> design unbiased estimators of $(E_2(z_j.))^2$ and $(E_2(z_D.))^2$. In fact,

$$E_2(z_j.^2) = (E_2(z_j.))^2 + \text{var}_2(z_j.),$$

while

$$E_2(z_D.^2) = (E_2(z_D.))^2 + \text{var}_2(z_D.).$$

These equations, however, hint towards alternative estimators for $(E_2(z_j.))^2$ and $(E_2(z_D.))^2$. For example, if v_{2j} and v_{2D} were design unbiased estimators of $\text{var}_2(z_j.)$ and $\text{var}_2(z_D.)$ respectively, then $z_j.^2 - v_{2j}$ would be a design unbiased estimator

of $E_2(z_j.^2)$, while $z_D.^2 - v_{2D}$ would be a design unbiased estimator of $E_2(z_D.^2)$.

Were there no differences between select and summary strata, rendering all the w_i within each summary stratum equal, then

$$\hat{var}_{2j}^{M} = \sum_{h=1}^{H} \left(\left\{ \sum_{i \in R_h} w_i^2 \right\} - T_h \right) \left[1/(v_h - 1) \right] \left[\left\{ \sum_{i \in S_j \cap R_h} y_i^2 \right\} - y_{jh}^2 / v_h \right]$$

and

$$\hat{var}_{2D}^{M} = \sum_{h=1}^{H} \left(\left\{ \sum_{i \in R_h} w_i^2 \right\} - T_h \right) \left[1/(v_h - 1) \right] \left[\left\{ \sum_{i \in S_D \cap R_h} y_i^2 \right\} - y_{Dh}^2 / v_h \right]$$

would be design unbiased estimators of $var_2(z_j.)$ and $var_2(z_D.)$ (Cochran, 2, p. 143, equation (5A.68)). Since there are such differences, we must instead be comforted that $\hat{var}_{2j}^{M}$ and $\hat{var}_{2D}^{M}$ are respectively almost design unbiased estimators of $var_2(z_j.)$ and $var_2(z_D.)$ under reasonable models.

These models may need a bit of explaining. In support of $\hat{var}_{2j}^{M}$, we first define the variable $y_i^{(j)}$ as equal to y_i when i is in segment j and zero otherwise, and then assume the following model: for each h, the $y_i^{(j)}$ are uncorrelated random variables with a common mean and variance. In support of $\hat{var}_{2D}^{M}$, we analogously define $y_i^{(D)}$ and proceed from there.

Putting It All Together

Plugging $z_j.^2 - v_{2j}^M$ and $z_D.^2 - v_{2D}^M$ respectively into $(E_2(z_j.))^2$ and $(E_2(z_D.))^2$ of equation (5), we have an estimator for the first stage variance of $\hat{X}$. This can then be added to (2) to yield (after some manipulation) this estimator:

$$\hat{var} = \sum_{D=1}^{L} n_D/(n_D-1) \left[\left(\sum_{j=1}^{n_D} z_j.^2\right) - z_D.^2/n_D\right] + \sum_{h=1}^{H} \left\{\left(\left[\sum_{i\in R_h} w_i^2\right] - T_h\right) \left[1/(v_h(v_h-1))\right] \cdot \left[\sum_{D=1}^{L} n_D/(n_D-1)\left(\left[\sum_{j=1}^{n_D} y_{jh}^2\right] - y_{Dh}^2/n_D\right) - y._h^2\right]\right). \quad (1)$$

for the total design variance of $\hat{X}$.

Unlike the variance estimator currently in use, $\hat{var}$ above would collapse to the first stage variance estimator if all the JES tracts were sampled with certainty by the Agricultural Survey (since all $w_i=1$, all $\left(\sum_{i\in R_h} w_i^2\right) - T_h=0$).

A collapsing to the first line of (1) would also occur if the second stage sampling design were conventional - that is, if the summary strata were nested within each of the JES segments - so that $y._h=y_{jh}=y_{Dh}$. This is as the textbooks say it should be when the first stage of sampling (here the JES

sample of frame units) is done with replacement (Cochran, 2, p. 307).

We will call the first line of equation (1) the nested variance estimator and the rest of (1) the non-nested adjustment. Although equation (1) provides a design unbiased estimator of the design variance of $\hat{X}$ under pristine conditions and a reasonably well-behaved variance estimator otherwise, there is no guarantee that $\hat{var}$ will be positive. This is because the non-nested adjustment is likely to be negative, while the nested variance estimator can, in theory, be as small as zero.

AN EMPIRICAL EXAMPLE

Gretchen McClung has graciously provided us with December 1986 Agricultural Survey NOL data for harvested acres of winter wheat in 41 states. Table 1 compares the estimated coefficients of variation (CV's) based on our new variance estimation formula (equation (1)) with those based on the "corrected" operational variance formula and reported in McClung (4).

The table also compares estimates of the share of the total variance contributed by the second stage of sampling (i.e., 100 x $\hat{var}_2/\hat{var}$, where equations (3) and (4) provide the alternative methods of calculating $\hat{var}_2$). In order to accomplish this task in a consistent and useful manner, both second stage variance estimates were divided by our total variance estimate.

Another question of some interest is the relative effects of the non-nested adjustment, since large adjustments raise the theoretical possibility of negative values in the right hand side of equation (1). The last column of Table 1 reports the relative sizes of the non-nested adjustments (i.e., 100 x $|\text{adjustment}|/\hat{var}$).

The geometric mean of our CV estimates is 11.8% higher than that of the operational estimates. Due to the nature of geometric means, our standard error estimates are also 11.8% higher on "average" than their operational counterparts. As for variances and relative variances, ours average 25% higher (100 x [1.118^2 -1]). These easy conversions are one reason we focus on geometric means.

Table estimates for December 1986 Agricultural harvested acres of winter wheat

State	New CV Estimator	Operational 2nd Stage Variance Share	New 2nd Stage Variance Share	Non-nested Adjustment Effect
AL	38.4	5.1	10.0	0.3
AR	22.9	33.1	74.8	2.8
AZ	33.6	15.1	26.6	0.8
CA	20.4	54.4	100.0[1]	2.3
CO	20.3	38.3	25.7	0.9
DE	26.8	10.2	10.9	1.0
FL	79.4	93.6	93.8	0.0[2]
GA	31.1	13.5	20.6	0.3
IA	73.6	10.6	2.0	0.0[2]
ID	13.7	35.0	52.7	5.2
IL	51.2	87.1	86.8	0.0[2]
IN	15.6	14.2	12.3	0.9
KS	13.4	32.9	22.5	1.6
KY	39.5	24.7	29.1	0.4
LA	35.8	29.7	91.1	1.3
MD	15.2	54.0	65.4	6.3
MI	18.6	12.6	8.2	0.5
MN	63.0	0.1	0.2	0.0[2]
MO	32.0	79.6	82.0	0.1
MS	29.9	36.2	86.7	1.2
MT	21.4	19.5	6.0	0.3
NC	22.2	42.9	60.9	2.8
ND	51.3	15.1	4.9	0.0[2]
NE	17.0	26.4	16.3	0.8
NJ	59.3	33.7	61.0	0.2
NM	23.7	20.8	17.9	1.3
NV	86.9	20.0	16.6	0.0[2]
OH	11.3	25.3	20.2	2.6
OK	10.7	79.6	69.4	2.4
OR	41.3	31.1	58.1	0.5

Table 1 (cont.)

State	"Corrected" Operational CV Estimator	New CV Estimator	Operational 2nd Stage Variance Share	New 2nd Stage Variance Share	Non-nested Adjustment Effect
PA	19.4	21.4	33.1	63.1	3.3
SC	25.5	31.4	39.2	50.3	1.2
SD	31.9	29.3	13.0	4.1	0.1
TN	36.1	35.3	23.5	21.4	0.5
TX	14.0	18.4	30.6	33.4	1.0
UT	29.8	25.6	32.0	50.0	2.1
VA	21.0	25.4	31.2	77.9	5.1
WA	16.5	15.6	33.1	24.1	2.3
WI	22.9	20.6	21.7	18.9	1.0
WV	38.1	59.4	3.6	9.2	0.2
WY	71.7	33.6	6.0	4.3	0.0[2]
Geometric Mean	25.2	28.2	21.8	23.5	0.2

Note: New York was excluded from the analysis because of problems with the data provided us.

[1]The actual estimate was higher, but the contribution to variance from the second stage of sampling is bounded by 100%.

[2]These numbers are all positive but round to 0.0%.

Let A_k be the proposed CV estimate for a given state and B_k be the operational CV estimate. If the difference between the two was just statistical noise, then the $C_k=\log(A_k/B_k)=\log(A_k) - \log(B_k)$ would be independent random variables with a common mean of zero. Furthermore, if the variances of the C_k were all bounded (but not necessarily equal), then the test statistic T, where $T^2=n(n-1)\bar{C}^2/\sum_{k=1}^{n}(C_k - \bar{C})^2$, $\bar{C}=\sum_{k=1}^{n} C_k/n$, and n=41, would have a standard normal distribution asymptotically.

This null hypothesis (that the C_k have mean zero and bounded variances) is rejected by the data because T equals 2.23. The probability that $|T|>2.23$ under the null hypothesis is less than 5%. (Note: T may be asymptotically normal, but 41 is not infinity. Consequently, it is reassuring to observe that the null hypothesis would still be rejected if T had a Student t distribution with as little as 10 degrees of freedom.)

This statistical result would be the same if we reverse the order of the CV estimators in C_k or use standard error or variance estimators in place of the CV estimators. Reversing the order of the CV estimators simply changes the test statistic from T

to -T, while employing standard error or variance estimators in place of the CV estimators produces the exact same value for T.

Our statistical result establishes that $\bar{C}$ is significantly greater than 0. Note that the antilog of $\bar{C}$ ($e^{\bar{C}}$) is the ratio of the geometric means of the A_k and B_k reported in the last line of Table 1. Thus, the 11.8% difference in the two geometric means is in some sense statistically significant.

Unfortunately, comparing the estimates of the variance shares contributed by the second stage of sampling does not lead to such satisfying results. We had expected that the operational estimates of these shares to be higher than ours. Instead, our estimates average 8.0% higher than the operational estimates. Furthermore, using the type of test statistic we used for comparing the total CV's yields statistically insignificant results even at the 25% level (treating T as a t variate with 40 degrees of freedom).

Given the nature of equations (1) and (3), it is possible for $\hat{var}_2^M$ to exceed $\hat{var}$. This happened only once (in California). In such circumstances, we recommend bounding the estimate of $var_2(\hat{X})$ by that of

$\widehat{var}(\hat{X})$, because the former is more dependent on model assumptions than the latter.

The effects of the non-nested adjustment are always rather small. They average 1.3% (straight arithmetic average) and never exceed 6.3%. Consequently, we have less reason to fear that our variance estimator will ever yield negative values in practice.

One final note; the effect of the non-nested adjustment on a CV estimate is roughly one half of its effect on the corresponding variance estimate reported in Table 1.

CONCLUSIONS

We have proposed a theoretically superior estimation formula for Agricultural Survey NOL variances than the one currently in use. Moreover, the current formula appears to be biased downward, which is a very undesirable characteristic. Although our estimator can, in principle, yield negative estimates of variance, this possibility seems remote in practice judging by the December 1986 winter wheat data.

We also proposed an alternative estimate for the variance share of the second stage of sampling. Our model-driven estimator differs from the operational estimator when, as is the case in most Agricultural Surveys, select strata are not identical to summary strata. The only justification for the operational estimator that we can see is ad hoc.

The empirical results here did not turn out as we had expected. This may be because farm tracts that overlap (OL) with the list frame were placed into summary strata alongside NOL tracts based on their JES responses. The way OL tracts were treated for the NOL portion of the Agricultural Survey (they were attributed winter wheat values of zero) appears to have had the effect of invalidating the assumptions of our model. Fortunately, the questionable practice of commingling OL and NOL tracts within summary strata was abandoned in the 1987 December Agricultural Survey.

Our estimators for the variance from both stages of sampling depend on models which may in reality fail. Happily, the effects of such models nearly cancel out in the total variance estimator; the consequence of modeling can be found only within the small non-nested adjustment term. As a result, we put more faith in our total variance estimator than our second stage variance estimator. The recommendations to follow reflect this favoritism.

RECOMMENDATIONS

1. The variance estimator for the NOL portion of Quarterly Agricultural Surveys proposed in equation (1) should be adopted by the agency (the nested variance estimator in the first line of (1) is a good, simplified alternative).

2. The method for estimating the variance contribution of the second stage of sampling (and implicitly the first) proposed in equation (3) should be adopted, except when the right hand side of (3) exceeds the right hand side of equation (1) (see recommendation 4).

3. In the unlikely event of equation (1) returning a negative value, the lesser of the second stage variance estimator in equation (3) and the nested variance estimator (the first line of (1)) should be used in its place.

4. The estimate of the second stage variance should not be allowed to exceed the estimate of the total Agricultural Survey variance (see recommendations 1 and 3).

5. Farm tracts overlapping the list frame should never be included in the same summary (or select) stratum as NOL tracts.

Let V be the variance estimator in equation (1), N be the nested variance estimator in the first line of (1), and S be the second stage variance estimator in equation (3). The recommendations for variance estimation described above can be re-expressed as follows:

If $V \geq 0$, let V estimate the total variance.

2) If $V<0$, let $\min(N,S)$ estimate the total variance.

3) If $S \leq V$, let S estimate the second stage variance.

4) If $S>V$, let the estimate for the total variance also be the estimate for the second stage variance.

REFERENCES

(1) Cochran, Robert and Huddleston, Harold. Unbiased Estimates for Stratified Subsample Designs. U.S. Department of Agriculture, Statistical Reporting Service, 1969(?).

(2) Cochran, William G. Sampling Techniques (3rd edition). New York: Wiley, 1977.

(3) Hartley, H.O. Estimation in the S.R.S. June and December Surveys. Technical Report #2. U.S. Department of Agriculture, Statistical Reporting Service, 1968(?).

(4) McClung, Gretchen. A Commodity Weighted Estimator. RAD Staff Report No. SRB-NERS-8802, U.S. Department of Agriculture, National Agricultural Statistics Service, January 1988.