A Strategy for Estimating the Number of Minority Operated Farms

Phil Kott, **Matt Fetter**
USDA/NASS

Abstract

The National Agricultural Statistics Service (NASS) uses a national stratified sample of area segments to measure production agriculture. It has supplemented this sample to estimate the total number of 2007 farms in specific categories, in particular the number farms operated by minorities. To meet a national accuracy criterion for the number of Asian-operated farms yet prevent unreasonably large sample sizes in some strata in California, NASS has designated specific segments for potential supplemental sampling based on Decennial Census information. We will show how this was accomplished without sacrificing the randomization consistency of the Asian estimators or increasing the anticipated mean squared errors of other estimators.

KEY WORDS: dual frame, re-stratification, anticipated variance

Introduction

The NASS Area Frame (AF) consists of the entire land area in the contiguous United States and Hawaii and thus encompasses every farm in those states. One sample is drawn annually for each state from the AF to supply all area-based sample needs for the survey year. This sample is primarily used for NASS's annual June Area Survey (JAS). It is also used to produce estimates for the population not on NASS's list sampling frame (LSF) of agricultural places and, in agricultural census years, the agency's census mail list (CML) of potential farms. Generally, farms that are not on the LSF or the CML tend to be smaller, and are also more likely to be minority-operated. Describing a method for improving the precision of the JAS-sample-based estimates of CML undercoverage for minority-operated farms is the focus of this paper.

The AF sample employs a stratified multi-stage design. The primary sampling units (PSUs) are units of land. Each PSU is placed in a land-use stratum. The stratum definitions are based on land-cultivation intensity. PSUs residing in the same stratum are, by design, similar with respect to cultivation intensity. Within each stratum, PSUs are separated into geographically-defined substrata. As a result, the agricultural characteristics of PSUs within the same substratum are likely to be even more homogeneous then at the stratum level. The PSUs are sampled with replacement from each substratum, using probability proportional to the acreage contained in the PSU. Once a PSU is selected, it is divided into relatively equal-sized secondary sampling units called "segments". One segment is then selected randomly without replacement from the PSU each time the PSU is selected.

In census years, NASS augments the annual JAS sample to boost the precision of the not-on-the-census-mail-list (NML) farm-count estimates. The supplemental AF segment sample drawn for this purpose is referred to as the Agricultural Coverage Evaluation Survey (ACES) sample. The ACES sample is selected from particular land-use strata where analysis suggests NML farms are more likely to be found. For many of the required NML estimates, this combined JAS/ACES sample is sufficient. Minority-operated farms, however, are relatively numerous in a handful of states. It would require unacceptably large PSU sample size allocations in these states to increase the precision of minority-operated farm estimates sufficiently using the standard land-use-defined strata. Consequently, a design more focused on targeting minority-operated farms called the "Minority Agricultural Area Coverage Evaluation Survey" (MACES) has been implemented. An example of it is described in what follows.

To sample, say, Asian-operated farms (hereafter called "Asian" farms) more efficiently, NASS targeted a few land-use strata in California. The universe of PSUs in a targeted land-use stratum were overlaid with Decennial Census blocks. PSUs that were at least partially overlaid by a census block containing 50 or more estimated Asians were designated as "Asian" PSUs. After the JAS/ACES sample was selected, an additional substratum was created. The new substratum was composed of all PSUs designated as Asian in the stratum. This was done without regard to whether the PSU was selected in the JAS/ACES sample. As a result, all designated Asian PSUs resided in two substrata -- the original JAS/ACES substratum containing it, and the new MACES Asian substratum. The MACES sample of PSUs was then selected with replacement from the Asian substratum, again using probability proportional to the acreage in the PSU. A segment in the was then randomly chosen from the selected Asian PSU in such a way that no segment was selected more than once (if possible) for the combined JAS/ACES/MACES sample.

Note that this sampling strategy resulted in the Asian PSUs being eligible for two samples—the JAS/ACES sample and the MACES sample. This will be accounted for in the weighting. (Hereafter the JAS/ACES sample will be referred to simply as the "JAS" sample, since the sub-stratification used for selecting these samples was the same for those strata receiving an ACES sample.)

A key property of the estimation strategy in this new proposal is that the same expansion factors are developed for all NML indications. Although the focus is on the accuracy in the estimate of NML Asian farms, other NML indications should not have to suffer.

In what follows, we limit our attention to a single land-use stratum in California. By increasing the number of segments selected from this land-use stratum, a much better national estimate of the number of Asian NML farms should result. An analogous methodology was applied to a number of other land-use strata in California and a few strata believed to contain many black-operated farms in Mississippi and Texas.

Notation (within a designated land-use stratum)

P_h – the set of PSU's in JAS substratum h (= 1, ..., H)
P_0 – the set of PSU's in the Asian supplemental substratum
P_{h0} – the set of Asian PSU's in substratum h ($P_0 \cap P_h$)
P_{h-} – the set of PSU's in substratum h excluding the Asian ones ($P_h - P_{h0}$)
N_h – the number of segments in a PSU from within P_h
N_0 – the number of segments in a PSU from within P_0
n_h – the JAS sample size in substratum h
n_0 – the sample size of the MACES supplement
S_{hj} – the set of N_{hj} segments in PSU j of substratum h
n_{hj-} – the number of times PSU j of substratum h is selected for the JAS sample
n_{hj0} – the number of times PSU j of substratum h is selected for the MACES supplement
n_{hj} = $n_{hj} + n_{hj-}$ is the number of times PSU j of substratum h is selected for the combined JAS/MACES sample.
n_{k-} – the number of times segment k is selected for the JAS subsample of segments (hereafter called the "JAS segment sample")
n_{k0} – the number of times k is selected for the supplemental MACES segment sample
n_k = $n_{k0} + n_{k-}$ is the number of times k is in the combined JAS/MACES segment sample.

For determining the size of the MACES supplemental sample, n_0, we can do the following. Suppose n_{ACES} is what the supplemental ACES sample would be if there were no MACES methodology. A reasonable MACES supplemental sample size is an integer close to $-$ and usually greater than $- n_{ACES}(N_0/N)$.

Weighting

The expected number of times PSU j from substratum h is selected for the JAS is

$$\mathrm{E}(n_{hj-}) = n_h\,(N_{hj}/N_h).$$

Similarly, the expected number of times *Asian* PSU j from substratum h is selected for the MACES supplement is:

$$\mathrm{E}(n_{hj0}) = n_0\,(N_{hj}/N_0).$$

If the PSU is not Asian, this expectation is zero.
From this we see that the expected number of times a PSU j is selected for the combined sample is

$$\begin{aligned}\mathrm{E}(n_{hj}) &= n_h\,(N_{hj}/N_h) + n_0\,(N_{hj}/N_0) && \text{when } j \in P_{h0}\\ &= n_h\,(N_{hj}/N_h) && \text{when } j \in P_{h-}.\end{aligned}$$

The expected number of times each of the N_{hj} segments in PSU j of substratum h is in the combined segment sample is $(1/N_{hj})\mathrm{E}(n_{hj})$. Thus,

$$\begin{aligned}\mathrm{E}(\mathrm{n_k}) &= n_h/N_h + n_0/N_0 && \text{when } k \in S_{hj} \text{ and } j \in P_{h0}\\ &= n_h/N_h && \text{when } k \in S_{hj} \text{ and } j \in P_{h-}.\end{aligned}$$

The sampling weight (expansion factor) for segment k is

$$w_k = n_k / \mathrm{E}(n_k)\,.$$

An unbiased estimator of a sum $T = \sum y_k$ taken across all the segments in the stratum (sampled or not) is the expansion estimator,

$$t = \sum_{h=1}^{H} \sum_{j \in P_h} \sum_{k \in S_{hj}} w_k\, y_k = \sum w_k\, y_k\ .$$

(note that w_k = 0 for those k never selected for the segment sample).

Let d_k = 1 when segment k is an Asian segment, 0 otherwise. One can show that the group-mean estimator, $t_{gm} = \sum a_k\, y_k$, where

$$\begin{aligned}a_k &= \frac{N_0}{\sum\sum\sum w_i d_i} w_k && \text{when } k \text{ is an Asian segment}\\ &= \frac{N - N_0}{\sum\sum\sum w_i\,(1 - d_i)} w_k && \text{otherwise, and}\end{aligned}$$

$$N = \sum_{h=1}^{H} N_h \ ,$$

will often produce more efficient estimators (i.e., ones with less variance). This is because with all the y_k set to 1, t_{gm} estimates the number of Asian segments in the stratum (N) perfectly. Similarly when all the y_k are set to d_k, t_{gm} estimates the number of Asian segments perfectly). The estimator t, although unbiased, does not. The usual expansion estimator based on *onl*y the JAS sample likewise estimates the number of segments in the stratum perfectly.

A Slightly Different Notation

Let us ignore the remote possibility that any segment is selected for the combined segment sample more than once. It is convenient to recast the notation a bit. Let *hk* denote one of the n_h segments selected from substratum *h*, where h can be 0 as well as an integer from 1 to *H*. The n_h sampled segments in substratum h are relabeled *h*1, *h*2,.., *hn*$_h$. In this notation, the expansion estimator becomes

$$t = \sum_{h=0}^{H} \sum_{k=1}^{n_h} w_{hk}\, y_{hk} \ ,$$

where

w_{0k} = 1/[(n_h /N_h) + (n_0 /N_0)] when segment 0*k* is in a PSU that is also in substratum *h*,

w_{hk} = 1/[(n_h/N_h) + (n_0/N_0)] for $h \geq 1$ when segment *hk* is from an Asian PSU, and

$w_{hk} = N_h / n_h$ when segment hk is from a non-Asian PSU.

We can write a$_{hk}$ within $t_A = \sum \sum a_{hk}\, y_{ak}$ as:

$$a_{hk} = \frac{N_0}{\sum_{g=0}^{H} \sum_{i=1}^{n_g} w_{gi} d_{gi}} w_{hk}$$

when segment *hk* is Asian

$$= \frac{N - N_0}{\sum_{g=1}^{H} \sum w_{gi} \left(1 - d_{gi}\right)} w_{hk} \ ,$$

otherwise.

Variance Estimation

One can estimate the variance of t with

$$v = \sum_{h=0}^{H} \left[n_h / \left(n_h - 1 \right) \right] \left\{ \sum_{k=1}^{n_h} \left(w_{hk}\, y_{hk} \right)^2 - \left(\sum_{k=1}^{n_h} w_{hk}\, y_{hk} \right)^2 / n_h \right\} .$$

Similarly, the following is a variance (mean squared error) estimator for the Hajek/ratio estimator t_A:

$$v_A = \sum_{h=0}^{H} [n_h / (n_h - 1)] \left\{ \sum_{k=1}^{n_h} (a_{hk} e_{hk})^2 - \left(\sum_{k=1}^{n_h} a_{hk} e_{hk} \right)^2 / n_h \right\},$$

where

$$e_{hk} = y_{hk} - \frac{\sum_{g=0}^{H} \sum_{i=1}^{n_g} w_{gi} d_{gi} y_{gi}}{\sum \sum w_{gi} d_{gi}}$$

when segment *hk* is Asian

$$= y_{hk} - \frac{\sum_{g=1}^{H} \sum w_{gi} \left(1 - d_{gi}\right) y_{gi}}{\sum \sum w_{gi} \left(1 - d_{gi}\right)}$$

otherwise.

The variance estimator *v* would be unbiased if the segments were subsampled *with* replacement within PSU's, and no segment happened to be selected more than once. To see why, one can write an unbiased variance estimator for t using the old notation as

$$v^* = \sum_{h=1}^{H} \left[\frac{n_h}{n_h - 1} \sum_{j \in P_h} \sum_{k \in S_{hk}} n_{k-} \left\{ \frac{y_k}{E(n_k)} - n_h^{-1} \sum_{g \in P_h} \sum_{i \in S_{hg}} n_{i-} \frac{y_i}{E(n_i)} \right\} \right]^2$$

$$+ \ \frac{n_0}{n_0 - 1} \sum_{j \in P_0} \sum_{k \in S_{0j}} n_{k0} \left\{ \frac{y_k}{E(n_k)} - n_0^{-1} \sum_{q \in P_0} \sum_{i \in S_{0q}} n_{i0} \frac{y_i}{E(n_i)} \right\}^2 ,$$

which (some work will reveal) is the same as the *v* above when n_k, n_{k0}, and n_k can only take on the values 0 or 1.

A Look at the Anticipated Variance for a Non-Asian Total

Is there any guarantee that combining the JAS and MACES samples will not produce a worse estimator for some total unrelated to the number of Asian farms? No, there is not. Nevertheless, under a model where the y_k across all segments in the stratum are uncorrelated random variables with a common mean, μ , and variance,

σ^2, the anticipated variance (model expected design mean squared error) of t_{gm} can be shown to be approximately,

$$\mathrm{AV}\left(t_{gm}\right) = \sigma^2 \sum \left\{ \left[1/E\left(n_k\right) \right] - 1 \right\},$$

where the summation is over all the segments in the stratum (sampled or not).
Similarly, the anticipated variance of the expansion estimator based only on the JAS sample can be shown to be

$$\mathrm{AV}\left(t_{JAS}\right) = \sigma^2 \sum \left\{ \left[1/\mathrm{E}\left(n_{k-}\right) \right] - 1 \right\}.$$

Since E(n_k) $\geq$ E(n_k) with strict inequality holding for all segments in Asian PSU's, we can conclude that the anticipated variance of the Hajek estimator based on the combined sample is less than the expansion estimator based on the JAS sample.

Sketched proof:

The estimators under both strategies are of the form: $\sum c_k y_k$, where the summation is over all the segments in the stratum, and $\sum c_k = N$. Thus, both are model unbiased, $\mathrm{E_M}\left\{\sum c_k y_k - \sum y_k\right\} = 0$, and have model variances of the form: $\mathrm{Var_M} = \sigma^2 \sum \left(c_k^2 - 2c_k + 1\right)$. Now $c_k \approx n_k/\mathrm{E}(n_k)$ for the estimator under the combined sample. If n_k is either 0 or 1, then $n_k^2 = n_k$, and the expectation under the design of c_k^2 and c_k are 1/E(n_k) and 1, respectively. Since the design expectation of the model variance/mean squared error is equal to the model expectation of the design variance/mean squared error, the anticipated variance of t_A is as written above. An analogous argument applies for the expansion estimator under the JAS sample.)

Reference

Skinner, C.J., Holmes, D.J. and Holt, D.(1994) Multiple Frame Sampling for Multivariate Stratification. *International Statistical Review*, 62 ,3, 333-347 .