United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research and
Applications
Division

SRB Staff Report
Number SRB-88-06

May 1988

Estimating Variances for the June Enumerative Survey

Phillip S. Kott

ESTIMATING VARIANCES FOR THE JUNE ENUMERATIVE SURVEY, by Phillip S. Kott, Research and Applications Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC 20250, May 1988. NASS Staff Report, Number SRB-88-06.

ABSTRACT

Under a reasonable model the variance estimation formula currently in use for the June Enumerative Survey (JES) is shown to have a slight downward bias. This bias can be reversed by removing finite population correction terms from the variance formula. The small upward bias in the adjusted formula is not dependent on the model. A slight overestimate of variance is generally considered preferable to a slight underestimate. Two minor modifications of the current JES sampling design are briefly discussed that would make strictly unbiased variance estimation possible.

KEY WORDS

Bias, finite population correction, model, sample design.

ACKNOWLEDGMENTS

The author would like to thank Ron Fecso, Read Johnston, George Hanuschak, Jack Nealon, and Charles Perry for their helpful comments on earlier drafts of this manuscript.

May, 1988

Washington, DC

TABLE OF CONTENTS

SUMMARY

This paper proposes eliminating finite population correction terms from variance estimates based on June Enumerative Survey (JES) area samples. Using design-based sampling theory, the revised variance estimation formula is shown to have (if anything) a small upward bias.

The direction of the bias of the current operational formula can not be determined unambiguously with conventional theory. The formula, however, is shown to have a slight tendency to be biased downward using a model-based analysis.

Since a slight overestimate of variance is generally considered preferable to a slight underestimate, the adjustment proposed here (removing finite population correction terms) would be an improvement over current procedures.

An alternative approach would be to modify the JES sampling design so that unbiased variance estimation is possible. Two such modifications are briefly discussed. One would transform the JES sampling design to the equivalent of simple random sampling without replacement, while the other would result in the equivalent of simple random sampling with replacement. The former strategy would slightly decrease the variances of JES estimators, the latter slightly increase them.

ESTIMATING VARIANCES FOR THE JUNE ENUMERATIVE SURVEY

By Phillip S. Kott

INTRODUCTION

The June Enumerative Survey (JES) area sample for a state is selected in a manner that is a cross between simple random sampling (srs) with replacement within strata and stratified without replacement srs. Because of the unusual nature of the JES sample design, deriving a simple expression for the variance of a JES area frame estimator is an elusive task.

Theoretical results from Raj (4) will be invoked to show that a JES area estimator is at least as efficient as one based on an analogous sample drawn using with replacement stratified srs. It will then be demonstrated that applying the with replacement stratified srs variance estimation formula to a JES area estimator produces a slightly conservative estimate of variance; i.e., one with (if anything) a small upward bias.

The National Agricultural Statistics Service (NASS) uses the without replacement stratified srs variance estimation formula for JES area estimates. Using a simple model, this practice will be shown to have a tendency to produce variances estimates with downward biases when segments within the same frame unit have correlated farm values. This approach is advocated because a slight overestimate of variance is generally considered preferable to a slight underestimate.

BACKGROUND

Here is a brief description of the JES area sample design. For more information on this process, consult Cotter and Nealon (2) or Houseman (3). After the land area of a state has been broken down into substrata (formerly called paper strata), a with replacement probability proportionate to size (pps) sample of frame units is chosen from each substratum. Each frame unit is, in principle, a cluster of area segments, the number of segments within a frame unit being the unit's measure of size. It must be noted, however, that NASS only delineates the area segments within those frame units selected for the sample (to do otherwise would be a tremendous drain on available resources).

Since frame units are selected with replacement, it is possible for a single unit to be selected more than once. Although this is unlikely to happen in a given substratum, there are so many substrata (1,832 in the 1987 JES, for example) that it can and does happen more often than one might think.

Suppose some frame unit **u** is selected **m** times. A without replacement simple random sample of m segments is then subsampled from frame unit **u**. (Note: in practice, m is very rarely, if ever, larger than the number of segments in **u**; we will assume that it never is).

For analytical purposes, let us direct our attention on a single substratum in a single state and on a single farm value of interest. Since sampling is independent across substrata, there is no loss in generality in treating a single substratum as if it were the entire universe.

Let x_{ij} be the farm value for segment j in frame unit i. Let m be the number of frame units in the substratum, N_i be the number of segments in unit i, and $N = \sum_{i=1}^{m} N_i$. We are interested in estimating $X = \sum_{i=1}^{m} \sum_{j=1}^{N_i} x_{ij}$ based on a sample, S, of n units. The estimator NASS uses is the (design) unbiased direct expansion estimator:

$$\hat{X} = (N/n) \sum_{ij \epsilon S} x_{ij}. \qquad (1)$$

This estimator is unbiased because the sample design is self-weighted; that is, each segment has an equal probability of selection. Given sample size n and population size N, that probability must be n/N.

An alternative expression for $\hat{X}$ is

$$\hat{X} = (N/n) \sum_{i=1}^{m} \sum_{j=1}^{N_i} n_{ij} x_{ij}, \qquad (1')$$

where n_{ij} is the number of times segment ij is in the sample. For the JES area sampling design, n_{ij} can only be zero or one.

THE VARIANCE OF $\hat{X}$

The variance of the estimator $\hat{X}$ depends on the sampling design used. Although $\hat{X}$ will be unbiased under any self-weighted design, its variance can vary from one such design to another.

Were the sample drawn using srs <u>without</u> replacement, the variance of $\hat{X}$ would be

$$\text{var}_{wtr}(\hat{X}) = (N^2/n)(1- n/N) \sum_{i=1}^{m} \sum_{j=1}^{N_i} (x_{ij}-\bar{x})^2/(N-1), \qquad (2)$$

where $\bar{x} = X/N$. On the other hand, were the sample drawn using srs <u>with</u> replacement, the variance of $\hat{X}$ would be a little larger:

$$\text{var}_{wr}(\hat{X}) = (N/n) \sum_{i=1}^{m} \sum_{j=1}^{N_i} (x_{ij}-\bar{x})^2. \qquad (3)$$

(Note that equation (1'), but not (1), defines $\hat{X}$ when some n_{ij} is greater than unity.)

As we have observed, the JES area sample is not drawn using either method. The variance of $\hat{X}$ based on the JES sampling design is

$$\text{var}_J(\hat{X}) = \sum_{gh\epsilon P} \sum_{ij\epsilon P} (p_{gh}p_{ij} - p_{ghij})(x_{gh}/p_{gh} - x_{ij}/p_{ij})^2/2$$

$$= \sum \sum (1 - [N^2/n^2]p_{ghij})(x_{gh} - x_{ij})^2/2, \quad (4)$$

where P is the population of N segments,

$p_{ij} = n/N$ is the selection probability of segment j in frame unit i, and

$$p_{ghij} = \begin{cases} n/N & \text{when } gh=ij \\ (n[n-1]/N^2)(N_i/[N_i-1]) & \text{when } g=i,\ h\neq j \\ n(n-1)/N^2 & \text{when } g\neq i \end{cases}$$

is the joint selection probability of segments gh and ij.

The first line of (4) is the variance of $\hat{X}$ based on an arbitrary without replacement sampling design employing a fixed sample size (Cochran (1), p. 260, eq. (9A.42) and elsewhere).

The expression for $\text{var}_{wr}(\hat{X})$ can be rewritten in a form similar to (4) as

$$\text{var}_{wr}(\hat{X}) = \sum_{gh\epsilon P} \sum_{ij\epsilon P} (1/n)(x_{gh} - x_{ij})^2/2. \quad (5)$$

This is a special case of a result in Raj (4; p. 49, eq. (3.23)).

Subtracting (4) from (5) yields:

$$\begin{aligned}\mathrm{var}_{wr}(\hat{X})-\mathrm{var}_{J}(\hat{X}) &= \sum_{i=1}^{m}\sum_{j<h}^{N_i}(n-1)/(n[N_i-1])(x_{ij}-x_{ih})^2 \\ &= \sum_{i=1}^{m}\sum_{h=1}^{N_i}(n-1)N_i/(n[N_i-1])(x_{ih}-\bar{x}_i)^2, \end{aligned} \tag{6}$$

where $\bar{x}_i = \sum_{j=1}^{N_i} x_{ij}/N_i$. Unless the segments in each respective frame unit have identical farm values, the right hand side of (6) will be positive. This means that the JES area sampling design is at least as efficient as (that is, produces an unbiased estimator with no more variance than) an analogous <u>with</u> replacement stratified sampling design.

WITHOUT REPLACEMENT COMPARISON

Comparing $\mathrm{var}_J(\hat{X})$ and $\mathrm{var}_{wtr}(\hat{X})$ is not as straightforward. The variance of $\hat{X}$ based on without replacement srs can also be written in the form of the first line of (4). Here, however, $p_{ijgh} = n(n-1)/(N[N-1])$ when $ij \neq gh$. Consequently,

$$\begin{aligned}\mathrm{var}_J(\hat{X})-\mathrm{var}_{wtr}(\hat{X}) = \\ [(n-1)/n]\Big[(N-1)^{-1}\sum_{gh\epsilon P}\sum_{ij\epsilon P}(x_{ij}-x_{gh})^2/2 \\ - \sum_{i=1}^{m}\sum_{j<h}^{N_i}(x_{ij}-x_{ih})^2/(N_i-1)\Big]\end{aligned}$$

$$= [(n-1)/n][(N/(N-1)) \sum_{i=1}^{m} \sum_{j=1}^{N_i} (x_{ij}-\bar{x})^2 - \sum_{i=1}^{m} (N_i/(N_i-1)) \sum_{j=1}^{N_i} (x_{ij}-\bar{x}_i)^2]. \quad (7)$$

The right hand side of (7) is neither unambiguously positive nor unambiguously negative in all situations. Its sign, however, can be analyzed by assuming a simple model.

Suppose x_{ij} obeys this stochastic equation:

$$x_{ij} = m + t_i + e_{ij}, \quad (8)$$

where t_i is a random variable with mean 0 and variance s_B^2 (for between frame units),

and e_{ij} is a random variable with mean 0 and variance s_W^2 (for within frame units).

The correlation between segments within the same frame unit under the model in (8) is

$$r = s_B^2/(s_B^2 + s_W^2). \quad (9)$$

This value is 0 (the segments are uncorrelated) if and only if $s_B^2=0$.

The model-expected value of the right hand side of (7) is

$$E^m(\hat{var}_J(X)-\hat{var}_{wtr}(X)) = (n-1)N^2(1-\sum_{i=1}^{m} N_i^2/N^2)s_B^2/(n[N-1]). \quad (10)$$

This means that the JES area sampling design has a tendency to be less efficient than an analogous without replacement stratified srs design when segment farm values within frame units are correlated.

The expectation of the difference between $\hat{var}_{wr}(\hat{X})$ and $\hat{var}_J(\hat{X})$ under the model can also be calculated (from (6)):

$$E^m\{\hat{var}_{wr}(\hat{X}) - \hat{var}_J(\hat{X})\} = (n-1)Ns_W^2/n. \qquad (11)$$

Observe that when the within frame unit variance, s_W^2, is zero, which is equivalent to r in (9) being one, this expected difference is zero. Thus, at one extreme, r=0, there is no model-expected difference between $\hat{var}_J(\hat{X})$ and $\hat{var}_{wtr}(\hat{X})$, while at the other, r=1, there is no difference (at all, see (6)) between $\hat{var}_J(\hat{X})$ and $\hat{var}_{wr}(\hat{X})$.

The model expectation of $\hat{var}_J(\hat{X})$ itself can be shown to be

$$E^m\{\hat{var}_J(\hat{X})\} = (N^2/n)[(1 - \sum_{i=1}^{m} N_i^2/N^2)s_B^2 + (1 - n/N)s_W^2)].$$

The right hand sides of both (10) and (11) are small compared to this when the sampling fraction n/N is small.

What do the revelations in (6) and (10) tell us about estimating the variance of $\hat{X}$ under the JES area sampling design? Nothing directly, because the $\hat{X}$ in the arguments of var_{wr}, var_{wtr}, var_J are based on different samples.

Let us now consider the standard with and without replacement srs variances estimators:

$$\hat{var}_{wr} = N^2 \sum_{i=1}^{m} \sum_{j=1}^{N_i} n_{ij}(x_{ij}-\bar{x}_s)^2/(n[n-1]),$$

and

$$\hat{var}_{wtr} = (1-f)N^2 \sum_{i=1}^{m} \sum_{j=1}^{N_i} n_{ij}(x_{ij}-\bar{x}_s)^2/(n[n-1]),$$

where $\bar{x}_s = \sum_{i=1}^{m} \sum_{j=1}^{N_i} n_{ij}x_{ij}/n$, and $f = n/N$.

The key is to evaluate the (design) expectation of $N^2 \sum_{i=1}^{m} \sum_{j=1}^{N_i} n_{ij}(x_{ij}-\bar{x}_s)^2/(n[n-1])$ under the JES (within substratum) sampling design; i. e.,

$$\begin{aligned} E\{N^2 \sum_{i=1}^{m} \sum_{j=1}^{N_i} n_{ij}(x_{ij}-\bar{x}_s)^2/(n[n-1])\} \\ &= N^2/(n[n-1])\{\Sigma\Sigma E(n_{ij}x_{ij}^2) - nE(\bar{x}_s^2)\} \\ &= (N/(n-1))\{\Sigma\Sigma x_{ij}^2 - N\bar{x}^2 - var_J(\hat{X})/N\} \\ &= nvar_{wr}(\hat{X})/(n-1) - var_J(\hat{X})/(n-1) \\ &\geq var_J(\hat{X}), \end{aligned}$$

where $var_{wr}(\hat{X})$ would be the variance of $\hat{X}$ had the

sample been drawn via with replacement srs (see (3)).

This suggests that $\hat{var}_{wr}$ has (if anything) a positive bias as an estimator of the variance of $\hat{X}$ under the JES area sampling design. Following similar reasoning,

$$E\{(1-f)N^2 \sum_{i=1}^{m} \sum_{j=1}^{N_i} n_{ij}(x_{ij}-\bar{x}_s)^2/(n[n-1])\}$$

$$= \{(1-f)N/(n-1)\}\{\Sigma\Sigma\ x_{ij}^2 - N\bar{x}^2 - var_J(\hat{X})/N\}$$

$$= n(N-1)var_{wtr}(\hat{X})/(N[n-1]) - (1-f)var_J(\hat{X})/(n-1)$$

$$= var_J(\hat{X}) + n(N-1)\{var_{wtr}(\hat{X}) - var_J(\hat{X})\}/(N[n-1]),$$

where $var_{wtr}(\hat{X})$ would be the variance of $\hat{X}$ had the sample been drawn via without replacement srs (see (2)). Since $var_{wtr}(\hat{X})$ has a tendency to be less than $var_J(\hat{X})$ when segment farm values within frame units are correlated, then $\hat{var}_{wtr}$ will tend to underestimate $var_J(\hat{X})$ under those conditions.

REPLICATE VARIANCE ESTIMATOR

There is an another well known variance estimator that merits a brief mention. NASS assigns each sampled segment to a replicate. Were the sample drawn totally <u>with</u> replacement, the unbiased estimate of X generated by each replicate would be

independent (note: $\hat{X}$ is the average of these replicate estimates). As a result, one unbiased estimator of $var_{wr}(\hat{X})$ would be

$$\hat{var}_{rep} = \sum_{r=1}^{R} (\hat{X}_{(r)} - \hat{X})^2/(n-1)$$

where $\hat{X}_{(r)}$ denotes the estimate of X based on replicate r; r=1, 2, ..., R.

Unfortunately, $\hat{var}_{rep}$ can be shown to be less efficient than $\hat{var}_{wr}$ (i.e., it has more variance as an estimator of $var_{wr}(\hat{X})$). Consequently, a a full treatment of $\hat{var}_{rep}$ will not be offered here.

CONCLUSION

The present version of the variance estimator for a JES area frame estimate uses the without replacement stratified srs formula. The result is a slight tendency to underestimate true variances (slight because the substrata sampling fractions, the n/N, are small).

If the substratum finite population correction terms (the (1-f)) were removed, the variance estimation formula would almost surely have a small upward bias.

It should be noted that if NASS were to change its sampling design so that segments from multiply selected hit frame units were sampled <u>independently</u> (which would allow the possibility of a segment being sampled more than once), the adjusted variance estimator would be unbiased (as would the replicate variance estimator).

Alternatively, Charles Perry has suggested that the sampling design be modified in the following manner. Each time a frame unit is selected, its measure of size should be decreased by a segment before the

next unit is selected. The resulting sample selection process is equivalent to simple random sampling without replacement, rendering the present variance estimation formula unbiased. This sampling design would, with all other things kept constant, yield JES estimators with a tendency to have slightly less variances than those produced currently (assuming the model in the text).

RECOMMENDATION

Barring a change in the sampling design for JES area samples, NASS should remove the substratum finite population correction terms from the JES variance estimations formula. Although this may result in a slight upward bias, it is generally preferable to overestimate variances rather than underestimate them.

Alternatively, the JES area sampling design can be slightly modified in one of two ways to make either the present variance estimation formula or the alternative proposed here exactly unbiased. The former approach would also result in estimators with a tendency to have slightly less variances than those produced currently.

REFERENCES

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

2. Cotter, Jim and Nealon, Jack. Area Frame Design for Agricultural Surveys, U.S. Dept. of Agr., Nat. Agr. Stat. Serv., August 1987.

3. Houseman, Earl E. Area Frame Sampling in Agriculture. SRS-20. U.S. Dept. of Agr., Nat. Agr. Stat. Serv., 1975.

4. Raj, Des. Sampling Techniques. New York: McGraw-Hill, 1968.