Survey Methodology, June 1987
Vol. 13, No. 1, pp. 45-62
Statistics Canada

Statistical Properties of Crop Production Estimators

CAROL A. FRANCISCO, WAYNE A. FULLER, and RON FECSO[1]

ABSTRACT

The National Agricultural Statistics Service, U.S. Department of Agriculture, conducts yield surveys for a variety of field crops in the United States. While field sampling procedures for various crops differ, the same basic survey design is used for all crops. The survey design and current estimators are reviewed. Alternative estimators of yield and production and of the variance of the estimators are presented. Current estimators and alternative estimators are compared, both theoretically and in a Monte Carlo simulation.

KEY WORDS: Crop surveys; Yield estimation; Two phase sample; Variance estimation.

1. INTRODUCTION

The National Agricultural Statistics Service (formerly known as the Statistical Reporting Service), U.S. Department of Agriculture, conducts objective yield surveys of corn, cotton, soybeans, rice, grain sorghum, sunflowers and wheat in states which are major producers of these field crops. Similar yield surveys are conducted in a number of other countries.

While field sampling procedures for each crop differ in terms of plot sizes, plot location methods, and vegetative and fruit measurement techniques, all surveys rely on the same basic design. A four-step sampling procedure is used. A description of this survey design is contained in Section 2. Section 3 describes the estimators of average crop yield and the variance estimators, evaluates them and explores alternative estimators. Conclusions and recommendations are presented in Section 4.

2. OBJECTIVE YIELD SURVEY DESIGN

The first two steps of sample selection produce the sample of area segments used in the June Enumerative Survey conducted by the National Agricultural Statistics Service (NASS). The area frame for each state is stratified by land use. For example, the State of California is divided into 12 land use strata. Each land use stratum is subdivided into areas called frame units. The size of a frame unit varies; the actual size of any given frame unit depends upon available boundary designations, available ancillary information, political boundaries, and so forth. Once frame units are established, the number of area segments in each frame unit is determined by dividing the total area of each frame unit by the target segment size. The target size is a function of the land use stratum into which the frame unit falls. For example, in California the target segment size is one half square mile in the orchard stratum and one square mile in all other cropland strata. Frame units typically contain between one and 30 area segments.

[1] Carol A. Francisco, Syntex Laboratories Inc., 3401 Hillview Avenue, Palo Alto, California 94304; Wayne Fuller, Department of Statistics, Iowa State University, Ames, Iowa 50011; and Ron Fecso, Survey Research Branch, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington D.C. 20250.

Each land use stratum is substratified on the basis of geography. To develop the geographic substrata, frame units within each land use stratum are ordered by county in such a manner that adjacent counties that are agriculturally similar are placed together (Fecso 1978). Substrata are formed from sequential groups of area segments. Thus, substrata contain area segments that are agriculturally similar and geographically close together. Within a given land use stratum, substrata have an equal number of segments and equal area (within rounding). Detailed information on the area frame design is available in Fecso and Johnson (1981) and Houseman (1975).

For purposes of variance estimation, it is the substrata within land use strata that are the sampling strata. Henceforth, the land use substrata will be referred to simply as strata.

The first step in sampling from the area frame is the selection of frame units within each stratum. The number of frame units allocated to a stratum depends on the agricultural nature of the stratum. Typically, eight to 15 frame units are drawn in cropland strata; whereas in agri-urban, city, and nonagricultural strata four to five frame units are drawn. Frame units within strata are selected at random with probability proportional to the number of area segments in the frame unit. At the second step, one area segment is chosen at random from each selected frame unit. Thus, each area segment within a stratum has an equal probability of selection.

Although the frame unit is the primary sampling unit for this design, because the frame units are selected with probability proportional to the number of segments and one segment is selected per sampled frame unit, the segment can be treated as the primary sampling unit. In our study, steps one and two in the sampling procedure are considered as one procedure, and the sample of segments will be treated as a stratified single stage simple random sample. Since the average sampling rate is about one percent, the finite population correction term will be ignored in our analysis.

The third and fourth steps in the sampling procedure involve the selection of fields and of plots within selected fields. As part of the June Enumerative Survey, all selected area segments are screened for fields which have been planted or are scheduled to be planted with the crop of interest. These fields are listed by segment number and order of enumeration within segment. A systematic sample of fields is selected with selection probabilities proportional to the product of the field area and the inverse of the probability of selection of the area segment in which the field is contained. Hence, the number of sampled fields per segment varies, and large fields within a segment can be selected more than once.

At the fourth and final step, two plots of roughly equal area are placed in each selected field using a random row and pace method of location. Where rows are not readily distinguishable, and in the case of wheat, a random number of paces along the field edge and a random number of paces into the field are used to locate plots. A further exception occurs in the wheat objective yield survey. For this survey the first plot is randomly located and the second plot is placed in a fixed position relative to the first plot. In the event that a large field is selected more than once during the third step of the sampling procedure, additional sets of two plots are independently sampled. Because plots are always sampled in pairs, we call the pair of plots the secondary unit. A maximum of eight plots (that is, four secondary units) per field is imposed.

3. ESTIMATION PROCEDURES

Formally, the sample is a two phase sample with subsampling in the second phase. Table 1 contains a schematic description of the sample. The phase one sample is a stratified simple

Table 1

Sampling Procedure for the Objective Yield Survey

Phase/Sampling Unit	Selection Procedure	Sampled Number[1]	Data Collected
Phase One			
Primary Sampling Unit: Segment	equal probability within strata	n_h	crop acres
Phase Two			
Primary Sampling Unit: Segment	unequal probability	K_h	crop acres, estimated production[2]
Secondary Sampling Unit: Pair of Plots	equal probability	m_{hk}	estimated production from plots

[1] Number is per stratum for primary sampling units and is per segment for secondary sampling units.
[2] Segment production is zero if the crop acreage is zero and is estimated from plot determinations if the crop acreage is positive.

random sample of segments. The phase two sample is composed of all segments with zero crop acres and a probability-proportional-to-crop-acres sample of segments with the crop. The sample of segments is the result of a probability-proportional-to-area systematic sample of first phase fields planted with the crop. A sample of secondary units, where each secondary unit is a pair of plots, is selected from the segments in the phase two sample that have the crop. Because the secondary unit is always a pair of plots, we will henceforth refer to secondary units and no longer speak of plots. We will also ignore the fact that the operational units used to locate the plots are fields and speak only of the sampled segments.

Notice that two types of segments are observed at phase two - those that have zero acres of the crop and those that have non-zero acres. The total number of second phase segments is K. The acres and the total production are known (both equal to zero) for an observed segment with zero acres. For second phase segments with positive acres, a subsample of secondary units is used to estimate production.

Let M_{hk} be the number of secondary units in segment k of the h-th stratum. Without loss of generality, M_{hk} could be assumed to be equal to A_{hk}, where A_{hk} is the crop area in segment hk. Equality requires only the choice of an appropriate scale for area.

Section 3.1 examines the yield estimator that is currently used. Conditions under which this estimator is unbiased for state average yield are investigated. A simple estimator of the variance of estimated yield is discussed in Section 3.2. Estimators of the unconditional variances of the yield and production estimators are developed in Section 3.3. A Monte Carlo study of estimators is given in Section 3.4.

3.1 Currently Used Yield and Production Estimators

Estimates of the state average yield are currently computed as though the sample were an equal probability simple random sample of secondary units. The estimator is the simple average yield of secondary units with positive acreages. That is, the estimated average yield per acre is

$$\bar{y} = D^{-1} \sum_{h=1}^{L} \sum_{k=1}^{n_h} \sum_{\ell=1}^{m_{hk}} Y_{hk\ell}\,\delta_{hk\ell}, \tag{3.1}$$

where

$$\delta_{hk\ell} = 1 \quad \text{if } A_{hk} > 0,$$

$$\delta_{hk\ell} = 0 \quad \text{if } A_{hk} = 0,$$

$$D = \sum_{h=1}^{L} \sum_{k=1}^{n_h} \sum_{\ell=1}^{m_{hk}} \delta_{hk\ell}, \tag{3.2}$$

m_{hk} is the number of sampled secondary units selected in segment hk, L is the number of strata, and $Y_{hk\ell}$ is the estimated yield per acre for secondary unit ℓ of segment hk. If the crop acreage in a segment, A_{hk}, is zero, then $m_{hk} = 1$ and $Y_{hk\ell} = 0$, by definition. The total number of observed secondary units for segments with positive acres is D.

Expression (3.1) can be written in the convenient operational form

$$\bar{y} = D^{-1} \sum_{t=1}^{D} Y_t, \tag{3.3}$$

where the subscript t replaces the triple subscript $hk\ell$ and the summation is over secondary units in segments with positive crop acres.

The estimator of average crop yield per acre (3.1) is a type of combined ratio estimator. This can be shown by using conditional selection probabilites to rewrite $\bar{y}$. In the NASS scheme, segments are selected systematically with probabilities proportional to expanded size, and segments with sufficiently large expanded acreage are included with certainty. The number of secondary units allocated to certainty segments is proportional to the size of the segment, up to rounding error. The rounding is performed by the systematic selection scheme. Let $\pi_{hk\ell}$ be the conditional probability that secondary unit ℓ in segment k of stratum h is selected, given the sample of segments selected at the first phase of the sampling procedure. We have

$$\pi_{hk\ell} = D \left(\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} M_{hk} \right)^{-1} N_h n_h^{-1} \tag{3.4}$$

for secondary units in segments with $A_{hk} > 0$, where N_h is the population number of segments in stratum h, M_{hk} is the number of secondary units in segment k of stratum h, and n_h is the number of segments in stratum h selected at the first phase. The conditional probability of observing a segment with zero acres at the second phase is one.

Then the mean estimator given in (3.1) can be written as

$$\bar{y} = \frac{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{K_h} \sum_{\ell=1}^{m_{hk}} \pi_{hk\ell}^{-1} Y_{hk\ell}}{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{K_h} \sum_{\ell=1}^{m_{hk}} \pi_{hk\ell}^{-1} \delta_{hk\ell}}, \tag{3.5}$$

where $N_h n_h^{-1}$ is the inverse of the first stage selection probability, K_h is the number of second phase segments drawn from stratum h, and $K = \Sigma K_h$. Given an appropriate scale, the numerator of (3.5) is an estimator of the total production and the denominator is an estimator of the total area. It can be shown that the numerator is an unbiased estimator by taking expectations, conditioning on the first phase sample units and then averaging over first phase samples. The denominator is a stratified estimator of the total number of secondary units. By the nature of the sampling, the number of sampling units is proportional to acreage and one can choose the scale so that the number of secondary units is equal to acreage. Hence, $\bar{y}$ can be viewed as the ratio of an unbiased estimator of the total production of the crop to an unbiased estimator of the total area under the crop.

To estimate total state production, NASS multiplies $\bar{y}$ *by* $\hat{A}$, where $\hat{A}$ is the estimator of total crop acreage defined by

$$\hat{A} = \sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} A_{hk}. \tag{3.6}$$

Thus, the estimated total production is

$$\hat{Y} = \hat{A}\,\bar{y}. \tag{3.7}$$

3.2 Simple Variance Estimators

Under the assumption of simple random sampling of secondary units from the entire set of secondary units available at the second phase, the estimated variance of $\bar{y}$ conditional on the second phase segments is

$$\hat{V}_2(\bar{y}) = D^{-1}\,(D-1)^{-1} \sum_{t=1}^{D} (Y_t - \bar{y})^2, \tag{3.8}$$

where the subscript 2 on $\hat{V}$ is used to denote conditional variance and the subscript t on Y replaces the triple subscript $hk\ell$. The sum over t is the sum over the D secondary units in segments with postive acres.

Because of the simplicity of expression (3.8), it has been suggested that it be used as an estimator of the unconditional variance. It has also been suggested that the variance of the estimated total state production be estimated with

$$\hat{V}_*(\hat{Y}) = \hat{A}^2 \hat{V}_2(\bar{y}) + \bar{y}^2 \hat{V}(\hat{A}) + \hat{V}(\hat{A})\hat{V}_2(\bar{y}), \tag{3.9}$$

where $\hat{A}$ is defined in (3.6) and $\hat{V}(\hat{A})$ is the usual variance estimator for a stratified estimated total,

$$\hat{V}(\hat{A}) = \sum_{h=1}^{L} N_h^2 n_h^{-1} (n_h - 1)^{-1} \sum_{k=1}^{n_h} (A_{hk} - \bar{A}_h)^2, \tag{3.10}$$

and

$$\bar{A}_h = n_h^{-1} \sum_{k=1}^{n_h} A_{hk}.$$

The estimator (3.9) is an estimator of the variance of a product based on an implicit assumption that $\bar{y}$ and $\hat{A}$ are uncorrelated.

Evaluation of the extent to which the estimator (3.9) tends to underestimate the variance of $\hat{Y}$ is difficult. We can express the unconditional variance of $\bar{y}$ as

$$\begin{aligned} V(\bar{y}) &= V_1 \{E_2(\bar{y})\} + E_1 \{V_2(\bar{y})\} \\ &= V_1 \{\hat{A}^{-1} \sum_{h=1}^{L} N_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.}\} + E_1 \{V_2(\bar{y})\}, \end{aligned} \tag{3.11}$$

where $Y_{hk.} = M_{hk} \bar{Y}_{hk.}$ is the total for the k-th segment in stratum h, and E_1 and V_1 denote the expectation and variance, respectively, with respect to first phase sampling.

The estimator $\hat{V}_2(\bar{y})$ is unbiased for the second component of expression (3.11) under simple random sampling of secondary units. Because sampling at phase two of the NASS scheme is done systematically, $\hat{V}_2(\bar{y})$ is a biased estimator of $V_2(\bar{y})$. The nature and extent of this bias depends upon the correlation structure of the list used in sample selection at the second phase. Also affecting the bias in $\hat{V}_2(\bar{y})$ as an estimator of the true variance is the fact that formula (3.8) was derived under an assumption of replacement sampling at phase two. To the extent that phase two sampling is actually done without replacement (because samples are drawn systematically from the list of expanded segment acreages, a segment is sampled more than once only if it is large), $\hat{V}_2(\bar{y})$ will overestimate $V_2(\bar{y})$.

The estimator $\hat{V}_*(\hat{Y})$ contains no estimator of $A^2 V_1 \{E_2(\bar{y})\}$, and this produces a negative bias. However, estimation of that component is not easy, even under the simplifying assumption of probability-proportional-to-size sampling at phase two. Because of these considerations, the performance of $\hat{V}_*(\hat{Y})$ will be studied by Monte Carlo methods in Section 3.4.

3.3 Alternative Estimators of Variance

An alternative approach to the estimation of $V(\bar{y})$ is to view the sample as a two phase sample, as shown in Table 1, and to assume that the unconditional probability of selecting a segment to receive a secondary unit is proportional to the conditional probability given the first phase segments.

Let π_{hk} be the conditional probability that segment k in stratum h is included in the second phase, given the first phase sample of segments. We have

$$\pi_{hk} = min(1, M_{hk} \pi_{hk\ell}), \tag{3.12}$$

where $\pi_{hk\ell}$ is a constant within segment hk. If $\pi_{hk} = 1$ and the segment is selected to receive more than one secondary unit, it is assumed that the secondary units are independently drawn.

Let π^*_{hk} be the unconditional probability that an observation is made on segment k in stratum h at phase two. If $A_{hk} = 0$, then π^*_{hk} is the unconditional probability that segment hk is selected to receive at least one secondary unit. If $A_{hk} = 0$, then π^*_{hk} is equal to the probability that segment hk is selected at the first phase of sampling. Let

$$\begin{aligned} \pi^*_{hk} &= \frac{n_h}{N_h} && \text{if } A_{hk} = 0\,, \\ \pi^*_{hk} &= \pi_{hk}\,\frac{n_h}{N_h} && \text{if } 0 < \pi_{hk} < 1\,, \end{aligned} \tag{3.13}$$

where π_{hk}, defined in (3.12), is the conditional probability that the hk-th segment is selected in phase two, given the first phase sample.

In our analysis we assume the π^*_{hk} to be fixed. This will be so and the probability π^*_{hk} will be the true unconditional probability if π_{hk} is a specified multiple of M_{hk} where the multiple is fixed before sample selection. Expression (3.13) will be an approximation if π_{hk} is a function of the segments selected at the first step of the selection procedure.

Expression (3.13) is proportional to M_{hk} for $M_{hk}\pi_{hk} \leq 1$. If $M_{hk}\pi_{hk\ell} > 1$, then the number of selected secondary units is greater than or equal to one. The correct number of secondary units to allocate to such segments to maintain a self-weighting sample of secondary units is $M_{hk}\pi_{hk\ell}$. In practice, the number of secondary units observed as a result of probability-proportional-to-size systematic sampling never differs from $M_{hk}\pi_{hk\ell}$ by more than one.

To simplify the remaining computations, we assume that the systematic sampling design contains no rounding error. In other words it is assumed that the number of secondary units observed per segment is equal to the number required for a self-weighting sample. Thus, it is assumed that the number of secondary units observed in a segment drawn as part of the second phase of sampling is

$$\begin{aligned} m_{hk} &= 1 && \text{if } 0 < \pi_{hk} < 1\,, \\ m_{hk} &= M_{hk}\pi_{hk\ell} && \text{if } \pi_{hk} = 1\,. \end{aligned} \tag{3.14}$$

Under this assumption, an unequal probability combined ratio estimator of the mean yield is equivalent to estimator (3.1). The combined ratio estimator is

$$\bar{y}_r = \hat{M}_r^{-1} \sum_{h=1}^{L} \sum_{k=1}^{K_h} \pi^{*-1}_{hk} M_{hk}\bar{y}_{hk.}\,, \tag{3.15}$$

where

$$\begin{aligned} \bar{y}_{hk.} &= m_{hk}^{-1} \sum_{\ell=1}^{m_{hk}} Y_{hk\ell} && \text{if } A_{hk} > 0\,, \\ \bar{y}_{hk.} &= 0 && \text{if } A_{hk} = 0\,, \end{aligned}$$

$$\hat{M}_r = \sum_{h=1}^{L} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} M_{hk} .$$

In expression (3.15) and the remaining expressions of this section, the reader can read $\hat{A}_r$ (total area) for $\hat{M}_r$ (total secondary units), if so desired.

In the following discussion, replacement sampling of segments with probabilities proportional to the area of a crop within the segment is assumed as an approximation to the probability-proportional-to-size systematic sampling scheme of the second phase. An estimator of the variance of $\bar{y}$ under the assumption of replacement sampling is

$$\hat{V}(\bar{y}_r) = \hat{M}_r^{-2} \sum_{h=1}^{L} K_h (K_h - 1)^{-1} \sum_{k=1}^{K_h} (\pi_{hk}^{*-1} u_{hk} - \bar{u}_{h.})^2 , \tag{3.16}$$

where

$$u_{hk} = M_{hk} (\bar{y}_{hk.} - \bar{y}_r) ,$$

$$\bar{u}_{h.} = K_h^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*-1} u_{hk} .$$

An estimator of the total production is

$$\hat{Y}_r = N \bar{M}_n \bar{y}_r , \tag{3.17}$$

where

$$\bar{M}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} M_{hk}$$

N is the total number of segments in the population and $W_h = N^{-1} N_h$. The Taylor approximation of the unconditional variance of the approximate distribution of $\hat{Y}_r$ is

$$V\{\hat{Y}_r\} = N^2 [\bar{M}_N^2 \, V\{\bar{y}_r\} + 2\bar{M}_N \bar{\bar{y}}_N \, C\{\bar{y}_r, \bar{M}_n\} + \bar{\bar{y}}_N^2 \, V\{\bar{M}_n\}] , \tag{3.18}$$

where $\bar{y}_r$ is given in (3.15), $\bar{M}_n$ is defined in (3.17),

$$\bar{M}_N = N^{-1} \sum_{h=1}^{L} \sum_{k=1}^{N_h} M_{hk} ,$$

$$\bar{\bar{y}}_N = \left(\sum_{h=1}^{L} \sum_{k=1}^{N_h} M_{hk} \right)^{-1} \sum_{h=1}^{L} \sum_{k=1}^{N_h} Y_{hk.} ,$$

$Y_{hk.} = M_{hk} \bar{Y}_{hk.}$ is the total for the k-th segment in stratum h, and $C\{\bar{y}_r, \bar{M}_n\}$ is the covariance between $\bar{y}_r$ and $\bar{M}_n$.

Under the unequal-probability-fixed-take procedure, the estimator $\bar{y}_r(\doteq \bar{y})$ is approximately conditionally unbiased for the mean yield for the $n = \Sigma n_h$ segments in the first phase sample. The mean yield of the n segments is

$$\bar{\bar{y}}_n = \bar{M}_n^{-1} \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.}\,.$$

Therefore, the covariance between $\bar{y}_r$ and $\bar{M}_n$ is the covariance between $\bar{M}_n^{-1}\,\bar{Y}_n$ and $\bar{M}_n$, where

$$\bar{Y}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} Y_{hk.}\,.$$

Using the common approximation for a ratio, the covariance between $\bar{y}_r$ and $\bar{M}_n$ can be approximated by

$$\begin{aligned} C\{\bar{M}_n^{-1}\,\bar{Y}_n,\, \bar{M}_n\} &\doteq C\{(\bar{Y}_n - \bar{\bar{y}}_N\,\bar{M}_n)\bar{M}_N^{-1},\, \bar{M}_n\} \\ &= \bar{M}_N^{-1}[C\{\bar{Y}_n,\, \bar{M}_n\} - \bar{\bar{y}}_N\, V\{\bar{M}_n\}]\,. \end{aligned} \tag{3.19}$$

If the probability of observing the pair $(Y_{hk.},\, M_{hk})$ is proportional to π_{hk}^*, an estimator of the covariance between $\bar{Y}_n$ and $\bar{M}_n$ is

$$\hat{C}\{\bar{Y}_n,\, \bar{M}_n\} = \sum_{h=1}^{L} W_h^2\, n_h^{-1}\, \hat{S}_{MYh} \tag{3.20}$$

where

$$\hat{S}_{MYh} = K_h (K_n^{-1})^{-1} \left(\sum_{j=1}^{K_h} \pi_{hj}^{*\,-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*\,-1}\, (M_{hk} - \bar{M}_h^*)(M_{hk}\bar{y}_{hk.} - \bar{y}_{h..}^*)\,,$$

$$\bar{M}_h^* = \left(\sum_{j=1}^{K_h} \pi_{hj}^{*\,-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*\,-1}\, M_{hk}\,,$$

$$\bar{y}_{h..}^* = \left(\sum_{j=1}^{K_h} \pi_{hj}^{*\,-1} \right)^{-1} \sum_{k=1}^{K_h} \pi_{hk}^{*\,-1}\, M_{hk}\bar{y}_{hk.}\,.$$

The estimator $\hat{S}_{MYh}$ is constructed as a degrees-of-freedom adjustment to a Horvitz-Thompson ratio estimator of the mean of the products $(M_{hk} - \bar{M}_h)(Y_{hk.} - \bar{Y}_{h..})$. The degrees-of-freedom adjustment, the factor $K_h(K_h - 1)^{-1}$, is introduced because it is necessary to replace the population means with sample means when constructing the product.

Substituting (3.15), (3.16), and (3.20) into (3.18) gives

$$\hat{V}\{\hat{Y}_r\} = N^2[\bar{M}_n^2\, \hat{V}\{\bar{y}_r\} + 2\bar{y}_r\hat{C}\{\bar{Y}_n,\, \bar{M}_n\} - \bar{y}_r^2\,\hat{V}\{\bar{M}_n\}]\,, \tag{3.21}$$

where $\hat{V}\{\bar{M}_n\}$ is the variance estimator for a stratified mean. Equation (3.21) is a stratified double sampling estimator of the variance of the estimated total state production. Unlike the estimator $\hat{V}_*(\hat{Y})$ of (3.9), estimator (3.21) does not assume that the yield and acreage estimators are uncorrelated. Equation (3.21) also uses an unconditional estimator of the variance of yield.

3.4. A Monte Carlo Comparison of Estimators

A Monte Carlo study was performed to illustrate the differences among alternative estimators. Cotton acreage data from the 1983 June Enumerative Survey in California and data from the corresponding 1983 objective yield survey were used as a basis for the study. For purposes of the Monte Carlo study, 28 strata were considered to have cotton.

Table 2 shows the distribution of cotton among the 28 strata as observed in the 1983 June Enumerative Survey. Fecso and Johnson (1981) describe the six different land uses, where land use is the first two digits of the stratum identification, as follows:

1300 - 50% or more cultivated land, primarily general crops with less than or equal to 10% fruit or vegetables;
1700 - 50% or more cultivated land, primarily fruit, tree nuts, or grapes mixed with general crops;
1900 - 50% or more cultivated land, primarily vegetables mixed with general crops;
2000 - 15-50% cultivated land with extensive cropland and hay;
3100 - residential mixed with agricultural lands, more than 20 dwellings per square mile;
4100 - less than 15% cultivated land, primarily privately owned rangeland.

A population was simulated from the results of the 1983 June Enumerative Survey. Table 2 compares the characteristics of the simulated population to the results of the survey. In the simulated population, cotton was determined to be present in segment k $(k = 1, \ldots, N_h)$ within stratum h $(h = 1, \ldots, 28)$ if $X_{hk} = 1$, where X_{hk} is an independent Bernoulli (p_h) random variable and p_h is the observed proportion of segments in stratum h found to have cotton in the 1983 June Enumerative Survey.

The next step in the creation of the population was the assignment of cotton acres to the segments for which $X_{hk} = 1$. A set of 1983 observed ratios of segment cotton acreages to the average segment acreage was compiled for land use substrata having more than one segment with cotton in the 1983 June Enumerative Survey. This set of observed ratios was used to generate the number of cotton acres in segments having cotton. If $X_{hk} = 1$, then a ratio, r_{hk}, was drawn from the set of observed ratios such that each observed ratio in the set had an equal probability of selection. The number of acres of cotton in segment hk, M_{hk}, was defined by

$$M_{hk} = r_{hk}\bar{M}_{h.}\,, \tag{4.1}$$

where $\bar{M}_{h.}$ was the observed average number of cotton acres for segments with cotton in stratum h in the 1983 June Enumerative Survey. (See Table 2.)

Results of the 1983 objective yield survey for cotton were used to simulate yield observations within segments. Since estimated yields were not readily accessible, an alternative variable - a major component of yield estimates - was used. This variable is the number of plants per 100 square feet. The estimated overall population mean number of plants per 100 square feet was 79.6 for the 1983 objective yield survey. Table 3 shows the average number of plants

Table 2

Cotton Acreage Estimates from the 1983 June Enumerative Survey in California and Cotton Acreages in the Simulated Population

Stratum	Target Segment Size (Acres)	Number of Segments in Stratum	Number of Segments Sampled in 1983	Percentage of Segments with Cotton		Mean Acres Cotton in Segments with Cotton	
				1983	Simulated Population	1983	Simulated Population
1314	640	291	10	60	60	197	200
1315	640	291	10	100	100	354	348
1316	640	291	10	90	89	167	173
1317	640	291	10	90	92	149	148
1318	640	291	10	50	53	481	422
1319	640	291	10	20	19	249[1]	260
1320	640	291	10	90	91	154	155
1321	640	291	10	60	61	270	274
1322	640	291	10	70	71	205	210
1323	640	291	10	80	79	288	279
1713	320	432	10	30	28	125	122
1714	320	432	10	30	31	58	57
1715	320	432	10	20	22	86[2]	84
1716	320	432	10	10	8	86[2]	89
1717	320	432	10	40	38	26	27
1718	320	432	10	30	29	144	144
1719	320	432	10	30	31	65	67
1720	320	432	10	30	30	38	35
1721	320	432	10	30	29	133	138
1722	320	432	10	50	47	130	131
1723	320	432	10	40	40	76	76
1906	640	362	10	70	73	117	127
1907	640	362	10	70	74	192	194
1908	640	362	10	80	83	253	246
2010	640	649	10	30	31	303	306
2011	640	649	10	40	41	175	165
3107	160	1,847	5	20	22	25[3]	25
4110	2,560	1,044	10	10	10	178	165

[1] Number of segments sampled was less than or equal to 2. Average of all segments in substrata within land use stratum 13 is shown.

[2] Number of segments sampled was less than or equal to 2. Average of all segments in substrata within land use stratum 17 is shown.

[3] Number of segments sampled was less than or equal to 2. Approximate acreages for this agri-urban stratum are shown.

per 100 square feet. The estimated overall population mean number of plants per 100 square feet was 79.6 for the 1983 objective yield survey. Table 3 shows the average number of plants per 100 square feet by stratum for the 1983 survey. The average for each stratum is based on all secondary units within the stratum that were drawn as part of the probability-proportional-to-estimated-size sampling scheme.

An analysis of variance of the 1983 plant data (Table 4) shows that 28 percent of the total variation among secondary units was due to between-segment differences within strata ($s_b^2 = 378.0$), whereas 58 percent of the total variation was due to variation among secondary units within segments ($s_w^2 = 776.6$). If the stratum component is treated as fixed, 67 percent of the within-segment variation is due to variance among secondary units.

Table 3

Average Number of Plants per 100 Square Feet from the 1983 Objective Yield Survey for Cotton in California and in the Simulated Population

Stratum	Average Number of Plants per 100 Square Feet	
	1983 Objective Yield Survey	Simulated Population
1314	78	76
1315	80	80
1316	67	68
1317	72	73
1318	80	80
1319	93	93
1320	92	91
1321	70	69
1322	84	84
1323	72	71
1713	118	117
1714	96[1]	95
1715	96[1]	93
1716	96[1]	86
1717	96[1]	96
1718	139	140
1719	96[1]	97
1720	96[1]	97
1721	89	86
1722	79	79
1723	84	85
1906	98	98
1907	67	67
1908	53	53
2010	118	118
2011	47	47
3107	80[2]	79
4110	60	59

[1] Number secondary units observed was less than or equal to 2. Secondary unit average for land use stratum 17 is shown.

[2] Number secondary units observed was less than or equal to 2. Secondary unit average for all strata is shown.

Table 4

Analysis of Variance for the 1983 Objective Yield Survey Data

Source	Degrees of Freedom	Sum of Squares	Mean Square	Variance Component	Percent of total
Stratum	26	80,193	3,084.3	187.3	14
Segment within Stratum	85	124,086	1,459.8	378.0	28
Residual	103	79,991	776.6	776.6	58
Total	214	284,270		1,341.9	100

When a segment had cotton, the mean number of plants per 100 square feet for segment *hk* was simulated by

$$\bar{c}_{hk} = \bar{c}_{h.} + e_{hk}, \tag{4.2}$$

where $\bar{c}_{h.}$ is the average number of plants per 100 square feet for stratum *h*, e_{hk} is distributed $N(0,\ s_b^2)$, and $s_b^2 = 378.0$. In the event that the simulated segment mean $(\bar{c}_{hk})$ was less than 10% of the stratum mean, then c_{hk} was set equal to $(.10)\bar{c}_{h.}$. Table 3 compares the simulated stratum means with those from the 1983 objective yield survey. The overall mean in the simulated population was $\bar{\bar{y}}_N = 79.6$.

From the simulated population 500 June Enumerative Survey samples were drawn using stratified random sampling. A total of 275 segments were drawn for each of the simulated samples. The number of segments drawn from each stratum was the same as that for the 1983 June Enumerative Survey (see Table 2). For each of the simulated samples, estimates of the mean number of acres per segment in the population, as well as the conditional probabilities π_{hk}, from (3.12), that the segments in the sample would receive plots in a draw, were calculated. These conditional probabilities were used at the second stage of sampling in the single start probability-proportional-to-estimated-size systematic sampling described in Section 2. Objective yield survey samples were simulated by selecting 220 secondary units using this systematic sampling scheme. Two objective yield survey samples were simulated for each of the 500 simulated June Enumerative Survey samples.

When a segment was selected to receive a secondary unit, the yield (number of plants per 100 square feet) observed within a field was simulated under the assumption that the coefficient of variation within each segment was constant. The observed number of plants was defined as

$$y_{hk\ell} = \bar{c}_{hk} + s_w \bar{\bar{y}}_N^{-1}\ \bar{c}_{hk} f_{hk\ell}, \tag{4.3}$$

where $y_{hk\ell}$ is the estimated average number of plants per 100 square feet for the ℓ-th secondary unit in segment *k* of stratum *h*, and $f_{hk\ell}$ is distributed $N(0,\ 1)$. The within-segment standard error is the square root of the $s_w^2 = 776.6$ of Table 4, and $\bar{\bar{y}}_N$ is the overall mean number of plants per plot. In the event that $y_{hk\ell}$ was less than 10% of the stratum mean, then $y_{hk\ell}$ was set equal to $(.10)\bar{c}_{hk}$. Similarly, if $y_{hk\ell}$ was greater than 190% of the stratum mean, then $y_{hk\ell}$ was set equal to $(1.9)\bar{c}_{hk}$.

Results of the simulations for cotton acreages are summarized in Table 5. The estimated mean acres per segment is

$$\bar{A}_n = \sum_{h=1}^{L} W_h n_h^{-1} \sum_{k=1}^{n_h} A_{hk}, \tag{4.4}$$

with estimated variance

$$\hat{V}(\bar{A}_n) = \sum_{h=1}^{L} W_h{}^2 n_h^{-1} (n_h - 1)^{-1} \sum_{k=1}^{n_h} (A_{hk} - \bar{A}_h)^2. \tag{4.5}$$

Table 5

Estimated Cotton Acreages from 500 Simulated June Enumerative Survey Samples

	$\bar{A}_n$	$\hat{V}(\bar{A}_n)$
Average	9.93	0.64
Range	8.13 - 12.21	
Variance	0.66	0.016

The average cotton acres per segment in the simulated population was 9.94, while the average of the 500 sample estimates was 9.93. The actual variance of the stratified estimator $\bar{A}_n$ was 0.63, while the average estimated variance for the 500 simulated samples was 0.64. Because the variation in estimated cotton acreage is small, π^*_{hk} provides a stable estimate of the unconditional probability that segment k in stratum h is selected to receive at least one secondary unit.

In addition to the estimators discussed previously, random group estimators of the variance were constructed. Two sets of random groups were formed for each objective yield survey sample. One set contained five groups ($\gamma = 5$) and one set contained ten groups ($\gamma = 10$). Random groups were created by dividing the primary sampling units, the segments, into subsets within each land use substratum. The first group in each set of groups was obtained by drawing a simple random sample without replacement of size $K_{h(\gamma)} = n_h/\gamma$ from the sample of segments selected from each stratum ($h=1, \ldots, 28$) of the parent June Enumerative Survey sample. The second random group was obtained in the same fashion by selecting $K_{h(\gamma)}$ segments from the remaining $n_h - K_{h(\gamma)}$ segments in each stratum. The remaining random groups were formed in a like manner. One land use substratum, stratum number 3107, had a sample size of $n_h = 5$ segments. Acreage and yield values of the observed five segments were repeated to form the ten observations required to create ten groups when $\gamma = 10$.

Let D_α be the number of secondary units with positive acres which were selected during the objective yield survey in random group α where $\alpha = 1, \ldots, \gamma$. Let $\bar{y}_{(\alpha)}$ denote the yield estimator obtained from the α-th random group:

$$\bar{y}_{(\alpha)} = D_\alpha^{-1} \sum_{t=1}^{D_\alpha} Y_{t(\alpha)}, \tag{4.6}$$

where $\bar{y}_{(\alpha)}$ is the analogue of equation (3.3) for the α-th group. The random group estimator of the variance of $\bar{y}$ is then given by

$$\hat{V}_{g\gamma}(\bar{y}) = \gamma(\gamma - 1)^{-1} \sum_{\alpha=1}^{\gamma} (\bar{y}_{(\alpha)} - \bar{y})^2. \tag{4.7}$$

This estimator is slightly biased for the ten group estimator because one stratum contained only five observations, and these observations were repeated in the groups.

Similarly, let $\hat{Y}_{(\alpha)}$ denote the total production estimator obtained from the α-th random group:

$$\hat{Y}_{(\alpha)} = N \bar{M}_{n(\alpha)} \bar{y}_{(\alpha)}, \tag{4.8}$$

where

$$\bar{M}_{n(\alpha)} = \sum_{h=1}^{L} W_h K_{h(\alpha)}^{-1} \sum_{k=1}^{K_{h(\alpha)}} M_{hk(\alpha)},$$

$M_{hk(\alpha)}$ is the number of acres of cotton in segment k of stratum h for random group α and $K_{h(\alpha)}$ is the number of segments in stratum h for the α-th group. The random group estimator of the variance of $\hat{Y}$ is then given by

$$\hat{V}_{g\gamma}(\hat{Y}) = \gamma(\gamma - 1)^{-1} \sum_{\alpha=1}^{\gamma} (\hat{Y}_{(\alpha)} - \hat{Y})^2. \tag{4.9}$$

Tables 6 and 7 summarize the results of the Monte Carlo study for yield and production estimators. Average values of the estimates and their variance estimates across the 1,000 simulated objective yield survey samples are shown in the tables. Simulation of two objective yield survey samples for each June Enumerative Survey sample made the estimation of between - and within - June Enumerative Survey variance components possible.

The estimator (3.1) currently used, $\bar{y}$, and the combined ratio estimator (3.15), $\bar{y}_r$, which is based on the π_{hk}^* calculated from June Enumerative survey results, provide estimates with similar accuracy (see Table 6). The equal efficiency is partly due to the accuracy with which the unconditional selection probabilities are estimated in each sample.

As was shown in Section 3.2, the conditional variance $\hat{V}_2(\bar{y})$ is an underestimate of $V(\bar{y})$. For this simulated population, $\hat{V}_2(\bar{y})$ underestimated the observed variance of $\bar{y}$ by 38%. The observed variance of $\bar{y}$ was 11.57 as compared to an average of 7.21 for $\hat{V}_2(\bar{y})$. This underestimation of the variance was consistent across samples. The estimated variance of $\hat{V}_2(\bar{y})$ was 0.99, with $\hat{V}_2(\bar{y})$ ranging from a low of 3.85 to a high of 11.24 in the 1,000 observations. Thus, the maximum observed estimate of the conditional variance was less than the true variance.

Table 6

Monte Carlo Properties of Yield per Acre Estimates and Estimated Variances[1]

	Estimator					
	$\bar{y}$	$\hat{V}_2(\bar{y})$	$\hat{V}_{g5}(\bar{y})$	$\hat{V}_{g10}(\bar{y})$	$\bar{y}_r$	$\hat{V}(\bar{y}_r)$
Average	79.74	7.21	12.62	12.39	79.76	12.39
Total Variance	11.57	0.99	74.58	36.86	11.56	12.51
Between JES	7.60	0.48	6.10	4.56	7.64	7.61
Within JES	3.97	0.51	68.48	32.30	3.92	4.90

[1] Two objective yield survey samples were simulated from each of 500 simulated June Enumerative Survey samples.

Table 7

Monte Carlo Properties of Production Estimates and Estimated Variances[1]

	Estimator[2]					
	$\hat{Y}$	$\hat{V}_*(\hat{Y})$	$\hat{V}_{g5}(\hat{Y})$	$\hat{V}_{g10}(\hat{Y})$	$\hat{Y}_r$	$\hat{V}(\hat{Y}_r)$
Average	73.04	40.85	48.99	48.53	73.07	48.73
Total Variance	49.69	82.52	1245.10	608.80	49.58	222.96
Between JES	46.35	78.17	50.82	208.48	46.30	199.58
Within JES	3.34	4.35	1194.28	400.32	3.28	23.38

[1] Two objective yield survey samples were simulated from each of 500 simulated June Enumerative Survey samples. There were $N = 92{,}240$ segments in the simulated population.

[2] The estimator $\hat{Y}$ is in millions of units and variances are in the corresponding units.

Assuming probability-proportional-to-size sampling with replacement of segments at the second phase, $\hat{V}_2(\bar{y})$ was shown in Section 3.2 to be unbiased for the variance of $\bar{y}$ conditional on the sample of segments selected at the first stage of sampling. The estimate of the expected value of the conditional variance of $\bar{y}$, $V_2(\bar{y})$, from the Monte Carlo study is 3.97. This large discrepancy (3.97 versus 7.21) can be attributed to the fact that the estimator $\hat{V}_2(\bar{y})$ ignores the effects of stratification in the population (see Tables 2 and 3) and to the fact that $\hat{V}_2(\bar{y})$ was derived under the assumption that segments are selected with replacement at the second stage of sampling.

The estimator (3.9), $\hat{V}_*(\hat{Y})$, underestimates the unconditional variance of $\hat{Y}$. While the observed variance of $\hat{Y}$ from the Monte Carlo simulations is 49.69 (million)^2, the average of the $\hat{V}_*(\hat{Y})$ is only 40.85 (million)^2. This 18% underestimate of the true variance occurs for a number of reasons. As was shown previously, there is a negative bias in $\hat{V}_2(\bar{y})$ as an estimator of $\hat{V}(\bar{y})$; another important factor contributing to the bias is the failure of $\hat{V}_*(\hat{Y})$ to take into account the covariance between $\bar{M}_n$ and $\bar{y}$. In this example, the bias caused by omitting the covariance term partially balances the bias associated with $\hat{V}(\bar{y})$.

Using expression (3.16), $\hat{V}(\bar{y}_r)$, as an estimator of the variance of $\bar{y}_r$ and expression (3.21), $\hat{V}(\hat{Y}_r)$, as an estimator of the variance of $\hat{Y}_r$ provided results which are much more satisfactory than those of the estimators currently used. The Monte Carlo average of the estimates $\hat{V}(\bar{y}_r)$ was 12.51, which overestimates the observed variance of $\bar{y}_r$ (11.57) by about 7%. About one-third of the overestimate (2-4%) can be attributed to the use of sampling without replacement at the first two stages of sampling. The remaining difference of about 4% is small relative to the standard error of the estimated difference. The variance of the difference was estimated by estimating the variance of the mean of z_{tj}, where

$$z_{tj} = (\bar{y}_{n,r(tj)} - 79.76)^2 - \hat{V}(\bar{y}_{n,r(tj)}) , \qquad (4.10)$$

for the j-th yield sample ($j = 1, 2$) within June Enumerative Survey sample t ($t = 1, \ldots, 500$). The estimated standard error of the difference was 0.58. Thus, the average value of $\hat{V}(\bar{y}_r)$ is within 1.5 standard errors of the estimated variance of $\bar{y}_r$. The average estimated variance of $\hat{Y}_r$ is within 2 percent of the variance observed in the Monte Carlo simulations.

Random group estimators of the variance of $\bar{y}$ displayed little bias. The Monte Carlo averages of estimators $\hat{V}_{g5}(\bar{y})$ and $\hat{V}_{g10}(\bar{y})$ were 9% and 7%, respectively, larger than the corresponding Monte Carlo variances. These differences are not significantly different from zero and are comparable to those obtained for the estimator $\hat{V}(\bar{y}_r)$. The variance estimator $\hat{V}(\bar{y}_r)$, however, is a much more stable variance estimator. The coefficient of variation for the estimator $\hat{V}(\bar{y}_r)$ is about 30%; it is 75% for $\hat{V}_{g5}(\bar{y})$. As expected (Wolter 1985), an increase in the number of random groups resulted in a decrease in the coefficient of variation of the random group variance estimator. The coefficient of variation for $\hat{V}_{g10}(\bar{y})$ was 50%. Differences among random groupings and yield samples within June Enumerative Surveys accounted for most of the variance in the random groups variance estimators.

4. CONCLUSIONS

Analyses show that the estimators of statewide average yield and total production currently used by the National Agricultural Statistics Service are satisfactory. However, the simple variance estimators $\hat{V}_2(\bar{y})$ and $\hat{V}_*(\hat{Y})$ were shown to have a negative bias, where the extent of the underestimation is a function of the within-segment variance and of the within-segment sampling rates. The estimator $\hat{V}_2(\bar{y})$ underestimated the true variance of $\bar{y}$ by nearly 40%, and $\hat{V}_*(\hat{Y})$ underestimated the true variance of $\hat{Y}$ by 18% for the simulated California cotton population.

The alternative estimators, $\bar{y}_r$ and $\hat{Y}_r$, were developed by viewing the yield sampling scheme as a two-phase process in which segments found to contain crop acreage during phase one (the June Enumerative Survey) are subsampled during phase two to estimate yield. The unconditional probability of selecting a segment to receive a secondary unit within a stratum, π^*_{hk}, is estimated by assuming that this probability is proportional to the conditional probability of selecting segments at the second phase of sampling. With this assumption, the unequal probability combined ratio estimator of the mean yield, $\bar{y}_r$, and the estimator of its variance, $\hat{V}(\bar{y}_r)$, were developed. The estimator of the total $\hat{Y}_r$ is a two-phase product estimator of the mean production per segment, where the estimator of the mean of the auxiliary variable (crop acreage) comes from the June Enumerative Survey (phase one of sampling). The variance estimator $\hat{V}(\hat{Y}_r)$ is a stratified double sampling (two-phase) estimator of the variance of $\hat{Y}_r$.

As shown by the Monte Carlo study, $\bar{y}_r$ and $\hat{Y}_r$ give estimates that are comparable to their currently used counterparts, $\bar{y}$ and $\hat{Y}$. Both $\hat{V}(\bar{y}_r)$ and $\hat{V}(\hat{Y}_r)$ are accurate variance estimators in samples of the size typically used by NASS. These results are due, in part, to the precision with which average crop acreages are estimated by the June Enumerative Survey. Precise acreage estimates produce estimates of selection probabilities that are close to the unconditional probabilities of selection. In addition, the ratio form of the estimator reduces the effect of replacing true unconditional probabilities with estimators.

Random group variance estimators are also essentially unbiased estimators of the variance of estimated yield and production. However, random group estimators are much less stable than $\hat{V}(\bar{y}_r)$ and $\hat{V}(\hat{Y}_r)$. Therefore, estimators $\hat{V}(\bar{y}_r)$ and $\hat{V}(\hat{Y}_r)$ are recommended over random group estimators.

The June Enumerative Survey forms phase one of the objective yield survey. Sampling procedures for the June Enumerative Survey are straightforward and, as was shown by the Monte Carlo study, provide accurate acreage estimates. Hence, no change in the overall design for phase one of the objective yield survey is recommended.

A number of modifications for phase two of the objective yield surveys should be investigated. The current procedure for estimating yield is a two phase procedure in which a combined ratio estimator is used. In states where the sample is relatively large, independent sampling at phase two within individual strata or for groups of strata, as well as the use of a separate ratio estimator should be considered.

Systematic sampling at phase two should be replaced if unbiased estimators of the variance are desired. Segments for yield sampling at phase two are now selected by computer at a national level so it should be relatively easy to change to a selection procedure with known joint selection probabilities. Estimators similar to those recommended for the current design would still be suitable if the same selection probabilities were retained. The scheme described by Fuller (1970) is one procedure that can be computerized, for which joint selection probabilities can be calculated, and which maintains specified selection probabilities and a degree of control similar to that of systematic sampling.

ACKNOWLEDGEMENTS

This research was supported in part by cooperative research agreement 58-319T-1-0054X with the National Agricultural Statistics Service, U.S. Department of Agriculture. We thank the referees for useful comments.

REFERENCES

FECSO, R. (1978). Cluster analysis as an aid in creating paper strata. Statistical Reporting Service, U.S. Department of Agriculture.

FECSO, R., and JOHNSON, V. (1981). The new California area frame: A statistical study. Statistical Reporting Service, U.S. Department of Agriculture.

FULLER, W.A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, Ser. B, 32, 209-226.

HOUSEMAN, E.E. (1975). Area frame sampling in agriculture. Statistical Reporting Service, U.S. Department of Agriculture.

PRATT, W.L. (1984). The use of interpenetrating sampling in area frames. Statistical Reporting Service, U.S. Department of Agriculture.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.