1992 CENSUS OF AGRICULTURE VARIANCE ESTIMATION

Richard Griffiths, David Hornick, Inez Chen, and Tony Eleftherakis, Bureau of the Census
Richard Griffiths, Agriculture Division, Bureau of the Census, Washington, D.C. 20233

KEY WORDS: Synthetic estimation; Integer weighting; Nonresponse survey.

1. INTRODUCTION

In order to study the United States agricultural system, an agriculture census is taken every five years. In the census, data are gathered for many aspects of agriculture in the U.S. For some aspects of the agricultural system data are gathered from all farms; for other aspects data are gathered only from a sample of farms. These items for which data are gathered only from a sample of farms are referred to as sample items. This paper will focus on the variance estimation methodology for the sample items.

In order to understand the census of agriculture's variance estimation methodology, it is necessary to understand the design and estimation procedures used. The next few sections discuss different aspects of the sample design and estimation.

2. NONRESPONSE ESTIMATION

In census of agriculture estimation two types of statistical estimation procedures are used: nonresponse estimation and sample estimation; the former, which accounts for nonresponse to the census, is the topic of this section.

To determine the number of nonrespondents which operate farms, a nonresponse survey is conducted independently for each of the states, except Alaska in which all farms receive 100% follow-up. 100% follow-up is the procedure by which response is obtained by "reported data acquired by telephone or mail, or secondary source information, or imputed from historic data or a combination of the above methods." (Source: 1987 Census of Agriculture Glossary.) Cases subject to 100% follow-up are not eligible for inclusion in the nonresponse survey.

The nonresponse survey in each state is a stratified systematic sample of all eligible nonresponse cases. From each case selected for the sample, enough information is collected to determine the farm status of that case. The information obtained from this sample provides us with an estimate of the proportion of nonrespondents which are farms in each stratum (called nonresponse stratum) at the state level. This naturally leads us to construct state-level estimates of the number of nonrespondent cases in each stratum which are actually farms. Synthetic estimation is used to construct county-level estimates of the number of nonrespondent farms.

The nonresponse strata are then collapsed to ensure that within each county the ratio of the estimated number of nonrespondents plus the number of eligible census respondents to the number of eligible census respondents is never greater than 2. The resulting strata are referred to as final nonresponse strata (FNRS).

After the county-level estimates are derived, a nonresponse weight is assigned to all respondent cases. A nonresponse weight of 1 is assigned to all cases which received 100% follow-up. Within each FNRS and each county a noninteger nonresponse weight is calculated and assigned to all eligible cases. This noninteger nonresponse weight is the ratio of the sum of the estimated number of nonrespondent farms and the number of eligible census respondent farms to the number of eligible census respondent farms. The noninteger nonresponse weight, ${}_{nin}W$, may be depicted as

$$ {}_{nin}W = \frac{\hat{N}_{ijk} + {}_{nf}T_{ijk}}{{}_{nf}T_{ijk}} $$

where

$\hat{N}_{ijk}$ = the estimated number of nonrespondent farms in state i, county j,

FNRS k (synthetic estimate),

$_{nf}T_{ijk}$ = the number of eligible census respondent farms in state i, county j, FNRS k.

For estimation purposes the $_{nin}W$ assigned to each case is randomly rounded to an integer. Since the $_{nin}W$ is no less than 1 and no greater than 2, all cases are assigned an integer nonresponse weight of 1 or 2. Section 5 details the procedures used for assigning these integer nonresponse weights.

3. SAMPLE DESIGN

The final 1992 Census of Agriculture mail list consists of 3.55 million addresses. Each of these addresses is designated to receive one of three forms: a nonsample form, a screener form, or a sample form. The nonsample form contains sections for the items for which data are collected from all farmers. The screener form is identical to the nonsample form except it contains some early questions that allow for quick identification of nonfarms. The sample form contains all the sections of the nonsample form plus some sections which contain items for which data are collected only from the farmers that receive this form; these items are the sample items.

In this paper our interest lies with the sample items; thus, we will describe the sample design used to obtain the addresses that the sample forms are mailed to.

For the 1992 Census of Agriculture, sample forms are mailed to all mail list addresses located in Alaska, Hawaii, and Rhode Island; in all other states, sample forms are mailed only to a sample of the mail list addresses.

Some addresses are selected to receive the sample form with certainty, that is, with probability equal to one. These addresses are referred to as certainty addresses; they are identified as operations expected to have a large total value for agricultural products sold, large acreage, multiunit operations, farms with special characteristics, or farms in counties with fewer than 100 farms in 1987.

All other addresses — referred to as noncertainty addresses — comprise the universe which is systematically sampled within each county: Addresses in counties containing 100 to 199 farms in 1987 are sampled at a rate of 1 in 2; addresses in counties containing at least 200 farms in 1987 are sampled at a rate of 1 in 6.

4. SAMPLE ESTIMATION

The first step in obtaining estimates from the sample item data is to poststratify all respondent sample farms. For this purpose, 33 initial sample poststrata (ISPS) are constructed. All certainty farms are assigned to ISPS 0. The noncertainty farms are assigned to ISPS 1 to 32 according to each farm's reported data for three items, total value of agricultural products sold, standard industrial classification, and land in farms.

Each certainty respondent farm is assigned a sample weight of one. Within each ISPS, each respondent noncertainty sample farm is assigned an initial sample weight equal to the sum of the $_{nin}W$ for all census respondents divided by the sum of the $_{nin}W$ for the sample form respondents. (Census respondents are those in-scope farms that respond to any of the forms.) The initial sample poststrata are then collapsed to ensure that each stratum contains a weighted total (adjusted for nonresponse) of at least 10 sample cases and each sample record has a sample weight no greater than two times the inverse of the mail sample rate for the county it is located in. The resulting strata are referred to as final sample poststrata (FSPS).

After the collapsing procedure is completed, a noninteger final sample poststratum weight is assigned to each noncertainty sample farm within each FSPS. This weight is equal to the ratio of the sum of the $_{nin}W$ for all census respondents to the sum of the $_{nin}W$ for the sample respondents.

Finally, the final real sample weight is assigned to all respondent sample farms. This weight is equal to the product of the noninteger final sample poststratum weight (1 for certainty farms) and the noninteger nonresponse weight for each farm. The integer portion of the final real sample weight is referred to as the base final sample weight.

For estimation purposes the final real sample weight is randomly rounded to an integer. This integer weight is called the final integer sample

weight and is equal to either the base final sample weight or the base final sample weight plus one. The integer weighting procedures are described in the next section.

5. INTEGER WEIGHTING PROCEDURES

For the purpose of estimating values for sample items, integer weighting procedures are performed on two weights, the final real sample weight and the noninteger nonresponse weight. A final integer sample weight is assigned to each noncertainty sample farm, and a final integer nonresponse weight is assigned to each certainty farm. (A final integer nonresponse weight is also assigned to each noncertainty sample farm, but it is not used in the estimation procedures for sample items.)

All noncertainty farms within each FSPS and each FNRS have identical final real sample weights. A systematic sample of farms is chosen to receive a final integer sample weight of the base final sample weight plus one. These farms are chosen with probability equal to the fractional portion of the final real sample weight. Farms not chosen in this sample receive a final integer sample weight equal to the base final sample weight. So, for example, if the final real sample weight is 7.2 in a particular FSPS and FNRS, 20% of the farms in this FSPS and FNRS are chosen to receive a final integer sample weight of 8; the other farms receive a final integer sample weight of 7.

The final integer nonresponse weight is assigned similarly. Within each FNRS the $_{nin}W$ for each farm is identical. A systematic sample of farms is chosen to receive a final integer weight of 2. These farms are chosen with probability equal to the fractional portion of the $_{nin}W$ for the FNRS in which they reside.

6. VARIANCE ESTIMATION

The variance estimation method to be used for the 1992 Census of Agriculture is a design-based variance estimator that may be best-classified as a Taylor series approximation (Taylor series only in the sense that the point estimator for which we obtained the variance estimator was assumed to be linear by assuming the sample weights to be constant).

Based on the description of the sample design and estimation procedures, we note that census of agriculture estimates are subject to three sources of sampling error: nonresponse, sample selection, and integer weighting. Our derivation of the variance formula assumes that the three sources of variability are independent.

Also, one needs to note that there is unignorable covariance induced by certain aspects of the design. There are two instances of covariance in our estimators; both are due to procedures used in adjusting for the number of nonrespondent farms. The synthetic estimation procedure used to obtain county-level estimates of the number of nonrespondent farms introduces a covariance among FSPS for county-level sample item estimates and among both FSPS and counties for state-level estimates.

The final sample item variance formulas to be used for 1992 Census of Agriculture estimated totals are given below, along with the notation necessary to understand the formulas. The first formula, (1a), is the county-level formula; the second formula, (1b), is the state-level formula.

$$\sum_{p_1=0}^{p'}\sum_{p_2=0}^{p'}\sum_{k=0}^{K}\left(\frac{{}_{nf}R_{ijp_1k}\cdot{}_{nf}R_{ijp_2k}}{{}_{nf}R_{ijp_1}\cdot{}_{nf}R_{ijp_2}}\right)\left({}_{nf}T_{ijp_1}\cdot{}_{nf}T_{ijp_2}\right)\left({}_{nf}\bar{X}_{ijp_1k}\cdot{}_{nf}\bar{X}_{ijp_2k}\right)\frac{\hat{V}(\hat{N}_{ijk})}{\left({}_{nf}T_{ijk}\right)^2}$$

$$+\sum_{p=0}^{p'}\sum_{k=0}^{K}{}_{nf}R_{ijpk}\,S^2_{ijpk}\left(1-\frac{{}_{nf}R_{ijpk}}{{}_{nf}T_{ijpk}}\right)\left(\frac{{}_{nf}T_{ijp}}{{}_{nf}R_{ijp}}\right)^2\left(1+\frac{(\hat{N}_{ijk})^2+\hat{V}(\hat{N}_{ijk})}{\left({}_{nf}T_{ijk}\right)^2}+2\frac{\hat{N}_{ijk}}{{}_{nf}T_{ijk}}\right) \qquad (1a)$$

$$+\sum_{p=0}^{p'}\sum_{k=0}^{K}\left[\left({}_{nf}R_{ijpk}-1\right)S^2_{ijpk}+{}_{nf}R_{ijpk}\left({}_{nf}\bar{X}_{ijpk}\right)^2\right]\left(1-\frac{1}{{}_{nf}R_{ijpk}}\right)\left(F_{ijpk}-F^2_{ijpk}\right)$$

$$\sum_{j=1}^{j'}\sum_{p_1=0}^{p'}\sum_{p_2=0}^{p'}\sum_{k=0}^{K}\left(\frac{{}_{nf}R_{ijp_1k}\cdot{}_{nf}R_{ijp_2k}}{{}_{nf}R_{ijp_1}\cdot{}_{nf}R_{ijp_2}}\right)({}_{nf}T_{ijp_1}\cdot{}_{nf}T_{ijp_2})\,({}_{nf}\bar{X}_{ijp_1k}\cdot{}_{nf}\bar{X}_{ijp_2k})\,\frac{\hat{V}(\hat{N}_{ijk})}{({}_{nf}T_{ijk})^2}$$

$$+\sum_{j=1}^{j'}\sum_{p=0}^{p'}\sum_{k=0}^{K}{}_{nf}R_{ijpk}\,S^2_{ijpk}\left(1-\frac{{}_{nf}R_{ijpk}}{{}_{nf}T_{ijpk}}\right)\left(\frac{{}_{nf}T_{ijp}}{{}_{nf}R_{ijp}}\right)^2\left(1+\frac{(\hat{N}_{ijk})^2+\hat{V}(\hat{N}_{ijk})}{({}_{nf}T_{ijk})^2}+2\frac{\hat{N}_{ijk}}{{}_{nf}T_{ijk}}\right)$$

$$+\sum_{j=1}^{j'}\sum_{p=0}^{p'}\sum_{k=0}^{K}\left[({}_{nf}R_{ijpk}-1)\,S^2_{ijpk}+{}_{nf}R_{ijpk}({}_{nf}\bar{X}_{ijpk})^2\right]\left(1-\frac{1}{{}_{nf}R_{ijpk}}\right)(F_{ijpk}-F^2_{ijpk}) \qquad \text{(1b)}$$

$$+\sum_{j_1=1}^{j'}\sum_{j_2\neq j_1}^{j'}\sum_{p_1=0}^{p'}\sum_{p_2=0}^{p'}\sum_{k=0}^{K}\left(\frac{{}_{nf}R_{ij_1p_1k}\cdot{}_{nf}R_{ij_2p_2k}}{{}_{nf}R_{ij_1p_1}\cdot{}_{nf}R_{ij_2p_2}}\right)({}_{nf}T_{ij_1p_1}\cdot{}_{nf}T_{ij_2p_2})\,({}_{nf}\bar{X}_{ij_1p_1k}\cdot{}_{nf}\bar{X}_{ij_2p_2k})$$

$$\cdot\left(\frac{\hat{se}(\hat{N}_{ij_1k})\,\hat{se}(\hat{N}_{ij_2k})}{{}_{nf}T_{ij_1k}\cdot{}_{nf}T_{ij_2k}}\right)$$

where

p' = the number of FSPS,

K = the number of FNRS,

j' = the number of counties in the state,

${}_{nf}R_{ijpk}$ = the number of sample respondent farms not receiving 100% follow-up in state i, county j, FSPS p, FNRS k,

${}_{nf}T_{ijpk}$ = the number of census respondent farms not receiving 100% follow-up in state i, county j, FSPS p, FNRS k,

${}_{nf}\bar{X}_{ijpk}$ = the sample mean of the sample respondent farms for item X in state i, county j, FSPS p, FNRS k,

S^2_{ijpk} = the sample variance of the sample respondent farms for item X in state i, county j, FSPS p, FNRS k,

F_{ijpk} = the fractional portion of the final real sample weight in state i, county j, FSPS p, FNRS k.

In these formulas each of the first three terms corresponds to one of the sources of error. The first term in both equations (1a) and (1b) corresponds to the variability in the estimator induced by the nonresponse survey estimation of the number of nonrespondent farms. Note that this term also contains the FSPS cross products of the variables; this is due to the covariance of the estimator among FSPS.

The second and third terms of equations (1a) and (1b) correspond to the variability in the estimator induced by the sample selection and integer weighting procedures, respectively.

The fourth term of equation (1b) corresponds to the covariance of estimators across counties due to the synthetic estimation of the number of nonrespondent farms at the county level.

These formulas required a good deal of time and determination to develop. A rather natural question to ask then is, to what degree could we have simplified the derivations yet obtained results nearly the same as those given by equations (1a) and (1b)? To study this question, we conducted an empirical study.

7. EMPIRICAL STUDY AND ANALYSIS

The empirical study was designed to examine the effects of certain assumptions on the variance estimation procedures. The three assumptions examined were (1) the equality of response rates for the sample and nonsample forms; (2) the independence of synthetic estimators among (a) geographical areas and (b) sampling strata; and (3) the treatment of nonresponse weighting factors as constants. Data from the 1987 Census of Agriculture were used to conduct the empirical study.

This study was intended as a precursor to a perhaps more theoretical treatment of the issue, with the hope being that it would give us an idea of the assumptions which may or may not be made in the derivation of our variance formulas. It is however limited in that the properties of our variance estimators are not well-known; it is for this reason that we plan a more theoretical study. Thus, in this paper, we must assume that

equations (1a) and (1b) are "good" variance estimators.

The first order of duty for the empirical study was to derive variance formulas using one or more of the assumptions. This was done so that variance results for these formulas could be compared to results for a formula which did not make use of any of these assumptions. The formulas are denoted as follows: F_0 is the 1992 Census of Agriculture formula which was derived using none of the assumptions (this is equation (1a) at the county level and equation (1b) at the state level); F_1 is the formula which used only assumption (1); F_2a is the formula which used only assumption (2a); F_2b is the formula which used only assumption (2b); and F_3 is the formula which used only assumption (3). Also, formulas such as F_13, which used both assumptions (1) and (3), were examined.

Next, variances were calculated for a set of items using each of these formulas. These variances were calculated for this set of items for the states of Iowa and Maine at both the county and state levels.

Assuming that F_0 is the most "correct" of the formulas, we will evaluate the harm each of the assumptions does to the variance estimates. The premise is that if a formula which contains one or more of the assumptions yields estimates nearly the same as F_0, then the assumptions made in deriving that formula had little adverse effect on our variance estimates. It is also assumed that use of these assumptions will result in some greater benefit, such as ease of development.

Table 1 below provides an example of the results obtained. In this table are the variance estimates for the six formulas for seven items in the state of Maine. Both county- and state-level variances are given. Only variance estimates for two counties and the state are given in order that space may be saved; however, these results are typical of the results overall.

Looking at the data from the two counties, we first observe that variances calculated using formulas F_13 and F_3 tend to be quite different from those calculated using F_0 for many items. This indicates that making the assumption of constant nonresponse weights is detrimental to the precision of the variance estimates. Results for formulas F_1 and F_2b differ somewhat from those for F_0 for some items, but probably not enough to be alarmed by. F_2a is exactly equal to F_0 for all items in the two counties; this is to be expected since the two formulas differ only at the state level.

At the state level we see that F_2b, while not wildly different from F_0, tends to be different enough to preclude its use as a good approximation to F_0. This shows the imprudence of ignoring the covariance among FSPS. Also at the state-level the variances calculated using F_2a become quite different from F_0, indicating that it is unwise to exclude the covariance among counties from the variance formulas.

Formula F_1 seems to be the closest approximation to F_0; hence, it may be possible to assume the equality of response rates for the sample and nonsample forms without adversely affecting the variance calculations.

8. CONCLUSIONS

All we'll venture to say is that it is best to make as few assumptions as possible. In our case, we assume the variance estimator that uses the fewest assumptions to be the most accurate of the variance estimators in the empirical study. If, however, we had not the time to dispense with all the assumptions, would it have been possible to derive a variance estimator that was nearly as accurate as the one we will employ for 1992? It appears the answer here is "yes." Making the assumption of equality of response rates for the sample and nonsample census forms would seem to have minimal effect on the final results. It is, however, probably unwise to make any of the other assumptions if we wish a variance estimator with a reasonable amount of accuracy.

Table 1

Item	F_0	F_1	F_2a	F_2b	F_3	F_13
County A						
1A	27.41	26.05	27.41	10.55	303.58	24.57
1B	$1.43x10^{11}$	$1.40x10^{11}$	$1.43x10^{11}$	$1.41x10^{11}$	$1.81x10^{11}$	$1.38x10^{11}$
2A	371.43	371.12	371.43	369.71	410.66	407.96
2B	$1.20x10^{9}$	$1.20x10^{9}$	$1.20x10^{9}$	$1.20x10^{9}$	$1.22x10^{9}$	$1.19x10^{9}$
3A	405.28	404.01	405.28	402.11	453.86	423.82
3B	$3.63x10^{10}$	$3.59x10^{10}$	$3.63x10^{10}$	$3.63x10^{10}$	$4.08x10^{10}$	$3.40x10^{10}$
4	$7.45x10^{7}$	$7.45x10^{7}$	$7.45x10^{7}$	$7.37x10^{7}$	$8.94x10^{7}$	$8.47x10^{7}$
County B						
1A	24.84	19.72	24.84	3.51	140.19	2.81
1B	$2.67x10^{11}$	$2.55x10^{11}$	$2.67x10^{11}$	$2.52x10^{11}$	$2.04x10^{11}$	$2.00x10^{11}$
2A	275.24	273.99	275.24	272.18	249.05	260.84
2B	$4.44x10^{8}$	$4.37x10^{8}$	$4.44x10^{8}$	$4.41x10^{8}$	$3.44x10^{8}$	$3.77x10^{8}$
3A	272.35	269.22	272.35	269.31	252.45	253.75
3B	$5.41x10^{9}$	$5.31x10^{9}$	$5.41x10^{9}$	$5.35x10^{9}$	$3.91x10^{9}$	$4.34x10^{9}$
4	$1.64x10^{7}$	$1.62x10^{7}$	$1.64x10^{7}$	$1.60x10^{7}$	$2.07x10^{7}$	$1.59x10^{7}$
State						
1A	3950.57	3665.75	497.27	2073.03	36,369.82	1950.06
1B	$4.57x10^{12}$	$4.26x10^{12}$	$3.72x10^{10}$	$3.90x10^{13}$	$1.12x10^{12}$	$4.00x10^{12}$
2A	5064.39	5027.35	4840.01	4931.40	7764.95	5247.05
2B	$6.50x10^{10}$	$6.47x10^{10}$	$6.44x10^{10}$	$6.45x10^{10}$	$7.72x10^{10}$	$6.06x10^{10}$
3A	6348.23	6152.02	5669.41	5852.24	11,879.72	6474.58
3B	$3.26x10^{11}$	$3.21x10^{11}$	$3.21x10^{11}$	$3.20x10^{11}$	$4.16x10^{11}$	$3.09x10^{11}$
4	$6.69x10^{8}$	$6.54x10^{8}$	$5.72x10^{8}$	$6.13x10^{8}$	$1.48x10^{9}$	$6.63x10^{8}$

Key

1A: total farm production expenses (farm count)
1B: total farm production expenses (dollars)
2A: livestock and poultry purchased (farm count)
2B: livestock and poultry purchased (dollars)
3A: hired farm labor (farm count)
3B: hired farm labor (dollars)
4: value of land and buildings (dollars)