

AN APPLICATION OF A TWO-PHASE RATIO ESTIMATOR AND THE DELETE-A-GROUP JACKKNIFE

John Amrhein USDA-NASS, Susan Hicks USDA-NASS, Phillip Kott USDA-NASS
John Amrhein, National Agricultural Statistics Service, 4818 South Bldg., Washington, D.C. 20250

KEY WORDS: Jackknife Variance, Ratio Estimator, Calibration Weights

Then we can define the rate of application for crop k, farm i as:

I. Introduction

The Vegetable Chemical Use Survey (VCUS) measures acreage receiving and amounts applied of pesticides to vegetable crops by State. The sample design for the 1996 VCUS consisted of two phases. The first-phase was a screening survey selected using stratified simple random sampling where some strata were completely enumerated. A second-phase sample was selected from the first-phase sample using a variation of probability proportional to size sampling. The second-phase sample was used to provide a detailed survey of chemical use practices by crop. This paper provides an evaluation of the estimators used for the VCUS and the model underlying them. It also contains a description of the delete-a-group jackknife variance estimator used in that evaluation and the finite population correction factors needed to adjust jackknife variance estimates in this context. Data for Michigan was employed in the evaluations.

II. A Model for Estimating Chemical Use by Crop

The design and sampling strategy for the second-phase sample are discussed in detail in Hicks et al. (1996). Following is a description of the model underlying the sample design and estimation. The VCUS estimates chemical use by crop. We assume that the total use of a particular chemical by farm i on crop k can be expressed with the following model:

$$y_{k,i} = a_{k,i} b_k + \epsilon_{k,i} a_{k,i} \quad (1a)$$

where:

- $y_{k,i}$ = chemical use for crop k, farm i
- $a_{k,i}$ = acres planted for crop k, farm i
- b_k = average chemical use per acre for crop k
(according to the model)
- $\epsilon_{k,i}$ = random error term for crop k, farm i
- $E(\epsilon_{k,i}) = 0$
- $E(\epsilon_{k,i}^2) = \sigma_{k,i}^2$
- $E(\epsilon_{k,i}\epsilon_{k,j}) = 0$ for $i \neq j$.

$$r_{k,i} = \frac{y_{k,i}}{a_{k,i}} = \mu_k + \epsilon_{k,i} \quad (1b)$$

Thus, the total rate of application for the population is:

$$R_k = \frac{\sum_{i \in P} y_{k,i}}{\sum_{i \in P} a_{k,i}} = \frac{\sum_{i \in P} a_{k,i} r_{k,i}}{\sum_{i \in P} a_{k,i}}$$

where P is the population. A design-consistent estimator of R_k , given a two-phase design is:

$$\hat{R}_k = \frac{\sum_{i \in S} f_i a_{k,i} r_{k,i} / \Pi_i}{\sum_{i \in S} f_i a_{k,i} / \Pi_i} = \frac{\sum_{i \in S} w_i^0 a_{k,i} r_{k,i}}{\sum_{i \in S} w_i^0 a_{k,i}} \quad (2)$$

where:

- S = the set of units in the second-phase sample
- f_i = the first phase weight for farm i
- π_i = the second-phase probability of selection for farm i
- $w_i^0 = f_i / \pi_i (n/u)$
- n = number of units selected
- u = number of usable units, or responding units

It can be shown from equations (1b) and (2) that the model variance of \hat{R}_k as an estimator of the superpopulation characteristic μ in equation (1b) is:

$$V_{model}(\hat{R}_k) = \frac{\sum_{i \in S} (w_i^0 a_{k,i})^2 \sigma_{k,i}^2}{(\sum_{i \in S} w_i^0 a_{k,i})^2} \quad (3)$$

However, we are more concerned with the variance of \hat{R}_k in equation (2) as an estimator of the finite population value of R_k . Assuming $\sigma_{k,i} = \sigma_k$ for all i, this is given approximately by:

$$\begin{aligned}
E_{model}[(\hat{R}_k - R_k)^2] & \approx \left[\frac{\sum_{i \in S} (a_{k,i} w_i^0)^2}{\left(\sum_{i \in S} a_{k,i} w_i^0\right)^2} - \frac{\sum_{i \in S} a_{k,i}^2 w_i^0}{\left(\sum_{i \in S} a_{k,i} w_i^0\right)^2} \right] \sigma_k^2 \\
& = V_{model}(\hat{R}_k) \left(1 - \frac{\sum_{i \in S} a_{k,i}^2 w_i^0}{\left(\sum_{i \in S} a_{k,i} w_i^0\right)^2} \right) \\
& = V_{model}(\hat{R}_k) \cdot FPC(\hat{R}_k) \quad (4)
\end{aligned}$$

The FPC is a finite population correction factor for the ratio R_k . Note that the FPC is a function of the denominator of the ratio in equation (2). For the ratio "rate of application," the denominator is total acres planted for crop k . For other ratios of chemical use the denominator can be, for example, total acres of crop k treated with a particular chemical. The FPC can vary with the type of ratio estimated and the chemical as well as the crop.

III. The Delete-a-Group Jackknife Variance Estimator

Kott (1997) shows how the delete-a-group jackknife variance estimator can be used with a variety of different estimators. This section describes how this variance estimator was used for the 1996 VCUS.

We first divide the first phase sample - both respondents and non-respondents - into 15 groups. The choice of 15 lengthens a traditional normality-based two-sided 95% confidence interval by about ten percent. To divide the first-phase sample into 15 groups, we randomly sort the units within strata and then select the first, sixteenth, thirty-first, ..., units for the first group, the second, seventeenth, thirty-second, ..., units for the second group and so on until all 15 groups are created.

Let S denote the final respondent sample used to compute \hat{R}_k . Let S_r denote that part of the final sample contained within originally sampled units in group r . The *jackknife replicate* $S_{(r)}$ is the whole final sample S with S_r removed. Similarly, $F_{(r)}$ is the whole first-phase sample F with F_r removed.

We create 15 sets of *replicate weights* $\{w_{i(r)}\}$, one for each r , such that $w_{i(r)} = 0$ for all units in S_r and $w_{i(r)} = f_{i(r)}/\pi_i$ otherwise, where $f_{i(r)} = f_i u_i / u_{i(r)}$, u_i is the number of usable units in the first-phase sample from the same first-phase stratum as i , and $u_{i(r)}$ is the number

of usable units in both the same first-phase stratum of i and $F_{(r)}$.

Observe that a $w_{i(r)}$ -value has been assigned to every element in S including those in S_r . Every unit then has 16 weights; a *full sample weight*, w_i , and 15 *replicate weights*, $w_{i(r)}$, for $r = 1, \dots, 15$. The delete-a-group jackknife variance estimator for \hat{R}_k is:

$$V_J = \frac{14}{15} \sum_{r=1}^{15} (\hat{R}_{k(r)} - \hat{R}_k)^2 \quad (5)$$

This is a good estimator for the model variance of \hat{R}_k in equation (3). It is also a good estimator of the randomization mean squared error of \hat{R}_k under certain conditions (see Kott 1997). Unfortunately, one of those conditions is that all the selection probabilities should be small which is often violated in the 1996 VCUS.

Equation (4) provides a sample-dependent FPC factor, $FPC(\hat{R}_k)$, for the model variance of \hat{R}_k as an estimator of R_k . V_J can be multiplied by this factor to produce a good estimator for the model variance of \hat{R}_k as an estimator of R_k . Observe that when all the weights are much greater than 1, $FPC(\hat{R}_k)$ will be approximately equal to 1 and the effect of multiplying V_J by this factor will be negligible.

How Important was the FPC Adjustment?

Table 1

Application per Planted Acre					
Crop	Active Ingredient	Estimate	FPC for Planted Acre	Jackknife CV (%)	Adjusted CV (%)
Asparagus	Carbaryl	0.88	0.52	11.3	8.2
Snap Beans	Acephate	0.79	0.30	17.4	9.6
Proc.	Linuron	0.84	0.28	34.7	18.2
Carrots	Chloro..	5.00	0.28	19.8	10.4
Celery	Atrazine	0.79	0.41	17.2	11.0
Sweet Corn	Chloro..	2.57	0.28	61.2	32.3
Fresh MKT	Ethal...	0.49	0.30	16.6	9.0
Cucumbers	Oxyflu...	0.13	0.28	26.1	13.7
Fresh MKT					
Cucumbers					
Proc.					
Onions, Bulb					

The jackknife CV is estimated as $\sqrt{V_J}/\hat{R}_k$, where V_J is estimated from equation (5). The adjusted CV is estimated as $\sqrt{V_J \cdot FPC(\hat{R}_k)}/\hat{R}_k$, where the FPC is estimated as in equation (4). As Table 1 shows, the FPC correction is significant for the VCUS. In fact, many crops were sampled completely in the second phase, especially if they were rare for that State. Without the FPC adjustment the jackknife variance estimator could over-estimate the CV by a factor 2.

IV. The Calibration Estimator

It is not necessary to calibrate the weights to produce unbiased estimates of chemical use. The original weights, w_i^0 , can be used. However, calibration provides a set of weights that reproduce a better estimate of total chemical used. In our case we calibrated the weights using the reported planting intentions from the screening survey so that they reproduce the first-phase estimated crop acreages. Note that this is not the same as calibrating the second-phase reported acreages to first-phase totals. Normally the responses from the first-phase will be the same as the responses from the second-phase. However, for the 1996 VCUS the screening survey was conducted before many of the crops were planted, thus acres planted from the screening survey represents planting intentions and not actual acres planted. It was not clear what effect this would have on the precision of the estimates of chemical use.

One of the properties of single-frame PPS sampling is that the conventional expansion estimator estimates the population total of the measure-of-size variable with zero variance. Let us define $w_i^0 = f_i/\pi_i$ as the original two-phase sampling weight for unit i in the absence of second-phase non-response. In our multiple frame design, $\sum w_i^0$ does not estimate any meaningful population total. Therefore, we use the least squares method to produce a single set of calibration weights, w_i^c , such that:

$$\sum_{i \in S} w_i^c \cdot a_{k,i} = \sum_{i \in F} f_i a_{k,i} \quad \forall k \quad (6)$$

where the right-hand-side is the estimate of total acres for crop k from the first-phase sample, based on planting intentions. The appeal for this approach is that each unit possesses only one weight for all the crops and the set of weights for all the units has the desirable property stated in (6). Equation (6) establishes a set of K calibration equations. The method we chose for computing the calibrated weights, w_i^c , that satisfy (6) is:

$$w_i^c = w_i^0 + \left(\sum_{j \in P} f_j \mathbf{a}_j - \sum_{j \in S} w_j^0 \mathbf{a}_j \right) \left(\sum_{j \in S} (w_j^0 z_j) \mathbf{a}_j' \mathbf{a}_j \right)^{-1} (w_i^0 z_i) \mathbf{a}_i \quad (7)$$

where \mathbf{a}_j is a row vector of crop acreages for unit j and z_j is initially set equal to 1. If any w_i^c falls outside an acceptable range (in our case, if any w_i^c is less than 1), we recompute all the calibration weights using equation (7) but with the "original" weight of any out-of-range unit i set to the range extreme (in our case, that means w_i^0 is set to 1) and the corresponding z_i set to 0. This

process (changing the original weights and z -values for out-of-range units and recomputing equation (7)) may need to be repeated more than once. It is even possible that no complete set of in-range calibration weights exists, but we have not faced that situation anywhere in this project.

We then create 15 sets of replicate calibrated weights $\{w_{i(r)}^c\}$ by first setting the initial weight $w_{i(r)}^0 = w_i^c$ and $w_{i(r)}^0 = 0$ for all units in S_r . We then calibrate the replicate weights using equation (7) with $F_{(r)}$ replacing F and $w_{i(r)}^0$ replacing w_i^0 .

Our estimator for the total rate of application, R_k is:

$$\hat{R}_k^c = \frac{\sum_{i \in S} w_i^c y_{k,i}}{\sum_{i \in S} w_i^c a_{k,i}} \quad (8)$$

where $a_{k,i}$ is the reported acreages from phase-two, not planting intentions as was used in the calibration, and S (here and in equation(7)) is the set of respondents in the second-phase sample. This estimator is model unbiased for estimating b_k in (1) and retains the design consistency of the estimator in (2) if w_i^c/w_i^0 is close to 1 for all i . The motivation for using this estimator under this design is discussed in Kott (1996). Also reference Brewer (1994) and Deville and Sarndal (1992) for discussions on calibration estimators.

Did Calibration Help Us or Hurt Us?

Table 2

Application Rate per Planted Acre					
Crop	Active Ingredient	Uncalibrated		Calibrated	
		Estimate	Jackknife CV (%)	Estimate	Jackknife CV (%)
Asparagus	Carbaryl	0.88	11.3	0.90	12.2
Snap Beans Proc.	Acephate	0.79	17.4	0.78	18.5
Carrots	Linuron	0.84	34.7	0.84	38.9
Celery	Chloro..	5.00	19.8	4.92	20.4
Sweet Corn Fresh MKT	Atrazine	0.79	17.2	0.73	35.7
Cucumbers Fresh MKT	Chloro..	2.57	61.2	2.24	99.6
Cucumbers Proc.	Ethal...	0.49	16.6	0.47	26.6
Onions, Bulb	Oxyflu...	0.13	26.1	0.13	29.7

The above Table shows results for the application rate per planted acre for the most common chemical applied by crop in Michigan. We focus on the unadjusted jackknife CVs because they are more robust to model failure than the adjusted CVs (all the $\epsilon_{k,i}$ for a given k need not have the same variance, and every second-phase sample farm need not be equally likely to respond). Moreover, since the target of both a

calibrated and an uncalibrated estimator is the same, the difference between their true unadjusted variance should roughly equal the difference between their true adjusted variances.

Looking at the Table 2, it appears that calibration decreases precision for many of the crop/chemical combinations, sometimes by a fairly large margin. This result, confirmed by other combinations not displayed, is not completely surprising since the main reason for calibrating the weights was to increase the accuracy of estimated chemical use totals not rates.

V. How Good is the Model?

The estimation strategy outlined in Section II was applied to NASS' 1996 VCUS. From the VCUS, NASS estimates five attributes, of which two are discussed here. These are 1) application per treatment acres (i.e., acres treated with the chemical in question counting multiple treatments multiple times), and 2) application per planted acre. These estimates are produced for each vegetable/chemical combination. Note that for the model expressed in equation (1), the dependent variable is pounds of chemical applied for each estimate, but the independent variable, $a_{k,i}$, can be either treatment acres or planted acres depending on what is being estimated.

Figures 1 and 2 display scatter plots of survey data for two case studies in the state of Michigan. A weighted regression was performed on all vegetable/chemical combinations with at least 30 observations. The two plots shown in the figures represent the "best" and the "worst" cases in terms of R^2 . These figures suggest that the fit of our model depends on the attribute of interest. The plots in Figure 1 show that the model appears to fit much better for rates per treatment acre since most producers merely follow the suggested label rate, which keeps the rate constant across all farms. It appears from the plot for Linuron that two application rates are present, possibly due to two chemical products containing Linuron having different suggested label rates.

By contrast, since not all acres are treated during any given treatment, the rate per planted acre varies, as shown in Figure 2 and as indicated by the reduction in the R^2 's. (Figure 1 shows more acres treated than Figure 2 shows planted because Figure 1 includes multiple treatments for the same acreage).

The fit of the model does not affect the (near) unbiasedness of our estimator in either the model or a randomization-based sense. Fit does, however, have a profound effect on the precision of an estimator. Table 3 shows that the higher the R^2 , the more precise the estimate as measured by the delete-a-group jackknife.

Figure 1 Application per Treatment

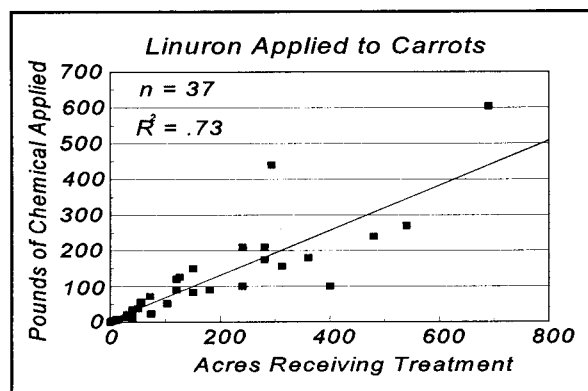
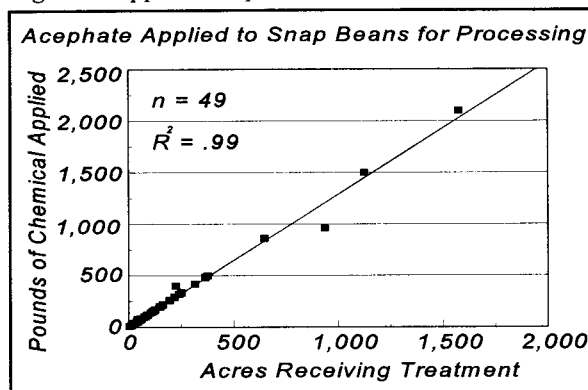


Figure 2 Application per Planted Acre

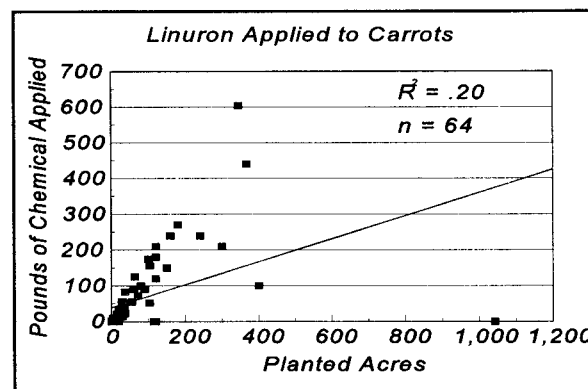
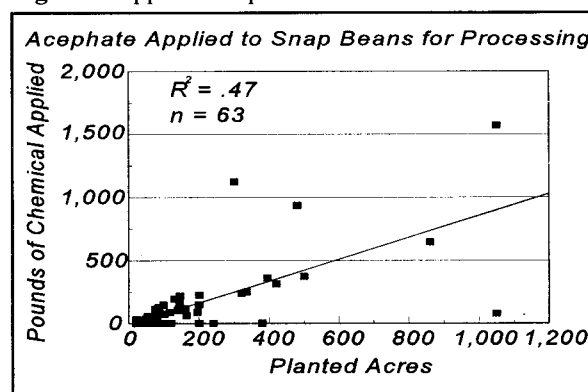


Table 3

	Jackknife CV (%)	R squared
Application rate per treatment acre:		
Acephate on snap beans:	3.0	0.99
Linuron on carrots:	7.9	0.73
Application rate per planted acre:		
Acephate on snap beans:	17.4	0.47
Linuron on carrots:	34.7	0.20

The jackknife CVs in Table 3 above are based on uncalibrated weights, unadjusted by the FPC. We used the unadjusted CVs because the fit of the model is strongly related to the variability of the estimation for the model parameter μ .

VI. Some Concluding Remarks

Although we used a model in designing the sample, it is important to remember that both the uncalibrated and calibrated estimators discussed in the text are nearly randomization unbiased in the absence of nonresponse (some modeling is needed to adjust for nonresponse). Moreover, the delete-a-group jackknife also has good randomization properties when all the sampling weights are large. Unfortunately, they are not, and we were forced to invoke a model to adjust CV estimates accordingly.

Our plans for the 1996 VCUS were to employ a two-phase design with sampling weights calibrated to first-phase acreage estimates. Variances would be estimated using a delete-a-group jackknife. We described this estimation strategy in the text.

We ran into two significant problems: 1) calibrating the weights increased the CVs of estimated ratios of particular interest to NASS; and 2) CV estimates from the delete-a-group jackknife were often much higher than the real CVs being estimated due to the lack of finite population correction in the jackknife.

We knew before we started that there was potential for calibrated ratio estimates to be less precise than uncalibrated ones. We had hoped, however, that these differences would either be small or to the advantage of the calibrated estimates. Unfortunately, that was often not the case.

Direct estimates of total chemical use, the target area where calibration is likely to do the most good, is of secondary interest to NASS. In the future, therefore, NASS will not calibrate VCUS weights.

We also knew that the delete-a-group jackknife would produce over-estimates of variance. We had hoped that the resulting conservative CV estimates would satisfy NASS's needs (they did provide enough information for our decision to drop calibration in the

future). NASS, however, determined that it required a more accurate measure of CV. We discussed one approach to finite population correction in the text. We are presently working on another one that requires fewer model assumptions and also has a randomization-based interpretation (see Kott 1997).

Acknowledgments

The authors wish to thank Susan Cowles for providing the survey data, estimates and documentation necessary to conduct these analyses.

References

- Brewer, K.R.W. (1994), "Survey Sampling Inference: Some Past Perspectives and Present Prospects", *Pakistani Journal of Statistics*, 10(1)A, 213-233.
- Deville, J-C. and C-E. Sarndal (1992), "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, 87, 376-382.
- Hicks, S. D., J. F. Amrhein and P.S. Kott (1996), "Methods to Meet Target Sample Sizes Under a Multivariate PPS Sampling Strategy", *ASA Proceedings of the Section on Survey Research Methods*.
- Hicks, S. D., J. F. Amrhein (1997), "An Evaluation of Different Estimators Used for the 1996 Vegetable Chemical Use Survey," forthcoming National Agricultural Statistics Service Staff Report.
- Kott, P.S. (1996), "Calibration Estimators Based on Several Separate Stratifications", *ASA Proceedings of the Section on Survey Research Methods*.
- Kott, P.S. (1997), "Using the Delete-A-Group Variance Estimator in NASS Surveys," forthcoming National Agricultural Statistics Service Staff Report.