METHODS TO MEET TARGET SAMPLE SIZES UNDER A MULTIVARIATE PPS SAMPLING STRATEGY

Susan Hicks USDA-NASS, John Amrhein USDA-NASS, Phil Kott USDA-NASS
Susan Hicks, National Agricultural Statistics Service, 3251 Old Lee Hwy, Fairfax, VA 22030

KEY WORDS: Target effective sample size, target realized sample size

I. Introduction

The National Agricultural Statistics Service's (NASS) 1996 Vegetable Chemical Use Survey (VCUS) will be sampled in two phases. The first-phase of the VCUS is a screening survey designed to collect data on total acreage of crops for as many as 25 agricultural commodities using stratified simple random sampling. Based on the results from the first-phase screening sample a second-phase sample is drawn from the first-phase sample for a detailed survey of chemical use practices by commodity. The traditional approach at NASS has been to employ a prioritized stratification scheme such that population units with the rarest items are grouped first into a stratum, then units with the next rarest items are grouped into the next stratum and so on. This technique for defining a single stratification scheme to meet the needs of multiple commodities works well enough when the number of commodities is small but becomes increasingly less efficient as the number of commodities increases. Because of the large number of commodities sampled for the VCUS, we needed an alternative sampling procedure which would allow us meet target sample sizes by commodity, provide adequate coverage of total farm acres by commodity, and minimize the total number of contacts.

Recently, Bankier (1986), Skinner (1991), and Skinner et al (1994) showed how an old method of combining independently drawn stratified random samples could be made more efficient than previously thought. Along those lines we explore a method for combining probability proportional to size sampling across multiple commodities in the second phase of sampling. This paper describes the proposed sampling scheme and explores a method for estimating the expected effective sample sizes under this design. Historical data from the VCUS is used to evaluate the efficiency of the sampling strategy at meeting target realized and effective sample sizes.

II. A Model for Estimating Chemical Use by Crop

The VCUS estimates chemical use by crop. We assume that chemical use per acre for crop k is roughly constant for all farms. Formally, we assume the total chemical use for farm i, crop k can be expressed as:

$$y_{k,i} = a_{k,i}\, b_k + \epsilon_{k,i}\, a_{k,i}$$

where:

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

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

$$r_{k,i} = \frac{y_{k,i}}{a_{k,i}} = \mu_k + \epsilon_{k,i}$$

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:

$$\hat{r}_k = \frac{\sum_{i\in S} f_i\, a_{k,i}\, r_{k,i} / \Pi_i}{\sum_{i\in S} f_i\, a_{k,i} / \Pi_i} = \frac{\sum_{i\in S} p_{k,i}\, r_{k,i} / \Pi_i}{\sum_{i\in S} p_{k,i} / \Pi_i}$$

where:

$$p_{k,i} = \frac{f_i \; a_{k,i}}{\sum_{i \in F} f_i \; a_{k,i}}$$

and,

F = the set of units in the first-phase frame
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

The model variance of this estimator can be expressed as:

$$E_{\epsilon}(\hat{r}_k - R_k)^2]$$

$$= E_{\epsilon}\left[\left(\frac{\sum_{i \in S} p_{k,i} \; \epsilon_{k,i} \; / \; \Pi_i}{\sum_{i \in S} p_{k,i} \; / \; \Pi_i} - \frac{\sum_{i \in P} a_{k,i} \; \epsilon_{k,i}}{\sum_{i \in P} a_{k,i}} \right)^2 \right]$$

$$= \sigma_k^2 \left[\frac{\sum_{i \in S} p_{k,i}^2 \; / \; \Pi_i}{(\sum_{i \in S} p_{k,i} \; / \; \Pi_i)^2} - 2\frac{\sum_{i \in S} p_{k,i} a_{k,i} \; / \; \Pi_i}{(\sum_{i \in S} p_{k,i} \; / \; \Pi_i)(\sum_{i \in P} a_{k,i})} \right.$$

$$\left. + \frac{\sum_{i \in P} a_{k,i}^2}{(\sum_{i \in P} a_{k,i})^2} \right]$$

The effective sample size of an estimation strategy is the population variance of the variable being estimated divided by the design variance of the estimator. The effective sample size for our design is:

$$n_k^* = \frac{\overset{2}{\sigma^{k}}}{E_{\epsilon}(\hat{r}_k - R_k)^2}$$

which can be estimated by:

$$\hat{n}_k^* =$$

$$\frac{1}{\dfrac{\sum_{i \in S} p_{k,i}^2 / \Pi_i^2}{\sum_{i \in S} (p_{k,i} / \Pi_i)^2} - 2 \; \dfrac{\sum_{i \in S} p_{k,i} \; a_{k,i} / \Pi_i}{(\sum_{i \in S} p_{k,i} / \Pi_i)(\sum_{i \in F} a_{k,i} \; f_i)} + \dfrac{\sum_{i \in F} a_{k,i}^2 \; f_i}{(\sum_{i \in F} a_{k,i} \; f_i)^2}}$$

With some work, we can see that the expected effective sample size before the second phase sample is selected is approximately:

$$E_2(\hat{n}_k^*) \approx \frac{1}{\sum_{i \in F} p_{k,i}^2 \left(\frac{1}{\Pi_i} - \frac{1}{f_i} \right)} \qquad (1)$$

Observe that as the sampling fraction approaches 1 for both the first-phase and the second-phase samples, the expected effective sample size approaches infinity and the variance approaches zero for crop k.

The first term in the denominator adjusts the effective sample size for inefficiencies in the design relative to the optimum design. The second term accounts for the effect of sampling in the first phase. Observe that, ignoring the possibility of certainty selections, the variance for any crop k will be minimized when Π_i is proportional to the size variable $p_{k,i}$. If we were to conduct separate surveys by crop, then Π_i would ideally be equal to $n_k p_{k,i}$(single variable PPS sampling) and the expected effective sample size would be equal to n_k -- the target sample size for crop k.

Thus, the expected effective sample size provides us with a measure of how well we have chosen the second-phase probabilities of selection with respect to crop k. We can also compute the expected realized sample size for crop k as:

$$E_2(\hat{n}_k^r) = \sum_{i \in F} I_{k,i} \cdot \Pi_i \qquad (2)$$

where $I_{k,i} = 1$ if unit i has the crop k and 0 otherwise.

III. Description of the Multivariate PPS Sampling Strategy

Under single variable PPS, the optimal probability of selection for crop k (ignoring the possibility of certainties) would be $n_k p_{k,i}$, where n_k is the target sample size for crop k. For the multivariable case, we wish to define a single probability, Π_i, for each unit that will satisfy the sample requirements for all crops (to be discussed later). Following is an algorithm that achieves that goal:

1. From the results of the first-phase screening, determine the number of positive reports available for sample in the second-phase for each crop.

2. Based on survey constraints and number available for sampling, set a target sample size for each crop.

3. Calculate a single draw selection probability for each commodity as follows:

$$p_{k,i} = \frac{f_i \, . \, a_{k,i}}{\sum_{i \in F} (f_i \cdot a_{k,i})}$$

where:

f_i = first-phase weight for farm i
$a_{k,i}$ = first-phase reported acres of crop k for farm i
$f_i \, a_{k,i}$ = first-phase expanded acreage estimate for crop k, farm i.

4. Calculate the maximum probability across all crops for each farm i:

$$MaxProb_i = MAX(P_{1,i}, \; P_{2,i}, \; \dots \; P_{k,i})$$

5. Calculate a single probability of selection for each farm as:

$$\Pi_i = MIN(1 \, , \; m \cdot MaxProb_i)$$

where the probabilities are truncated to one to ensure that no farm can have a selection probability greater than one and m is chosen to insure that the target sample sizes are met for all crops. (See Section IV).

6. Identify additional certainty farms. If the targets are realized sample sizes and the target sample size is equal to the number available for any crop, then designate all farms with that crop as certainties.

7. Generate descriptive statistics by crop:

i. Expected effective sample size
ii. Expected realized sample size
iii. Number of certainties
iv. Percentage of crop covered by certainties

8. Modify targets and rerun steps 2-7 until expected counts are acceptable.

9. Order the first phase sample by presense and absense of the rarest crop first, then the next rarest crop, etc....

10. Randomly order the population units within each sort variable.

11. Generate a random start, RS, between 0 and 1. Calculate the cumulative probability for unit i as:

$$Cum_i = Cum_{i-1} + \Pi_i$$

12. Unit i is selected for the sample if:

$$Cum_{i-1} < RS + j \leq Cum_i \quad for \; any \; j = 0,1,2,....n$$

where n is the total number of units selected for the sample.

IV. Methods for Choosing m to Meet the Specified Targets

Choosing the best value for m depends on whether the targets are effective sample sizes or realized sample sizes. Although the effective sample size provides a better measure of the efficiency of our design relative to the optimum design for crop k, the realized sample size may be a more practical target for some surveys.

Choosing m to meet the expected effective sample size targets:

In our model for estimating chemical use we derived an estimate of the expected effective sample size, equation (1). If we ignore the effect of certainties on the final sample, we can approximate the level of m required to meet the target effective sample sizes by defining:

$$\Pi_i' = m \cdot MaxProb_i$$

Substituting Π_i' for Π_i in equation (1) we can solve for m and obtain:

$$m \geq \frac{n_k \sum_{i \in F} p_{k,i}^2 \; / \; MaxProb_i}{1 + n_k \sum_{i \in F} p_{k,i}^2 \; / \; f_i} \qquad for \; all \; k$$

We calculate the level of m needed to meet the expected effective sample size targets for each crop and take the maximum across all crops. Of course, when second phase certainties are factored in the expected effective sample size can fall short of the targets for some crops.

Choosing m to meet the expected realized sample size targets:

For equation (1) to hold we must assume that the frame, F, is complete for each crop. The VCUS samples from a list frame that is known to be incomplete for some crops. Thus, the survey designers prefer to sample rare crops with certainty and all other crops at a rate necessary to obtain approximately 100 positive reports.

If the goal is simply to obtain n_k positive reports for each crop k, then the optimum m is simply:

$$m \geq \frac{n_k}{\sum_{i \in F} I_{k,i} \cdot MaxProb_i} \qquad for\ all\ k$$

where $I_{k,i}$ is 1 if unit i has crop k and 0 otherwise. The final procedure we chose attempted to meet both constraints.

V. Results

For the 1996 VCUS, we wrote a program in SAS which provides the sampling statistician with interactive screens for setting target sample sizes, reviewing expected effective and realized sample sizes and rerunning the program with new targets if necessary. The first module of the program calculates the required Π_i using the steps in Section III. Output from that module using historical data from 1994 for Michigan is included Figure 1.

Figure 1

Expected Sample Sizes for State =MI

Crop	Avail	Target	Prob 1's	Expected Realized	 Effective	Phase 1 Acres	Percent Acres in Prob 1s
ASPARGUS	241	100	50	112	193	10874.5	42.8%
BPEPPER	224	100	102	135	813	1826.6	81.4%
CABBAGE	160	100	100	119	2023	2064.2	88.6%
CANTALOU	177	100	80	106	184	457	74.6%
CARROTS	99	99	99	99	INF	6574.5	100.0%
CAULIFLO	78	78	78	78	INF	352.5	100.0%
CELERY	35	35	35	35	INF	2077.3	100.0%
CUKES	294	100	118	176	473	21065.7	61.4%
DRYONION	104	100	69	77	2565	4704.2	92.3%
SNPBNS	233	100	95	133	648	15272.1	70.0%
STRAWBRY	220	100	50	98	296	1461.5	44.3%
SWCORN	354	100	122	166	409	9901.4	68.7%
TOMATOS	300	100	120	160	869	4863.5	84.8%

The expected total number of farms to be contacted is 548

Note that the expected effective sample sizes are greater than the realized sample sizes for all crops. This is due to the small first-phase sampling fractions, f_i. Also note that for many crops a large percentage of the sampled acres are coming from certainty farms.

If the expected sample sizes and other descriptive statistics are acceptable to the sampling statistician, then the sample units are selected using systematic PPS sampling (steps 9-12). Sample output using the same data is included below:

Figure 2

Actual Sample Sizes for State =MI

Crop	Avail	Target	Prob 1s	Number Selected	Phase 1 Acres	Percent Acres in Phase 2
ASPARGUS	241	100	50	114	10874.5	75.8%
BPEPPER	224	100	102	134	1826.6	91.3%
CABBAGE	160	100	100	119	2064.2	96.8%
CANTALOU	177	100	80	104	457	85.8%
CARROTS	99	99	99	99	6574.5	100.0%
CAULIFLO	78	78	78	78	352.5	100.0%
CELERY	35	35	35	35	2077.3	100.0%
CUKES	294	100	118	172	21065.7	84.3%
DRYONION	104	100	69	78	4704.2	96.0%
SNPBNS	233	100	95	132	15272.1	87.6%
STRAWBRY	220	100	50	99	1461.5	72.1%
SWCORN	354	100	122	167	9901.4	83.3%
TOMATOS	300	100	120	159	4863.5	92.9%

The total number of farms to be contacted is 548

The actual sample that was selected in 1994 for Michigan using the old prioritized sampling procedure consisted of 742 total contacts. Thus, with the new procedure we were able to obtain adequate representation by crop with fewer contacts.

VI. Evaluation and Future Work

The current project focused on providing a procedure for selecting the phase II sample for the VCUS that would meet the target sample sizes by crop, ensure adequate coverage of crop acreage, and minimize the total number of contacts. We also provided more diagnostic statistics for the sampling statistician to evaluate the performance of the sample than has been provided in the past. Thus, we felt that the proposed sampling procedure would meet the goals of the survey.

Although the proposed sampling procedure focuses on determining the optimum value for m to compute the probability of selection, another option would be to search for the optimum probability of selection, Π_i, for each farm that would satisfy the target sample size requirements and minimize the total number of hits.

References

Bankier, M.D. (1986), "Estimators based on several stratified samples with applications to multiple frame surveys," *Journal of the American Statistical Association*, 81, 1074-1079.

Skinner, C. J. (1991), "On the efficiency of raking ratio estimation for multiple frame surveys," *Journal of the American Statistical Association*, 86, 779-784.

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