SSC Annual Meeting, June 1999
Proceedings of the Survey Methods Section

USING MULTI-PHASE SAMPLING TO LIMIT RESPONDENT BURDEN ACROSS AGRICULTURE SURVEYS

Phillip S. Kott and Matthew J. Fetter[1]

ABSTRACT

The National Agricultural Statistics Service is developing strategies that limit the amount of sample overlap across unrelated surveys by using multi-phase sampling principles. In its simplest form, Sample A is selected first, and then Sample B is chosen from among those members of the population not selected for Sample A. Effectively, Sample B is selected in two-phases. This two-phase approach extends easily to the coordination of more than two samples, although meeting accuracy and/or sample-size targets while maintaining strict sample exclusivity is not always possible. Variation of the basic approach address this problem, but lead to some theoretical difficulties. Sampling weights may be based on products of conditional selection probabilities rather than on unconditional selection probabilities. Randomization-based variance estimation likewise may depend on the product of conditional joint selection probabilities. In practice, variance estimates will be reasonable but may not always be randomization consistent.

KEY WORDS: Conditional selection probability; unconditional selection probability; Poisson permanent random number.

1. INTRODUCTION

We will describe a potential methodology for drawing a particular sample – the Agricultural Resource Management Study (ARMS) screening sample – in a manner that limits overlap with other surveys. Unlike the "Perry-Burt" technique (see Perry et al. 1994) currently used at the National Agricultural Statistics Service (NASS), this new methodology handles unstratified, unequal probability sampling designs properly. Such a design is already being used in the new Crops/Stocks Survey (CS) and may be used in drawing the ARMS screening sample in the future. Bailey and Kott (1997) provide a description of the CS sample design.

The basic concepts of the new methodology in an idealized environment will be outlined first. We will then show how it can be applied to a particular NASS application. Unfortunately, the basic concept as outlined can not always be applied in practice. Consequently, a few variations are proposed. A discussion follows.

2. THE BASIC CONCEPT

Suppose we want to draw R samples sequentially. Our principle interest is the sample design for the last survey (denoted R). For now, assume that the first R – 1 samples are drawn independently. Our goal is to limit the possibility that a farm selected for (at least) one of the previous R – 1 samples is selected again for Sample R. It is not necessary that the populations of interest be the same for all the samples. We do require, however, that we can identify whether or not a unit (farm) in the population of interest for one sample, say Sample r (= 1, ..., R), is also in the population of interest for another sample, say Sample s, s ≠ r.

The *conditional selection probability* of a farm for Sample r is the probability of selecting the farm for Sample r at the time of selection. We denote this probability by p_r (we suppress the subscript denoting the farm for convenience). By convention, a farm not eligible for selection in r has $p_r = 0$.

The *effective unconditional selection probability* of a farm for Sample r is its conditional selection probability times the probability that it is available for sampling (more on "availability" later). We denote this by π_r. It is this value we use in estimation, albeit often in adjusted form.

[1] Phillip S. Kott, Matthew J. Fetter, National Agricultural Statistics Service, Room 305, 3251 Old Lee Hwy, Fairfax, VA 22030, USA, pkott@nass.usda.gov.

For all Samples r < R, a farm's conditional and effective unconditional selection probabilities are equal; that is, $p_r = \pi_r$. Ideally for Sample R, we want to set $p_R = 0$ when the eligible farm has been selected in a previous sample. Such a farm is said to be *unavailable* for Sample R. Notice the distinction between the *eligibility* of a farm for Sample R – meaning that it meets the requirements for sampling – and its *availability* – meaning it has not been selected for one of the other samples.

The effective unconditional selection probability for a farm being in Sample R is

$$\pi_R = (1 - p_1)(1 - p_2) \cdots (1 - p_{R-1})p_R; \tag{1}$$

that is, the probability of the farm not being selected for Sample 1, not being selected for Sample 2, ..., not being selected for Sample R – 1, and then being selected for Sample R. Equation (1) also defines the *true* unconditional selection probability of the farm when p_R is set independently of the first R – 1 samples. In practice, this may not be the case.

Equation (1) tells us that, for a farm not selected in a previous survey, the Sample-R conditional selection probability is

$$p_R = \pi_R / [(1 - p_1) \cdots (1 - p_{R-1})] \tag{2}$$

as long as no $p_r = 1$ for r < R; otherwise, p_R would be undefined. This is an important but obvious restriction. If a farm was a certainty in a previous sample, there is no way to avoid the possibility it will also be in Sample R without violating randomization-based principles (observe that π_R in equation (1) would be zero, an unacceptable value, no matter to what value p_R is set). This means as a practical matter, *the "1 – p_r" term must be removed from "(1 – p_1) $\cdots$ (1 – p_{R-1})" in equations (1) and (2) when $p_r = 1$.* For ease of exposition, we will assume that all $p_r < 1$ from now on.

3. AN APPLICATION AND A SMALL EXTENSION

Suppose we want to co-ordinate the ARMS screening sample with R – 1 other independently drawn samples in the same survey year. If possible, we want no farm in one of the other samples to be drawn into the ARMS sample. As long as there are no certainties is any of the other samples, this is a simple matter. Calling the ARMS screening Sample R and drawing it last, we need only set the p_R and use equation (1) to determine the π_R.

Setting the p_R–values makes sense when our only concern is assuring that a required number of farms of various types get in the ARMS screening sample. If we are more concerned with the efficiency of the ARMS estimation strategies, we may want to target farm π_R–values. This can cause additional complications we will consider later.

It is a simple matter to extend this framework to include previous ARMS screening samples by allowing the farm's conditional and effective unconditional selection probabilities for Sample r < R to be unequal. Our focus remains the final ARMS screening sample (R). The new potential inequality of selection probabilities has no effect on equations (1) and (2), since each p_r, r < R, clearly denotes a conditional selection probability.

The CS is made up of three separate and dependent modules (two yield modules and a crops/stocks module). For our purpose, these can be looked at as three different samples (among the first R – 1) where, as with previous ARMS screening samples, the farm conditional and effective unconditional selection probabilities need not be equal.

4. TARGETING THE UNCONDITIONAL PROBABILITIES

As noted previously, one way to design the ARMS screening sample is to target the conditional selection probabilities among farms available for sampling (i.e., after removing selections for any of the other R – 1 samples). Unfortunately, this can lead to farms with very large sampling weights (a farm's unadjusted weight is $1/\pi_R$). An alternative approach is to target π_R – values and let equation (2) determine the conditional probabilities for the ARMS screening sample.

A problem with this approach is that it may be difficult to accurately target the number of farms with particular control characteristics that will fall into the ARMS screening sample. For example, suppose we want n farms in the sample to have positive cattle-control data. We could set the π_R-values for all *eligible* list-frame farms with positive cattle-control data so that the sum of these π_R-values equals n (recall that "eligible" means meeting the requirement for the ARMS screening sample as opposed to being available for sampling). This only assures that the *expected number* of farms in the sample with positive cattle-control data is n. The actual number may vary, which can be a cause of concern.

Systematic (unequal) probability sampling can be used to mitigate this concern. For example, suppose we separate the available sample into mutually exclusive groups along the lines of the present ARMS screening strata and choose a constant π_R-value for each group as we do now for each stratum. Drawing a systematic probability sample within each group using p_R-values derived from equation (2) should yield a number of hits per group very close to its expected value.

In the future, NASS may select the ARMS screening sample in a manner similar to the CS; that is, draw dependent samples for a number of survey items by setting a farm's effective unconditional selection probability for an item equal to the item's target sample size times the item's frame-specific "measure of size" raised to the 3/4'th power. The combined sample often produces counts that exceed each item-specific sample-size target (see Bailey and Kott, 1997).

4.1 A "Generalized" Solution

A problem with targeting π_R-values and applying equation (2) is that there is no guarantee that the resulting conditional selection probability, p_R, will less than or equal to 1, a requirement for a probability. To assure that this requirement is satisfied, we can generalize the method of bit. Let us suppose that there is a target π_R for each eligible farm (available or not). When the resulting p_R is greater than 1, we allow the possibility that the farm is in both (some) r and R and drop "$1 - p_r$" from "$(1 - p_1)\cdots(1 - p_{R-1})$" in equation (2).

Dropping $1 - p_r$ may not be enough to assure p_R is no greater than 1. We may be forced to allow a farm into Sample R that is in two (or more) other samples.

4.2 A "Modified" Solution

Rules for potential sample overlap have to be determined before looking at the particular farms affected. We may also want to require p_R be less than 1/m since we need to control the probability that a farm will be in one of the next $m - 1$ ARMS screening samples.

We suspect that in practice some combination of setting the p_R and π_R-values will evolve from trial-and-error. For example, we may first set the Sample-R target effective unconditional probability for a farm at $\pi_R^{(t)}$ and then let

$$p_R = \min\{1/m,\ \pi_R^{(t)}/[(1 - p_1)\cdots(1 - p_{R-1})]\}.$$

Consequently, the effective unconditional selection probability of the farm would be modified to

$$\pi_R = \min\{(1/m)(1 - p_1)\cdots(1 - p_{R-1}),\ \pi_R^{(t)}\}.$$

This could, of course, defeat whatever purpose we had for setting the original target effective unconditional selection probabilities. Some work is definitely needed in this area.

Observe that even with an m as small as 3, the largest value π_R can take is (1/3)(2/3)(2/3) = 4/27, and that assumes the farm is not eligible for any other sample but the ARMS screening samples. This suggests that we may want to determine m on a farm-by-farm basis when using this approach with larger farms getting a small m and smaller farms a larger m.

4.3 Using Permanent Random Numbers

One way to keep the target π_R-values but limit the potential for sample overlap when *Samples 1 through R-1 are selected independently* is to do the following. Let $\pi_{R-1} = 1 - (1 - p_1)\cdots(1 - p_{R-1})$ be the probability that the farm is a selection in at least one previous (to R) sample (we are again, for simplicity, ignoring the possibility that the farm is a certainty selection in a previous sample). If the farm is eligible for Sample R, choose a uniform random number, ρ^*, from the unit interval, [0, 1). Assign the farm the permanent random number (PRN):

$\rho = \rho^{*}\pi_{R-1}$ if it has been selected in a previous sample,
$\rho = \pi_{R-1} + \rho^{*}(1 - \pi_{R-1})$ otherwise.

Observe that the probability density for each possible PRN in [0, 1) is the same, which is a requirement for PRN's.

The farm is selected for Sample R if it is not a selection in a previous sample, and its PRN is less than $\pi_{R-1} + \pi_R$. If the farm *is* in a previous sample, it is also selected for Sample R when ρ is less than $\pi_{R-1} + \pi_R - 1$.

This method of sample selection is a form of Poisson PRN sampling. A farm is chosen for R if

$\rho \in [\, \pi_{R-1}, \pi_{R-1} + \pi_R)$ when $\pi_{R-1} + \pi_R \le 1$, or
$\rho \in [\, \pi_{R-1}, 1) \cup [0, \pi_{R-1} + \pi_R - 1)$ otherwise.

Thus, the farm's selection probability is π_R. The sets $[\, \pi_{R-1}, \pi_{R-1} + \pi_R)$ and $[\, \pi_{R-1}, 1) \cup [0, \pi_{R-1} + \pi_R - 1)$ are called Poisson PRN *sampling ranges*.

With this sampling method, we do not explicit calculate conditional selection probabilities for the farms in Sample R. Nevertheless, note that the old requirement to assure that Sample R not overlap a previous sample, namely, that p_R in equation (2) be less than or equal to 1, is equivalent to the requirement $\pi_{R-1} + \pi_R \le 1$.

Observe that if R=2, the probability of the farm being in both samples when $\pi_2 + \pi_1 > 1$ is $\pi_2 + \pi_1 - 1$. Using the "generalized" method described in Section 4.1, it is $\pi_2\, \pi_1$, which is greater than $\pi_2 + \pi_1 - 1$ unless π_2 or π_1 equal 1. This is because $(1 - \pi_2)(1 - \pi_1) > 0$ implies $\pi_2\, \pi_1 > \pi_2 + \pi_1 - 1$.

5. DISCUSSION

The Poisson PRN method for choosing Sample R requires that the R−1 previous samples be independently drawn (effectively, the methods treats the union of Samples 1 through R−1 as if it were drawn using a Poisson PRN process). This is not always the case for the ARMS screening sample.

The problem of non-independent previous samples disappears when R=2. In fact, it is a simple matter to co-ordinate any number of samples by creating farm-specific Poisson PRN sampling ranges for the current sample that begin where the previous ranges end. Moreover, when some overlap across samples becomes unavoidable, we can order the samples in such a way that the probability of being in two particular samples is minimized (e.g., making them adjacent in the order of sample selection).

Extending the "generalized" and "modified" approaches of Sections 4.1 and 4.2 in a similar manner is more difficult. We need to keep track of a variety of conditional probabilities depending on which previous samples we allow to overlap the one currently being selected. These methods do have the advantage of being better able to meet sample-size targets because it is always possible to adjust the latest conditional selection probability as needed.

With the kinds of sampling designs we have discussed here, a calibration technique should be used to estimate an item-specific mean or total (when estimating ratios, by contrast, calibration often provides little of value). A reasonable estimator for the model variance and randomization mean squared error of a calibration estimator based on Sample R, say, is

$$v = \sum_{j \in S_R} (a_j e_j)^2 (1 - \pi_{jR}), \tag{3}$$

where a_j is the calibrated weight for farm j,
e_j is the item-specific residual for farm j, and
π_{jR} is the effective unconditional selection probability of farm j for Sample R.

We are now allowing the possibility that Sample R is co-ordinated with a number of previous samples and not just the union of all previous samples taken as a whole. The variance estimator v is missing terms of the form $(a_j e_j)(a_k e_k)(\pi_{jk} - \pi_j \pi_k)$, where π_{jk} is the product of joint conditional probabilities of selection. Unless the R samples are all Poisson (so $\pi_{jk} = \pi_j \pi_k$ when $j \ne k$), this omission renders v in equation (3) biased as an estimator for randomization mean squared error. It is doubtful, however, that the bias will be of practical importance.

NASS actually uses a jackknife to estimate variances, which is asymptotically equivalent in expectation to ignoring the finite population correction terms (the $(1 - \pi_{jR})$) in equation (3). Kott (1997) discusses the need for the model on which the calibration is based to include an intercept when Poisson sampling is used.

In practice, all three techniques, generalized, modified, and Poisson PRN, may be used in combination.

Allocation using the multivariate schemes described in Bailey and Kott (1997) is not an exact science, so it may not be imprudent to truncate effective unconditional selection probabilities liberally if not universally (i.e., often, but not always).

REFERENCES

Bailey, J. T. and P.S. Kott (1997). An application of multiple list frame sampling for multi-purpose surveys. *ASA Proceedings of the Section on Survey Research Methods*, 496-500.

Kott, P.S. (1998). *Using the Delete-A-Group Jackknife Variance Estimator in NASS Surveys.* National Agricultural Statistics Service Research Report RD-98-01.

Perry, C.R., Burt, J.C., and Iwig, W.C. (1994). “Redrawing the 1993 Farm Costs and Returns Survey list sample to reduce its overlap with three other 1993 surveys and the 1992 FCRS.” *ASA Proceedings of the Section on Survey Research Methods*, 632-637.