International Conference on Establishment Surveys, II
June 2000, Buffalo, New York

THE THEORY AND PRACTICE OF MAXIMAL BREWER SELECTION WITH POISSON PRN SAMPLING

Phillip S. Kott And Jeffrey T. Bailey, National Agricultural Statistics Service
Phillip S. Kott, NASS, Room 305, 3251 Old Lee Hwy, Fairfax, VA 22030, USA
pkott@nass.usda.gov

K.R.W. Brewer suggests that when estimating the total of a single item for which there is control (auxiliary) data, one employ a ratio or regression estimator and draw the sample using probabilities proportional to the control values raised to a power between 1/2 and 1. Brewer's sample selection scheme can be expanded to multiple targets by drawing overlapping Poisson samples for a number of items simultaneously using permanent random numbers (PRN's). We can call the result an example of "Maximal Brewer Selection" (MBS). This paper develops the theory behind MBS and the calibration estimator rendering it practical. It goes on to describe how this estimation strategy is being used at the National Agricultural Statistics Service.

Key Words: Model; Model-assisted; Randomization; Calibration; Delete-a-group jackknife

1. INTRODUCTION

K.R.W. Brewer's (1963) article in the *Australian Journal of Statistics* is one of the truly remarkable works in the survey sampling literature. It discusses a model-based approach to survey sampling theory, contrasts that approach with the conventional randomization paradigm, and shows how the two can be used in tandem. All this seven years before Royall (1970) set the survey world buzzing with prediction theory (another name for the model-based approach) and almost three decades before the publication of Särndal et al. (1992) made model-assisted survey sampling (which uses models and randomization in tandem) the new conventional wisdom.

This paper builds on one small result in Brewer's impressive opus and some of his work since then. Suppose we are interested in estimating a population (P) total, $T = \sum_P y_i$, with a random sample (S) of size n. We suspect that the y_i follow the model

$$y_i = \beta x_i + k\epsilon_i, \tag{1}$$

where $E(\epsilon_i | x_i) = E(\epsilon_i \epsilon_j | x_i, x_j) = 0$ $(i \neq j)$, and $Var(\epsilon_i | x_i) = \sigma_i^2$ is known for all i (but k need not be known).

Equation (1) is a useful model for many establishment surveys. Whether or not it is correct, the following estimator is nearly randomization unbiased for large n (and randomization consistent under a number of sampling designs),

$$t = (\textstyle\sum_P x_i)\sum_S (y_i/\pi_i)/\sum_S (x_i/\pi_i),$$

where π_i is the selection probability of unit i. Of course, in order for t to be practical, the population sum $\sum_P x_i$ must be known, and the individual x_i must also be known for all units in the *sample*. In what follows, we further require x_i to be known for all units in the *population*. Such an x is called a "control" variable for the *target* variable y.

Brewer showed that when $\pi_i \propto \sigma_i$ the randomization-expected model variance of t was (asymptotically) minimized for fixed sample size n. In this sense, $\pi_i = n\sigma_i / \sum_P \sigma_k$ – if less than or equal to 1 for all i – is the optimal selection scheme given sample size n and estimator t . Godambe (1955) has a similar result for randomization unbiased estimators.

It is sometimes assumed that the σ_i have the form x_i^g, where $0 \leq g \leq 1$. If that is the case, then when g = 1, the optimal selection scheme (i.e., randomization-expected model-variance minimizing) is probability proportional to size, $\pi_i = nx_i / \sum_P x_k$, and t collapses into the Horvitz-Thompson mean-of-ratios estimator ($n^{-1} \sum_S (y_i/\pi_i)$; this is Godambe's 1955 result). When g = 0, the optimal sampling scheme is self-weighting, $\pi_i = n/N$. For establishment surveys, however, g is usually between ½ and 1. Brewer has said (out loud, if not in print) that a sensible value for g in many surveys is 3/4.

Sadly, Brewer's suggestion that the unit selection probabilities be in proportion to some known control value, x_i, raised to the 3/4 power has not been implemented much in practice. One problem is that many real establishment surveys have multiple targets of interest with varying relevant control values. Recently, however, several survey organizations have come to use calibration estimators in place of traditional expansion and ratio estimators. This has allowed the National Agricultural Statistics Service (NASS) to begin implementing a multivariate version of Brewer's suggestion in its Crops/Stocks Survey (CS). Internally, NASS calls this procedure "multivariate probability proportional to size" sampling. A better name would be "Maximal Brewer Selection" (MBS). This method of sample selection has proven more flexible than the stratification approaches NASS has traditionally used (see Bosecker 1989).

Section 2 fills out the theory of Brewer selection when there is a single target and control. Section 3 describes a simple extension for multiple targets each with its own control variable. Briefly, a Brewer selection probability is assigned to each population unit for every target variable of interest. The largest of these for each unit is then used for the actual sample selection. Section 4 addresses a several questions that NASS needed to resolve before making MBS practical to use. Poisson PRN sampling allows the agency to focus on different combinations of target variables in different survey periods. Section 5 further describes NASS's experience with this new sampling scheme. Section 6 offers some comments including one that describes a method for co-ordinating samples to minimize overlap.

2. BREWER SELECTION

2.1. Some Theory

Suppose we have target values, y_i, which we believe (roughly) obey the model in equal (1). We will call $t_C = \sum_S a_i y_i$, based on a sample S with n members, a *calibration estimator* for T if the *calibration equation*

$$\sum_S a_i x_i = \sum_P x_i \qquad (2)$$

is satisfied, and each $a_i = \pi_i^{-1}[1 + O_P(1/\surd n)]$, where π_i is (again) the selection probability of unit i, and O_P refers to an asymptotic probability order *with respect to the randomization rather than the model* (see Isaki and Fuller 1982 for a development of asymptotics in a finite population context). This is a bit of a generalization of the definition of a calibration estimator in Deville and Särndal (1992).

One obvious choice for the a_i is $\pi_i^{-1}(\sum_P x_k / \sum_S [x_k/\pi_k])$. This renders t_C equal to t in Section 1. The choice satisfies the calibration equation, and the a_i are sufficiently close to the π_i^{-1} as long as the design and population are such that $(\sum_S [x_k/\pi_k] - \sum_P x_k)/\sum_P x_k$ is $O_P(1/\surd n)$.

The model variance of t_C as an estimator of T is

$$\begin{aligned} E_\epsilon[(t_C - T)^2] &= E_\epsilon[(\textstyle\sum_S a_i y_i - \sum_P y_i)^2] \\ &= E_\epsilon[(\textstyle\sum_S a_i \epsilon_i - \sum_P \epsilon_i)^2] \\ &= \textstyle\sum_S a_i^2 \sigma_i^2 - 2\sum_S a_i \sigma_i^2 + \sum_P \sigma_i^2. \end{aligned} \qquad (3)$$

Since each $a_i \approx 1/\pi_i$, $E_\epsilon[(t_C - T)^2] \approx \sum_S \sigma_i^2/\pi_i^2 - 2\sum_S \sigma_i^2/\pi_i + \sum_P \sigma_i^2$

Technically, the relative difference between the left and right hand sides of the above equation is $O_P(1/\surd n)$. For our purposes, this defines when the two sides of an equation are approximately equal.

The randomization expectation (denoted using the subscript "P") of the model variance of t_C is

$$E_P\{ E_\epsilon[(t_C - T)^2]\} \approx \sum_P \sigma_i^2/\pi_i - \sum_P \sigma_i^2. \qquad (4)$$

Under mild conditions, this is the same as the model expectation of the randomization means squared error of t_C. Isaki and Fuller called that last quantity the "anticipated variance" of t_C, presumably meaning “the anticipation under the model of the randomization mean squared error or variance” (randomization mse and variance are virtually identical under the designs Isaki and Fuller had in mind). We will use their term here, but keep in mind an alternative meaning for “anticipated variance:” the model variance anticipated before sampling.

If we restrict ourselves to a randomization estimator like t_C, a sensible policy is to choose selection probabilities so that the right hand side of equation (4) is minimized for a given sample size n. Since $n = \sum_P \pi_i$, it is a simple matter to set up a Langrangian equation, the solution to which is $\pi_i = n\sigma_i/\sum_P \sigma_k$. For this solution to be valid each π_i must be no greater than 1. We assume that to be the case for the time being.

The anticipated variance of t_C is (asymptotically) minimized by setting the unit probabilities of selection equal to $n\sigma_i/\sum_P \sigma_k$ *no matter which method it used to draw the sample.* In fact, the same minimum variance is obtained if the sample size itself is allowed to be random with an *expected value* equal to n. Poisson sampling is a simple example of a sampling scheme with a random sample size.

2.2. The Selection Scheme

Suppose we have a working assumption about the σ_i in equation (1). In particular, suppose σ_i is believed to be proportional to x_i^g for some g between 1/2 and 1. Let us reparameterize the model as

$$y_i = \beta(x_i + [\textstyle\sum_P x_k / \sum_P x_k^g] x_i^g \epsilon_i), \qquad (5)$$

where (again) $E(\epsilon_i \,|\, x_i) = E(\epsilon_i \epsilon_j \,|\, x_i, x_j) = 0$ $(i \neq j)$, and (now) $\text{Var}(\epsilon_i \,|\, x_i) = \sigma^2$. We have chosen this parameterization so that σ is invariant to changes in scale (units of measurement) of the y_i and x_i. Notice that when $g = 1$, σ^2 is the relative variance of y_i under the model. Thus, σ^2 for any g is in some sense a generalized relative variance for y_i.

Observe that σ_i^2 in equation (4) now equals $\beta^2 [\sum_P x_k / \sum_P x_k^g]^2 x_i^{2g} \sigma^2$. Since under the model $T \approx \beta \sum_P x_k$, the relative anticipated variance of t_C is

$$\frac{E_P\{\, E_\epsilon[(t_C - T)^2]\}}{E_\epsilon(T^2)} \approx \frac{\sum_P x_i^{2g}(\pi_i^{-1} - 1)}{(\sum_P x_i^g)^2}\, \sigma^2.$$

Similarly, the *asymptotic anticipated coefficient of variance* for t_C under the model in equation (5) can be defined as

$$\text{ACV}(t_C) = \frac{[\sum_P x_i^{2g}(\pi_i^{-1} - 1)]^{1/2}}{\sum_P x_i^g}\, \sigma. \qquad (6)$$

Observe that $\text{ACV}(t_C)$ decreases, all other things held constant, as any of the π_i increases.

The right hand side of equation (6) attains its minimum for a fixed expected sample size, $n_E = \sum_P \pi_i$, when $\pi_i = n_E x_i^g / \sum_P x_k^g$ if all these selection probabilities are bounded by 1. Furthermore, at that minimum, $\text{ACV}(t_C) \leq \sigma/\sqrt{n_E}$. Near equality holds when all $n_E x_i^g / \sum_P x_k^g << 1$.

Equation (6) further tells us that if we knew σ, we could be assured of meeting meeting an ACV target, say, C. We do this by setting $\pi_i = \min\{1, n_T x_i^g / \sum_P x_k^g\}$ and $n_T \geq (\sigma/C)^2$.

We can call n_T the “targeted sample size.” The expected sample size, $n_E = \sum_P \pi_i$, is less than or equal to n_T. Equality holds only when all the $n_T x_i^g / \sum_P x_k^g$ are bounded by 1, which we are *not* requiring. Nevertheless, *setting the selection probabilities at* $\pi_i = \min\{1, n_T x_i^g / \sum_P x_k^g\}$ *assures* $ACV(t_C) \leq \sigma/\sqrt{n_T}$ under the model in equation (5) .

In practice σ^2 must be guessed at or estimated from previous data, say by

$$s^2 = \frac{\sum_f w_i x_i^g e_i^2}{\sum_f w_i x_i^g},$$

where f denotes the previous sample, w_i is the weight for unit i in that sample, $e_i = [\sum_F x_k^g / \sum_F x_k](y_i - bx_i)/(bx_i^g)$, $b = \sum_f w_i y_i / \sum_f w_i x_i$, and F is the previous population. Alternatively,

$$s^2 = \frac{(\sum_f w_k x_k^g) \sum_f w_i (y_i - bx_i)^2 / x_i^g}{(\sum_f w_i y_i)^2} .$$

When the model holds, $e_i \approx \epsilon_{i.}$ That is one justification of our choices for the e_i and s^2. Another follows. If the selection probabilities were $\pi_i = n^* x_i^g / \sum_P x_k^g << 1$ for all i, then the relative randomization variance of t_C as an estimator for $\sum_F y_i$ under Poisson sampling (which is what NASS uses) would be roughly $[\sum_F (y_i - Bx_i)^2/\pi_i]/(\sum_F y_i)^2$, where $B = \sum_F y_i / \sum_F x_i$. This can be reasonably estimated with the sample actually drawn by $[\sum_f w_i (y_i - bx_i)^2/\pi_i]/(\sum_f w_i y_i)^2 = s^2/n^*$. Thus, our choice for defining s^2 is in some way robust to model failure.

We will call the a sample selection procedure where each $\pi_i = \min\{1, n_T x_i^g / \sum_P x_k^g\}$ and $½ \leq g \leq 1$ "Brewer selection." This name applies whether or not the choice of n_T depends on σ in equation (5).

3. MULTIPLE TARGETS

Suppose we have M target variables, and y_{im} denotes the unit i y—value for the m'th target. Each target has its own (maybe unique, maybe not) control variable, and x_{im} denotes the unit i x—value for the m'th control. Furthermore, suppose each target/control pair is believed to obey the following model:

$$y_{im} = \beta_m (x_{im} + [\sum_P x_{km} / \sum_P x_{km}^g] x_{im}^g \epsilon_{im}), \tag{7}$$

where $E(\epsilon_{im} | x_{im}) = E(\epsilon_{im}\epsilon_{jm} | x_{im}, x_{jm}) = 0$ $(i \neq j)$, and $\mathrm{Var}(\epsilon_{im} | x_{im}) = \sigma_m^2$ for all m.

A set of weights, $\{a_i\}$, can often be constructed for a sample S that satisfies the M calibration equations

$$\sum_S a_i x_{im} = \sum_P x_{im}, \quad m = 1, ..., M,$$

such that every $a_i = \pi_i^{-1}[1 + O_P(1/\surd n)]$, where π_i is (again) the selection probability of unit i. Each calibration estimator $t_{C(m)} = \sum_S a_i y_{im}$ provides a model unbiased estimator for $T_m = \sum_P y_{im}$ under the model in equation (7).

One potential way to construct these weights is with the formula inspired by linear regression:

$$a_i = \pi_i^{-1} + (\sum_P \mathbf{x}_k - \sum_S \pi_k^{-1}\mathbf{x}_k)(\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} c_i \pi_i^{-1} \mathbf{x}_i', \tag{8}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{iM})$ is a row vector, and the choice for the c_i is arbitrary as long as $\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k$ is invertible. Popular choices are $c_i = 1/x_{i1}$ when M = 1 (so t_C becomes t from Section 1), and $c_i = 1$ (when one x_{im} is constant across i). Brewer (1994) suggests $c_i = (1 - \pi_i)/z_i$, where z_i is some composite measure of size across the M controls. We will return to this question of setting the c_i in Section 4.

Given target ACV's (denoted C_m) for all M target variables under the model in equation (7) and known σ-values (σ_m) for each variable, we can be assured of meeting these target ACV's when every

$$\pi_i = \min\{1, \max\{n_{T1} h_{i1}^{(g)}, ..., n_{Tm} h_{iM}^{(g)}\}\}, \tag{9}$$

where $n_{Tm} = C_m / \sigma_m$, and $h_{im}^{(g)} = x_{im}^g / \sum_k x_{km}^g$.

Observe that π_i in equation (9) can also be expressed as

$$\pi_i = \max\{\pi_{i1}, ..., \pi_{iM}\}, \qquad (10)$$

where $\pi_{im} = n_{Tm} h_{im}^{(g)}$ is Brewer selection for variable m. Consequently, the selection scheme in equation (10) can be called "Maximal Brewer Selection (MBS)." This name applies whether or not each target sample size n_{Tm} is set equal to C_m / σ_m.

4. APPLYING MBS

4.1. Poisson PRN Sampling

Brewer selection can be shown to minimize ACV(t_C) for a fixed n_E under the model in equation (5) and conversely to minimize the expected sample size given a target ACV. Maximal Brewer selection when M > 1 does *not* necessarily minimize the expected overall sample size given M target ACV's. Sigman and Monsour (1995) sketch a method for determining selection probabilities that are optimal (i.e., expected-sample-size minimizing) in this sense.

Although not optimal, MBS is relatively simple and conveniently flexible when combined with Poisson Permanent-Random-Number (PRN) sampling (Ohlsson 1995 uses the term "PRN;" the concept can be found in Brewer et al. 1972). In such a design, every population unit i is independently assigned a random number p_i – a PRN – from the uniform distribution on the interval [0, 1). Unit i is selected for the sample if and only if $p_i < \pi_i$.

Poisson sampling, whether employing PRN's or not, has the well-known property that the joint selection probability of two distinct units i and k is equal to the product of their individual selection probabilities; that is, $\pi_{ik} = \pi_i \pi_k$. This greatly eases randomization variance estimation. This method of sampling also assures that $\sum_S z_i/\pi_i \approx \sum_P z_i$, since the relative variance of $\sum_S z_i/\pi_i$ is less than $(\sum_P z_i^2/\pi_i)/(\sum_P z_i)^2$, which is O(1/n) under very mild restrictions on the z_i and π_i (see Isaki and Fuller 1982).

Poisson PRN sampling furthermore allows us to think of a sample drawn with MBS inclusion probabilities as the union of M Poisson PRN samples each drawn using the same PRN's and individual Brewer selection probabilities. This is convenient when we are interested in estimates of different combinations of target variables in different surveys.

For example, NASS makes estimates for potatoes in Minnesota in June and December, row crops (e.g., soybeans and corn) in March, June, and December, and small grains (e.g., wheat and barley) in March, June, September, and December. It wants to contact the same farms throughout the year, but has little interest in sampling a farm for the September survey if it has not historically had small grains. Thus, Poisson PRN samples of farms using the same PRN's can be drawn for potatoes, row crops, and small grains, each with its own Brewer selection probabilities. The union of all three is the overall sample in June. Similarly, the union of the row-crops and small-grains samples is the overall sample in March. Bailey and Kott (1997) discuss NASS's use of MBS and Poisson PRN sampling in Minnesota in greater detail.

Two additional points should be made at this time. One is that NASS actually draws the row-crops sample itself using MBS with individual row crops (soybeans, corn, etc.) serving as the target variables. The other is that MBS as practiced by the agency is the result of individual Brewer selections and Poisson PRN sampling. MBS is the cart and the individual Brewer selections the horse.

The overall MBS sample may not be the most efficient (expected-sample-size minimizing) way to meet multiple ACV targets. It is, however, the most efficient way of combining individual Brewer-selected samples.

4.2. A Count Control Variable

One potential target variable in an establishment survey is the number of units in P that still exist during the survey period. An obvious control variable for this target is unity, which can be assigned to each unit in P. Such a control is called a "count variable."

Whether or not the number of units still in existence is really of direct interest to survey managers, setting one component of $\mathbf{x}_i$, say x_{i0} equal to 1 for all i is a sensible policy. For one thing, it assures us that $t_{C(m)}$ will be randomization unbiased when $y_{im} > 0$, but $x_{im} = 0$; that is to say, when survey managers are surprised that unit i has a positive quantity of target variable m.

4.3. Calibration and Variance Estimation

NASS determines its calibration weights by first employing equation (8) with $c_i = 1 - \pi_i$. Brewer (1994) calls such a weighting scheme "cosmetic calibration," because the estimator can be put in prediction form ($t_{C(m)} = \sum_S y_{im} + (\sum_P \mathbf{x}_i - \sum_S \mathbf{x}_i)\mathbf{b}_m$, where $\mathbf{b}_m$ is defined below equation (11)) when $\mathbf{x}_i$ contains a count-variable component. He argues that with cosmetic calibration individual weights rarely fall below unity. Weights below unity are deemed undesirable by many.

Under the weighting that results from employing equation (8) with $c_i = 1 - \pi_i$, when π_i is 1, a_i is also 1. Cosmetic calibration weights lower than unity, although rare, can still occur. NASS uses an iterative process described below that has, so far, successfully eliminated all weights less than unity. When plugging $c_i = 1 - \pi_i$ into equation (8) produces an $a_j < 1$, π_j in the equation is set equal to unity, and the equation run again for all i. This process is continued until all $a_i \geq 1$.

The estimator $t_{C(m)}$ is model unbiased not only under the model in equation (7), but also under the more general model:

$$y_{im} = \mathbf{x}_i\gamma_m + u_{im},$$

where γ_m is an unspecified M-vector, and $E(u_{im}|\mathbf{x}_i) = 0$.

In order to be able to estimate the model variance of $t_{C(m)}$, we need to add the assumptions $E(u_{im}u_{jm}|\mathbf{x}_i, \mathbf{x}_j) = 0$, and $E(u_{im}^2|\mathbf{x}_i) = \sigma_{im}^2 < \infty$. In sharp contrast to the design stage, we are allowing the unit variances to be unspecified as long as they are finite.

Following the same reasoning that produced equation (3) leads to

$$E_\epsilon[(t_{C(m)} - T_m)^2] = \sum_S a_i^2\sigma_{im}^2 - 2\sum_S a_i\sigma_{im}^2 + \sum_P \sigma_{im}^2. \qquad (3')$$

When n is large, we can make use of the near equalities $\sum_S a_i\sigma_{im}^2 \approx \sum_S \sigma_{im}^2/\pi_i \approx \sum_P \sigma_{im}^2$, and conclude

$$E_\epsilon[(t_{C(m)} - T_m)^2] = \sum_S (a_i^2 - a_i)\sigma_{im}^2 < \sum_S a_i^2\sigma_{im}^2.$$

For a Poisson sample, the randomization mean squared error of $t_{C(m)}$ is

$$E_P[(t_{C(m)} - T_m)^2] \approx \sum_P \ddot{e}_{im}^2(\pi_i^{-1} - 1),$$

where $\ddot{e}_{im} = y_{im} - \mathbf{x}_i\mathbf{B}_m$, and $\mathbf{B}_m = (\sum_P c_k\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum_P c_k\mathbf{x}_k'y_{km}$ (since $\sum_S a_iy_{im} - \sum_P y_{im} = \sum_S a_i\ddot{e}_{im} - \sum_P \ddot{e}_{im} = \sum_S \ddot{e}_{im}/\pi_i + (\sum_P \mathbf{x}_k - \sum_S \pi_k^{-1}\mathbf{x}_k)(\sum_S c_k\pi_k^{-1}\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum_S c_i\pi_i^{-1}\mathbf{x}_i'\ddot{e}_{im} - \sum_P \ddot{e}_{im} \approx \sum_S \ddot{e}_{im}/\pi_i + (\sum_P \mathbf{x}_k - \sum_S \pi_k^{-1}\mathbf{x}_k)(\sum_S c_k\pi_k^{-1}\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum_P c_i\mathbf{x}_i'\ddot{e}_{im} - \sum_P \ddot{e}_{im} = \sum_S \ddot{e}_{im}/\pi_i - \sum_P \ddot{e}_{im}$). When the c_k are all equal, the vector $\mathbf{B}_m$ is often called the "finite-population" or "census" regression coefficient.

An estimator for both the model variance and randomization mean squared of $t_{C(m)}$ is

$$v(t_{(m)}) = \sum_S (a_i^2 - a_i)e_{im}^2. \tag{11}$$

where $e_{im} = y_{im} - \mathbf{x}_i\mathbf{b}_m$, and $\mathbf{b}_m = (\sum_S c_k\pi_k^{-1}\mathbf{x}_k{}'\mathbf{x}_k)^{-1} \sum_S c_k\pi_k^{-1}\mathbf{x}_k{}'y_{km}$ are the sample analogues of $\ddot{e}_{im}$ and $\mathbf{B}_m$, respectively.

4.4. The Delete-a-Group Jackknife

The problem with v in equation (11) is that is requires e_{im} to be calculated separately for each target variable. That is one reason why NASS uses a delete-a-group (DAG) jackknife variance estimator (Kott 1998). The DAG jackknife is also convenient when estimating the variances of domain totals and of ratios.

The Poisson sample is randomly divided into 15 replicate groups, denoted $S_1, S_2, ..., S_{15}$ (some groups can have one more member than others). The complement of each S_r is called the jackknife replicate group $S_{(r)} = S - S_r$. NASS then creates 15 sets of replicate weights. For the rth set: $a_{i(r)} = 0$ when $i \in S_r$; and

$$a_{i(r)} = a_i + (\sum_P \mathbf{x}_k - \sum_{S(r)} a_k\mathbf{x}_k)(\sum_{S(r)} c_k a_k \mathbf{x}_k{}'\mathbf{x}_k)^{-1} c_i a_i \mathbf{x}_i{}'$$

otherwise. This choice assures $a_{i(r)} \approx a_i$ for $i \in S_{(r)}$ when 15 is deemed large. Moreover, these two equalities will prove helpful. Under the model, because the ϵ_{im} are uncorrelated across units,

$$\sum_S a_{i(r)}\epsilon_{im} - \sum_S a_i\epsilon_{im} = -\sum_{Sr} a_i\epsilon_{im} + (\sum_P \mathbf{x}_k - \sum_{S(r)} a_k\mathbf{x}_k)(\sum_{S(r)} c_k a_k \mathbf{x}_k{}'\mathbf{x}_k)^{-1}\sum_{S(r)} c_i a_i \mathbf{x}_i{}'\epsilon_{im} \approx -\sum_{Sr} a_i\epsilon_{im}.$$

Even without the model,

$$\begin{aligned}
\sum_S a_{i(r)}\ddot{e}_{im} - \sum_S a_i\ddot{e}_{im} &= -\sum_{Sr} a_i\ddot{e}_{im} + (\sum_P \mathbf{x}_k - \sum_{S(r)} a_k\mathbf{x}_k)(\sum_{S(r)} c_k a_k \mathbf{x}_k{}'\mathbf{x}_k)^{-1}\sum_{S(r)} c_i a_i \mathbf{x}_i{}'\ddot{e}_{im} \\
&\approx -\sum_{Sr} a_i\ddot{e}_{im} + (\sum_P \mathbf{x}_k - \sum_{S(r)} a_k\mathbf{x}_k)(\sum_{S(r)} c_k a_k \mathbf{x}_k{}'\mathbf{x}_k)^{-1}\sum_{S(r)} c_i \pi_i^{-1} \mathbf{x}_i{}'\ddot{e}_{im} \\
&\approx -\sum_{Sr} a_i\ddot{e}_{im} + (\sum_P \mathbf{x}_k - \sum_{S(r)} a_k\mathbf{x}_k)(\sum_{S(r)} c_k a_k \mathbf{x}_k{}'\mathbf{x}_k)^{-1}\sum_{S} c_i \pi_i^{-1} \mathbf{x}_i{}'\ddot{e}_{im} \\
&\approx -\sum_{Sr} a_i\ddot{e}_{im} + (\sum_P \mathbf{x}_k - \sum_{S(r)} a_k\mathbf{x}_k)(\sum_{S(r)} c_k a_k \mathbf{x}_k{}'\mathbf{x}_k)^{-1}\sum_{P} c_i \mathbf{x}_i{}'\ddot{e}_{im} \\
&= -\sum_{Sr} a_i\ddot{e}_{im}
\end{aligned}$$

when 15 is deemed large.

The DAG variance estimator for $t_{C(m)}$ is :

$$v_J(t_{C(m)}) = (14/15) \sum^{15} (\sum_S a_{i(r)}y_{im} - t_{C(m)})^2, \tag{12}$$

which WESVAR (Westat 1997) calls JK1.

It is easy to see that under the model in equation (7) and the error structure assumed above, the model expectation of $v_J(t_{C(m)})$ when 15 (as well as n) is assumed to be large is approximately $\sum_S a_i^2\sigma_{im}^2$ (since $\sum_S a_{i(r)}y_{im} - t_{C(m)} = \sum_S a_{i(r)}y_{im} - \sum_S a_i y_{im} = \sum_S a_{i(r)}\epsilon_{im} - \sum_S a_i\epsilon_{im} \approx -\sum_{Sr} a_i\epsilon_{im}$).

We sketch below a proof that the randomization expectation of $v_J(t_{C(m)})$ is approximately $\sum_P \ddot{e}_{im}^2\pi_i^{-1}$ when $\sum_P \ddot{e}_{im} \approx 0$. This last near equality obtains exactly when $c_i = 1/(\gamma\mathbf{x}_i{}')$ for some row vector γ (since then $\sum_P \ddot{e}_{im} = \sum_P \gamma\mathbf{x}_i{}'c_i\ddot{e}_{im} = \gamma\sum_P c_i\mathbf{x}_i{}'\ddot{e}_{im} = 0$). In practice, NASS does not deliberately choose a c_i with this property, however. This can cause the DAG jackknife to be randomization-biased. NASS sets $c_i = 1 - \pi_i$ and includes within $\mathbf{x_i}$ (for calibration purposes) a component $x_{i0} = 1$. Thus, when all the π_i are small, $c_i \approx 1 = 1/(\gamma\mathbf{x}_i{}')$ for $\gamma = (1, 0, ..., 0)$. When some π_i are large, the randomization mean squared error is smaller than $\sum_P \ddot{e}_{im}^2\pi_i^{-1}$, so whatever small bias in $v_J(t_{C(m)})$ is caused by $\sum_P \ddot{e}_{im}$ not being near zero is likely to be overwhelmed by $\sum_P \ddot{e}_{im}^2\pi_i^{-1}$ being larger than $\sum_P \ddot{e}_{im}^2(\pi_i^{-1} - 1)$.

Let n_r be the size of S_r. When 15 is large, $n/n_r \approx 15$ and $14/15 \approx 1$. The replicate group S_r can be viewed as a random subsample of S, and $q_r = (n/n_r)\sum_{Sr}\ddot{e}_{im}/\pi_i$ is a nearly randomization-unbiased estimator of $\sum_S\ddot{e}_{im}/\pi_i$, which is approximately 0 when $\sum_P\ddot{e}_{im} = 0$. The randomization variance of q_r with respect to the subsampling is approximately $(n/n_r)\sum_S\ddot{e}_{im}{}^2/\pi_i$ for each r. Now $\sum_S a_{i(r)}y_{im} - t_{C(m)} \approx \sum_S a_{i(r)}\ddot{e}_{im} - \sum_S a_i\ddot{e}_{im} \approx -\sum_{Sr}\ddot{e}_{im}/\pi_i \approx -q_r/15$. We can conclude that the randomization expectation of $v_J(t_{C(m)})$ in equation (12) with respect to the subsampling when $\sum_P \ddot{e}_{im} = 0$ is approximately $\sum_S \ddot{e}_{im}{}^2/\pi_i$.

5. MORE ON THE NASS EXPERIENCE

NASS prepares different samples in the various US states. NASS integrated its crops, stocks, and livestock surveys in the mid 1980s. Stratified simple random samples were drawn using a *priority* stratification scheme. For example, Stratum 1 might be large hog farms, Stratum 2 large crop farms that are not large hog farms, and so on, depending on the priorities of the target variables. Simple expansion estimates were generated from the sample data. Livestock variables were removed from the integrated Crops/Stocks (CS) Survey in the mid 1990s.

In the 1997/98 growing year, NASS drew a MBS Poisson PRN sample for the CS in one state, Minnesota. This proved very successful (Bailey and Kott 1997) . In 1998/99, this selection method was used in four states. By 1999/2000, 14 states had MBS Poisson PRN samples. Plans are to use MBS exclusively in the following year.

Rather than explicitly adding x_{i0} to the other x_{im} in selection equation (9), NASS has set a minimum value for π_i at roughly 0.01. For the most part, the same control variables have been used in the selection equation and the calibration (equation (8)), although a count (intercept) variable has been added to every calibration. Figure 1 provides a chart of how many control variables were used in each of the 14 1999/2000 CS states.

NASS has set g in equation (9) equal to 0.75. Brewer (1999) seems to show a slight preference for g = .6. Table 1 reports estimated s-values in one state (PA) based on June 1999 survey data and various values for g. Crop and stock target variables for a single commodity (e.g., corn) use the same control value. One thing to notice is the s seems to increase as the fraction of the sample with positive x-values (called "the commodity population") and positive y-values decreases. The second is that NASS's choice of g = .75 everywhere needs to be explored more thoroughly. In principle, the best choice for g minimizes s asymptotically.

Nonresponse has been handled using the pre-existing imputation scheme, which relies on the old priority stratification. DAG jackknife variances are estimated treating non-response as a second phase of sampling and pretending that respondents were reweighted using the priority strata as the reweighting groups. If the models supporting the imputation scheme are correct, this will (if anything) bias mean squared error estimates upward.

6. COMMENTS

The change in the Crops/Stocks Survey from an estimation strategy featuring stratified simple random sampling and a simple expansion estimator to Poisson PRN sampling with maximal Brewer selection probabilities and a (cosmetic) calibration estimator has proven very successful at NASS. The Agency is currently exploring the use of the new strategy in other surveys as well. In the interest of honest disclosure, NASS actually uses collocated sampling (Brewer et al., 1972) sampling rather than Poisson sampling. This modestly reduces the sample-size variability. Mean squared errors are estimated as if Poisson samples were drawn.

Kott and Fetter (1999) show how Poisson PRN sampling can easily to adapted to limit the number of times a single unit is selected across co-ordinated surveys. Let $\pi_i{}^{(q)}$ be the unit i selection probability for survey q (= 1, 2, ... Q). Unit i is in the sample for survey q when its PRN, p_i, is in the interval $[\tau_{i,q-1}, \tau_{i,q})$, where $\tau_{i,0} = 0$, and $\tau_{i,f} = \pi_i^{(1)} + ... + \pi_i^{(f)}$. For this *sequential interval Poisson* (SIP) sampling methodology described above to work, $\tau_{i,q}$ cannot exceed unity. Fortunately, it is a simple matter to generalize SIP sampling a bit. We can redefine $\tau_{i,f}$ as $\pi_i^{(1)} + ... + \pi_i^{(f)} - I_{(i)}$, where $I_{(i)}$ is the largest integer less than $\pi_i^{(1)} + ... + \pi_i^{(f)}$. When $\tau_{i,q-1} > \tau_{i,q}$, the interval $[\tau_{i,q-1}, \tau_{i,q})$ is similarly redefined as the union of $[\tau_{i,q-1}, 1)$ and $[0, \tau_{i,q})$.

The larger $I_{(i)}$ the greater the number of survey samples in which unit i can find itself (that number will either be $I_{(i)}$ or

$I_{(i)} + 1)$. This is another reason NASS needs to explore the value at which g in equation (9) is set. The smaller the value, the less likely a particular unit with large control values with be selected for a sample.

It may also be that the best choice for g varies by target variable. Worse, $\mathrm{Var}(\epsilon_{im}) \propto x_{im}^{g}$ may not even be the appropriate specification. Oddly, this widely used specification began as an approximation of $ax_{im} + bx_{im}^{2}$ (see Cochran, 1963, p. 256), which has prompted the belief that ½ must be the lower bound of g in practice. In the NASS application, the quality of control information is better for larger values. Consequently, it is possible that the best g for some target variables is, in fact, less than ½.

REFERENCES

Bailey, J. T. and P.S. Kott (1997), "An Application of Multiple List Frame Sampling for Multi-purpose Surveys, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 496-500.

Bosecker, R. (1989), *Integrated Agricultural Surveys*, NASS Research Report Number SSB-89-05, Washington, DC: National Agricultural Statistics Service.

Brewer, K.R.W. (1963), "Ratio Estimation and Finite Populations: Some Results Deductible from the Assumption of an Underlying Stochastic Process," *Australian Journal of Statistics*, **5**, pp. 93-105.

Brewer, K.R.W. (1994), "Survey Sampling Inference: Some Past Perspectives and Present Prospects," *Pakistan Journal of Statistics,* **10(1)A**, pp. 213-233.

Brewer, K.R.W. (1999), "Cosmetic Calibration with Unequal Probability Sampling," *Survey Methodology*, forthcoming.

Brewer, K.R.W., Early, L.J., and Joyce, S.F. (1972), "Selecting Several Samples from a Single Population," *Australian Journal of Statistics*, **14**, pp. 231-239.

Cochran, W.G. (1963), *Sampling Techniques*, 2nd Edition, New York: Wiley.

Deville, J-C. and Särndal, C-E. (1992), "Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, **87**, pp. 376-382.

Godambe, V.P. (1955), "A Unified Theory of Sampling from Finite Populations,' *Journal of the Royal Statistical Society*, **B17**, pp. 269-278.

Isaki, and Fuller, W.A. (1982), "Survey Design Under the Regression Superpopulation Model," *Journal of the American Statistical Association*, **77**, 89-96.

Kott, P. S. (1998). *Using the Delete-A-Group Jackknife Variance Estimator in NASS Surveys*, RD Research Report No. RD-98-01, Washington, DC: National Agricultural Statistics Service.

Kott, P.S. and Fetter, M.J. (1999), "Using Multi-phase Sampling to Limit Respondent Burden Across Agriculture Surveys," *Proceedings of the Survey Methods Section, Statistical Society of Canada*, forthcoming.

Ohlsson, E. (1995), "Coordination of Samples Using Permanent Random Numbers," in Cox, B.G., Binder, D.A., Chinnappa, N., Christianson A, Colledge, M.J., and Kott, P.S. (eds) *Business Survey Methods*, New York: Wiley, pp. 153-169.

Royall, R. M. (1970), " On Finite Population Sampling Under Certain Linear Regression Models," *Biometrika* **57**, pp. 377-387.

Särndal, C-E, Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer.

Sigman, R.S. and Monsour, N. J. (1995), " Selecting Samples from List Frames of Businesses," in Cox, B.G., Binder, D.A., Chinnappa, N., Christianson A, Colledge, M.J., and Kott, P.S. (eds) *Business Survey Methods*, New York: Wiley, pp. 133-152.

Westat, Inc. (1997), *A User's Guide to WesVarPC®, Version 2.1*, Rockville, MD: Westat**.**

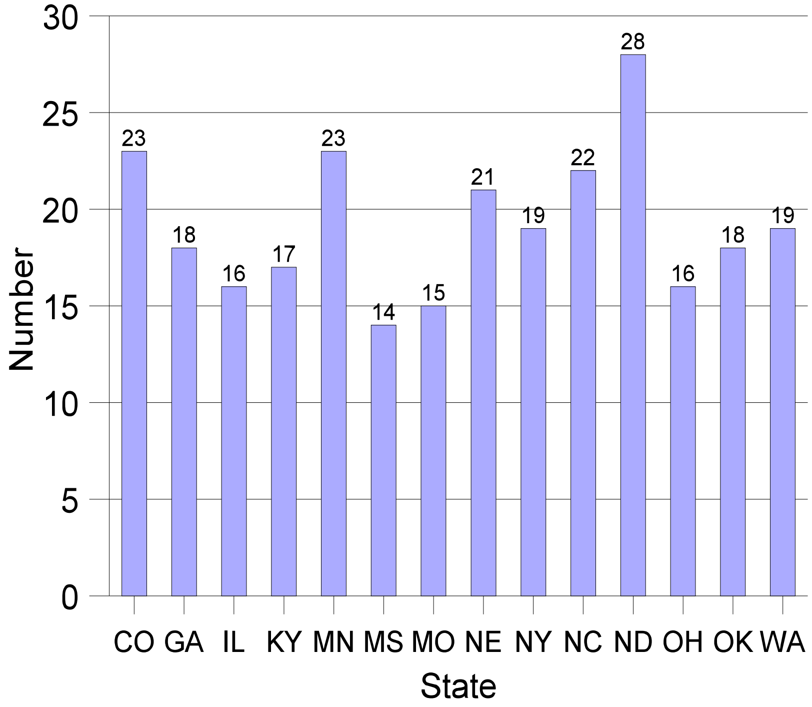

Figure 1: Number of Control Variables in Each State

Table 1: Target Variable Calculations in PA With Different Values for g

Commodity	Commodity Population	Survey: Response Rate	Survey: Positive Reports	Survey: % with Item	s with $g = 0.5$	s with $g = 0.6$	s with g =0.75	s with $g = 0.9$
Alfalfa Acres	18006	84.3%	372	60.7%	1.26	1.26	1.27	1.31
Wheat Stocks	8079	84.4%	29	6.2%	16.88	14.93	12.59	10.82
Barley Acres	5206	84.3%	122	46.0%	1.39	1.40	1.45	1.55
Corn Stocks	21268	82.4%	314	36.1%	2.64	2.51	2.43	2.47
Corn Acres	21268	84.3%	559	78.5%	0.75	0.74	0.76	0.81
Oat Stocks	11824	84.4%	114	22.2%	2.81	2.85	2.95	3.13
Oat Acres	11824	84.3%	250	54.8%	1.39	1.41	1.47	1.55
Other Hay	19478	84.3%	446	65.5%	1.30	1.28	1.27	1.31
Potato Acres	829	84.3%	67	59.4%	0.82	0.82	0.87	0.99
Rye Acres	4210	84.3%	103	40.5%	1.95	1.98	2.08	2.24
Soybean Stocks	7030	83.9%	79	18.4%	3.90	3.73	3.56	3.48
Soybean Acres	7030	84.3%	234	67.1%	0.95	0.95	0.96	1.00
Tobacco Acres	979	84.3%	9	33.3%	1.33	1.35	1.41	1.48
Wheat Acres	3836	84.3%	230	67.9%	0.98	1.03	1.14	1.29