Poisson Sampling, Regression Estimation, and the Delete-a-Group Jackknife

Phillip S. Kott

When coupled with the simple expansion estimator, Poisson sampling leads to estimators with higher-than-necessary variances. That problem vanishes when the expansion estimator is replaced by a randomization-consistent regression estimator. A simultaneous estimator for the model variance and randomization mean squared error of this estimation strategy is developed. It is nearly identical to the weighted residual variance estimator, but can be slightly better at estimated the model variance when finite population correction matters. When finite population correction can be ignored, an appropriately-defined delete-a-group jackknife variance estimator is shown to have desirable asymptotic properties making it a practical alternative in many applications.

KEY WORDS Asymptotic, finite population correction, model variance, randomization consistent, randomization mean squared error.

Phillip S. Kott is Chief Research Statistician, Research and Development Division, National Agricultural Statistics Service, 3251 Old Lee Highway, Room 305, Fairfax, Virginia, 22030. This paper was originally prepared for the Joint Statistical Meetings, August 2000, Indianapolis, Indiana.

1. Introduction

Poisson sampling is perhaps the simplest form of unequal probability selection. Its use often leads to inefficient estimation, which is why it is not more widely used. When combined with a regression-type estimator, however, the advantages of Poisson sampling can be realized. That is why the National Agricultural Statistics Service (NASS) has recently overhauled its major crop survey program and adopted Poisson sampling (see Kott and Bailey, 2000).

This paper reviews and extends the theory supporting the use of Poisson sampling coupled with a randomization-consistent regression estimator. Section 2 introduces the basic setup. Section 3 explores the randomization and model-based properties of the estimation strategy. Since the large-sample and large-population properties of variance estimators will be a main focus here, some care is taken in the development of the asymptotics. Section 4 proposes a simultaneous estimator of model variance and randomization mean squared that is slightly better than the one proposed in Särndal *et al.* (1989). Section 5 addresses the issue of small-sample bias in variance estimation. Section 6 discusses the applicability of the delete-a-group jackknife variance estimator when the sample size is large and the population even larger. Finally, Section 7 offers a broader discussion of the topics covered in the text.

2. Background

Suppose we want to estimate a population (U) total, $T = \sum_U y_k$ based on a sample (S) of y-values. If the probability that population unit k is in the sample is π_k, then the simple expansion of T is $t = \sum_S y_k/\pi_k$. Another useful way to render t is as $t = \sum_U y_k I_k/\pi_k$, where I_k is a random variable equal to 1 when $k \in S$ and 0 otherwise. This means $E(I_k) = \pi_k$. Under randomization-based inference the y_k are fixed constants, while the I_k are random variables. It is easy to see that t is a randomization-unbiased estimator of T; that is $E_p(t) = T$, where the subscript p denotes the expectation with respect to the I_k (this is a convention; the p derives from “probability”).

The randomization variance of t is

$$Vr_p(t) = E_p[(t - T)^2] = \sum_U (y_k/\pi_k)(y_g/\pi_g)(\pi_{kg} - \pi_k\pi_g),$$

where $\sum_U$ denotes $\sum_{k\in U} \sum_{g\in U}$ in this context, and $\pi_{kg} = E(I_k I_g)$ is the joint selection probability of units k and g. When k = g, $\pi_{kg} = \pi_k$. The randomization variance of t very much depends on how exactly the sample is drawn, and in particular of the joint selection probabilities.

Under Poisson sample, each unit k is sampled independently of every other unit in the population. Consequently, $\pi_{kg} = \pi_k\pi_g$ when $k \neq g$. This simplifies the randomization variance of t immensely:

$$Var_p(t) = \sum_U (y_k/\pi_k)^2(\pi_k - \pi_k^2) = \sum_U (y_k^2/\pi_k)(1 - \pi_k),$$

and leads to the simple unbiased randomization variance estimator:

$$var_p(t) = \sum_S (y_k/\pi_k)^2(1 - \pi_k).$$

The problem with Poisson sampling in this context is that it can lead to a larger-than-necessary randomization variance. This is because the sample size of a Poisson sample is random. It has an expected value of $n^* = E(\sum_U I_k) = \sum_U \pi_k$, and a variance of $Var(\sum_U I_k) = \sum_U \pi_k (1 - \pi_k)$.

Under a sample design where the sample size is fixed at $n = n^*$, the unit selection probability of each unit k set equal to π_k, and each y_k is proportional to π_k, the randomization variance of t would be zero (because if, say, $y_1/\pi_1 = b$, then $t \equiv bn = T$). Under Poisson sampling, by contrast, this variance would be $b^2 \sum_U \pi_k (1 - \pi_k)$.

The problem caused by random sample size can be eliminated when Poisson sampling is coupled with this regression estimator:

$$t_R = t + (\sum_U \mathbf{x}_k - \sum_S \pi_k^{-1}\mathbf{x}_k)(\sum_S c_k\pi_k^{-1}\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum_S c_k\pi_k^{-1}\mathbf{x}_k'y_k, \quad (1)$$

where $\mathbf{x}_k = (x_{k1}, ..., x_{kQ})$ is a row vector of values known for all S, c_k is a constant, $\sum_U \mathbf{x}_k$ is known, and $\sum_S c_k\pi_k^{-1}\mathbf{x}_k'\mathbf{x}_k$ is invertible. For the simple example given above where y_k is proportional to π_k , we can now let $\mathbf{x}_k$ be the scalar π_k. It is easy to see that t_R will always be $n^*(y_1/\pi_1)$, which is also what T is.

The regression estimator in equation (1) is a very slight variation of the so-called general regression estimator (GREG). See, for example, Särndal, Swensson, and Wretman. (1992). A good review of regression estimators in the survey sampling context is Brewer (1994). The GREG is poorly named because it does not include purely model-based regression estimators.

The regression estimator in equation (1) can be rewritten as $t_R = \sum_S a_k y_k$, where a_k is the regression weight of k:

$$a_k = \pi_k^{-1} + (\sum_U \mathbf{x}_i - \sum_S \pi_i^{-1}\mathbf{x}_i)(\sum_S c_i\pi_i^{-1}\mathbf{x}_i'\mathbf{x}_i)^{-1} c_k\pi_k^{-1}\mathbf{x}_k'. \quad (2)$$

It is well known (and easy to see) that the a_k satisfy the *calibration equation*: $\sum_S a_k\mathbf{x}_k = \sum_U \mathbf{x}_k$ (Deville and Särndal 1992; t_R is identical to the regression estimator in their equation (1.6) when their q_k is set to equal our c_k).

3. Properties of the Estimation Strategy

The regression estimator, t_R, under Poisson sampling has both desirable randomization-based and model-based properties under mild conditions.

3.1. Randomization-based Properties

The randomization-based properties of t_R are asymptotic (we use the more accurate modifier “randomization” in place of the often-used “design”). That is to say, they depend on the expected sample size, n*, being large. A sufficient condition for an

estimation strategy (an estimator coupled with a sampling design) to be randomization consistent is that its relative mean squared error should approach 0 as n* grows arbitrarily large.

Let N be the population size of U. We want to entertain the possibility that O(n*) is less than O(N). Consequently, we assume the following as N and n grow arbitrarily large and *Q remains fixed*:

$$0 < L_y \le \sum_U y_k^{\delta}/N < B_y < \infty, \qquad \delta = 1, ..., 8; \tag{3.1}$$

$$0 < L_{xq} \le \sum_U x_{kq}^{\delta}/N < B_{xq} < \infty, \qquad q = 1, ..., Q;\ \delta = 1, ..., 8; \tag{3.2}$$

$$0 < L_c \le \sum_U c_k^{\delta}/N < B_c < \infty, \qquad \delta = 1, ..., 8 \tag{3.3}$$

$$0 < L_\pi \le \sum_U [(N/n^*)\pi_k]^{-\delta}/N < B_\pi < \infty, \quad \delta = 1, ..., 8. \tag{3.4}$$

The relative randomization mean squared error of the expansion estimator, t, under Poisson sampling is $\sum_U (y_k^2/\pi_k)(1-\pi_k)/(\sum_U y_k)^2 < \sum_U (y_k^2/\pi_k)/(\sum_U y_k)^2$. Equations (3.1), (3.4), and Scharwz's inequality tell us that the numerator of this last expression is $O(N^2/n^*)$, while its denominator is $O(N^2)$. Thus, the relative randomization mean squared of t under Poisson sampling is $O(1/n^*)$, and the estimation strategy is randomization consistent. Furthermore, since $E_P[(t - T)^2]/T^2 = O(1/n^*)$, $(t - T)/T = O_p(1/\sqrt{n^*})$, and $t - T = O_p(N/\sqrt{n^*})$

The regression estimator, t_R, from equation (1) under Poisson sampling and the assumptions in equation (3) is equal to $t + O_p(N/\sqrt{n^*})$. This is because, using similar argument as above (and Scharwz's inequality repeatedly), the components of $\sum_U \mathbf{x}_k - \sum_S \pi_k^{-1}\mathbf{x}_k$ are $O_p(N/\sqrt{n^*})$, while the absolute values of the components of $\sum_S c_k\pi_k^{-1}\mathbf{x}_k'\mathbf{x}_k$ and $\sum_S c_k\pi_k^{-1}\mathbf{x}_k'y_k$ are $O_p(N)$. Thus, like t, t_R is randomization consistent. Furthermore, $(t_R - T)/T = O_p(1/\sqrt{n^*})$, and the relative mean squared error of t_R is $O(1/n^*)$.

Assuming, as we will from now on, that $N^{-1}(\sum_U c_k\mathbf{x}_k'\mathbf{x}_k)$ is invertible, let $\mathbf{B} = (\sum_U c_k\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum_U c_k\mathbf{x}_k'y_k$, and $e_k = y_k - \mathbf{x}_k\mathbf{B}$, so that $\sum_U c_i\mathbf{x}_i'e_i = 0$. We can now express the error of t_R as

$$
\begin{aligned}
t_R - T &= \textstyle\sum_S a_i y_i - \sum_U y_i \\
&= \textstyle\sum_S a_i(\mathbf{x}_i\mathbf{B} + e_i) - \sum_U (\mathbf{x}_i\mathbf{B} + e_i) \\
&= \textstyle\sum_S a_i e_i - \sum_U e_i \\
&= \textstyle\sum_S e_i/\pi_i + (\sum_U \mathbf{x}_k - \sum_S \pi_k^{-1}\mathbf{x}_k)(\sum_S c_k\pi_k^{-1}\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum_S c_i\pi_i^{-1}\mathbf{x}_i'e_i - \sum_U e_i \\
&= \textstyle\sum_S e_i/\pi_i + (\sum_U \mathbf{x}_k - \sum_S \pi_k^{-1}\mathbf{x}_k)(\sum_S c_k\pi_k^{-1}\mathbf{x}_k'\mathbf{x}_k)^{-1}(\sum_U c_i\mathbf{x}_i'e_i + O_P(N/\downarrow n^*)) - \sum_U e_i \\
&= \textstyle\sum_S e_i/\pi_i + (\sum_U \mathbf{x}_k - \sum_S \pi_k^{-1}\mathbf{x}_k)(\sum_S c_k\pi_k^{-1}\mathbf{x}_k'\mathbf{x}_k)^{-1}\, O_P(N/\downarrow n^*) - \sum_U e_i \\
&= \textstyle\sum_S e_i/\pi_i - \sum_U e_i + O_p(N/n^*) .
\end{aligned}
$$

This tells us that the randomization mean squared error of t_R under Poisson sampling is dominated by $\mathrm{Var}_P(\sum_S e_k/\pi_k) = \sum_U (e_k^2/\pi_k)(1-\pi_k)$. This is identical to the variance of the expansion estimator under Poisson sampling except that e_k has replaced y_k.

3.2. Model-based Properties

Suppose the y_k were random variables that satisfied the following model:

$$y_k = \mathbf{x}_k\beta + \epsilon_k, \tag{4}$$

where β is an unknown column vector, $E(\epsilon_k|\mathbf{x}_k, I_k) = E(\epsilon_k\epsilon_g|\mathbf{x}_k,\mathbf{x}_g, I_k, I_{g)}) = 0$ for $k \neq g$, and $E(\epsilon_k^2|I_k) = \sigma_k^2 = f(\mathbf{x}_k, \mathbf{z}_k) < \infty$, where $\mathbf{z}_k$ is a vector of values associated with k. The σ_k^2 *need not be known*. Moreover, there is no reason that I_k cannot be a function of the components of $\mathbf{x}_k$ and $\mathbf{z}_k$.

It is easy to see that as long as the regression weights satisfy the calibration equation, $\sum_S a_k\mathbf{x}_k = \sum_U \mathbf{x}_k$, t_R will be model unbiased in the sense that $E_\epsilon(t_R - T) = 0$. Moreover, its model variance is

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

When σ_i^2 has the form $\mathbf{x}_i\mathbf{h}$, for some not-necessarily-specified vector $\mathbf{h}$, then $\sum_S a_i\sigma_i^2 = \sum_U \sigma_i^2$, and the model variance of t_R collapses to $\sum_S (a_i^2 - a_i)\sigma_i^2$. Alternatively, if we add to ours asymptotic assumptions the following:

$$0 < L_\sigma \leq \sum_U \sigma_k^{2r}/N < B_\sigma < \infty, \quad r = 1, ..., 4, \tag{3.5}$$

then one can see that the model variance of t_R is $O(N^2/n^*)$, while $\sum_S a_i\sigma_i^2 - \sum_U \sigma_i^2$ is $O_P(N/\sqrt{n^*})$. Observe that although we are interested in model-based expectations here, we plan to invoke a large-sample, randomization-based equality. Model-based theory does not deny the applicability of the law of large numbers to random samples. It simply resists taking averages (expectations) across all possible samples.

Our last equality suggests the following asymptotic approximation for the model variance of t_R:

$$E_\epsilon[(t_R - T)^2] \approx \sum_S a_i^2\sigma_i^2 - \sum_S a_i\sigma_i^2, \tag{5}$$

which drops $O_P(N/\sqrt{n^*})$ terms. In so doing, it assumes that O(N) is greater than O(n).

What about likewise replacing a_i^2 by π_i^{-2} (and a_i by π_i^{-1}) in equation (5)? Such a substitution would effectively drop $O_P(N^2/[n^*]^{3/2})$ term. To see why, observe that

$$\begin{aligned}\sum_S a_i^2\sigma_i^2 &= \sum_S \pi_i^{-2}\sigma_i^2 + 2\,(\sum_U \mathbf{x}_i - \sum_S \pi_i^{-1}\mathbf{x}_i)(\sum_S c_i\pi_i^{-1}\mathbf{x}_i'\mathbf{x}_i)^{-1}\sum_S c_i\pi_i^{-2}\mathbf{x}_i'\sigma_i^2 + \\ &\quad (\sum_U \mathbf{x}_i - \sum_S \pi_i^{-1}\mathbf{x}_i)(\sum_S c_i\pi_i^{-1}\mathbf{x}_i'\mathbf{x}_i)^{-1}\sum_S \sigma_i^2c_i^2\pi_i^{-2}\mathbf{x}_i'\mathbf{x}_i\,(\sum_S c_i\pi_i^{-1}\mathbf{x}_i'\mathbf{x}_i)^{-1}\,(\sum_U \mathbf{x}_i - \sum_S \pi_i^{-1}\mathbf{x}_i)' \\ &= \sum_S \pi_i^{-2}\sigma_i^2 + O_P(N^2/[n^*]^{3/2}).\end{aligned}$$

In subsequent analyses, such asymptotic arithmetic will often be left to the reader.

Suppose finite population correction really matters. At the extreme, O(N) = O(n), and $O_P(N^2/[n^*]^{3/2})$ is of the same asymptotic order as the $O_P(N/\sqrt{n^*})$ term dropped by equation (5). Finite population correction still matters somewhat when $O(N) = O([n^*]^{3/2})$; that is when the population is *relatively large* (Kott 1990). Under this

setup equation (5) appropriately drops a $O_P(n^*)$. term. Observe, however, that replacing a_i^2 by π_i^{-2} would effectively drops a larger, $O_P([n^*]^{3/2})$ term.

3.3. Anticipated Variance

The model variance of t_R is a function of the realized sample and does not depend at all on the sampling design. As noted in the previous section, it is $O_P(N^2/n^*)$ under the (extended) asymptotic assumptions of equation (3). In fact, if we are willing to drop $O_P(N^2/[n^*]^{3/2})$ terms, the model variance can be approximated by

$$E_\epsilon[(t_R - T)^2] \approx \sum_S (\sigma_i^2/\pi_i^2)(1 - \pi_i).$$

The randomization expectation of the model variance of t_R is then

$$E_p\{E_\epsilon[(t_R - T)^2]\} \approx \sum_U (\sigma_i^2/\pi_i)(1 - \pi_i). \quad (6)$$

This quantity can be called the "anticipated variance" of t_R; that is, the model variance anticipated before random sampling . The term is due to Isaki and Fuller (1982), although equation (6) goes back considerably further in the literature. They use use it to mean $E_\epsilon\{E_p[(t_R - T)^2]\}$, what that model anticipates the randomization mean squared error to be. The expectation operators can be switched, and the two concepts of anticipated variance coincide, when ϵ_k and ϵ_k^2 are uncorrelated with I_k given $\mathbf{x}_k$ and $\mathbf{z}_k$, where $\sigma_k^2 = f(\mathbf{x}_k, \mathbf{z}_k)$, as we have assumed. This is weaker than the requirement that the ϵ_i and I_i be independent, as stated in Isaki and Fuller. Maintaining the latter condition would rule out designs where $\pi_k \propto \sigma_k$ for some hypothesized σ_k^2. This selection probability rule minimizes the asymptotic anticipated variance on the right hand side of equation (6) for a fixed expected sample size, $n^* = \sum_U \pi_i$. Brewer (1963) makes a similar point.

From equation (6), we can also see that the anticipated variance of the randomization-consistent regression estimator is (asymptotically) a function of the unit

selection probabilities but not the joint selection probabilities. Every design with the same unit selection probabilities produces an regression estimator with the same anticipated variance. If minimizing anticipated variance is the goal, then *there is no penalty from using Poisson sampling*.

4. Simultaneous Variance Estimation

It is a simple matter to estimate the (approximate) model variance of t_R expressed in equation (5):

$$v = \sum_S (a_i^2 - a_i) r_i^2, \tag{7}$$

where $r_i = y_i - \mathbf{x}_i\mathbf{b}$, and $\mathbf{b} = (\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} \sum_S c_k \pi_k^{-1} \mathbf{x}_k' y_k$. Now

$$r_i = \epsilon_i - \mathbf{x}_i(\mathbf{b} - \beta) = \epsilon_i - \mathbf{x}_i(\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} \sum_S c_k \pi_k^{-1} \mathbf{x}_k' \epsilon_k,$$

so

$$E(r_i^2) = \sigma_i^2 + 2\mathbf{x}_i(\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} c_i \pi_i^{-1} \mathbf{x}_i' \sigma_i^2 + \mathbf{x}_i(\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} (\sum_S c_k^2 \sigma_k^2 \pi_k^{-2} \mathbf{x}_k' \mathbf{x}_k)^{-1} (\sum_S c_k \pi_k^{-1} \mathbf{x}_k' \mathbf{x}_k)^{-1} \mathbf{x}_i',$$

After a little work, we can conclude that v is asymptotically model unbiased:

$$E_\epsilon(v) = \sum_S (a_i^2 - a_i)\sigma_i^2 + O_p(N^2/[n^*]^2). \tag{8}$$

Observe that the terms we are ignoring in equation (8) are smaller than the $O_p(N^2/[n^*]^{3/2}$ terms we would have ignored had we replaced a_i with π_i^{-1}.

We can likewise show that v is an asymptotically unbiased estimator for the randomization mean squared error of t_R under Poisson sampling. In this context, however, we are willing to drop $O_p(N^2/[n^*]^{3/2})$ terms. The equalities

$$r_i = e_i - \mathbf{x}_i(\mathbf{b} - \mathbf{B}) = e_i - \mathbf{x}_i(\textstyle\sum_S c_k\pi_k^{-1}\mathbf{x}_k'\mathbf{x}_k)^{-1}\sum_S c_k\pi_k^{-1}\mathbf{x}_k'e_k$$
$$= e_i - \mathbf{x}_i\,\mathbf{O}_{(q\,x\,q)p}(1/N)\mathbf{O}_{(q\,x\,1)p}(N/\downarrow n^*),$$
$$= e_i - O_p(1/\downarrow n^*) \qquad (9)$$

ultimately imply that $v = \sum_S (a_i^2 - a_i)r_i^2 = \sum_S (\pi_i^{-2} - \pi_i^{-1})e_i^2 + O_p(N^2/[n^*]^{3/2})$. From which, we conclude

$$E_p(v) = \sum_U (\pi_i^{-1} - 1)e_i^2 + O(N^2/[n^*]^{3/2}). \qquad (10)$$

The *relative* model bias of v (as an estimator of $E_\epsilon[(t_R - T)^2] \approx \sum_S a_i^2\sigma_i^2 - \sum_S a_i\sigma_i^2$) is $O_p(1/n^*)$; see equation (8). Its relative randomization bias (as an estimator of $E_p[(t_R - T)^2] \approx \sum_U (\pi_i^{-1} - 1)e_i^2$) is $O(1/[n^*]^{1/2})$; see equation (10). Empirical analyses like that in Wu and Deng (1983) have showed that this emphasis on the model bias can lead to superior coverage estimates.

4.1. Two Alternatives

The variance estimator v in equation (7) is very close to the weighted residual variance estimator of Särndal, Swensson, and Wretman (1989), which is $v_{SSW} = \sum_S a_i^2(1 - \pi_i)r_i^2$ for t_R under Poisson sampling. Like our simultaneous variance estimator, the weighted residual variance estimator was designed to estimate both model variance and randomization mean squared error with an emphasis on getting the model variance more-nearly unbiased.

Kott (1990) offers another estimator of both model variance and randomization mean squared error. The idea there is to multiply the traditional "randomization" mean squared error estimator, $v_{RB} = \sum_S (r_k/\pi_k)^2(1 - \pi_k)$ it this context, by $E_\epsilon[(t_R - T)^2]/E_\epsilon(v_{RB})]$. Kott's estimator is cumbersome to compute and requires the σ_i^2 be specified up to a constant. To my knowledge, it has never been used in practice.

4.2. An Example

The following simple example will clarify the differences among v, v_{SSW}, and v_{RB}. Suppose all the selection probabilities where equal, $\pi_k = \pi$, and all $\mathbf{x}_k = 1$. Let n be the realized sample size and N the population size. The expected sample size, n*, equals $N\pi$. When all the c_k are 1, a_k becomes N/n the randomization consistent regression estimator is $t_R = \sum_S (N/n)y_k$. This is exactly the same as the expansion estimator under simple random sampling given a fixed sample size n.

In our asymptotic environment, n* is large, and the probability that n is zero is itself zero. The variance estimator proposed here is $v = \sum_S (N/n)^2(1 - n/N)(y_k - \sum_S y_i/n)^2$. It is easy to show that $v_E = [n/(n-1)]v$ is model unbiased when all the σ_k^2 are equal. Moreover, the model unbiased variance estimator corresponds exactly the randomization estimator conditioned on the realized sample size n. This is the best variance estimator for confidence interval construction.

The traditional (unconditional) randomization mean squared error estimator for t_R under Poisson sampling is $v_{RB} = \sum_S (N/n^*)^2(1 - n^*/N)(y_k - \sum_S y_i/n)^2$. This estimator has an unfortunate property. As the sample size increases making t_R more accurate, v_{RB} gets larger rather than smaller. The near randomization unbiasedness of v_{RB} refers to an average squared difference between t and T taken all possible samples. For particular sample sizes, v_{RB} need not be very good.

In order-of-probability notation, $v = v_E (1 + O_P(1/n^*))$. In contrast, $v_{RB} = v_E(1 + O_P(1/\sqrt{n^*}))$. Treating v_E as the gold standard, v is easily seen to be superior to v_{RB}.

The weighted residual mean squared error estimator for t_R under Poisson sampling is $v_{SSW} = \sum_S (N/n)^2(1 - n^*/N)(y_k - \sum_S y_i/n)^2$. This is approximately equal to v when N is relatively large (i.e., $N = O([n^*]^{3/2})$) , but v_{SSW} can be noticeably different from v when n is not that close to n* and n/N is not negligible.

The example is extremely simple but it makes an important point. All three estimators estimate the unconditional randomization mean squared error adequately, but v does the best job at estimating the conditional randomization mean squared error.

Like all model-based variance estimators, it is conditioned on the realized sample not averaged over all possible samples. In the context of this example, it is a simple matter to construct a conditional randomization mean squared error estimator (it would be v). That is not the case in general. Moreover, although the literature on conditional randomization-based theory for regression estimators is growing, without a model one does not really know on what ancillary statistics to condition.

5. A Note on Adjusting for Small-sample Bias

It is a tempting to scale v in equation (7) by $n/(n - Q)$ to account for the fact that r_i^2, a squared residual from a Q-variate regression, is a slightly biased estimator for σ_i^2. This *ad hoc* adjustment would only be reasonable in our context when n is relatively small (technically, $O(n) = 1$), and either N is large (so that the model variance of t_R is approximately $\sum_S a_i^2\sigma_i^2$) or $\sigma_i^2 = \mathbf{x}_i\mathbf{h}$ for some $\mathbf{h}$ (so that $\sum_S a_i\sigma_i^2 - \sum_U \sigma_i^2$ can be still ignored even when n is small).

A better approach than the ad hoc adjustment of v would be to replace the r_i^2 with unbiased estimators for the components of $\sigma^2 = (\sigma_1^2, ..., \sigma_n^2)'$, namely, $\mathbf{r}_{(2)} = \mathbf{M}^{-1}(r_1^2, ..., r_n^2)'$, where the i,jth element of then n x n matrix $\mathbf{M}$ is $m_{ij} = [1 - \mathbf{x}_i(\sum_S c_k\pi_k^{-1}\mathbf{x}_k'\mathbf{x}_k)^{-1}c_j\pi_j^{-1}\mathbf{x}_j]^2$. See Chew (1970). Calculating $\mathbf{r}_{(2)}$ involves inverting an n x n matrix, but we are presuming n is relatively small. The *ad hoc* adjustment to v can also produce a model-unbiased variance estimator, but only if additional assumptions are made (e.g., when $c_k\pi_k^{-1}$ is constant across k, and σ_k^2 is likewise constant across k as in the example in Section 3.2).

6. Delete-a-group Variance Estimation

Many surveys have multiple variables of interest. The problem with v in equation (7) is that is requires r_k to be calculated separately for each such variable, even when a common regressor vector, $\mathbf{x}_k$, is employed. That is one reason why a *delete-a-group jackknife variance estimator* can prove helpful in practice. The term can be found in

Kott (2001), while the variance estimator itself in some form has long been used, not always with theoretical justification. A NASS research report, Kott (1998), discusses a wide variety of uses for the delete-a-group jackknife.

In this section, *we assume that finite-population correction can be ignored*. Formally, $1/N \leq O(1/[n^*]^2)$. This means when $O_P(N)$ terms are dropped, the model variance in equation (5) is approximately $V_0 = \sum_S a_i^2\sigma_i^2$, which is $O_p(N^2/n^*)$. Moreover, $v_0 = \sum_S a_i^2 r_i^2$ becomes asymptotically indistinguishable from the simultaneous variance estimator, v.

Let the Poisson sample be randomly divided into G replicate groups, denoted $S_1, S_2, ..., S_G$ (some groups can have one more member than others). The complement of each S_g is called the *jackknife replicate group* $S_{(g)} = S - S_g$. A sets of replicate weights is computed for each replicate group. For the gth set: $a_{i(g)} = 0$ when $i \in S_g$; and

$$a_{i(g)} = a_i + (\sum_U \mathbf{x}_k - \sum_{S(g)} a_k\mathbf{x}_k)(\sum_{S(g)} c_k a_k \mathbf{x}_k'\mathbf{x}_k)^{-1} c_i a_i \mathbf{x}_i' \qquad (11)$$

otherwise. The $a_{i(g)}$ have been computed to be reasonably close to the corresponding a_i for $i \in S(g)$ and to satisfy the calibration equation $\sum_S a_{k(g)}x_k = \sum_U x_k$ for all g. We return to equation (11) in the concluding section.

The delete-a-group variance estimator for t_R is :

$$v_J = (G - 1/G) \sum^G (\sum_S a_{i(g)}y_i - t_R)^2, \qquad (12)$$

which WESVAR (Westat 1997) calls JK1. This can be re-expressed as

$$v_J = (G - 1/G) \sum^G (\sum_S a_{i(g)}u_i - \sum_S a_i u_i)^2$$

where u_i may be either ϵ_i or e_i.

For ease of exposition, let us assume that n/G equals an integer, d. To do otherwise, complicates the subsequent formulae without adding insight.

Result are again asymptotic. We assume G grows arbitrarily large along with n*,

but allow the possibility that $O_p(G) < O(n^*)$. As a result, $O(d) \geq O(1)$. Nevertheless, O(d) must be less than O(n*) because $n = O_p(n^*)$ and $O_p(G = nd)$ must be greater than O(1).

The sets that S_g and $S_{(g)}$ can be viewed as simple random subsamples of S. With that in mind:

$$\begin{aligned}
\textstyle\sum_S a_{i(g)}u_i - \sum_S a_i u_i &= -\sum_{Sg} a_i u_i + (\sum_U \mathbf{x}_k - \sum_{S(g)} a_k \mathbf{x}_k)(\sum_{S(g)} c_k a_k \mathbf{x}_k' \mathbf{x}_k)^{-1} \sum_{S(g)} c_i a_i \mathbf{x}_i' u_i \\
&= -\sum_{Sg} a_i u_i + (\sum_{Sg} a_k \mathbf{x}_k)(\sum_{S(g)} c_k a_k \mathbf{x}_k' \mathbf{x}_k)^{-1} \sum_{S(g)} c_i a_i \mathbf{x}_i' u_i \\
&= -\sum_{Sg} a_i u_i + \{[d/n] \sum_S a_k \mathbf{x}_k + \mathbf{O}_{(1 \times q)p}(\downarrow d)\} \mathbf{O}_{(q \times q)p}(1/N) \sum_{S(g)} c_i a_i \mathbf{x}_i' u_i \\
&= -\sum_{Sg} a_i u_i + \mathbf{O}_{(1 \times q)p}(d/n^*) \sum_{S(g)} c_i a_i \mathbf{x}_i' u_i. \qquad (13)
\end{aligned}$$

Consequently,

$$\begin{aligned}
E_\epsilon[(\textstyle\sum_S a_{i(g)}\epsilon_i - \sum_S a_i \epsilon_i)^2] &= \sum_{Sg} a_i^2 \sigma_i^2 + \mathbf{O}_{(1 \times q)p}(d/n^*) \sum_{S(g)} c_i^2 a_i^2 \sigma_i^2 \mathbf{x}_i' \mathbf{x}_i \mathbf{O}_{(q \times 1)p}(d/n^*) \\
&= \sum_{Sg} a_i^2 \sigma_i^2 + O_p(N^2 d^2/[n^*]^3) \\
&= \sum_{Sg} a_i^2 \sigma_i^2 + O_p(G^{-2} N^2/n^*),
\end{aligned}$$

where the dominant term is $O_p(N^2 d/[n^*]^2) = O_p(G^{-1} N^2/n^*)$. From this, we can see that the delete-a-group is asymptotically model unbiased estimator for $V_0 = \sum_S a_i^2 \sigma_i^2$:

$$\begin{aligned}
E_\epsilon(v_J) &= ([G - 1]/G)\,[\textstyle\sum^G \sum_{Sg} a_i^2 \sigma_i^2 + O_p(G^{-2} N^2/n^*)] \\
&= \textstyle\sum_S a_i^2 \sigma_i^2\,[1 + O_p(1/G)]
\end{aligned}$$

Establishing the asymptotic randomization-based properties of v_J is a bit more difficult. From equation (13):

$$\begin{aligned}
\textstyle\sum_S a_{i(g)}e_i - \sum_S a_i e_i &= -\sum_{Sg} a_i e_i + \mathbf{O}_{(1 \times q)p}(d/n^*) \sum_{S(g)} c_i a_i \mathbf{x}_i' e_i. \\
&= -\sum_{Sg} a_i e_i + \mathbf{O}_{(1 \times q)p}(d/n^*)\{[n/(n - d)] \sum_S c_i a_i \mathbf{x}_i' e_i + \mathbf{O}_{(q \times 1)p}(N/\downarrow [n - d])\} \\
&= -\sum_{Sg} a_i e_i + \mathbf{O}_{(1 \times q)p}(1/G)\{[1 + O(1/G)] \sum_S c_i \pi_i^{-1} \mathbf{x}_i' e_i + \mathbf{O}_{(q \times 1)p}(N/\downarrow n^*)\} \\
&= -\sum_{Sg} a_i e_i + \mathbf{O}_{(1 \times q)p}(1/G)\,\{[1 + O_p(1/G)] \sum_U c_i \mathbf{x}_i' e_i + \mathbf{O}_{(q \times 1)p}(N/\downarrow n^*)\}
\end{aligned}$$

$$= -\sum_{Sg} a_i e_i + O_p(G^{-1} N/\sqrt{n^*}), \qquad (14)$$

where the dominant term is $O_p(dN/n^*) = O_p(G^{-1} N)$.

Combining equations (12) and (14):

$$\begin{aligned} v_J &= ([G-1]/G) \sum^G (\sum_S a_{i(g)} e_i - \sum_S a_i e_i)^2 \\ &= ([G-1]/G) \sum^G [-\sum_{Sg} a_i e_i + O_p(G^{-1} N/\sqrt{n^*})]^2 \\ &= ([G-1]/G) [\sum^G (\sum_{Sg} a_i e_i)^2 + O_p(G^{-1} N^2/\sqrt{n^*})] \\ &= \sum^G (\sum_{Sg} a_i e_i)^2 + O_p(N^2/G^2) + O_p(G^{-1} N^2/\sqrt{n^*}). \qquad (15) \end{aligned}$$

We now turn our attention to the randomization expectation of v_J under the random subsamplings of sample S. *We need an addition assumption;* namely, $c_i = 1/(\mathbf{x}_i\gamma)$ for some vector γ. Under this assumption, $\sum_U e_i = \sum_U \gamma'\mathbf{x}_i' c_i e_i = \gamma' \sum_U c_i \mathbf{x}_i' e_i = 0$

The replicate group S_g can be viewed as a random subsample of S. In fact, dropping $O_p(N^2/G^2)$ and $O_p(G^{-1} N^2/\sqrt{n^*})$ terms, equation (15) implies that for any g, $E_2(v_J) \approx E_2(w_g^2/G)$, where $w_g = G\sum_{Sg} a_i e_i = (n/d)\sum_{Sg} a_i e_i$, and the subscript 2 refers to the subsampling.

Now

$$\begin{aligned} E_2(w_g^2) &= Var_2(w_g) + [E_2(w_g)]^2 \\ &= \{n^2/[d(n-1)]\}[\sum_S (a_i e_i)^2 - (\sum_S a_i e_i)^2/n] + (\sum_S a_i e_i)^2 \\ &= G(n/[n-1])\sum_S (a_i e_i)^2 + (1 - n/[d(n-1)]) (\sum_S a_i e_i)^2 \\ &= G(n/[n-1])\sum_S (a_i e_i)^2 + (1 - n/[d(n-1)]) (\sum_U e_i + O_p(N/\sqrt{n^*}))^2 \\ &= G \sum_S (a_i e_i)^2 + O_p(N^2/n^*) . \qquad (16) \end{aligned}$$

From which we can conclude $E_2(v_J) = \sum_S (a_i e_i)^2 + O_p(G^{-1}N^2/n^*)$, which is asymptotically indistinguishable from v_0.

From the derivation of equation (16), we see that when $c_i \neq 1/(\mathbf{x}_i\gamma)$, so that $\sum_S a_i e_i \neq O_p(N/\sqrt{n^*})$, v_J can have an upward bias as an estimator of the randomization mean squared error of t_R.

7. Discussion

Let us first review the main results discussed so far.

Under the linear model in equation (4), the regression estimator, t_R, in equation (1) has the same (asymptotic) anticipated variance under Poisson sampling as it does under any sampling design having the same set of first-order selection probabilities (equation (6)). "Anticipated variance" here means the randomization expectation of the model variance, but under mild conditions that value is identical to the model expectation of the randomization variance.

When the expected sample size, n*, is large, *both* the model variance and randomization mean squared error of t_R under Poisson sampling can be estimated by $v = \sum_S (a_i^2 - a_i) r_i^2$. Under mild asymptotic assumptions (in equation (3)), the relative bias of v as an estimator of model variance is $O_p(1/n^*)$ if N is relatively large (i.e., $O([n^*]^{3/2})$ or if $\sigma_i^2 = \mathbf{x}_i\mathbf{h}$ for some not-necessarily-specified $\mathbf{h}$. The relative bias of v as an estimator of randomization mean squared error is $O(1/[n^*]^{½})$.

When n is not large, randomization-based properties lose much of their relevance. Section 4 describes a method of modifying v to produce an unbiased estimator for the model variance of t_R when $\sigma_i^2 = \mathbf{x}_i\mathbf{h}$. This method is also effective when N is large, and the model variance asymptotically approximated by $V_O = \sum_S a_i^2\sigma_i^2$.

When both n and N are large, but N is so large that finite-population correction is ignorable ($N \geq O([n^*]^2)$, the delete-a-group jackknife variance estimator (equation (12)) can be used to estimate both the model variance and randomization mean squared error of t_R provided that G, the number of jackknife replicate groups, is also large. The asymptotic unbiasedness of the latter requires on additional assumption: $c_i = 1/(\mathbf{x}_i\gamma)$ for some vector γ. *The* γ *need bear no relationship to the* ***h*** *discussed above.* The simplest way for this assumption to be satisfied is for c_i to be set equal to 1 and $\mathbf{x}_i$ to contain unity as one of its components.

Several observations are in order.

The choice of c_k in equation (1) has no impact on the anticipated variance expressed in equation (6). Consequently, there is no reason to prefer on set of c_k over

another on efficiency grounds. A secondary concern is that the a_k all be bounded from below by unity. Following Brewer (1994), this suggests the choice $c_k = 1 - \pi_k$ (although this limits the possibility that some $a_i < 1$, it does not remove it). When N is relatively large, the resultant estimator is asymptotically indistinguishable from that when $c_k = 1$.

The variance of any model-unbiased weighted estimator ($t_R = \sum_S a_k y_k$, where $\sum_S a_k \mathbf{x}_k = \sum_U \mathbf{x}_k$) can be estimated with v. This variance estimator is asymptotically unbiased when $\sigma_i^2 = \mathbf{x}_i\mathbf{h}$. A strictly model-unbiased estimator can be computed following the suggestion in Section 4. In addition, when finite-population correction is ignorable, the delete-a-group jackknife, v_J, with replicate weights calculated using equation (11) is an asymptotically model-unbiased variance estimator provided that G is large.

It is also tempting to use v or v_J to estimate the randomization mean squared error for a randomization consistent t_R based on a non-Poisson sample even though there may be a nonignorable randomization bias. This is especially true when there are $O[(n^*)^2]$ cross terms in a plug-in mean-squared-error estimator such as $v^* = v + \sum_{i,i\in S\ (i\neq j)} [(\pi_{ij} - \pi_i\pi_j)/\pi_{ij}](r_i/\pi_i)(r_j/\pi_j)$. In that case, the asymptotic randomization unbiasedness of v* has not been clearly established except under strong conditions. See Breidt and Opsomer (2000). Moreover, with so many terms, the variance of v* itself becomes an issue. The potential bias in v may not be as much a practical concern as the variance of v*.

Finally, observe that just as equation (11) can be used to compute replicate weights for any model-unbiased estimator, it can also be used to compute replicate weights for a randomization-consistent, model-unbiased estimator when

$$a_k = \pi_k^{-1} + (\sum_U \mathbf{x}_i - \sum_S \pi_i^{-1}\mathbf{x}_i)(\sum_S c_i\pi_i^{-1}\mathbf{x}_i'\mathbf{x}_i)^{-1} c_k\pi_k^{-1}\mathbf{x}_k' + O_p(1/n). \qquad (17)$$

The resultant delete-a-group jackknife will have the usual model and randomization-based properties. Deville and Särndal (1992) and Singh and Mohl (1998) discuss a number of calibration estimators with weights satisfying equation (17).

8. References

Breidt, F.J. and Opsomer, J. D. (2000). Local polynomial regression in survey sampling. Annals of Statististics, 28, 1026-1053.

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 93-105.

Brewer, K.R.W. (1994). Survey sampling inference: some past perspectives and present prospects. Pakistan Journal of Statistics, 10(1)A 213-233.

Chew, V. (1970). Covariance matrix estimation in linear models. Journal of the American Statististical Association, 65 173-181.

Deville, J-C. and Särndal, C-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical. Association, 87 376-382.

Isaki, and Fuller, W.A. (1982). Survey design under the regression superpopulation model. Journal. American. Statistical. Association, 77, 89-96.

Kott, P.S. (1990). Estimating the conditional variance of a design consistent regression estimator. Journal of Statistical Planning and Inference, 24 287-296.

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. (2001). The delete-a-group jackknife. Journal of Official Statistics, forthcoming.

Kott, P.S. and Bailey, J. T. The theory and practice of maximal Brewer selection. Proceedings of the Second International Conference on Establishment Surveys, Invited papers, 269-278.

Särndal, C-E, Swensson, B., and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of a finite population total. Biometrika, 76, 527-537,

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

Singh, A.C. and Mohl, C.A. (1996). Understanding calibration estimators in survey sampling. Survey. Methodology, 22 2, 107-115.

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

Wu, C.F.J. and Deng, L.Y. (1983). Estimation of variance of the ratio estimator: an empricial study. In Scientific Inference, Data Analysis and Robustness, (Eds. G.E.P. Box, *et al*.) New York: Academic Press, 245-277.