Recent Developments in Multiple Frame Surveys

Sharon Lohr[1]
Department of Mathematics and Statistics, Arizona State University, Tempe AZ 85287-1804

Abstract

With increasing demographic and technological diversity, it is becoming more difficult for a single sample selected from a single sampling frame to adequately represent the population. Multiple frame surveys are increasingly used in situations where several sampling frames may provide better coverage or cost-efficiency for estimating population quantities of interest. Examples include combining a list frame of farms with an area frame or using two frames to sample landline telephone households and cellular telephone households. We review the history of multiple frame surveys including some of J.N.K. Rao's many contributions to the subject. We then discuss some recent work on internally consistent and efficient estimators for three or more frames, and resampling methods for variance estimation in multiple frame surveys. Connections between multiple frame surveys and Rao's contributions to other areas of statistics are discussed.

Keywords: Bootstrap, Complex Survey, Dual Frame Survey, Jackknife, Sampling for Rare Events, Variance Estimation.

1. Introduction

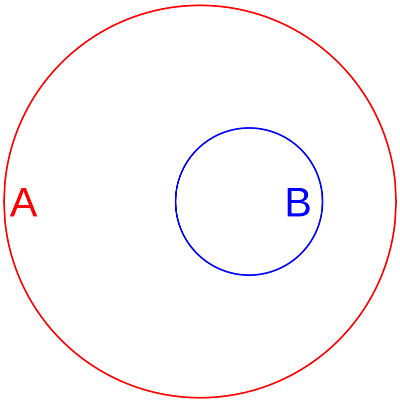
It is a privilege to be invited to participate in this session honoring J.N.K. Rao's contributions to statistics on the occasion of his 70th birthday. The most difficult task for me was deciding on one topic to focus on, since he has contributed to so many areas of statistics. I decided to speak about his contributions to multiple frame surveys, partly because I have direct knowledge of his contributions in this area, and partly because his work in multiple frame surveys goes far beyond the immediate topic of multiple frame surveys. Like so many of Rao's contributions to statistics, his multiple frame survey work has implications for many areas, including survey design, small area estimation, empirical likelihood, computer-intensive inference, misclassification, measurement errors, calibration, and imputation.

First, what is a multiple frame survey? In classical sampling theory, there is one sampling frame. This frame can be a list of sampling units, or a set of geographic regions, or even a sequential procedure specifying how units are to be located and selected. A probability sample is taken from the frame, and the inclusion probabilities in the sampling design can be used to make inferences about the population in the sampling frame. Let y_i be a measurement on unit i in the population of N units, let $\mathcal{S}$ denote the set of units in the sample, and let $\pi_i = P(\text{unit } i \text{ is included in the sample})$. Then the Horvitz-Thompson estimator of the population total $Y = \sum_{i=1}^{N} y_i$ is

$$\hat{Y} = \sum_{i \in \mathcal{S}} w_i y_i,$$

where $w_i = 1/\pi_i$ is the sampling weight.

Figure 1: Frame B is a subset of Frame A.



In many cases, however, a frame that covers the entire population is very expensive to sample from. An alternate frame may be available that does not cover the entire population but is cheaper to sample from. For example, in an agricultural survey on rice, an area frame would include all farms that produce rice but would be expensive to sample; in addition, relatively few of the farms sampled would be rice producers (Fecso et al., 1986). A list frame, providing the contact information for known rice producers, will be inexpensive to sample from but most likely does not include all farms that produce rice. In a dual frame survey, independent probability samples are taken from frame A (the area frame) and frame B (the list frame); this is depicted in Figure 1.

Rare populations can often be sampled more efficiently using a multiple frame sample (Kalton and Anderson, 1986). In an epidemiologic study, for example, frame A might be that used for a general population health survey, while frame B might be a list frame of clinics specializing in a certain disease.

In other situations, all frames are incomplete; for example, frame A in Figure 2 might be a frame of landline telephones and frame B might consist of cellular telephone numbers (Tucker et al., 2007). It is unknown in advance whether a household member sampled using one frame also belongs to the other frame (Brick et al., 2006).

Figure 2: Overlapping frames for land and cell telephones. Tucker et al. (2007) estimated that 46.4% of households have only landlines, 6% have only cell phones, 42.2% have both, and 5.4% have neither.

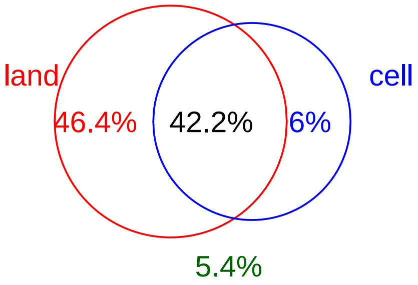
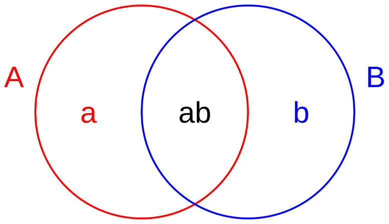


Figure 3: Overlapping frames A and B and three domains.



The general situation for two overlapping frames is displayed in Figure 3. There are three domains: domain a consists of units in frame A but not in frame B, domain b consists of units in frame B but not in frame A, and domain ab consists of units in both frames.

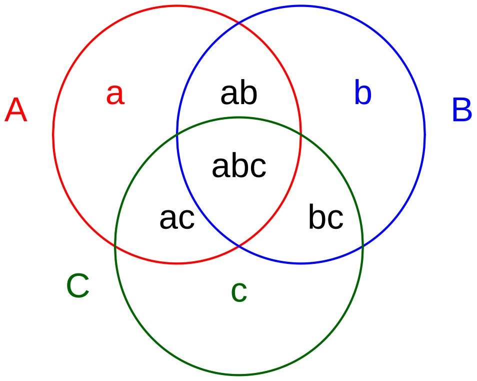
More than two frames can be employed as well, as illustrated in Figure 4 for a three-frame survey in which all frames are incomplete. In this situation, there are seven domains. Iachan and Dennis (1993) give an example of a three-frame survey used to sample the homeless population, where frame A is a list of soup kitchens, frame B is a list of shelters, and frame C consists of street locations. To enable reliable estimation of health characteristics for California residents of Vietnamese or Korean ethnicity, the California Health Interview Survey supplements a random digit dialing sample with samples from lists of households with surnames thought to belong to those ethnic groups (Cervantes and Brick, 2007).

One goal in analyzing data from a multiple frame survey is often to estimate the population total Y, using information from independent samples taken from the frames. In a dual frame survey, we can write

$$Y = Y_a + Y_{ab} + Y_b,$$

where Y_a is the total of the population units in domain a, Y_{ab} is the total of the population units in domain ab, and Y_b is the total of the population units in domain b.

Figure 4: Frames A, B, and C are all incomplete and overlap.



A special case of this is estimating the population size

$$N = N_a + N_{ab} + N_b,$$

as discussed in Haines and Pollock (1986). As multiple frame surveys become more prevalent, however, the goals expand to include estimation of general population characteristics, fitting models thought to describe the superpopulation, and employing multiple frame surveys in small area estimation. In Sections 2 and 3 we review early uses of and point estimators for dual frame surveys. In Section 4, we discuss variance estimation, and introduce two bootstrap methods—developed in joint research with Rao—for constructing interval estimates from multiple frame surveys. Section 5 outlines some connections between multiple frame surveys and other problems in statistics.

2. Some History of Multiple Frame Surveys

Hansen, Hurwitz, and Madow (1953) describe what is often considered to be one of the earliest examples of a dual frame survey. The Sample Survey of Retail Stores was conducted by the U.S. Census Bureau in 1949. In this survey, a probability sample of primary sampling units (psus) was chosen. Within each psu, a census of retail firms on a list compiled from the records of the Old Age and Survivors Insurance Bureau was taken; and an area sample was taken of firms not on the list. In this case, a *screening* dual frame design was employed within each selected psu, so called because units in the list frame were screened out of the area frame before sampling. Thus, the estimator of total sales summed the two estimators within each psu—in essence, a screening dual frame survey is a stratified sample, in which frame A is one stratum and frame B is the second stratum. To my knowledge, Rao was not involved in the design of this survey.

However, Rao has been involved with many of the subsequent developments in multiple frame surveys. Rao completed his Ph.D. in statistics in 1961 from Iowa State University with advisor H. O. Hartley. Hartley worked through the theory of dual frame estimators during that time with Rao and Jack Graham, carefully drawing the Venn diagrams on the board with colored chalk. Hartley (1962) proposed estimators for the general dual-frame situation depicted in Figure 3, with results for Figure 1 following as a special case. He used a weighted average of the estimators in the overlap domain ab, with

$$\hat{Y}(\theta) = \hat{Y}_a^A + \theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b^B, \qquad (1)$$

where $\hat{Y}_a^A$ is the estimated population total for units in domain a, $\hat{Y}_{ab}^A$ is the estimated population total in domain ab using the sample from frame A, $\hat{Y}_{ab}^B$ is the estimated population total in domain ab using the sample from frame B, $\hat{Y}_b^B$ is the estimated population total for domain b, and $0 \le \theta \le 1$.

Hartley (1962, 1974) proposed choosing θ in (1) to minimize the variance of $\hat{Y}_H(\theta)$. Because the frames are sampled independently, the variance of $\hat{Y}_H(\theta)$ is

$$V[\hat{Y}(\theta)] = V[\hat{Y}_a + \theta\hat{Y}_{ab}^A] + V[(1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b^B].$$

Thus, for general survey designs, the variance-minimizing value of θ is

$$\theta_{\text{opt}} = \frac{V(\hat{Y}_{ab}^B) + \text{Cov}\,(\hat{Y}_b^B, \hat{Y}_{ab}^B) - \text{Cov}\,(\hat{Y}_a^A, \hat{Y}_{ab}^A)}{V(\hat{Y}_{ab}^A) + V(\hat{Y}_{ab}^B)}. \qquad (2)$$

Note that if one of the covariances in (2) is large, it is possible for θ_{opt} to be smaller than 0 or greater than 1. When frame A and frame B are the same, i.e., domains a and b are empty, however, θ_{opt} is between 0 and 1.

Hartley (1974) referred several times to a personal communication from Rao, who derived maximum likelihood estimators for dual frame surveys using the scale-load approach pioneered in Hartley and Rao (1968). See Rao (1983) for a brief description of these methods, which he had presented at the 1973 International Statistical Institute meeting in Vienna. Rao (1983) derived the estimator

$$\hat{Y} = N_a\bar{y}_a + N_b\bar{y}_b + N_{ab}\bar{y}_{ab}$$

where $\bar{y}_{ab}$ is the mean of $n_{ab}^A + n_{ab}^B - d$ distinct units and d is the number of units in both samples, using maximum likelihood. He also showed that $\hat{Y}$ is the posterior expected value under a noninformative prior distribution. In fact, Rao, as co-editor of *Sankhyā* Series C in 1974, was the person who encouraged Hartley to submit the 1974 paper to *Sankhyā* based on his work in multiple frame surveys. Hartley (1974) was later reprinted in the IASS Jubilee Commemorative Volume *Landmark Papers in Survey Statistics* as one of the nineteen papers selected for publication in that volume. (Another of the papers in the *Landmark Papers* volume was Rao and Scott (1981), to which we shall return in Section 5.)

One can argue that Rao has worked on dual frame surveys for his entire career. One finds dual frame ideas early in Rao's work. Rao and Graham (1964) developed composite estimators for a rotation sample. Their estimator for the population mean of a characteristic of interest for the current month is

$$\bar{y}_0' = Q(\bar{y}_{-1}{}' + \bar{d}) + (1-Q)\bar{y}_0,$$

where $\bar{y}_0$ is the estimator for the current month, $\bar{d}$ estimates the difference between the current and previous months using units measured at both times, and $\bar{y}_{-1}'$ is the composite estimator for the previous month. You can see the basic features of a dual frame estimator here: The two frames are those of the current month and the previous month and the overlap is used to improve estimation of the population mean of interest.

Rao (1968) studied a dual frame survey of beef cattle producers, where a list frame of persons thought to be producers was combined with an area frame. One problem with the list frame is that some of the persons listed were in partnership with other persons on the list; such partnerships had a higher probability of being selected. Respondents were asked to list all the persons in their operation, and reweighted to reflect the multiplicity.

Graham and Rao (1978), in an MAA volume intended to introduce mathematicians to important aspects of statistics, wrote a paper summarizing the state of survey sampling. They discussed multiple frame surveys in Section 9 on recent developments in sampling, and discussed their potential for improving survey practice. This paper was one of the first review papers to note the importance of multiple frame surveys in sampling, and it presaged a number of later developments in the area.

3. Estimating Population Quantities

Many estimators have been proposed for estimating population totals and other quantities. In this section, we look at optimal estimators and then the pseudo-maximum-likelihood estimator developed by Rao and collaborators.

3.1 Optimal Estimators

Hartley's (1962, 1974) estimator is optimal among all estimators of the form $\hat{Y}_a^A + \hat{Y}_b^B + \theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B$. Fuller and Burmeister (1972) proposed modifying Hartley's estimator by incorporating additional information regarding the estimation of N_{ab}. The estimator is:

$$\hat{Y}_{FB}(\beta) = \hat{Y}_a^A + \hat{Y}_b^B + \beta_1\hat{Y}_{ab}^A + (1-\beta_1)\hat{Y}_{ab}^B + \beta_2(\hat{N}_{ab}^A - \hat{N}_{ab}^B). \qquad (3)$$

Rao (1983) and Skinner (1991) showed that $\hat{Y}_{FB}$ can be derived from maximum likelihood principles when a simple random sample is taken in each frame. As with Hartley's estimator, the parameters β_1 and β_2 are chosen to minimize the variance of $\hat{Y}_{FB}(\beta)$; the optimal values are

$$\begin{bmatrix} \beta_{1,opt} \\ \beta_{2,opt} \end{bmatrix} = -\text{Cov}\begin{bmatrix} \hat{Y}^A_{ab} - \hat{Y}^B_{ab} \\ \hat{N}^A_{ab} - \hat{N}^B_{ab} \end{bmatrix}^{-1} \begin{bmatrix} \text{Cov}\,(\hat{Y}^A_a + \hat{Y}^B_b + \hat{Y}^B_{ab}, \hat{Y}^A_{ab} - \hat{Y}^B_{ab}) \\ \text{Cov}\,(\hat{Y}^A_a + \hat{Y}^B_b + \hat{Y}^B_{ab}, \hat{N}^A_{ab} - \hat{N}^B_{ab}) \end{bmatrix} \quad (4)$$

In practice, the covariances used in (2) and (4) are unknown, so the optimal values of the parameters must be estimated from the data. Let $\hat{\theta}_{\text{opt}}$ be the estimator of θ_{opt} that results when estimates of the covariances are substituted into (2).

Rao has long viewed survey estimators in terms of weights and, in fact, his approach to dual frame estimation through weight modifications is reminiscent of his work on calculating jackknife and bootstrap variance estimators through modifying the weight vectors. Skinner and Rao (1996) wrote the optimal estimators in terms of weight modifications in addition to the representation as linear combinations of the estimated domain totals. The weight of each sampled unit in the intersection domain ab is reduced to compensate for the multiplicity. Let $\delta_i(a) = 1$ if unit i is in domain a and 0 otherwise, and define $\delta_i(ab)$ and $\delta_i(b)$ similarly. The adjusted weights for Hartley's method become

$$\tilde{w}^A_{i,H} = \delta_i(a)w^A_i + \hat{\theta}_{\text{opt}}\delta_i(ab)w^A_i$$

and

$$\tilde{w}^B_{i,H} = \delta_i(b)w^B_i + (1 - \hat{\theta}_{\text{opt}})\delta_i(ab)w^B_i.$$

3.2 Pseudo-Maximum-Likelihood Estimation

Skinner and Rao (1996) pointed out that since $\hat{\theta}_{\text{opt}}$ depends on the covariances of the particular response studied, the weight adjustments may differ for each response studied. This can lead to inconsistencies among estimates. For example, suppose $\hat{Y}_1(\hat{\theta}_{\text{opt},1})$ estimates total medical expenses in the population over age 65, $\hat{Y}_2(\hat{\theta}_{\text{opt},2})$ estimates total medical expenses in the population aged 65 or less, and $\hat{Y}_3(\hat{\theta}_{\text{opt},3})$ estimates total medical expenses in the entire population. If the surveys have complex design, it is likely that $\hat{Y}_1(\hat{\theta}_{\text{opt},1}) + \hat{Y}_2(\hat{\theta}_{\text{opt},2}) \neq \hat{Y}_3(\hat{\theta}_{\text{opt},3})$.

Skinner and Rao (1996) proposed modifying the simple random sample estimator to obtain a pseudo-maximum-likelihood (PML) estimator for a complex design. The PML estimator uses the same set of weights for all response variables and has the form

$$\hat{Y}_{PML}(\theta) = \frac{N_A - \hat{N}^{PML}_{ab}(\theta)}{\hat{N}^A_a}\hat{Y}^A_a + \frac{N_B - \hat{N}^{PML}_{ab}(\theta)}{\hat{N}^B_b}\hat{Y}^B_b + \frac{\hat{N}^{PML}_{ab}(\theta)}{\theta\hat{N}^A_{ab} + (1-\theta)\hat{N}^B_{ab}}[\theta\hat{Y}^A_{ab} + (1-\theta)\hat{Y}^B_{ab}]. (5)$$

where $\hat{N}^{PML}_{ab}(\theta)$ is the smaller of the roots of the quadratic equation

$$[\theta/N_B + (1-\theta)/N_A]x^2 + \theta\hat{N}^A_{ab} + (1-\theta)\hat{N}^B_{ab} - [1 + \theta\hat{N}^A_{ab}/N_B + (1-\theta)\hat{N}^B_{ab}/N_A]x = 0.$$

Skinner and Rao (1996) suggested using the value θ_P that minimizes the asymptotic variance of $\hat{N}^{PML}_{ab}(\theta)$:

$$\theta_P = \frac{N_a N_B V(\hat{N}^B_{ab})}{N_a N_B V(\hat{N}^B_{ab}) + N_b N_A V(\hat{N}^A_{ab})}. \quad (6)$$

The estimator in (5) adjusts the estimators of the three domain totals Y_a, Y_{ab}, and Y_b by the optimal estimator of N_{ab}.

In practice, N_a, N_b, $V(\hat{N}^A_{ab})$, and $V(\hat{N}^B_{ab})$ are estimated from the data so that an estimator $\hat{\theta}_P$ of θ_P is substituted into (5). The adjusted weights are

$$\tilde{w}^A_{i,P} = \begin{cases} \dfrac{N_A - \hat{N}^{PML}_{ab}(\hat{\theta}_P)}{\hat{N}^A_a}\, w^A_i & \text{if } i \in a \\ \dfrac{\hat{N}^{PML}_{ab}(\hat{\theta}_P)}{\hat{\theta}_P\hat{N}^A_{ab} + (1-\hat{\theta}_P)\hat{N}^B_{ab}}\hat{\theta}_P\, w^A_i & \text{if } i \in ab \end{cases}$$

and

$$\tilde{w}^B_{i,P} = \begin{cases} \dfrac{N_B - \hat{N}^{PML}_{ab}(\hat{\theta}_P)}{\hat{N}^B_b}\, w^B_i & \text{if } i \in b \\ \dfrac{\hat{N}^{PML}_{ab}(\hat{\theta}_P)}{\hat{\theta}_P\hat{N}^A_{ab} + (1-\hat{\theta}_P)\hat{N}^B_{ab}}(1-\hat{\theta}_P)\, w^B_i & \text{if } i \in ab \end{cases}$$

Although $\hat{\theta}_P$ depends on the estimated variances of the overlap domain size, it does not depend on covariances of other response variables. The PML estimator thus uses the same set of weights for each response variable. Lohr and Rao (2006) extended the PML approach to more than two frames. Skinner and Rao (1996), Rao and Skinner (1999), and Lohr and Rao (2006) found that the PML estimator has small mean squared error and works well in a wide variety of survey designs.

4. Variance Estimation

For screening dual frame surveys, variance estimation is straightforward: standard methods for stratified samples can be used to estimate variances. Variance estimation can be more complicated for other estimators. The adjusted weights for the Hartley estimator of the population total depend on $\hat{\theta}_{\text{opt}}$, which is a function of the estimated covariances from both frames. Functions of totals, or other statistics such as percentiles, also rely in a more complex way on estimators from both samples. Several methods can be used to estimate variances of estimated population quantities in general multiple frame surveys. These methods include Taylor linearization techniques, jackknife, and bootstrap.

4.1 Linearization and Jackknife Methods

The Taylor linearization and jackknife methods, discussed in Lohr and Rao (2000), assume that a population characteristic of interest τ can be expressed as a twice continuously differentiable function of population totals from the frames. For Taylor linearization, the partial derivatives of this function are used together with the estimated covariance matrix of the population totals estimated from frame A, and the estimated covariance matrix of the population totals estimated from frame B, to give a linearized estimator of the variance of the estimator $\hat{\tau}$. For example, $\tau = Y/X$ might be a ratio of two population totals from a dual frame survey, with

$$\hat{\tau} = \frac{\hat{Y}(\frac{1}{2})}{\hat{X}(\frac{1}{2})} = \frac{\hat{Y}_a^A + \frac{1}{2}\hat{Y}_{ab}^A + \frac{1}{2}\hat{Y}_{ab}^B + \hat{Y}_b^B}{\hat{X}_a^A + \frac{1}{2}\hat{X}_{ab}^A + \frac{1}{2}\hat{X}_{ab}^B + \hat{X}_b^B},$$

for $\hat{Y}(\frac{1}{2})$ and $\hat{X}(\frac{1}{2})$ as defined in (1). The estimated totals from frame A are $(\hat{Y}_a^A, \hat{Y}_{ab}^A, \hat{X}_a^A, \hat{X}_{ab}^A)$ with estimated covariance matrix S_A, and the estimated totals from frame B are $(\hat{Y}_b^B, \hat{Y}_{ab}^B, \hat{X}_b^B, \hat{X}_{ab}^B)$ with estimated covariance matrix S_B. The linearization estimator of the variance is then

$$g_A^T S_A g_A + g_B^T S_B g_B,$$

where $g_A^T = g_B^T = [\hat{X}(1/2)]^{-1}(1, 1/2, -\hat{\tau}, -\hat{\tau}/2)^T$ for this example comes from the vector of derivatives used in the linearization. Under regularity conditions, Skinner and Rao (1996) showed that the linearization estimator of the variance is consistent. It requires, however, that the derivatives be calculated separately for each estimator that is considered.

Demnati et al. (2007) derived linearization estimators of the variance by taking derivatives of a function of the weights rather than of the means. These are similar to the linearization framework in Demnati and Rao (2004), but allow for multiple frames.

The jackknife estimator of the variance relies on the property that independent samples are taken from the two frames (Lohr and Rao, 2000). Suppose a stratified cluster sample is taken from frame A, and an independent stratified cluster sample is taken from frame B. A jackknife variance estimator carries out the jackknife separately in frames A and B. Let $\hat{\tau}_{(hi)}^A$ be the estimator of the same form as $\hat{\tau}$ when the observations of sample psu i of stratum h from the frame-A sample are omitted from the data. Similarly, let $\hat{\tau}_{(lj)}^B$ be the estimator of the same form as $\hat{\tau}$ when the observations of sample psu j of stratum l from the frame-B sample are omitted. Then, if $\tilde{n}_h^A$ is the number of primary sampling units in stratum h of the sample in frame A, and $\tilde{n}_l^B$ is the number of primary sampling units in stratum l of the sample in frame B, the jackknife estimator of the variance is

$$v_J(\hat{\tau}) = \sum_{h=1}^{H} \frac{\tilde{n}_h^A - 1}{\tilde{n}_h^A} \sum_{i=1}^{\tilde{n}_h^A} (\hat{\tau}_{(hi)}^A - \hat{\tau})^2 + \sum_{l=1}^{L} \frac{\tilde{n}_l^B - 1}{\tilde{n}_l^B} \sum_{j=1}^{\tilde{n}_l^B} (\hat{\tau}_{(lj)}^B - \hat{\tau})^2. \qquad (7)$$

The jackknife estimator of the variance is consistent for smooth functions of population means.

The jackknife estimator of the variance has many advantages, but cannot necessarily be used for statistics such as medians. A confidence interval for a population quantity τ is calculated using the jackknife as $\hat{\tau} \pm t\sqrt{v_J(\hat{\tau})}$. In addition, the number of replicate weights needed for the jackknife is fixed at $\sum_h \tilde{n}_h^A + \sum_l \tilde{n}_l^B$. If one of the designs is a simple random or stratified random sample, the number of replicates required for the jackknife can be very large.

4.2 Bootstrap

A bootstrap estimator of the variance can be more flexible than the jackknife, since it can work with nonsmooth functions and the number of bootstrap iterations is determined by the user. In a single frame survey, the rescaling bootstrap of Rao and Wu (1988) works as follows: Suppose stratum h has $\tilde{n}_h$ primary sampling units. Sample $m_h = \tilde{n}_h - 1$ psu's from the psu's in stratum h using simple random sampling with replacement. Let $m_{hi}(b)$ be the number of times psu i of stratum h is selected in bootstrap sample b. Then the bootstrap weights for unit k within stratum h and psu i for bootstrap sample b are

$$w_{hik}(b) = w_{hik} \frac{\tilde{n}_h}{m_h} m_{hi}(b).$$

In this section, we present two bootstrap methods for multiple frame surveys that have been developed in joint research with Rao: a *separate* bootstrap, in which the bootstrap is applied separately to the samples from frames A and B, and a *combined* bootstrap, in which the psu's from both frames are resampled together. We present the methods here for dual frame surveys; the bootstrap works similarly with more than two frames.

First, let's define the bootstrap weights for the frames. In frame A, for bootstrap sample b we sample $\tilde{n}_h^A - 1$ psu's with replacement from stratum h and define $w_{hik}^A(b) = [\tilde{n}_h^A/(\tilde{n}_h^A - 1)] m_{hi}^A(b) w_{hik}^A$, where $m_{hi}^A(b)$ is the number of times psu i of stratum h is selected in the bootstrap sample. Similarly, for frame B we sample $\tilde{n}_l^B - 1$ psu's with replacement from stratum l and define $w_{ljk}^B(b) = [\tilde{n}_l^B/(\tilde{n}_l^B - 1)] m_{lj}^B(b) w_{ljk}^B$.

We can express the estimator $\hat{\tau}$ as a function of the weights for the samples from frames A and B. To ease the notation, we write $\mathbf{w}^A$ to be the vector of w_{hik}^A weights from frame A, and $\mathbf{w}^B$ to be the vector of w_{ljk}^B weights from frame B. Note that $\mathbf{w}^A$ and $\mathbf{w}^B$ are the original weights for the two frames, before any of the adjustments for multiplicity outlined in Section 3. Following Demnati and Rao (2004), who viewed linearization variance estimators as a function of the weights, we express

$$\hat{\tau} = h(\mathbf{w}^A, \mathbf{w}^B)$$

as a function h of the two vectors of weights. To calculate bootstrap estimates, then, we substitute the bootstrap weights for iteration b for the original vector of weights. We consider three bootstrap estimators: $\hat{\tau}^{*A}(b)$ and $\hat{\tau}^{*B}(b)$ replace the original weights by the bootstrap weights for just one of the frames, while $\hat{\tau}^{*}(b)$ replaces both sets of weights.

$$\hat{\tau}^{*A}(b) = h(\mathbf{w}^A(b), \mathbf{w}^B)$$

$$\hat{\tau}^{*B}(b) = h(\mathbf{w}^A, \mathbf{w}^B(b))$$

$$\hat{\tau}^{*}(b) = h(\mathbf{w}^A(b), \mathbf{w}^B(b))$$

To use the jackknife to estimate the variance, we had to remove one psu at a time from each frame. The bootstrap, employing resampling, allows more flexibility. We propose two bootstrap estimators. The separate bootstrap estimator is similar in form to the jackknife, performing the bootstrap in each sample separately and then combining the variance terms:

$$v_s = \frac{1}{B_1}\sum_{b=1}^{B_1}(\hat{\tau}^{*A}(b) - \hat{\tau})^2 + \frac{1}{B_2}\sum_{b=1}^{B_2}(\hat{\tau}^{*B}(b) - \hat{\tau})^2 \quad (8)$$

With the separate bootstrap estimator, the number of bootstrap iterations can differ for the two frames.

The combined bootstrap estimator does bootstrap for both frames simultaneously:

$$v_c = \frac{1}{B}\sum_{b=1}^{B}(\hat{\tau}^{*}(b) - \hat{\tau})^2. \quad (9)$$

This has the advantage of essentially halving the amount of replicate estimators needed for the bootstrap. If an agency releases replicate weights in a public use data file, the combined bootstrap reduces the number of replicate weight columns needed. If replicate weights are released for the separate bootstrap or the jackknife, a data user can easily discover which observations came from the same frame. If one of the frames is small, as in the U.S. Survey of Consumer Finances, where frame B consists of wealthy households likely to own assets such as tax-exempt bonds, frame identification might increase disclosure risk. The combined bootstrap, with each bootstrap iteration resampling from both frames, helps to maintain frame confidentiality.

Both bootstrap estimators are asymptotically equivalent to the linearization variance estimator when τ is a smooth function of population means, under regularity conditions on the sampling designs. In addition, as with the single frame bootstrap estimator studied by Shao and Chen (1998), the bootstrap is consistent for estimating the variance of some nonsmooth statistics such as the median.

Table 1 presents partial results from a simulation study comparing variance estimators and interval estimators. We used a factorial design with factors: (1) two or three frames, (2) simple random sample or cluster sample in frame A, (3) sample sizes of 100 or 200 for each frame. The population size was set to 10,000 in each domain. For the bootstrap, we used either 100 or 500 iterations. Five thousand replications were performed for each simulation runs in R version 2.1.1. We examined the relative bias, calculated as 100(average variance estimate - EMSE)/EMSE with EMSE the Monte Carlo estimate of mean squared error, and the relative standard deviation, calculated as (standard deviation of the variance estimates)/$\sqrt{\text{EMSE}}$, for each setting of the simulation design factors.

Table 1: Results from simulation with two frames for estimating the population total Y, the population size N, and the population median m. When Cl=Yes, a cluster sample was drawn from frame A. Bsep 100 refers to the separate bootstrap with 100 bootstrap iterations in each frame; Bc 500 refers to the combined bootstrap with 500 bootstrap iterations.

				Relative Bias				
n^A	Cl?	n^B		JK	Bsep	Bsep	Bc	Bc
					100	500	100	500
100	No	100	Y	2.2	2.0	1.8	1.8	2.1
			N	-2.3	-1.8	-2.0	-2.3	-2.3
			m		11.4	5.8	11.6	6.3
200	No	100	Y	-1.8	-1.9	-2.2	-2.2	-2.3
			N	-2.0	-2.2	-2.0	-1.9	-1.8
			m		16.1	10.9	16.5	11.3
200	Yes	100	Y	-0.3	-1.0	-0.2	-1.4	-0.8
			N	-2.6	-2.8	-1.8	-2.7	-1.8
			m		10.4	7.1	10.3	7.1

The relative bias for all methods is quite small for estimating the population total and size. It is somewhat larger, and usually positive, for estimating the population median. This bias lessens if more bootstrap iterations are performed—we found that $B = 500$ works well for estimating the variance of $\hat{Y}$ and $\hat{N}$, but that $B = 1000$ performs better for estimating the variance of the median. We can also reduce the relative bias by using interpolated values for the population and sample medians.

Confidence intervals for the linearization method, the jackknife, and both bootstraps can be calculated as

$$\hat{\tau} \pm 1.96\sqrt{v}.$$

If desired, a t critical value, using the smallest value of degrees of freedom from the frames, can be substituted for 1.96 to obtain a more conservative interval. This interval relies on the approximate normality of the statistic $\hat{\tau}$.

The combined bootstrap allows confidence intervals to be formed directly from the bootstrap distribution using either the percentile bootstrap or the bootstrap t method. All confidence intervals are consistent, and in our simulation studies where data were generated from a normal distribution, performed similarly. All empirical coverage probabilities for nominal 95% intervals were between 0.93 and 0.97, and the average lengths of intervals were about the same for all methods. There was a large difference in stability depending on the number of bootstrap iterations, however: the bootstrap with 500 iterations was more stable than the bootstrap with 100 iterations.

5. Connections

In this paper, I have highlighted some of Rao's contributions to multiple frame surveys and shown how they have ties to related results in survey sampling and other areas of statistics such as replication variance methods. The connections go far beyond those mentioned here, however.

Lu (2007) recently examined the problem of chi-square tests in multiple frame surveys. She estimated population proportions using pseudo-maximum-likelihood and derived chi-square tests based on Wald tests and Rao-Scott (1981) approximations. Hypotheses of interest in dual frame surveys include hypotheses about domain probabilities as well as those about probabilities for the population as a whole. Lu (2007) found that a dual frame approach sometimes allows testing of hypotheses that are untestable in a single frame survey, and can allow more flexibility in modelling missing data mechanisms.

Small area estimation, in which reliable estimates are desired for population subgroups in which the sample size is small, As Rao (2003) points out, multiple frame surveys can be used to improve the accuracy of small area estimates in subgroups of interest. Rao (2006) and Rao and Wu (2007) show how empirical likelihood methods can be used in forming estimates from dual frame surveys, providing a link with the work in that area.

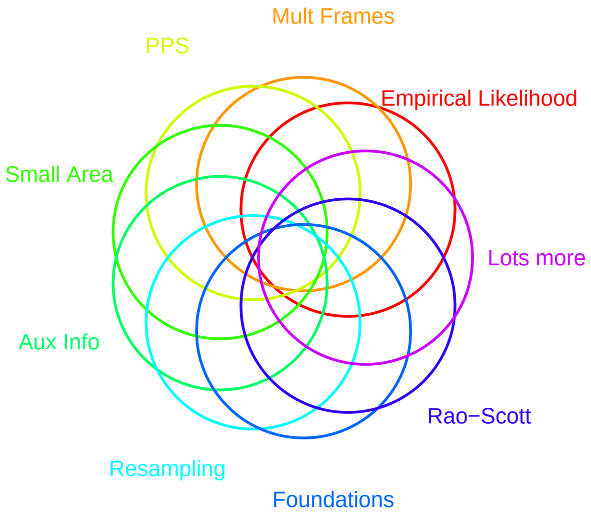
Bellhouse (2001) reviewed some of Rao's contributions to survey sampling up to that point, and gave a timeline outlining the major themes of his work. One can think of Rao's publications on multiple frame surveys as one frame out of many; these overlap with the frames of his publications on small area estimation, empirical likelihood, survey weights, resampling methods in sampling, unequal probability sampling, ratio estimation, foundations, interpenetrating samples, chi-square tests, and many other areas (Figure 5). All of Rao's work has been grounded in solving real problems, which is part of the reason it has been so influential in the discipline of survey sampling.

Acknowledgements

This work was partially supported by the National Science Foundation under grant 0604373. I would also like to thank J.N.K. Rao for our collaboration; some of the results from that collaboration are reviewed in this paper, and the bootstrap estimators of the variance presented here were developed in joint work with him.

Rao frequently refers to H.O. Hartley as his "guru." According to www.thefreedictionary.com, a guru is "a teacher and guide in spiritual and philosophical matters; a trusted counselor and adviser; a mentor; a recognized leader in a field." It is clear from this definition that Rao himself has been a guru to many, many statisticians. I feel privileged to number myself among these.

Figure 5: Rao's work in sampling



References

Bellhouse, D. (2001), "J.N.K. Rao: An Appreciation of His Work," in *Proceedings of the Survey Methods Section, Statistical Society of Canada.*

Brick, J. M., Dipko, S., Presser, S., Tucker, C., and Yuan, Y. (2006). "Nonresponse Bias in a Dual Frame Survey of Cell and Landline Numbers," *Public Opinion Quarterly* **70**, 780–793.

Cervantes, I. F. and Brick, J. M. (2007). *California Health Interview Survey: Sample Design*, CHIS Methodology Series, Report 1, www.chis.ucla.edu/pdf/CHIS2005_method1.pdf.

Demnati, A. and Rao, J. N. K. (2004), "Linearization Variance Estimators for Survey Data," *Survey Methodology,* 30, 17–26.

Demnati, A., Rao, J. N. K., Hidiroglou, M. A., and Tambay, J.-L. (2007), "Linearization Variance Estimators for Dual Frame Survey Data," Paper presented at the Joint Statistical Meetings, Salt Lake City.

Fecso, R., Tortora, R. D., and Vogel, F. A. (1986), "Sampling Frames for Agriculture in the United States," *Journal of Official Statistics*, **2**, 279–292.

Fuller, W. A., and Burmeister, L. F. (1972), "Estimators for Samples Selected From Two Overlapping

Frames," in *ASA Proceedings of the Social Statistics Section*, 245–249.

Graham, J. E. and Rao, J. N. K. (1978), "Sample Surveys: Theory and Practice," in *Studies in Statistics*, ed. R. V. Hogg, Washington, D.C.: Mathematical Association of America, 107–167.

Haines, D. E. and Pollock, K. H. (1998), "Combining Multiple Frames to Estimate Population Size and Totals," *Survey Methodology*, **24**, 79–88.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953). *Sample Survey Methods and Theory, Volume 1.* Wiley, New York.

Hartley, H. O. (1962), "Multiple Frame Surveys," *Proceedings of the Social Statistics Section, American Statistical Association*, 203–206.

Hartley, H. O. (1974), "Multiple Frame Methodology and Selected Applications," *Sankhyā*, Ser. C, **36**, 99–118.

Hartley, H. O., and Rao, J. N. K. (1968), "A New Estimation Theory for Sample Surveys," *Biometrika*, **55**, 547–557.

Iachan, R. and Dennis, M. L. (1993), "A Multiple Frame Approach to Sampling the Homeless and Transient Population," *Journal of Official Statistics*, **9**, 747–764.

Kalton, G., and Anderson, D. W. (1986), "Sampling Rare Populations," *Journal of the Royal Statistical Society*, Ser. A, **149**, 65–82.

Lohr, S. L. and Rao, J. N. K. (2000), "Inference in Dual Frame Surveys," *Journal of the American Statistical Association*, **95**, 271–280.

Lohr, S. L. and Rao, J. N. K. (2006), "Estimation in Multiple-frame Surveys," *Journal of the American Statistical Association*, **101**, 1019-1030.

Lu, Y. (2007), "Longitudinal Estimation in Dual Frame Surveys," Ph.D. Dissertation, Arizona State University.

Rao, J. N. K. (1968), "Some Nonresponse Sampling Theory when the Frame Contains an Unknown Amount of Duplication," *Journal of the American Statistical Association*, **63**, 87–90.

Rao, J. N. K. (1983), "H.O. Hartley's Contributions to Sample Survey Theory and Methods," *The American Statistician*, **37**, 344–350.

Rao, J. N. K. (2003), *Small Area Estimation*, New York: Wiley.

Rao, J. N. K. (2006), "Empirical Likelihood Methods for Sample Survey Data: An Overview," *Austrian Journal of Statistics*, **35**, 191–196.

Rao, J. N. K. and Graham, J. E. (1964), "Rotation Designs for Sampling on Repeated Occasions," *Journal of the American Statistical Association*, **59**, 492–509.

Rao, J. N. K. and Scott, A. J. (1981), "The Analysis of Categorical Data from Complex Sample Surveys: Chi-square Tests for Goodness of Fit and Independence in Two-way Tables," *Journal of the American Statistical Association*, **76**, 221–230.

Rao, J. N. K., and Skinner, C. J. (1999), "Dual Frame Surveys: Pseudo Maximum Likelihood and Single Frame Estimators," in *Statistical Inference and Design of Experiments*, ed. U.J. Dixit and M.R. Satam, New Delhi: Narosa Publishing House, 63–71.

Rao, J. N. K., and Wu, C. F. J. (1988), "Resampling Inference With Complex Survey Data ," *Journal of the American Statistical Association*, **83**, 231–241.

Rao, J. N. K., and Wu. C. (2007), "Empirical Likelihood Methods," to appear in *Sample Surveys: Theory, Methods and Inference, Handbook of Statistics, Vol. 29*, ed. D. Pfeffermann and C. R. Rao, Amsterdam: North Holland.

Shao, J. and Chen, Y. (1998), "Bootstrapping Sample Quantiles Based on Complex Survey Data under Hot Deck Imputation," *Statistica Sinica*, **8**, 1071–1086.

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., and Rao, J. N. K. (1996), "Estimation in Dual Frame Surveys With Complex Designs," *Journal of the American Statistical Association*, 91, 349–356.

Tucker, C., Brick, J. M., and Meekins, B. (2007). "Household Telephone Service and Usage Patterns in the United States in 2004: Implications for Telephone Samples," *Public Opinion Quarterly*, **71**, 3–22.