NOTES ON A COMPOSITE SIZE MEASURE FOR SELF-WEIGHTING SAMPLES IN MULTIPLE DOMAINS

R.E. Folsom, F.J. Potter, and S.R. Williams, Research Triangle Institute

1. Introduction

In recent years, RTI has used a composite size measure procedure for achieving self-weighting samples for multiple domains in multi-stage designs. The procedure requires some knowledge of population counts, either estimated or actual count of the numbers of elements in the population, for each domain in each sampling unit. This procedure can also be used in a two phase design in which the domain membership information is collected in the first phase sampling units and used in the second phase. The procedure is well suited to sampling from large list frames with clusters of elementary units that can serve as primary sampling units to reduce data collection costs. RTI has used this procedure to sample customers for personal interview from the billing file of a major Southeast utility, to sample Medicaid recipients in state-level surveys (Lynch et al. 1986), for a Medicaid household survey (Folsom and Iannacchione, 1980), and to sample patient records at a sample of hospitals in Florida (Williams et al., 1978).

A basic two domain version of the composite size measure was used at RTI by Dr. Walt Hendricks to select primary sampling units for the National Assessment of Educational Progress (NAEP) in the early 1970's. While we are aware of no earlier published references to the composite size measure, use of the basic two domain version has undoubtedly been part of survey practice for many years. The general multiple domain version described here was developed at RTI in early 1978 by Dr. Ralph Folsom. Our first application of the multi-domain composite size measure was for a study to obtain information on the use of selected health care services in short-term hospitals (Williams et al. 1978 and Drummond et al. 1984). The purpose of this paper is:

a) to demonstrate the robustness of the procedure to accommodate changes in domain membership criteria and domain sampling fractions <u>after</u> primary sampling units have been selected

b) to describe a procedure to ensure adequate domain frames for final sample selection

c) to describe the use of these procedures in 2 recent surveys conducted by RTI.

2. Basic Procedures

Consider a conceptual list frame of N units consisting of J domains. Define the following quantities:

N_j = The total count of elementary units in domain j, j=1,...,J,

$$N = \sum_{j=1}^{J} N_j,$$

n_j = the desired sample size from domain j,

$$n = \sum_{j=1}^{J} n_j, \text{ and}$$

f_j = the desired sampling fraction for domain j, $f_j = n_j/N_j$.

Assume that the list frame is or can be divided into I primary sampling units (PSUs). Let

N_{ij} = the count of elementary units in domain j in PSU i,

n^* = the desired sample size from all domains in each PSU, that is, a fixed workload is desired in each PSU.

To select a sample of m PSUs, a composite size measure S_i is computed for each PSU where

$$S_i = f_1N_{i1} + f_2N_{i2} + \ldots + f_JN_{iJ}$$

$$= \sum_{j=1}^{J} f_jN_{ij} \qquad (1)$$

Make m PSU selections with the expected selection frequency for PSU-i strictly proportional to the composite size measure S_i. This can be performed by using a procedure such as Chromy's probability minimal replacement sequential selection procedure (Chromy 1981).

In the following presentation, we will assume that the strictly proportional to size expected selection frequency

$$E(m_i) = m\, S_i/S_+$$

is ≤ 1 for all primary frame units-i. For this case, we will envision m nonreplacement PSU selections with $E(m_i)$ becoming the sample inclusion probability for primary unit-i; that is, the probability of selecting PSU (i) is

$$P(\text{PSU}i) = m\, S_i/S_+$$

where $S_+ = \sum_{i=1}^{I} S_i$

Given our definition of the composite size measure

$$S_+ = \sum_{i=1}^{I} S_i = \sum_{i=1}^{I} \sum_{j=1}^{J} f_jN_{ij}$$

$$= \sum_{j=1}^{J} f_j \sum_{i=1}^{I} N_{ij}$$

$$= \sum_{j=1}^{J} f_j N_j$$

$$= \sum_{j=1}^{J} n_j = n$$

The desired sample size of elementary units (n^*_{ij}) in PSU i from domain j is computed as

$$n^*_{ij} = n^*\, f_j\, N_{ij}/S_i. \qquad (2)$$

It is clear by construction that these domain allocations sum to the desired common workload n^* for each sample PSU-i. Because the n^*_{ij} are usually noninteger numbers, they are stochastically rounded such that over repeated roundings $E(n_{ij}^R)=n^*_{ij}$ and the sum of the n_{ij}^R equals n^*, the desired common workload for PSU-i. To achieve n_j exactly, compute $n^*_j = \sum_i n^*_{ij}$ and adjust the original n^*_{ij} by the (n_j / n^*_j) prior to rounding. This modification has an effect similar to the rate change. That is, let $f'_j = (n_j/n^*_j)f_j$. Using the rounded sample size for domain j in PSU i (n_{ij}^R), a simple random sample of n_{ij}^R units is selected from the N_{ij} elementary units.

To demonstrate the self-weighting nature of the procedure, the sampling weights will be computed. To compute the sampling weights for elementary units selected in domain j, consider first the conditional probability of selecting n^*_{ij} units in domain j in PSU i, given the PSU i is selected,

P(domain j unit selected in PSU i|PSU i)

$$= n^*_{ij}/N_{ij}$$

$$= (n^* f_j N_{ij}/S_i)/N_{ij}$$

$$= n^* f_j/S_i.$$

The use of the unrounded sample size n^*_{ij} in this solution is equivalent to using the unconditional second stage selection probability averaged over the possible stochastic roundings. The overall unconditional probability of selecting an elementary unit in domain j is the product of the probability of selecting PSU i and the probability of selecting an elementary unit of domain j in PSU i given that PSU i was selected. That is,

P(a domain j unit is selected)

$$= P(\text{domain j unit selected in PSU i}|\text{PSU i}) \times P(\text{PSU i})$$

$$= (n^* f_j/S_i) \times (m S_i/S_+)$$

$$= (n^* m f_j/S_+)$$

$$= f_j$$

recalling that $S_+ = \sum n_j = n$ and $n^* m = n$. Therefore, the unconditional probability of selecting an elementary unit in domain j is f_j and the sampling weights are $1/f_j$. For each domain j, the design results in a self weighting sample. Since the expected number of domain j selections is $N_j f_j$ when all domain members have the same unconditional inclusion probability, f_j, the desired sample size $n_j=N_j f_j$ is clearly achieved in expectation. In the following section, a modification is proposed that achieves the n_j exactly, while remaining self-weighting by domain.

3. Modifications to the Basic Procedure

The basic procedure results in a self weighting sample in each domain and a fixed workload in each PSU. In some surveys conducted by RTI that employed this procedure, the sponsor has requested changes in the sampling rates in selected domains or has redefined the domain membership by changing eligibility rules for elementary units after the PSUs were selected and field work had begun. The following describes the procedures to accommodate these changes to the sampling design.

3.1 Changing the Sampling Rates

One rationale for changing f_j is to achieve n_j exactly as indicated in Section 2. Otherwise one only achieves n_j in expectation; i.e.,

$E\{\sum_i n^*_{ij}\}=n_j$ but $\sum_i n^*_{ij} \neq n_j$ unless one adjusts f_j to f'_j as shown before.

Let f_j and f'_j denote the original and revised sampling rate in domain j, respectively. Let n_j and n'_j denote the original and revised sample sizes desired in domain j. Assume that m PSUs were selected with probability strictly proportional to the composite size measure S_i and

$$S_i = \sum_{j=1}^{J} f_j N_{ij}$$

Let n^* denote the sample count initially desired from each PSU. To achieve a self-weighting design in each domain, the unrounded revised sample size in domain j from PSU i is computed as

$$n'_{ij} = n^* f'_j N_{ij}/S_i$$

The n'_{ij} are then stochastically rounded to achieve the revised sample size of n'_j. The rounded sample size in each domain is selected in each PSU using simple random sampling from the N_{ij} units.

Because the n'_{ij} are stochastically rounded, the conditional probability of selecting an elementary unit in domain j from PSU i given that PSU i was selected is

P(domain j unit selected in PSU i|PSU i)

$$= n'_{ij}/N_{ij}$$

$$= n^* f'_j/S_i.$$

The unconditional probability of selecting an elementary unit in domain j is then

P(domain j unit selected)

$$= P(\text{domain j unit selected in PSU i}|\text{PSU i}) * P(\text{PSU i})$$

$$= (n^* f'_j/S_i) \times (m S_i/S_+)$$

$$= (n^* m) f'_j/S_+.$$

Since n^* was the original sample count desired from each PSU,

$$n^* m = n = \sum n_j = S_+.$$

Therefore,

$$P(\text{domain } j \text{ unit selected}) = f'_j$$

The design is again self-weighting in each domain.

However, to achieve a self-weighting design after changing sampling rates, the restriction of equal workload in each PSU was removed. To see this, note that the unrounded sample size in PSU i for domain j is

$$n'_{ij} = n^* f'_j N_{ij}/S_i$$

and

$$\sum_j n'_{ij} = n^* \sum_j f'_j N_{ij}/S_i.$$

However, $\sum_j f'_j N_{ij}$ may not equal S_i, so $\sum_j n'_{ij}$ may not equal n^*.

3.2 Changing Domain Membership

Procedures to accommodate changes in domain membership can be drawn from the procedures just described to accommodate changes in the sampling fraction.

Let N_{ij} and N'_{ij} denote the original and revised domain counts, respectively, for the jth domain of PSU i. Also let f_j and f'_j and n_j and n'_j denote the original and revised sampling fractions and sample sizes desired for the jth domain. Once again, assume that m PSUs were selected with probability strictly proportional to the composite size measure S_i.

Let n^* denote the sample count initially desired from each PSU. The unrounded sample size for the jth domain in the ith PSU is computed as

$$n'_{ij} = n^*(f'_j N'_{ij})/S_i.$$

The n'_{ij} are stochastically rounded to the revised sample size and the sample for each domain is selected with equal probability in each PSU from the N'_{ij} eligible domain units.

Following the algebra described in Section 3.1, it is easy to show that the probability of selection is f'_j for each member of the jth domain. The sampling weight is then $1/f'_j$. Once again, the number of elementary units selected in each PSU will vary.

3.3 Comments on the Modified Procedure

The key requirement that permits these modifications is that the sum of the PSU composite size measures equals the product of the sample size initially desired from each PSU (n^*) and the number of sample PSUs (m). That is,

$$S_+ = n^* m.$$

Because these numbers are generally fixed prior to selecting the PSUs, this procedure is very flexible for accommodating changes in the sampling fractions or domain membership rules.

It should be noted that in some situations, the domain-specific sample size indicated for the ith PSU may exceed the number of domain eligible units in the PSU. When this occurs, the number of times that a particular unit is selected may be greater than one. To accommodate this situation, the sampling weight for the unit is multiplied by the number of selections or the data for the unit are included in the analysis file multiple times. In the latter case, each occurrence of the unit's data will receive the appropriate sampling weight. One can also assure that this problem does not occur by requiring that primary frame units meet the following minimum size requirement

$$S_i > n^* f_{max}$$

where f_{max} is the largest of the J domain sampling rates f_j. When this is not the case, one collapses contiguous primary frame units until the combined unit meets the minimum size requirement.

The procedure can be generalized for stratified designs. For stratified designs, the sampling fractions for individual domains can be different for some strata. This allows for oversampling of some domains in selected strata.

When domain counts are not known or cannot be reasonably estimated prior to PSU selection, one can still achieve self-weighting samples by domain. What one sacrifices when the composite size measure cannot be reasonably approximated is variation in the PSU level sample sizes n_i. To illustrate this shortcoming of typical primary samples based on total population size measures, let N_i denote the total count of units in the ith PSU and

$$N = \sum_{i=1}^{I} N_i$$

Define $S_i = n N_i/N$

where n is the total sample size desired from all domains. Select m PSUs with probability strictly proportional to S_i. In each of the m PSUs, the elementary units are classified into the domains and within-PSU domain counts are generated. This classification can be based on developing a domain classified list frame within each sample PSU, or by conducting an initial large screening sample which provides domain classification. A self-weighting sample is still obtained if all $N_{ij} \geq n^R_{ij}$ where n^R_{ij} is the stochastically rounded sample size.

Let N_{ij} denote the count of elementary units in the jth domain.

Compute the estimated domain count as

$$\hat{N}_j = \sum W_i N_{ij}$$

where $W_i = 1/P(\text{PSU i Selected})$

$$= S_+/mS_i.$$

The domain sampling fractions are then computed,

$$f_j = n_j/\hat{N}_j$$

where n_j is the desired sample size from domain j. The domain sample size is then computed as

$$n_{ij} = n^* f_j N_{ij}/S_i$$

where n^* is the desired workload in each PSU. Because, in general,

$$\sum_j f_j N_{ij} \neq S_i,$$

the actual number of units selected in each PSU will vary. This variation will be substantial when the rates f_j vary widely and primary units have very different mixes of domain members.

4. Examples

Two examples of the use of these procedures are:

a) a survey of customers of a major utility in the Southeast; and

b) a survey of Medicaid recipients in two southwestern states.

During 1984, RTI conducted the survey of residential customers of a utility in the Southeast. The utility sought information from two types of customers: customers participating in an energy conservation program, and non-participating customers. The eligible study populations were customers in owner-occupied dwellings, including single family dwellings, multi-family dwellings, and mobile homes. To achieve precision required, RTI estimated a sample size of 1,780 responding non-participants and 400 responding program participants distributed across three geographic service regions with an oversampling of non-participants in one region.

The utility provided RTI a billing file of all customers, approximately 2.4 million, with indicators of housing unit type and program participation. All program participants were in owner-occupied dwellings but owner occupancy could not be determined from the billing file for non-participants.

RTI developed first stage units from meter reading routes by processing the 2.4 million record data file and generating customer counts by program participation and type of dwelling. To account for renter-occupied dwellings, RTI used the zipcode of the meter reading route to merge the 1980 Census estimates of dwelling-specific owner-occupancy rates at the 5-digit zipcode level. RTI estimated the number of non-participants in owner-occupied dwellings in each meter reading route and for the service area and regions. Using the count of participants and the estimated count of non-participants, RTI computed the sampling fractions for each study population and service region. To facilitate sampling and data collection, RTI combined some of the meter reading routes with small counts to form PSUs and split a few of the large routes. From 6,695 meter reading routes, 4,993 PSUs were generated for the first stage sampling.

Using the counts of eligible participants and nonparticipants, RTI computed a PSU size measure in the following form:

$$S_i = f_p \times N_{ip} + f_{np} \times N_{inp},$$

where f_p = sampling fraction for program participants,

f_{np} = sampling fraction for non-participants,

N_{ip} = count of program participants in PSU(i),

N_{inp} = estimated count of nonparticipants in PSU(i).

Based on clustering effect and data collection costs, RTI set the respondent sample size at 6 responding customers per PSU. Therefore, an estimated 364 PSUs were required to achieve the desired sample size. RTI selected a sample of 364 PSUs and a supplemental sample of 364 PSUs with probability strictly proportional to the size within the three service regions using a sequential selection algorithm (Chromy 1981).

After selecting the PSUs, RTI estimated the sample size of customers required in each route by program participation and dwelling unit type. For each PSU, RTI computed the sample size using the following equation where the domains, denoted by j, are (1) program participants, (2) nonparticipants in single family dwellings, (3) nonparticipants in multi-family dwellings, and (4) nonparticipants in mobile homes:

$$n^*_{ij} = n^* \times f_k \times occ_j \times N_{ij} / S_i,$$

where n^*_{ij} = sample size for domain j in PSU i,

n^* = desired sample size from each PSU inflated for nonresponse,

f^* = sampling fraction for program participants (k=p) or nonparticipants (k=np)

occ_j = estimated proportion of owner-occupied dwellings in population j,

N_{ij} = count of customers in PSU i in population j, and

S_i = size measure for PSU i.

Because the n^*_{ij} were generally non-integer numbers, RTI stochastically rounded the sample size for each combination of program participation, dwelling unit type, and region. RTI selected 4,507 customers from the 364 PSUs.

Because RTI was concerned that the eligibility rates based on the 1980 Census data may not be sufficiently accurate for all regions, RTI randomly partitioned the 364 PSUs into a subsample of early reporting PSUs and of other PSUs. The screening data from these early reporting PSUs was used to determine the accuracy of the estimated eligibility rates so that supplemental samples from the PSUs held in reserve could be selected. Based on the data from the early reporting PSUs, a supplemental sample from 16 PSUs in one region was selected for this study.

From the equations given above, it is easy to show that this sampling procedure resulted in self-weighting samples. Nonresponse sampling weight adjustments for such a self-weighting sample is the simple ratio of the sample count to the number of respondents for weighting classes corresponding to the sampling strata.

The second example of application of the composite size measure is a study of Medicaid recipients in two southwestern States. The main study objective was to evaluate access to and satisfaction with services of a specific Medicaid program. Basic information for this study was obtained through a sample survey of

recipients. The sampling frame was developed from the program's membership file, which contains information for identifying study eligibles and useful for stratifying and for developing size measures for the frame.

Several factors were specified as important controls in sample selection, both to increase survey precision and to ensure that selected subclasses would be adequately represented in the sample to develop separate statistics. The sample sizes were to be controlled for:

- urban and rural residence
- age (3 age groups)
- sex
- race (2 groups)
- type of Medicaid plan (2 groups).

Also, recipients were to be selected in geographic clusters to reduce field cost.

To accomplish these objectives, 5-digit ZIP areas were stratified geographically into urban and rural. Requisite sample sizes were determined for each of 24 cells within geographic stratum (3 age groups X 2 sex groups X 2 racial groups X 2 plan types) and size measures were calculated according to (1). Some ZIP areas were combined to form primary sampling units (PSUs) with adequate subclass (or domain) counts. Specifically, adjacent areas were combined until the combined size measure was greater than or equal to n^*f_j for the largest sampling fraction. This ensures that $N_{ij} \geq n^*_{ij}$.

PSUs were selected with probability proportional to S_i and n^* recipients were selected from each sample PSU. The n^* recipients were obtained by selecting n^R_{ij} from subclass (j) of PSU (i) (n^R_{ij} is the randomly rounded sample size; refer to discussion on stochastic rounding).

After the PSUs were selected and field work had begun, eligibility designations were revised and it became necessary to change sampling rates in selected domains. The modification described earlier in this paper for changing domain membership was used. Recall that this can be accomplished without loosing the desired self-weighting sample feature.

5. Discussion

The concepts described in this paper provide an efficient method for obtaining subclass (domain) estimates with controlled precision when subclasses are of different size and subclass sampling frames do not exist. Some of the more important properties of this method are:

- by selecting clusters with probabilities proportional to a specifically-structured composite size measure, clusters needed in the sample in order to supply information about rare domains tend to be oversampled,
- the sample sizes per cluster can be controlled so that field work and respondent burden is approximately the same for each cluster,
- precision of study estimates can be controlled for each study domain,
- self-weighting samples can theoretically be obtained for each study domain, although, in practice, slight departures from epsem samples may be desirable and often are unavoidable,
- the final sample selection can be conducted efficiently in the field at the time of screening and according to a simple protocol,
- simple to adjust sampling weights for nonresponse, and
- flexible for mid-survey changes in sample sizes, sampling rates, and population counts.

REFERENCES

Chromy, J. R. (1981). "Variance Estimators for a Sequential Sample Selection Procedure" in D. Krewski, R. Platek, and JNK Rao, eds. Current Topics in Survey Sampling. New York: Academic Press, Inc.

Drummond, D. J. et al. (1984). "A Design for Achieving Prespecified Levels of Representation for Multiple Domains in Health Record Samples," in Health Survey Research Methods Proceedings of the Fourth Conference on Health Survey Research Methods. DHHS Pub. No. (PHS) 84-3346. Washington DC: National Center for Health Services Research.

Folsom, R. E. and V. G. Iannacchione (1980). "NMCUES State Medicaid Household Survey Sample Design Statement," an RTI final report for Health Care Financing Administration and the National Center for Health Statistics.

Lynch, J. T., D. S. DeWitt, S. R. Williams, and J. T. Lessler (1986). "Evaluation of the Arizona Health Care Cost Containment System: ACCESS Consumer Survey," an RTI final report for the SRI International.

Williams, S. R. et al., (1978). "Development of a Feasible Data Collection Plan for Hospital Data in Florida," an RTI project report.