A Generalized Epsem Two-Phase Design for Domain Estimation

Avinash C Singh[1] and Rachel Harter[2]
[1]NORC at the University of Chicago, 55 East Monroe St., Chicago, IL 60603
[2]RTI International, P.O. Box 12194, Research Triangle Park, NC 27709

Abstract
Composite measures of size can be used to select primary sampling units in a two-stage design such that multiple subdomains are self-weighting. This method can be generalized to situations such as two-phase designs where the PSU probabilities are prescribed so that the sample allocations to domains within PSUs are adjusted to achieve self-weighting domains. The method is illustrated for an area probability sample of housing units.

Keywords: two-phase, self-weighting, EPSEM, domain estimation

1. Introduction

Some survey designs are intended to support analytical goals for multiple domains within the target population. Folsom, Potter, and Williams (1987) presented a method for allocating a sample of units within primary sampling units (PSUs) such that the sample is self-weighted[1] or equal probability within each domain while controlling for an equal number of cases per PSU. The method involves a composite measure of size for selecting the PSUs and defining the allocation. The same basic method can be inverted to produce self-weighted samples by domain when the PSUs are already selected, if equal workload per PSU is not a requirement. In this form it can be used for two-phase samples as well as two-stage samples, requiring less advance data about the domains. The method can be further generalized to stratified PSUs.

Section 2 reviews the original composite measure of size method as discussed by Folsom et al. (1987), along with a simple example of the method. Section 3 presents the inverted method with pre-selected PSUs and known population totals. Section 4 extends the method to two phase designs where the phase two frame counts must be estimated in phase one. Section 5 gives the expansion of the two-phase method to stratified PSUs. Section 6 gives a hypothetical example based on the study for which the two phase design was developed. Section 7 discusses limitations of the method when allocations to domains in PSUs exceed the available frame counts. We conclude with a few remarks in Section 8.

2. Self-Weighted Samples in Multiple Domains with Equal Workloads in PSUs

Consider a two-stage design where the first stage units are schools, for example, and the second stage units are individual teachers. Two domains of interest may be male and

[1] Self-weighted samples are often called EPSEM samples (equal probability of selection method) using terminology from Kish (1965).

female teachers. A composite measure of size can be used to allocate the number of sample teachers within schools in such a way that the selected teachers are self-weighting (or Epsem, using Kish (1965) terminology) for both male and female teachers.

Let i denote a sample school, $i = 1,\ldots,m$, where m is the total number of sample schools. Suppose that the number of male teachers (N_{i1}) and the number of female teachers (N_{i2}) is known for every school in the frame. Further, suppose the desired number of male sample teachers ($n_{.1}$) and female sample teachers ($n_{.2}$) are specified for precision requirements. Then we know the desired sample rate for male teachers $f_1 = n_{.1}/N_{.1}$ and for female teachers $f_2 = n_{.2}/N_{.2}$, where $N_{.d}$ is the sum of N_{id} for domain d across all schools.

Now suppose we want equal sample sizes in all schools, $n_{i.} = n^*$ for all schools i. We first determine the number of schools required.

$$m = (n_{.1} + n_{.2}) / n^* \qquad (1)$$

We will select m schools with probability proportional to size, where a school's measure of size combines the known number of male and female teachers.

$$S_i = f_1 N_{i1} + f_2 N_{i2} \qquad (2)$$

The general result in Folsom et al. (1987) assumes that there are D domains, $d = 1,\ldots,D$, and the N_{id} is known for every primary sampling unit i. The desired number of cases n_d in each domain is known, as well as the sampling fractions f_d and the total desired sample size $n = \sum_d n_{.d}$. Furthermore, we want an equal number of cases, n^*, in each primary sampling unit. Then the number of primary sampling units to select is

$$m = n/n^* \qquad (3)$$

Define a composite measure of size associated with primary sampling unit i by

$$S_i = \sum_d f_d N_{id} \qquad (4)$$

Then the allocation of sample teachers for domain d in school i is

$$n_{id} = n^* f_d N_{id} / S_i \qquad (5)$$

It is easy to verify that the total sample size across all sample schools is $n_{.1} + n_{.2}$, or more generally that

$$n = \sum_d n_{.d} \qquad (6)$$

Furthermore, it is also easy to show that the probability of an individual teacher being selected is equal for all sample teachers in domain d.

$$\text{P(teacher in domain } d \mid \text{school } i) \text{ P(school } i) = f_d$$

The known facts, specifications, and results for the general problem according to Folsom et al. (1987) are summarized in Table 1.

Table 1. Overview of Original Problem to Equalize Probabilities and Probabilities for All Domains

Known or Estimated	
N_{id}	Populations totals for each domain d in PSU i, for all PSUs in the frame
Specified	
$n_{.d}$	Desired sample for each domain d, from precision requirements
n and f_d	Total sample size n and domain sampling rates f_d are also known from $n_{.d}$ and N_{id}
n^*	Desired constant number of sample cases in each PSU i
Epsem domain samples	Equal probability within domains across all sample PSUs
Derived	
m	Number of PSUs to select
S_i	Composite measure of size for each PSU in the population
n_{id}	After PSUs are selected, allocation of sample cases by domain and PSU

3. Self-Weighted Samples in Multiple Domains with Preselected PSUs

Now suppose that the schools have already been sampled, and their probabilities of selection are known. It is still possible to select equal probability samples of male and female teachers using a composite frame size for stage two, but not with equal sample sizes per school.

More generally, given the m PSUs with selection probabilities π_i, the N_{id} population totals by domain d and sample PSU i, and the specified domain sample sizes $n_{.d}$ (or, equivalently, the domain sampling rates f_d across the sample PSUs), it is possible to select an equal probability sample for each domain d. The quantity S_i is defined by equation (4) as before. Technically S_i is a measure of size only for allocating secondary sample units and not for selecting the PSUs as before; in this situation it is intuitively an expected sample size for the PSU over all domains. Then the allocation of sample units for domain d in PSU i is given by

$$n_{id} = (n_{i.} f_d N_{id}) / S_i \qquad (7)$$

where

$$n_{i.} = n \pi_i^{-1} S_i / (\textstyle\sum_i \pi_i^{-1} S_i) \qquad (8)$$

Equations (7) and (8) are derived by assuming equal probabilities within domain d, using the fact that $n = \sum_i n_{i.}$, and solving backwards. It is straightforward to show with these allocation values that all sample cases in the same domain have the same probability.

$$\text{P(unit in domain } d \text{ in PSU } i) = \text{P(unit in domain } d \mid \text{PSU } i) \text{ P(PSU i)}$$

$$= (n_{id} / N_{id}) \pi_i$$

$$= \{[(n_{i.} f_d N_{id}) / S_i] / N_{id}\} \pi_i$$

$$= (n_{i.}) f_d \pi_i / S_i$$

$$= [n \pi_i^{-1} S_i / (\textstyle\sum_i \pi_i^{-1} S_i)] \; [f_d \pi_i / S_i]$$

$$= [\, n \;/ (\textstyle\sum_i \pi_i^{-1} S_i)] \;\; f_d$$

which is a constant times f_d for all sample units in domain d.

4. Extension to Two-Phase Designs

With the situation inverted so that the PSUs are preselected with probabilities known, it is no longer necessary that the N_{id} population values be known for all PSUs prior to PSU selection, as long as the N_{id} values are known or estimated for the sample PSUs prior to secondary sample allocation. While the original problem involved a two-stage design, this situation can be carried out with a two-phase design as well as a two-stage design.[2] That is, a first-phase survey can be carried out on the primary sampling units to estimate the N_{id} values.

Alternatively, the unweighted first phase sample totals themselves, denoted N'_{id} , can be used in place of N_{id} in the formulas as frame totals for phase two. But in this scenario, f'_d and $n_{.d}$ cannot both be specified. Assuming that the $n_{.d}$ values are specified, the conditional second phase sampling rates are derived as

$$f'_d = n_{.d} \;/ N'_{.d} \tag{9}$$

To keep notation consistent, when N'_{id} is used, the quantity

$$S'_i = \textstyle\sum_d f'_d \, N'_{id} \tag{10}$$

is intuitively an expected size of the sample totals across domains for PSU i.

The allocation formulas also differ because we need to take the first phase sampling probabilities into account. Let g_i denote the conditional probability of selection for any unit in the first phase sample in PSU I, assuming equal probabilities within each PSU. The phase one samples are not pre-determined, so the phase one samples should be as large as the schedule and budget allow to maximize the frames for phase two sampling, especially in PSUs with higher numbers and concentrations of the rarer domains. The unconditional probability of selection for a unit to be in the phase one sample is $\pi_i \, g_i$.

The allocation of phase two sample units to PSU i is

$$n_{i.} = n \, (\pi_i \, g_i)^{-1} \, S'_i \,/\, (\textstyle\sum_i (\pi_i \, g_i)^{-1} S'_i) \tag{11}$$

The allocation of phase two sample units to domain d in PSU i is

$$n_{id} \;\; = \;\; (n_{i.} \, f'_d \, N'_{id}) \,/\, S'_i \tag{12}$$

The conditional probability of a unit being selected into phase 2 is

$$n_{id} \;/ N'_{id} \;\; = (n_{i.} \, f'_d) \,/\, S'_i$$

$$= [\, n \, (\pi_i \, g_i)^{-1} \;\; f'_d \,] \,/\, [\, \textstyle\sum_i (\pi_i \, g_i)^{-1} S'_i \,]$$

Then the overall probability of selection for a unit in domain d is

[2] Two stage designs and two phase designs are discussed in standard sampling books such as Kish (1965), Cochran (1977), and Lohr (1999).

P(domain d in PSU i in phase 2) = P(domain d | PSU i and phase 1) P(PSU i and phase 1)

$$= [\, n\, (\pi_i\, g_i\,)^{-1}\ f'_d\,] \,/\, [\, \textstyle\sum_i (\pi_i\, g_i\,)^{-1} S'_i\,]\ (\pi_i\, g_i\,)$$

$$= [\, n\, f'_d\,] \,/\, [\, \textstyle\sum_i (\pi_i\, g_i\,)^{-1} S'_i\,]$$

which is a constant multiple of f'_d . The equal probability criterion for each domain is satisfied.

Table 2 summarizes what is known, specified, and derived in a two-phase design with preselected PSUs and observed frame counts N'_{id} for phase two sampling.

Table 2. Overview of Two-Phase Design to Equalize Probabilities for All Domains with Preselected PSUs

Known or Estimated	
m	Number of PSUs pre-selected
π_i	Probability of selection for each sample PSU i
N'_{id}	First phase sample totals for all PSUs i in the sample (determined after first phase sampling)
Specified	
$n_{\cdot d}$	Desired sample for each domain d, from precision requirements
n	Total sample size n
g_i	Sampling probabilities (and sample sizes) for phase one sample of units within PSU i
Epsem domain samples	Equal probability within domains across all sample PSUs
Derived	
f'_d	Domain sampling rates, conditioned on the first phase sample PSUs
S'_i	Composite measure of size for each PSU in the sample, strictly for allocating secondary sample units
n_{id}	Allocation of secondary sample units by domain and PSU

5. Generalization of the Two-Phase Method for Stratified PSUs

Now suppose that the PSUs are stratified. For example, the schools may be stratified by school district in our simple example. Then the formulas are slightly more complex with the additional subscript h representing the stratum, but the concept and the approach are the same. In this generalization of the two-phase method we still assume that the PSU probabilities are known and that equal probability samples by domain are required. Again, there is no requirement for equal workload by PSU. We again assume that the phase two frame counts, N'_{hid}, are obtained from the phase one sample.

The domain sampling rates, conditioned on the phase one survey results, are defined by

$$f'_d = n_{\cdot\cdot d}\ / N'_{\cdot\cdot d} \qquad (13)$$

The expected sample size for PSU i across domains, used only for phase two allocations, is defined by

$$S'_{hi} = \textstyle\sum_d f_d\, N'_{hid} \qquad (14)$$

The allocations are defined here in top-down order. The allocation of $n_{h\cdot\cdot}$ sample units to stratum h is

$$n_{h..} = n \sum_i \pi_i^{-1} S'_{hi} / (\sum_h \sum_i \pi_i^{-1} S'_{hi}) \quad (15)$$

The allocation to PSU *i* in stratum *h* is

$$n_{hi.} = n_{h..} \pi_{hi}^{-1} S'_{hi} / (\sum_i S'_{hi}) \quad (16)$$

Finally, the allocation to domain *d* in PSU *i* in stratum *h* is

$$n_{hid} = (n_{hi.} f'_d N'_{hid}) / S'_{hi} \quad (17)$$

Table 3 summarizes what is known, specified, and derived for this slightly more complex scenario.

Table 3. Overview of Two Phase Design to Equalize Probabilities for All Domains with Preselected PSUs Within Strata

Known or Estimated	
m_h	Number of PSUs pre-selected in each stratum *h*
π_{hi}	Probability of selection for each sample PSU *i* in stratum *h*
N'_{hid}	Phase one sample counts in domain *d* for the phase two frame, for all PSUs i in the sample
Specified	
$n_{..d}$	Desired sample for each domain *d*, from precision requirements
n	Total sample size *n*
g_{hi}	Sampling probabilities (and sample sizes) for phase one sample of units within PSU *i* in stratum *h*
Epsem domain samples	Equal probability within domains across all sample PSUs
Derived	
f'_d	Domain sampling rates, conditioned on the first phase sample PSUs
S'_{hi}	Composite measure of size for each PSU in the sample for phase two sampling
n_{hid}	Allocation of sample cases by domain, PSU, and stratum

6. Example

A team of university researchers developed a set of tests for physical and cognitive functions. They desired to "norm" the tests, establishing typical ranges of results for the general population, by measuring the results on children recruited to take the tests. Because the test results vary by age and gender, the goal was to recruit male and female children by year of age. Furthermore, the researchers wanted Spanish-speaking children as well as English-speaking children. Assuming equal completion rates, the specified initial sample sizes for twelve age/gender/language cells are shown in Table 4.

Table 4. Desired Completed Tests by Demographic Domain

	English-speaking		*Spanish-speaking*	*Hispanic*
Age	*male*	*female*	*male*	*female*
3	200	200	200	200
4	200	200	200	200
5	200	200	200	200

Originally the researchers desired a probability sample representative of the U.S. population for each of these domains (as well as many additional age groups, which we omit here for simplicity). Once recruited, the sample children were required to be brought

to a test site to take the tests in person. Therefore, an area probability design with a limited number of test sites was an efficient design of choice. NORC proposed to select a subsample of the PSU geographies in NORC's National Frame (Harter et al., 2010). The National Frame is a multi-stage cluster sample of geographies, with housing unit addresses compiled for the smallest level of geography in the sample. The geographies are sampled and the address lists are compiled following the decennial census to support face-to-face interviews throughout the decade.

For norming the tests, 16 of the National Frame's 79 highest level geographies were selected as PSUs. The PSUs were stratified the same way the National Frame had been stratified, basically by MSA status and size. The strata and PSU sample sizes are shown in Table 5. For the National Frame, stratum 1 MSAs had been selected with certainty. The PSUs were subsampled systematically with probability proportional to size (PPS) where the measure of size (MOS) was the number of Spanish-speaking households, since the Spanish-speaking children and elderly cells would be the hardest to fill. Probabilities of selection for the PSUs were the product of the original National Frame probabilities and the subsampling probabilities. Some of the stratum 1 PSUs were subsampled with certainty.

Table 5. Subsampling of PSUs from NORC's National Frame

	Stratum h	*Population*	*National Frame*	*Sample PSUs*
1.	Largest MSAs	24	24	12
2.	Other MSAs	607	17	2
3.	Non-MSA Counties	1,852	38	2
	Total	2,483	79	16

Each PSU was to be divided into smaller geographical "site areas." Each site area would contain a testing site, and the site areas were to be approximately 10 x 10 miles in urban areas and 30 x 30 miles in rural areas to provide reasonable driving distances for children to be brought to a test site. Figures 1 and 2 illustrate the process of defining site areas. In Figure 1, a 10 x 10 mile grid is placed over the Chicago MSA. Then each census tract in the Chicago MSA is assigned to a grid cell based on the geographic location of the tract centroid. The resulting site areas are shown in Figure 2.

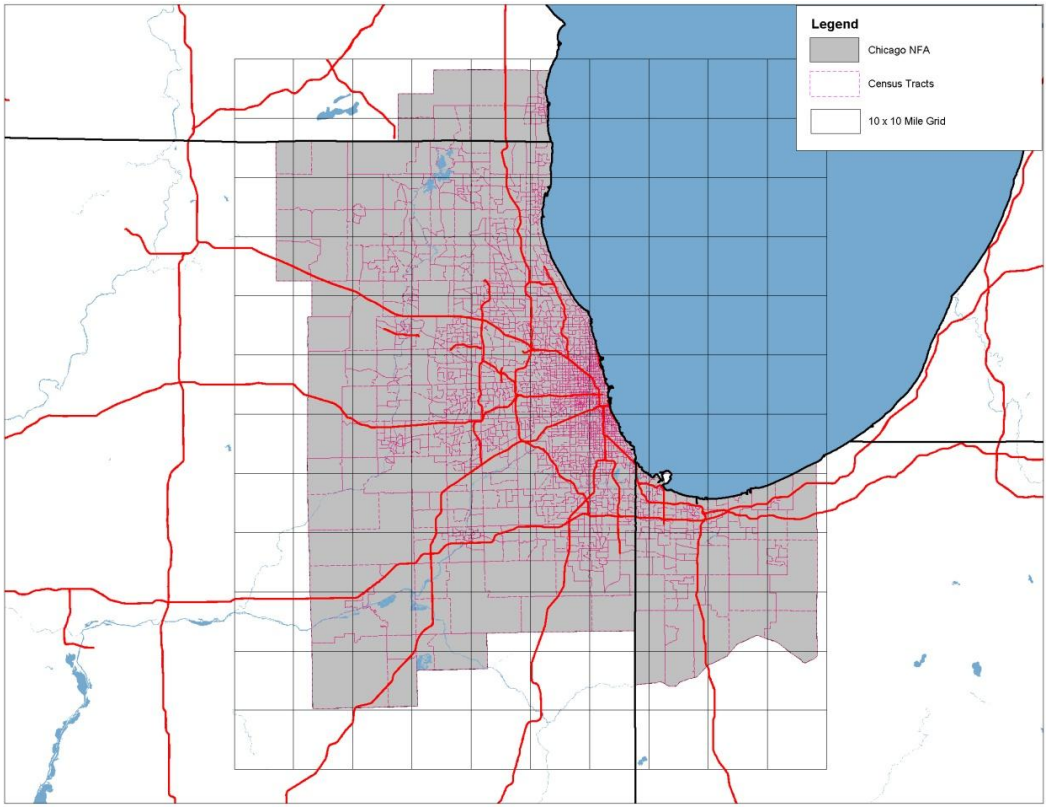

Figure 1. 10 x 10 Mile Grid over Chicago MSA

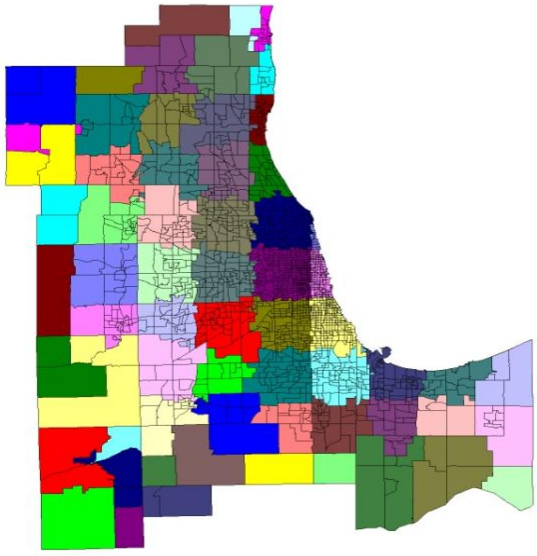

Figure 2. Site Areas in Chicago MSA with Tracts Assigned to Grid Cells

One site area was to be selected per PSU, using systematic PPS sampling where the measure of size was the number of Spanish-speaking households. Therefore, in subsequent notation, subscript *i* denotes both the PSU and the site area.

Using the U.S. Postal Service's Delivery Sequence File, we planned to select a large sample of housing units for a mail screener to roster the households' children by gender, age, and language. The screener also would solicit telephone numbers for soliciting parental cooperation for testing. In this way we planned to obtain the phase two domain frame totals N'_{hid} for each site area *i* in stratum *h*. The screener was the first phase of the study.

With the N'_{hid} frame totals in hand, and the specified sample sizes by domain, we were prepared to allocate the desired samples sizes by domain and geography for the second phase of the study to conduct the cognitive and neurological tests. We would recruit by telephone, with incentives for the sample participants to be brought to the test site.

Ultimately the sample design was never implemented, although we had subsampled the PSUs from the National Frame. Limitations in grant funding led the researchers to revert to convenience sampling near their network of cooperating universities. Nevertheless, the original plan for a probability sample allowed the original Folsom et al. (1987) result for equal probability domain samples to be generalized in a concrete way. For the sake of illustration, we continue the two-phase example with hypothetical probabilities and results.

Table 6 shows illustrative probabilities of selection for 16 test sites. These hypothetical probabilities reflect the initial National frame probabilities, the subsampling probabilities for PSUs, and the selection of one test site per PSU.

Table 6. Probabilities of Phase One Selection

Stratum	*National Frame Probability*	*Subsampling Probability*	*Conditional Site Probability*	*Unconditional Site Area Probability*	*Conditional Phase One Probability*	*Phase One Probability* $\pi_i\, g_i$
1	1	1	0.001239	0.001239	0.60	0.000743
1	1	1	0.000972	0.000972	1.00	0.000972
1	1	1	0.003408	0.003408	0.60	0.002045
1	1	1	0.003561	0.003561	0.60	0.002137
1	1	1	0.001985	0.001985	0.60	0.001191
1	1	1	0.002083	0.002083	0.60	0.001250
1	1	0.955785	0.003287	0.003142	0.60	0.001885
1	1	0.905962	0.005583	0.005058	0.60	0.003035
1	1	0.566914	0.005294	0.003001	0.60	0.001801
1	1	0.512026	0.003166	0.001621	0.60	0.000973
1	1	0.297088	0.003637	0.001081	0.60	0.000648
1	1	0.151207	0.003524	0.000533	1.00	0.000533
2	0.365373	0.221107	0.008489	0.000686	1.00	0.000686
2	0.031206	0.008483	0.099389	0.000026	1.00	0.000026
3	0.038964	0.051664	0.082108	0.000165	1.00	0.000165
3	0.031257	0.798539	0.102352	0.002555	0.60	0.001533

Suppose that we mailed questionnaires to households in the site areas to collect household rosters and telephone numbers. Table 7 shows the resulting counts N'_{hid} by domain across all 16 test sites. These counts are not actually population totals, but they are illustrative frame totals for our phase two sampling.

Table 7. Eligible Children by Domain
Phase One Frame Totals for Phase Two Sampling

Stratum	*Site*		*English-Speaking*		*Spanish-speaking Hispanics*	
h	*Area i*	*Age*	*male*	*female*	*male*	*female*
Total	*Total*	3	1,033	975	252	240
		4	1,024	975	235	237
		5	1,185	1,148	242	239
1	1	3	148	138	25	23
		4	132	147	25	28
		5	150	149	27	24
	2	3	8	9	31	33
		4	9	7	37	35
		5	12	5	30	38
	3	3	123	130	28	27
		4	125	122	24	25
		5	146	150	22	27
	4	3	62	60	24	18
		4	67	69	19	23
		5	73	77	21	22
	5	3	61	34	23	13
		4	75	39	18	13
		5	85	62	18	7
	6	3	83	77	13	10
		4	81	72	10	11
		5	102	110	14	14
	7	3	50	33	40	42
		4	66	20	32	35
		5	79	87	38	36
	8	3	98	100	5	6
		4	80	94	3	5
		5	104	103	5	5
	9	3	93	88	16	15
		4	85	94	12	16
		5	112	109	18	17
	10	3	112	105	20	24
		4	98	107	19	22
		5	104	96	17	19

	11	3	84	85	8	10
		4	88	90	11	8
		5	91	89	12	13
	12	3	44	49	2	3
		4	50	43	3	0
		5	52	45	1	2
2	13	3	23	28	5	6
		4	28	30	7	5
		5	29	27	8	5
	14	3	25	22	0	0
		4	21	27	0	0
		5	28	20	0	0
3	15	3	16	15	0	0
		4	17	13	0	0
		5	14	18	0	1
	16	3	3	2	12	10
		4	2	1	15	11
		5	4	1	11	9

The desired initial sample sizes in Table 4 divided by the frame totals in Table 7 give us the conditional overall sampling rate f'_d for each domain, as shown in Table 8.

Table 8. Sampling Rates By Domain Across Site Areas and Strata

	English-Speaking		*Spanish-speaking Hispanics*	
Age	*male*	*female*	*male*	*female*
3	0.194	0.205	0.794	0.833
4	0.195	0.205	0.851	0.844
5	0.169	0.174	0.826	0.837

From the composite measure of size in equation (14) and equations (15)-(17) we determine the allocations for each stratum, each site area, and each domain within each site area. The resulting allocations are shown in Table 9. The allocations are not integers, but controlled, probabilistic rounding can be used to preserve the probabilities while converting the allocations to integers. Alternatively, simple rounding will lead to an approximately Epsem sample design.

Table 9. Phase Two Sample Allocations by Stratum, Site Area, and Domain

Stratum	*Site*		*English-Speaking*		*Spanish-speaking Hispanics*	
h	*Area i*	*Age*	*male*	*female*	*male*	*female*
1436.82	295.52	3	29.15	28.80	20.19	19.50
		4	26.23	30.68	21.65	24.04
		5	25.76	26.41	22.70	20.43
	139.44	3	1.21	1.44	19.14	21.40
		4	1.37	1.12	24.50	22.98

		5	1.58	0.68	19.29	24.74
	102.69	3	8.81	9.86	8.22	8.32
		4	9.03	9.25	7.55	7.80
		5	9.11	9.66	6.72	8.35
	64.64	3	4.25	4.36	6.74	5.31
		4	4.63	5.01	5.72	6.87
		5	4.36	4.75	6.14	6.51
	90.58	3	7.50	4.43	11.59	6.88
		4	9.30	5.08	9.72	6.96
		5	9.11	6.86	9.44	3.72
	95.94	3	9.72	9.56	6.24	5.04
		4	9.57	8.94	5.15	5.62
		5	10.42	11.59	7.00	7.09
	99.07	3	3.88	2.72	12.73	14.04
		4	5.17	1.65	10.92	11.85
		5	5.35	6.08	12.60	12.08
	33.38	3	4.73	5.11	0.99	1.25
		4	3.89	4.80	0.64	1.05
		5	4.37	4.47	1.03	1.04
	78.89	3	7.56	7.58	5.33	5.25
		4	6.97	8.10	4.29	5.67
		5	7.94	7.97	6.25	5.97
	170.35	3	16.86	16.74	12.34	15.55
		4	14.88	17.06	12.57	14.43
		5	13.64	13.00	10.92	12.36
	177.05	3	18.97	20.34	7.41	9.72
		4	20.05	21.53	10.92	7.87
		5	17.91	18.08	11.57	12.69
	89.28	3	12.09	14.26	2.25	3.55
		4	13.86	12.52	3.62	0.00
		5	12.45	11.13	1.17	2.38
849.47	67.65	3	4.91	6.33	4.38	5.51
		4	6.03	6.79	6.57	4.65
		5	5.40	5.19	7.29	4.61
	781.82	3	139.11	129.70	0.00	0.00
		4	117.88	159.18	0.00	0.00
		5	135.82	100.14	0.00	0.00
113.71	84.62	3	14.17	14.08	0.00	0.00
		4	15.19	12.20	0.00	0.00
		5	10.81	14.35	0.00	3.83
	29.09	3	0.29	0.20	4.70	4.11
		4	0.19	0.10	6.30	4.58
		5	0.33	0.09	4.48	3.72

7. Phase Two Allocations Exceeding Frame Totals

In the above example, some of the allocations are actually larger than the frame counts. This is a potential problem with all versions of the problem. Folsom et al. (1987) suggested that a case can be selected more than once, inflating the weight by the number of times the case is selected. Alternatively, in the unstratified version of the original problem, domains or site areas can be collapsed until the following condition is met:

$$S_i \geq n^* f_d \tag{18}$$

for all domains *d*. That is, for all PSUs and all domains, the total measure of size for the PSU must be larger than the approximate share of the PSU's sample in each domain. Since n^* can be expressed as n/m, this condition is equivalent to

$$f_d \leq m\, S_i \,/\, n = \pi_i \tag{19}$$

for all *i* and *d*. By collapsing domains with large and small values of f_d, or by collapsing PSUs with large and small values of π_i, the extreme values can be shrunk toward the middle to help satisfy the above condition.

In the revised, two-phase application, the above condition is somewhat modified. In the case of a single stratum, given *n* and the PSU selection probabilities π_i's, we need $n_{id} \leq N'_{id}$ which implies from equation (10) that

$$n_{i.}\, f'_d \leq S'_i$$

That is, the expected total phase one sample for each PSU must be larger than the approximate share of the PSU's phase two sample in each domain. Restating the condition,

$$n_{i.}\, f'_d \leq \sum_d f'_d\, N'_{id}$$

This implies further from (11) that

$$nf'_d \leq \pi_i g_i \sum_i \frac{\sum_d f'_d N'_{id}}{\pi_i g_i} = \pi_i g_i \sum_d f'_d \sum_i \frac{N'_{id}}{\pi_i g_i} = \pi_i g_i \sum_d f_d \widehat{N}_d \text{ or}$$

$$f'_d \leq \pi_i g_i \frac{\sum_d f'_d \widehat{N}_d}{n}, \tag{20}$$

where $\widehat{N}_{d'}$ is an estimate of the domain population size from the first phase sample. When the N_{id} values are known, and there is no need for a phase one survey, then $\sum_d f_d\, N_{id} = n$, resulting in a condition identical to (19) of Folsom, et al. However, $\sum_d f'_d\, N'_{id} \neq n$, so condition (20) does not simplify.

With a stratified two-phase design, the above condition in (20) is

$$f'_d \leq \pi_{hi} g_{hi} \frac{\sum_d f'_d \widehat{N}_d}{n} \tag{21}$$

Collapsing domains or PSUs to reduce extreme values is an option. Alternatively, or in addition, the phase one sample size can be increased. An increased sample size overall

will increase the phase two frame size, reducing f'_d and increasing the numerator on the right-hand side of (21). The phase one sample sizes can be increased selectively in domains with small values of π_i or π_{hi}, because those are the PSUs most likely to have disproportionately large allocations in an attempt to equalize the probabilities overall. In our example, we sampled all frame households into the phase one sample to help equalize the probabilities. Even so, collapsing and increasing phase one samples may not fully rectify the problem, as was the case in this example. It is always an option to select all cases in a PSU domain and live with the inequality of weights, or to allow multiple selections and weight accordingly.

For some studies, whether two-phase or not, it may be more appropriate to set bounds on the n_{id} allocations, and equalize the weights as much as possible within domain using an optimization routine, without insisting on full equality. See, for example, Gabler et al. (2009).

8. Concluding Remarks

The use of a composite measure of size to allocate equal probability samples for multiple domains in two-stage samples is a useful technique currently available in the SUDAAN software system (http://www.rti.org/sudaan/page.cfm?objectid=FA210070-7FAE-4E83-B127D764B8C274B2) and employed successfully at RTI International for many years for studies such as The National Survey of Child and Adolescent Well-Being (http://www.rti.org/page.cfm?objectid=D688C979-8B27-456E-AD0AF638862E7365). The generalizations presented here extend the technique to additional situation of multiple domains where the population totals are not known in advance for all PSUs, and where PSUs are pre-selected. It is clear that sometimes the allocations to PSUs and domains can exceed the available frame counts. Sometimes the situation can be remedied, or at least ameliorated somewhat, but the technique may not work perfectly in all situations, especially where the PSU probabilities are quite diverse. Nevertheless, these techniques are useful tools to sampling statisticians in a variety of situations.

References

Cochran, W. (1977). *Sampling Techniques, 3rd Ed*. John Wiley & Sons: New York.

Folsom, R.E., Potter, F.J., and Williams, S.R. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. In *Proceedings of the American Statistical Association*, Section on Survey Research Methods, 792-796.

Gabler, S., Ganninger, M. and Münnich, R. (2009). Optimal allocation of the sample size to strata under box constraints. *Metrica*, Springer-Verlag.

Harter, R., Eckman, S., English, N., and O'Muircheartaigh, C. (2010). Applied Sampling for Large-Scale Multi-Stage Area Probability Designs. In *Handbook of Survey Research, Second Edition*, P. Marsden and J. Wright, eds. Emerald.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons: New York.

Lohr, S. (1999). *Sampling: Design and Analysis, 2nd Ed*. Brooks/Cole: Boston.