SAMPLE ALLOCATION AND SELECTION METHODS FOR OVERSAMPLING SUBPOPULATIONS

K.P. Srinath, Abt Associates Inc.
4800 Montrgomery Lane, Bethesda, MD 20814

Key Words: Domain, Varying Probability

1 Introduction

Generally, in sample surveys the estimates for certain small subpopulations whose members cannot be identified in advance of sampling are not precise because the number of sampled units belonging to these subpopulations is small. Often, there is a need to improve the precision of these estimates or sometimes there is a requirement for a predetermined expected number of units belonging to a subpopulation in the overall sample for purposes of data analysis. Several techniques are available for increasing the expected sample size belonging to a subpopulation or subpopulations. For example, a simple but expensive method of achieving an expected sample size for a subpopulation is to increase the size of the overall sample. Another method is to stratify the population according to the density of the subpopulation and then use nonproportional allocation. A third method is to use a two-phase design in which a screening interview is conducted on a large sample in the first phase to identify the members of the subpopulation and retain them in the sample, and then select a subsample of the sample from the group not belonging to this subpopulation. Strategies for improving subpopulation estimates with differential sampling rates when there are two strata in the population have been given by Waksberg (1973) and Weller, Huggins and Singh (1991). A discussion of some of the techniques for increasing the sample size in the subpopulations can be found in Massey, Judkins and Waksberg (1993).

Some of the techniques used to achieve the desired sample size from the subpopulations might make the overall estimates very inefficient due to allocation or selection methods that are very different from the optimum methods needed to obtain precise overall estimates. Therefore, there is a need to balance the need for improving the precision of the subpopulation estimates with the loss in efficiency of the estimates for the general population.

In this paper, some methods of sample allocation are proposed which attempt to keep the strata sample sizes close to the sample sizes which are considered optimum from the point of view of efficiency of the overall estimates. The stratification boundaries are the same as those created for maximizing the efficiency of the overall estimates. A method of revising the probabilities of selection of primary sampling units which maximizes the expected proportion of sampled units belonging to a subpopulation in the overall sample subject to certain constraints is also suggested.

2. Sample Allocation Methods

2.1 Allocation with an increase in sample size

Let a population of **N** units be stratified into **L** strata. Let $\mathbf{N_h}$ be the number of units in the population in the hth stratum. Let **n** be the number of units in the sample that is required to obtain an overall estimate of some characteristic of interest with a prespecified precision.

Let $\mathbf{N_{dh}}$ be the number of units in the population in the hth stratum belonging to the subpopulation or domain of interest. Then, N_{dh} / N_h is the proportion of the population in the hth stratum which belongs to the subpopulation or domain. Let $\mathbf{n_h}$ be the sample number of units allocated to the hth stratum using proportional, optimum, Neyman or some other allocation for purposes of estimating the overall population parameters like totals, means, ratios or proportions efficiently. The expected number of units in the sample belonging to the domain is given by

$$E(n_{dh})=n_h\frac{N_{dh}}{N_h} \quad .$$

The expected number of units belonging to the domain in the overall sample is given by

$$E(n_d)=\sum_{h=1}^{L} E(n_{dh}) \quad .$$

Suppose $E(n_d)$ ressulting from an allocation like proportional or Neyman of the overall sample is too small for obtaining reasonably precise estimates for the domain of interest and there is a requirement that this be number be $\mathbf{n_d^*}$. To meet this requirement, the sample size in each stratum may have to be increased resulting in an increase in the overall sample size. Since the proportions of the population belonging to the domain of interest could be very different in different strata, there are several ways of allocating the sample to different strata to get the required domain sample size. Allocations that strictly minimize the overall increase in sample size may make the overall estimates more imprecise than necessary.

One criterion in determining the new allocation is to make the differences between the new sample allocation and the old allocation as small as possible and at the same time achieve the desired expected sample size for the domain. Let $\mathbf{n_h^*}$ be the number of units that are required to be selected from the hth stratum.

We want n_h^* to be such that

$$\sum_{h=1}^{L} (n_h^* - n_h)^2$$

is a minimum subject to the constraint

$$\sum_{h=1}^{L} n_h^* \frac{N_{dh}}{N_h} = n_d^* \quad .$$

The allocation which satisfies the above criterion is as follows.

$$n^*_h = n_h + [n^*_d - E(n_d)] \frac{P_{dh}}{\sum_{h=1}^{L} P^2_{dh}}$$

where $P_{dh} = \frac{N_{dh}}{N_h}$.

2.2 Example

We use a slightly modified example from Cochran (1977) to illustrate the allocation method. The data are derived from a stratified sample of tire dealers. The dealers are assigned to strata according to the number of new tires held. The population means m_h and variances v_h of the number of new tires held are shown. If we are interested in the estimating the overall average number of new tires held, then using Neyman allocation, the number of tire dealers to be sampled(n_h) from each stratum is shown. Suppose we are interested in tire dealers who belong to a specific domain, for example, those who hold a specific brand of tire and we want to estimate the total number of tires belonging to this brand. In this case, there may be a requirement for a certain number of sampled units belonging to this domain. The proportion of dealers belonging to the subpopulation or domain P_{dh} in each stratum is also shown. The stratum boundaries are denoted by B_h.

St.	B_h	N_h	P_{dh}	m_h	v_h	n_h
1	1-9	19,850	0.05	4.1	34.8	2832
2	10-19	3,250	0.10	13.0	92.2	755
3	20-29	1,007	0.15	25.0	174.2	321
4	30-39	606	0.20	38.2	320.4	262
T		24,713	0.064			**4170**

Using n_h and P_{dh} given above, we see that $E(n_d) = 318$. Suppose $n_d^* = 432$. We want the domain expected sample size to be 432 instead of 318. The new sample allocation and the expected domain sample size are shown below along with old sample size and the expected domain sample size.

St.	n_h	$E(n_{dh})$	n_h^*	n_{dh}^*
1	2832	142	2908	145
2	755	76	907	91
3	321	48	549	83
4	262	52	566	113
T	4170	318	4930	432

The increase in sample size is 760 units. If we were drawing a simple random sample, then we would require a sample of 6750 units to get an expected domain sample size of 432.

To get an idea of the loss in efficiency due to this allocation, we compare n_h^* with n_{opt} using Neyman allocation. The two are shown below.

Stratum	n_h^*	n_{opt}
1	2908	3347
2	907	892
3	549	380
4	566	311
Total	4930	4930

Comparing the variances of the sample mean $\bar{y}_{st}$ for the two allocations, we see that the loss in efficiency due to using n_h^* instead of n_{opt} is around 6 %. Of course, using n_{opt} will not meet the requirement of getting a sample 432 belonging to the domain. The expected sample size will be 375.

The new sample allocation depends on the initial allocation. If the initial allocation has been different say as shown in the following table, then the new allocation to meet the sample requirements follows closely the initial allocation. Both the allocations are shown below.

Stratum	n_h	$E(n_{dh})$	n_h^*	n_{dh}^*
1	3000	150	3083	154
2	600	60	767	77
3	340	51	590	88
4	230	46	563	113
Total	4170	307	5003	432

2.3 Allocation with a fixed sample size

Sometimes it is possible to achieve the domain sample size without an increase in the overall sample size, if this is important because of a fixed budget. In this case, as before we want to minimize the quantity

$$\sum_{h=1}^{L} (n_h^* - n_h)^2$$

but subject to two constraints which are

$$\sum_{h=1}^{L} n_h^* \frac{N_{dh}}{N_h} = n_d^*$$

and

$$\sum_{h=1}^{L} n_h^* = n \quad .$$

The allocation that satisfies these conditions is n_h^*

$$= n_h + \frac{[n_d^* - E(n_d)]}{\sum_{h=1}^{L} P_{dh}^2 - \frac{(\sum_{h=1}^{L} P_{dh})^2}{L}} \left(P_{dh} - \frac{\sum_{h=1}^{L} P_{dh}}{L}\right)$$

Turning to the previous example, and using the above expression and values of n_h as those that were obtained under Neyman allocation we get the new allocation as follows.

St.	N_h	n_h Neyman	n^*_h	n^*_h
1	19,850	2832	2148	1893
2	3,250	755	527	688
3	1,007	321	549	983
4	606	262	946	606
T	24,713	4170	4170	4170

The initial allocation yields a sample of 946 in stratum 4, but the population in this stratum is only 606. Therefore, all the units in stratum 4 are included in the sample. The remaining number of units to be sampled which is 3564 is reallocated using the same formula given, but now applied to three strata, which results in the final allocation. The values of n_d^* resulting from the new allocation are give below.

St.	n_h^*	n_{dh}^*
1	1893	95
2	688	69
3	983	147
4	606	121
Total	4170	432

Generally, the initial overall sample size is not large enough to provide a sample of the desired size belonging to a domain. In such cases, the overall sample will have to be necessarily increased.

2.4 Allocation for two domains

Suppose there are two domains for which we need prespecified expected number of units in the overall sample. Then, one method of meeting this requirement is to do the allocation sequentially. First the overall sample is allocated for maximizing the precision of the overall estimate. Then this allocation is changed to accommodate the first requirement. The resulting allocation is again changed to satisfy the second requirement.

This procedure is illustrated by going back to the original example and assuming two domains of interest. The proportions of the two domains in each stratum are shown below. Let P_{dh} denote the proportion of units in domain d in stratum h and $P_{d'h}$ denote the proportion of units in domain d' in stratum h.

St.	N_h	n_h (Neyman)	P_{dh}	$P_{d'l}$	n_{dh}	$n_{d'h}$
1	19,850	2832	0.05	0.02	142	56
2	3,250	755	0.10	0.05	76	38
3	1,007	321	0.15	0.10	48	32
4	606	262	0.20	0.15	52	39
T	24,713	4170	0.064	0.03	318	165

There are two domains of interest. 6.4% of the total units in the population belong to domain 1 and 3% of the total units belong to domain 2. We want say 432 units in the sample which belong to domain 1 and 215 units in the sample belonging to domain 2. That is we want $n_d^* = 432$ and $n_{d'}^* = 215$.

Using the allocation given earlier for one domain with an increase in sample size ,we now get a new allocation to satisfy the requirement for domain 2. This is shown in the table below. We also compute the expected domain sample size for domain 1 with this new allocation.

	n_h^*	$n_{d'h}^*$	n_{dh}^*
1	2860	57	143
2	825	41	82
3	462	46	69
4	474	71	95
T	4621	215	389

We see from the above table that the requirement for domain 2 is satisfied but not domain 1.
We need 432 but we only have 389. The allocation is now changed to meet the requirement for domain 2. This results in the final allocation as follows.

St.	$n_h^*(final)$	n_{dh}^*	$n_{d'h}^*$
1	2889	144	58
2	882	88	44
3	548	82	55
4	589	118	88
T	4908	432	245

In the final allocation both the requirements are met.

An alternative is to minimize the quantity

$$\sum_{h=1}^{L} (n_h^* - n_h)^2$$

under the constraints

$$\sum_{h=1}^{L} n_h^* \frac{N_{dh}}{N_h} = n_d^* \quad \text{and} \quad \sum_{h=1}^{L} n_h^* \frac{N_{d'h}}{N_h} = n_{d'}^*.$$

The allocation n_h^* which satisfies the above conditions is give below.

$$n_h^* = n_h - \frac{(n_d^* - E(n_d))}{(\sum_{h=1}^{L} P_{dh}P_{d'h})^2 - \sum_{h=1}^{L} P_{dh}^2 \sum_{h=1}^{L} P_{d'h}^2} (\sum_{h=1}^{L} P_{d'h}^2) P_{dh}$$

$$+ \frac{[n_{d'}^* - n_{d'}]}{(\sum_{h=1}^{L} P_{dh}P_{d'h})^2 - \sum_{h=1}^{L} P_{dh}^2 \sum_{h=1}^{L} P_{d'h}^2} (\sum_{h=1}^{L} P_{dh}P_{d'h}) P_{dh}$$

$$+ \frac{(\sum_{h=1}^{L} P_{dh}P_{d'h})(n_d^* - E(n_d)) - (\sum_{h=1}^{L} P_{dh}^2)(n_{d'}^* - E(n_{d'}))}{(\sum_{h=1}^{L} P_{dh}P_{d'h})^2 - \sum_{h=1}^{L} P_{dh}^2 \sum_{h=1}^{L} P_{d'h}^2} P_{d'h}$$

This allocation will give the exact expected sample sizes in the domains but may not be necessarily be efficient.

3. Revision of selection probabilities.

In large scale household surveys, the sample is usually selected in several stages. The primary sampling units at the first stage of selection are usually selected with probability proportional to the total population in each primary sampling unit in order to obtain efficient estimates of population totals and means. If the primary sampling units have very different proportions of the total population which belong to a subopulation of interest and if it is desired to maximize the expected proportion of the sample which belong to this population, then it is possible to revise the original probabilities of selection to achieve this goal and at the same time keep the revisions to a minimum so as not to affect the efficiency of the overall estimates. In this section, a simple procedure of revising the probabilities is suggested.

Suppose we are selecting a sample of n primary sampling units (PSU) from N units. Let the number of elementary units in the ith psu be N_i. Let the number of elementary units belonging to the jth domain in the ith psu be N_{ij}.

Let $P_{ij} = \frac{N_{ij}}{N_i}$ be the proportion of the jth domain in the ith psu.

The size of the domain in the population is

$$N_j = \sum_{i=1}^{M} N_{ij}.$$

Let n_j be the sample size from domain j in the overall sample. Let π_i be the probability of inclusion of ith psu in the sample. Generally, these inclusion probabilities are proportional to some measure of size to maximize the efficiency of overall estimates. Procedures for revising the measures of size or deriving composite measures of size to reflect domain sizes and also achieve desired domain sample sizes can be found in Folsom, Potter and Williams (1987) and Fahimi and Judkins (1991).

In this section, the idea is to revise the optimum probabilities of selection so as to maximize the expected proportion of thesample which belongs to a specific domain. The objective is to keep the revised probabilities of selection as close to the original as possible and at the same time maximize the expected proportion of domain units in the sample.

Let π_i^* be the revised probability of selection of the ith psu. We want to maximize the quantity

$$\sum_{i=1}^{M} \pi_i^* P_{ij} - \sum_{i=1}^{M} (\pi_i^* - \pi_i)^2$$

subject to the constraint that

$$\sum_{i=1}^{M} \pi_i^* = n.$$

Using this criterion, the revised probabilities for the selection of psus when we are interested in one domain is

$$\pi_i^* = \pi_i + \frac{1}{2}[P_{ij} - \frac{\sum_{i=1}^{M} P_{ij}}{M}]$$

Consider the following example in which we want to select 3 psus out of 7.

The initial probabilities of inclusion and also the revised probabilities are shown.

PSU	N_i	P_{ij}	π_i	π_i^* Revised
1	150	0.2	0.375	0.32
2	50	0.4	0.125	0.17
3	300	0.10	0.750	0.65
4	250	0.50	0.625	0.72
5	150	0.15	0.375	0.30
6	100	0.60	0.250	0.40
7	200	0.20	0.500	0.44
Total	1200		3	3

References

Massey, J.T., Judkins D. and Waksberg J. (1993) " Collecting Health Data on Minority Populations," Proceedings of the Section on Survey Research Methods, American Statistical Association, 75-84.

Waksberg, J. (1973) "The Effect of Stratification With Differential Sampling Rates on Attributes of Subsets of the Population, " Proceedings of the Section on Social Statistics Section, American Statistical Association, 429-434.

Weller, G.D., Huggins, V.J. and Singh, R.P. (1991) " Oversampling The Low Income Population In the Survey of Income and Program Participation, " Proceedings of the Section on Survey Research Methods, American Statistical Association, 544-549.

Fahimi, M. and Judkins, D. (1991) " PSU Probabilities Given Differential Sampling at Second Stage, " Proceedings of the Section on Survey Research Methods, American Statistical Association, 538-543.

Folsom, R.E., Potter, F.J. and Williams, S.R. (1987) " Notes on a Composite Measure for Self-Weighting Samples in Multiple Domains ", Proceedings of the Section on Survey Research Methods, Amereican Statistical Asssociation, 792-796.

Cochran, W.G. Sampling Techniques, John Wiley & Sons, 1977, New York.