Allocation to Strata When Sample Selection is Through Screening a Larger Sample

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

Key Words: Eligibility Rate, Screening Cost

1. Introduction

In sample surveys, sometimes, the sample from a population of interest is selected through screening a larger sample because the target population cannot be identified in advance of sampling. For example, if we are interested in sampling households with children, a large sample of households is screened to select a sample of households with children. The size of the screening sample depends on the desired sample from this subpopulation of interest and the eligibility rate which is the proportion of this subpopulation in the general population. When the general population is divided into strata and the proportion of the subpopulation varies widely among strata, it is of interest especially when screening costs are high, to look at allocations which minimize the screening sample without too much loss in precision of the estimates relating to the subpopulation. Disproportionate allocations may also be of interest when we want to increase the domain sample size for a given screening sample size. We examine such allocations and provide an example.

The number of sampling units that need to be screened to obtain a specified domain sample size depends on the eligibility rate. If we are selecting a simple random sample, then the number of sampling units that need to be screened is simply obtained by dividing the desired domain sample size by the eligibility rate. An alternate strategy is to stratify the population according to the density of the domain or the subpopulation and then use nonproportional allocation. This disproportional allocation while decreasing the screening costs, may increase the variance of the overall estimate. Disproportional allocation results in widely varying sampling weights due to thin samples from strata containing a low proportion of the domain population of interest. We want to minimize the increase in the variability in the weights and at the same try to minimize the screening costs. In this paper, we are assuming that we are not interested in producing estimates for the general population. If this is also of interest, then the allocation of the screener sample should try to minimize the loss in efficiency of the estimates for the general population.

A discussion of some of the techniques to minimize this loss can be found in Srinath (1996).

2. Notation

Let M_h denote the total number of sampling units in the screening population in the hth stratum. Assume that there are L strata. That is, h= 1,2,3......... L. Let the number of sampling units in the population that belongs to the domain of interest be N_h. Then the eligibility rate in the hth stratum is defined as

$$e_h = \frac{N_h}{M_h}.$$

The overall eligibility rate in the population is

$$e = \frac{N}{M} = \frac{\sum_{h=}^{L} M_h e_h}{M} \quad \text{where} \quad M = \sum_{h=1}^{L} M_h \quad \text{and}$$

$$N_h = \sum_{h=1}^{L} N_h.$$

Let the number of sampling units selected in the hth stratum for screening be m_h. Let the expected number of sampling units falling into the domain resulting from screening m_h units
be n_h. The total expected sample size in the domain is denoted by

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

Note that $n_h = e_h m_h$. The total number of households to be screened is given by

$$m = \sum_{h=1}^{L} m_h.$$

If we want a total sample of n units belonging to the domain of interest, then if we are drawing a simple random sample, the size of the screening sample is $m = \frac{n}{e}$. We will assume that there is no nonresponse to the survey either at the screening stage or at the data collection stage.

3. Sample Allocation

In a stratified population in which eligibility rates vary by strata, we want m_h such that

$$\sum_{h=1}^{L} e_h\, m_h = n.$$

When the number of screening sampling units is fixed, a common allocation is proportional allocation, which is $m_h = m\,\frac{M_h}{M}$. This is the same as allocating the total domain sample size using

$$n_h = n\frac{N_h}{N}.$$

This allocation results in the same weight for all selected units in the sample and the number of sampling units screened is the same as in simple random sampling.

Another allocation which uses the information on variability of a characteristic within strata is Neyman allocation. If we are estimating some characteristic like the proportion of sampling units in the domain in the hth stratum possessing a certain characteristic and if this proportion in the population P_h in the hth stratum, then an allocation which minimizes the variance of the overall sample proportion p where p is

$$p = \frac{\sum_{h=1}^{L} N_h\, p_h}{N},$$

and p_h is the sample proportion in the hth stratum is given by

$$n_h = n\,\frac{W_h\,\sqrt{P_h(1-P_h)}}{\sum_{h=1}^{L} W_h\,\sqrt{P_h(1-P_h)}}.$$

Here $W_h = \frac{N_h}{N}$.

These allocations do not attempt to minimize the screening sample or for a given screening sample size, maximize the domain sample size. We want to find an allocation of a fixed screener sample size that minimizes the conditional variance of the subpopulation estimate.

To determine this allocation, we must first look at the variance of the sample proportion p defined earlier. The conditional variance of p (conditional on getting the same number of subpopulation units n_h in the hth stratum in repeated sampling) is given by

$$V(p) = \sum_{h=1}^{L} \frac{W_h^2\; P_h(1-P_h)}{n_h}.$$

ignoring the finite population correction. If we put $n_h = e_h m_h$ in the variance, then we get

$$V(p) = \sum_{h=1}^{L} \frac{W_h^2\; P_h(1-P_h)}{e_h m_h}.$$

We want to minimize $V(p)$ subject to the condition $\sum_{h=1}^{L} m_h = m$. This leads to allocation of the total number of households that we want to screen as given below. The number of households that need to be screened in the hth stratum is given by

$$m_h = m\,\frac{\dfrac{W_h\sqrt{P_h(1-P_h)}}{\sqrt{e_h}}}{\sum_{h=1}^{L}\dfrac{W_h\sqrt{P_h(1-P_h)}}{\sqrt{e_h}}}.$$

If we assume that the variances are the same within strata, then this allocation reduces to

$$m_h = m \frac{\frac{W_h}{\sqrt{e_h}}}{\sum_{h=1}^{L} \frac{W_h}{\sqrt{e_h}}}$$

which is similar to proportional allocation except for the eligibility rates. This allocation can also be written as

$$m_h = m \frac{M_h \sqrt{e_h}}{\sum_{h=1}^{L} M_h \sqrt{e_h}}. \qquad (1)$$

The expected number of subpopulation units in the sample resulting from this allocation may be greater than what we get under proportional allocation or simple random sampling and gives a higher precision for the same number of screened households.

If we want exactly "n" completed interviews, then we set the number of households to be screened under this allocation equal to

$$m = n \frac{\sum_{h=1}^{L} \frac{W_h}{\sqrt{e_h}}}{\sum_{h=1}^{L} W_h \sqrt{e_h}}. \qquad (2)$$

This will result in a decrease in the number of households that need to be screened for the same expected subpopulation sample size. There is an increase in the variance of the estimates but this increase is more than offset by the decrease in screening costs.

4. Example

Assume that we want to select a sample of 1,000 persons belonging to a minority group, say Hispanic population through random digit dialing (RDD). If it is known that 10.7% of the general population belongs to this group, then we need to screen 9,346 persons to obtain a sample of 1,000 persons belonging to this group. We stratify telephone exchanges according the percent of Hispanic population out of the total in the exchange. The distribution of telephone exchanges by strata is as given in Table 1.

If we allocate the screener sample of size 9,346 persons to each stratum proportionately, then we would get an expected sample of 1,000 persons belonging to the Hispanic group. To use the allocation based on eligibility rates, we first determine the screener sample size that is required to achieve a sample of 1,000 persons. Using (2) we compute this to be 4,450 persons. Now, using (1) we allocate this sample to each stratum. Table 2 shows both the proportional allocation and the new allocation of the screener sample size and the expected sample size from the minority population.

We see from Table 2 that we need only a sample of 4,450 to obtain a sample of 1,000 belonging to the subpopulation of interest. If we assume equality of within strata variances, then we can compute the product of the screener sample size and the variance of the estimate achieved under each allocation. The ratio of this product under proportional allocation to the product under the new allocation is 1.54. There are other arbitrary allocations which one could do to reduce the screener sample size but those allocations will have either a higher variance that is not sufficiently offset by the decrease in screening costs or a higher screening cost that is not offset by the decrease in variance. One such example is seen in Table 3. The increase in the variance due to disproportional allocation of the screener sample is not sufficiently offset by the decrease in screening costs. The ratio of the product under this allocation relative to the eligibility rate allocation is again 1.54. The efficiency as measured by ratio of the product of the two quantities defined above is maximum for the allocation based on the eligibility rates.

In conclusion, the allocation based on eligibility rates does reasonably well in balancing screening costs and the loss in precision of the estimates.

References:

Srinath, K.P. (1996) " Sample Allocation Methods for Oversampling Subpopulations," Proceedings of the Section on Survey Research Methods, American Statistical Association.

Table 1: Percentage of Telephone Exchanges by Percent Hispanic

Stratum (% Hispanic)	Percentage of Telephone Exchanges	Percentage of the Total Population	Percentage of the Hispanic Population
0-<5	60.7	59.1	9.6
5-<20	23.7	24.1	24.1
20-40	9.3	9.7	25.2
40-<60	3.7	3.8	17.4
60 and >	2.6	3.3	23.7
Total	100.0	100.0	100.0

Table 2: Allocation of the Screener Sample and the Expected Subpopulation Sample Size

Stratum	Eligibility Rate (%)	Proportional Allocation	Expected Minority Sample	Allocation Based on Eligibility Rate	Expected Minority Sample
0-<5	1.74	5,524	97	1,316	23
5-<20	10.7	2,253	241	1,331	142
20-<40	27.9	906	253	864	241
40-<60	49.5	355	176	449	222
60 and >	75.8	308	233	490	372
Total	10.7	9,346	1,000	4,450	1,000

Table 3: Arbitrary Allocation

Stratum	Arbitrary Allocation	Expected Sample Size
0-<5	287	5
5-<20	664	71
20-<40	692	193
40-<60	479	237
60 and >	652	494
Total	2,774	1,000