Advances and Applications in Statistics
Volume 15, Number 1, 2010, Pages 49-59
Published Online: May 22, 2010
This paper is available online at http://www.pphmj.com

A SIMULTANEOUS DETERMINATION OF THE OPTIMAL SAMPLE SIZE AND SAMPLING ALLOCATIONS FROM A STRATIFIED FINITE POPULATION

SHAUL K. BAR-LEV and BENZION BOUKAI

Department of Statistics
University of Haifa
Haifa 31905, Israel

Department of Mathematical Sciences
IUPUI, Indianapolis, IN 46202, U. S. A.

Abstract

A procedure for the simultaneous determination of the optimal sample size and the optimal sampling allocations from a stratified population is presented. Motivated by a real-life practicable application, the procedure allows, with a pre-assigned confidence level, the determination of the overall sample size required for achieving a prescribed level of (proportional) accuracy of the weighted sample average, while at the same time it provides the optimal sampling allocations to the various population's strata. The objective is to draw sample of the entire population in a manner that would faithfully reflect, with a high confidence, a known characteristic of the population for the purpose of a personal interview or other such matters. The procedure is illustrated on a real-life data and an extensive numerical validation study is also provided.

1. Introduction

The present study is motivated by a real-life practical problem related to sampling units from a stratified finite population. The objective is to draw a small

2010 Mathematics Subject Classification: 62D55.

Keywords and phrases: sampling, optimal allocation, stratified sampling.

Received November 20, 2009

subset of the entire population in a manner that would faithfully reflect, with a high confidence, a known characteristic of the population (on a variable of interest) for the purpose of a personal interview. To fix idea, consider a private banker or a brokerage firm that is interested in inviting for a personal interview a sample of its clients, based on the known values of the 'size' of the clients' portfolio or account, appropriately stratified, to discuss the firm's services or other such matters. We stress that unlike accounting audit problems (see for example, Smith [6] or Knight [3]), the application at hand does not involve any statistical estimation problems *per say*, as the entire population values of the underlying variable of interest are considered to be known, but rather, it is a re-visitation of the 'classical' problem of determining sampling allocations in the stratified settings (see Evans [2]). Such a problem is considered nowadays to be a standard problem in many textbooks on the subject, either as proportional allocations (proportional to the relative sizes of the strata), or by some means of cost considerations. However, all the available approaches assume that the overall sample size is given *a priori*. This is of little use to those, like our Bankers, who wish to minimize the sampling costs while at the same time maintain a desired precision or accuracy with a high level of overall confidence that the sample drawn would faithfully reflect the underlying population values. Henceforth, we propose a procedure that allows the simultaneous determination of the overall sample size required for achieving a prescribed level of (proportional) accuracy of the sample weighted average – all at a given level of statistical confidence, while at the same time it provides the optimal sampling allocations to the various population's strata.

The basic notation and setup needed to describe this particular application is provided below. In Section 2, we present the proposed procedure for the simultaneous determination of the optimal sample size and the optimal sampling allocations to the various strata. In Section 3, we illustrate this procedure by implementing it on a 'real-life' data, though coded to protect some proprietary information. While the 'story' on the brokerage firm we provide as an example might be construed as contrived, we feel nonetheless it provides the appropriate motivation to the problem. In Section 4, we present the results of some numerical studies, where extensive re-sampling and simulations are used to validate the proposed procedure and the results obtained.

We begin with some standard notation and a well known setup. Consider a finite population of N units comprised of k strata, each with N_i relatively homogeneous

units with respect to the variable or the characteristic of interest, X. Thus, the corresponding set of $N = \sum_1^k N_i$ fixed values, $\{x_{ij},\ i = 1, ..., k;\quad j = 1, ..., N_i\}$ completely determines the population distribution X. Let μ_i and σ_i^2 denote the mean and the variance of X in the ith stratum;

$$\mu_i = \frac{1}{N_i}\sum_1^{N_i} x_{ij}, \qquad \sigma_i^2 = \frac{1}{N_i}\sum_1^{N_i}(x_{ij} - \mu_i)^2. \tag{1}$$

Ignoring the given stratification, the overall mean and variance of X in the entire population are:

$$\mu_x = \frac{1}{N}\sum_1^{k}\sum_1^{N_i} x_{ij}, \qquad \sigma_x^2 = \frac{1}{N}\sum_1^{k}\sum_1^{N_i}(x_{ij} - \mu_x)^2. \tag{2}$$

Let $w_i = N_i/N,\ \ i = 1, ..., k.$ It is well known that

$$\mu_x = \sum_1^k w_i\mu_i, \tag{3}$$

so that the population mean of X, is a weighted average of the k strata means, $\mu_1, \mu_2, ..., \mu_k,$ with the corresponding weights of $w_1, w_2, ..., w_k$. Note that the variance of X in (2) may be decomposed as $\sigma_x^2 = \sigma_W^2 + \sigma_B^2,$ where $\sigma_W^2 = \sum_1^k w_i\sigma_i^2,$ is the variance component measuring the within-strata variability and where $\sigma_B^2 = \sum_1^k w_i(\mu_i - \mu_x)^2,$ is the variance component measuring the between-stratum variability. The ratio, $\sigma_B^2/(\sigma_B^2 + \sigma_W^2),$ is considered as a measure of the proportion of the total variability in the values of X that may be attributed to the k strata identified in the population.

To facilitate the sampling procedure for the purpose outlined above, we begin by formulating it at first, in a surrogate context of a standard estimation problem of μ_x using sample data as obtained from the recorded values of n units drawn by some fashion from the entire population. One of the most commonly used sampling schemes in such stratified case is to draw from each of the k strata, a simple random

sample without replacement (SRSWOR) of n_i observations $\left(\text{where } \sum n_i = n\right)$ and use it to obtain an estimator, $\hat{\mu}_i$, of the *i*th stratum's mean, μ_i, and to combine these *k* estimates in an estimator for (3) as:

$$\hat{\mu}_{\text{stra}} = \sum_{1}^{k} w_i \hat{\mu}_i . \tag{4}$$

The simplest approach in (4) is to take, $\hat{\mu}_i \equiv \bar{x}_{n_i}$, the *i*th sample average, which under the SRSWOR is an unbiased estimator of the corresponding stratum mean μ_i, so that $E(\bar{x}_{n_i}) = \mu_i$, and has a variance of

$$V(\bar{x}_{n_i}) = E(\bar{x}_{n_i} - \mu_i)^2 = \frac{\sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right), \tag{5}$$

where the expectation is taken with respect to the governing SRSWOR design (see for example, Cochran [1]). Accordingly, it follows from (3)-(5) that

$$\hat{\mu}_{\text{stra}} = \sum_{1}^{k} w_i \bar{x}_{n_i}, \tag{6}$$

is an unbiased estimator of the population mean μ_x, so that $E(\hat{\mu}_{\text{stra}}) = \mu_x$ and its variance is:

$$V(\hat{\mu}_{\text{stra}}) = E(\hat{\mu}_{\text{stra}} - \mu_x)^2 \equiv \sum_{1}^{k} w_i^2 V(\bar{x}_{n_i}) = \sum_{1}^{k} \frac{w_i^2 \sigma_i^2}{n_i}\left(\frac{N_i - n_i}{N_i - 1}\right). \tag{7}$$

Since $\hat{\mu}_{\text{stra}}$ is a linear combination of $\bar{x}_{n_1}, \bar{x}_{n_2}, ..., \bar{x}_{n_k}$, it follows that (under very mild conditions; see for example, Cochran [1] or Levy and Lemeshow [4] for details) its sampling distribution may be approximated by the normal distribution, having mean μ_x and variance $V(\hat{\mu}_{\text{stra}})$ as in (7). This fact will be used in Section 2 below to facilitate the sample size determination we seek here.

2. Determining the Optimal Sample Size and Allocations

The situation described above is not uncommon, when a sample of a total of *n* units from the stratified population is desired. To determine the sampling allocation

of the n_i units to be sampled from each stratum, we consider a cost-function which accounts for the sampling cost (per sampled units) as well as the cost that might be realized by a magnitude of the estimation error. Specifically, assuming that the cost of sampling each unit is the same across the strata, the average total cost $\overline{C}$ incurred for sampling *n* units is given by

$$\overline{C} = c \cdot n + E(\hat{\mu}_{\text{stra}} - \mu_x)^2, \tag{8}$$

where *c* represents the units sampling cost relative to that of the estimation error. By substituting (7) in (8) together with $n = \sum_1^k n_i$, it follows that $\overline{C}$ can be written as

$$\overline{C} = \sum_1^k \left[c \cdot n_i + \frac{w_i^2 \sigma_i^2}{n_i} \left(\frac{N_i - n_i}{N_i - 1} \right) \right]. \tag{9}$$

Accordingly, one seeks to determine the values of $n_1, n_2, ..., n_k$, that would minimize the total average cost $\overline{C}$ in (9), thus resulting with the optimal sampling allocations. It is straightforward to show (see for example, Levy and Lemeshow [4]) that for a given *n*, the optimal sampling allocation from each stratum $(\text{w.r.t. } \overline{C})$ is

$$n_i = \frac{w_i \sigma_i}{\sum_1^k w_l \sigma_l} \times n. \tag{10}$$

Note that the right-hand-side of (10) may not be an integer and hence the n_i on the left-hand-side of it would have to be an integer approximation of it. Further, if all the within-stratum variances are equal, then this optimal allocation will be reduced to the usual proportional allocation in which, $n_i = w_i \times n$, for $i = 1, ..., k$.

Now, it is clear from (9) and (10) that for a given *n*, the stratified sampling procedure which uses these optimal sampling allocations will result, on the average, with the smallest total cost as well as with the smallest standard error for the estimator $\hat{\mu}_{\text{stra}}$. In fact, for a given total sample size *n* and with these optimal allocations, as in (10), the variance (7) of $\hat{\mu}_{\text{stra}}$ is minimized and is $(\text{for large } N_i)$ given by

$$V_{\min}(\hat{\mu}_{\text{stra}}) = \frac{U^2}{n} - \frac{\sigma_W^2}{N}, \tag{11}$$

where $U^2 = \sum_1^k w_i\sigma_i$ and $\sigma_W^2 = \sum_1^k w_i\sigma_i^2$ is the within-strata variance component discussed in Section 1. Note that (11) does not depend on $n_1, n_2, ..., n_k$, but only on n. Using expression (11) for $V_{\min}$ and the normal approximation to the sampling distribution of $\hat{\mu}_{\text{stra}}$, one can determine the nominal value of n such that for a given (and desired) accuracy level ε, the magnitude of the estimation error, $|\hat{\mu}_{\text{stra}} - \mu_x|$, would be less $\varepsilon \cdot \mu_x$ with a desired confidence level $1 - \alpha$. More explicitly, the value of n is determined so as to satisfy

$$Pr(|\hat{\mu}_{\text{stra}} - \mu_x| \le \varepsilon \cdot \mu_x) \ge 1 - \alpha, \tag{12}$$

for given $\varepsilon > 0$ and $0 < \alpha < 0.5$. It can be easily verified that with given μ_x, U^2, σ_W^2 and N, (12) holds with a sample size n^*, satisfying

$$n^* \ge \frac{z^2U^2}{\varepsilon^2\mu_x^2N + z^2\sigma_W^2} \times N, \tag{13}$$

where z is the appropriate percentile value as determined from the standard normal distribution in order to achieve the desired confidence level $1 - \alpha$. Typically, n^* is chosen to be the smallest integer satisfying (12). By combining (10) and (13), we arrive at the procedure for simultaneously determining the 'optimal' sample size and the 'optimal' sampling allocations that would achieve a prescribed proportional accuracy at a desired confidence level and that would minimize an underlying sampling costs (assuming a fixed per-unit cost structure). Such a simultaneous procedure could be summarized as follows:

(I) For a desired accuracy level ε and a confidence level $1 - \alpha$, determine, as in (13), the 'optimal' sample size n^* that satisfies (12);

(II) Use n^* to determine the 'optimal' (w.r.t. (8)) sampling allocations to the k strata, $n_1^*, n_2^*, ..., n_k^*$, where (as in (10)),

$$n_i^* = \frac{w_i\sigma_i}{\sum_1^k w_l\sigma_l} \times n^*.$$

3. An Illustration

We illustrate the proposed sampling procedure with an example of a 'real-life' data - though sufficiently masked to protect some propriety information. As in the introduction, consider a brokerage firm that wishes to draw a random sample from its clients' population for the purpose of conducting personal interviews. The firm identified seven strata in its clients' population according to some measure of the client's portfolio 'size' X – see Table 1. The firm is cognizant of the relatively high fixed cost of the personal interviews (irrespective of the client's portfolio 'size'), and wishes to determine to 'optimal' size of a SRSWOR and the various sampling allocations it would need so as to faithfully reflect the stratified population of clients' portfolios it has.

Table 1. The brokerage data summarized into the 7 strata according to the measure of the portfolio 'size'

Stratum	N_i	μ_i	σ_i
# 1	572	3.10	2.60
# 2	315	16.66	4.98
# 3	201	37.51	6.69
# 4	150	63.46	8.30
# 5	141	99.15	11.47
# 6	93	144.80	13.48
# 7	69	219.74	53.22
Total	1541	43.28	57.10

Table 2 provides the calculated optimal sample size, n^* and the corresponding optimal sampling allocations, $n_1^*, n_2^*, ..., n_7^*,$ for several choices of confidence and accuracy levels. The table also provides in each instance the calculated standard error, $SE^* = \sqrt{V_{\min}^*}$, as obtained from (11) upon using this simultaneous sampling scheme with the indicated allocations as in (I) and (II) above. For instance, for a confidence level $1-\alpha = 0.99$, and a relative accuracy level of 6.5% (i.e., $\varepsilon = 0.065$), the stratified sampling scheme will require optimally only $n^* = 48$, with the following sampling allocations: $n_1^* = 6,\ n_2^* = 6,\ n_3^* = 5,\ n_4^* = 5,\ n_5^* = 6,\ n_6^* = 5$ and $n_7^* = 15$ from the seven given strata of this population – all with a standard error of $SE^* = 1.098$.

To validate the results on the sample size calculations, a large number, $M = 100{,}000$, of re-samples (as simulated experiments using the actual data) was drawn at random from the population according to the allocation prescription given by the stratified sampling scheme outlined in Table 2. In each instance, the observed confidence level, CL^* (that is, the proportion of samples among the M re-samples that achieved the prescribed level of accuracy), was calculated. The results are also included in Table 2. The close agreement in each instance, between the observed and the nominal confidence levels is highly indicative of the applicability of the simultaneously optimal sampling procedure as presented here.

Table 2. Optimal allocations and sample sizes for various confidence and accuracy levels along with the observed confidence level obtained from $M = 100{,}000$ re-samples, each, and the corresponding estimation standard error

Stratum: →			# 1	# 2	# 3	# 4	# 5	# 6	# 7	Observed	
CL	**ε**	$\boldsymbol{n}^*$	572	315	201	150	141	93	69	$SE^\star$	$CL^\star$
90%	0.05	**33**	4	4	4	3	5	3	10	1.347	89.61%
	0.045	**42**	5	5	5	4	6	4	13	1.181	90.23%
	0.04	**52**	6	7	6	5	7	5	16	1.051	90.06%
	0.035	**66**	8	8	7	7	9	7	20	0.919	90.10%
	0.03	**86**	11	11	9	9	11	9	26	0.789	90.03%
95%	0.06	**33**	4	4	4	3	5	3	10	1.347	94.72%
	0.055	**39**	5	5	4	4	5	4	12	1.228	94.74%
	0.05	**47**	6	6	5	5	6	5	14	1.110	95.02%
	0.045	**57**	7	7	6	6	8	6	17	0.999	95.05%
	0.04	**70**	9	9	8	7	9	7	21	0.889	94.79%
	0.035	**89**	11	11	10	9	12	9	27	0.773	94.96%
97.5%	0.07	**32**	4	4	4	3	4	3	10	1.367	97.33%
	0.065	**38**	5	5	4	4	5	4	11	1.246	97.64%
	0.06	**44**	5	6	5	4	6	5	13	1.153	97.55%
	0.055	**51**	6	7	6	5	7	5	15	1.062	97.48%
	0.05	**60**	7	8	7	6	8	6	18	0.970	97.43%
	0.045	**73**	9	9	8	7	10	8	22	0.868	97.55%
	0.04	**89**	11	11	10	9	12	9	27	0.773	97.52%
99%	0.08	**32**	4	4	4	3	4	3	10	1.367	98.90%
	0.075	**38**	5	5	4	4	5	4	11	1.246	99.11%
	0.07	**42**	5	5	5	4	6	4	13	1.182	99.00%
	0.065	**48**	6	6	5	5	6	5	15	1.098	99.04%
	0.06	**56**	7	7	6	6	7	6	17	1.008	99.03%
	0.055	**66**	8	8	7	7	9	7	20	0.920	99.07%
	0.05	**76**	9	10	8	8	10	8	23	0.848	98.94%
	0.045	**91**	11	12	10	9	12	9	28	0.763	98.99%

To further demonstrate this point, we present in Table 3 the results obtained in an extensive numerical study based on $M = 1{,}000{,}000$ re-samples. In each instance, the resulting observed confidence level, CL^*, is provided along with the *effective* confidence level (the coverage probability) as was calculated directly from the normal distribution having mean 0 and a standard deviation of SE^*/μ_x.

Table 3. Optimal sample sizes for various confidence and accuracy levels along with the effective and observed confidence level obtained from $M = 1{,}000{,}000$ re-samples, each

CL	ε	n^*	$SE^\star$	Effective CL	Observed $CL^\star$
90%	0.05	**33**	1.347	89.190%	89.208%
	0.045	**42**	1.181	90.069%	90.168%
	0.04	**52**	1.051	90.053%	90.092%
	0.035	**66**	0.919	90.053%	90.056%
	0.03	**86**	0.789	90.009%	89.976%
95%	0.06	**33**	1.347	94.617%	94.673%
	0.055	**39**	1.228	94.740%	94.796%
	0.05	**47**	1.110	94.874%	94.920%
	0.045	**57**	0.999	94.891%	94.944%
	0.04	**70**	0.889	94.846%	94.888%
	0.035	**89**	0.773	94.993%	95.014%
97.5%	0.07	**32**	1.367	97.330%	97.407%
	0.065	**38**	1.246	97.606%	97.632%
	0.06	**44**	1.153	97.576%	97.620%
	0.055	**51**	1.062	97.497%	97.555%
	0.05	**60**	0.970	97.429%	97.479%
	0.045	**73**	0.868	97.511%	97.555%
	0.04	**89**	0.773	97.486%	97.525%
99%	0.08	**32**	1.367	98.867%	98.896%
	0.075	**38**	1.246	99.083%	99.111%
	0.07	**42**	1.182	98.965%	98.988%
	0.065	**48**	1.098	98.962%	99.997%
	0.06	**56**	1.008	98.999%	99.012%
	0.055	**66**	0.920	99.037%	99.059%
	0.05	**76**	0.848	98.928%	98.941%
	0.045	**91**	0.763	98.931%	98.952%

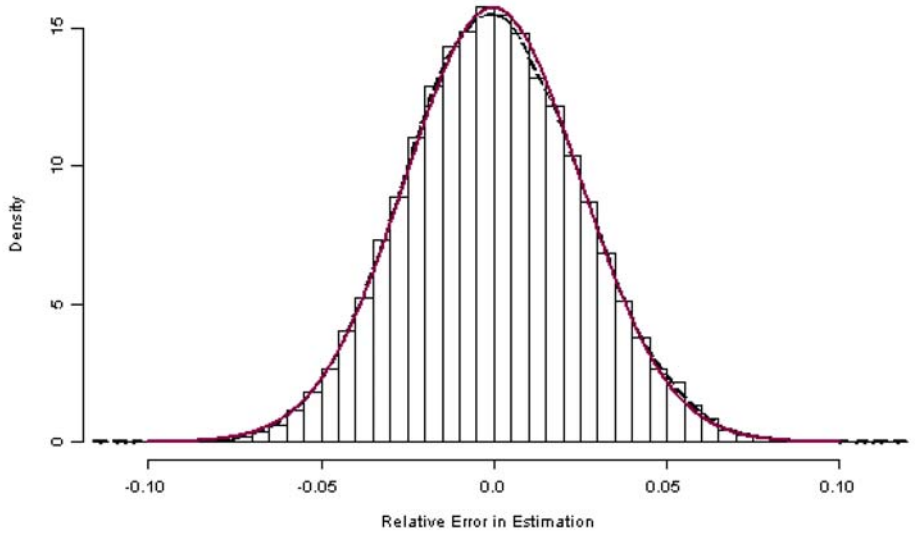

Figure 1. Sampling distribution of the re-sampled relative errors when $CL = 99\%$ and $\varepsilon = 0.065$, and with a sample size of $n^* = 48$.

Figure 1 illustrates the applicability of this procedure and the normal approximation used. It displays, in terms of a histogram and a kernel-estimated density curve, the sampling distribution of the relative error of estimation, $r = (\hat{\mu}_{\text{stra}} - \mu_x)/\mu_x$, as observed in $M = 50{,}000$ re-samples, each with $n^* = 48$ observations as needed to meet a 99% confidence level with a proportional accuracy of $\varepsilon = 0.065$. Indeed as expected, 99% of all these re-sampled instances, the relative error values fall between – 0.065 and + 0.065.

Also displayed in the figure is the theoretical density curve of the normal distribution having mean 0 and a standard deviation of $SE^*/\mu_x = 0.0254$. As can be seen, the two density curves are almost indistinguishable from one another - attesting to the extent of agreement with the normal approximation theory and the validity of the results as presented.

The results above demonstrate well the applicability of the proposed procedure for the simultaneous determination of the optimal sample size and optimal sampling allocations in a stratified population. Clearly, the determination of the sampling sizes requires the knowledge of the relative size of each stratum as well as the within-stratum means and variability. However, this procedure is well suited in situations

where a faithful representation, via sampling, of the known stratified population is being sought under both, cost and accuracy considerations. As was illustrated above, this proposed procedure effectively and efficiently accomplishes both.

References

[1] W. G. Cochran, Sampling Techniques, John Wiley and Sons, New York, 1977.

[2] W. D. Evans, On stratification and optimal allocation, J. Amer. Statist. Assoc. 46 (1951), 95-104.

[3] P. Knight, Sampling in auditing: an auditor's viewpoint, The Statistician 28(4) (1979), 253-266.

[4] P. S. Levy and S. Lemeshow, Sampling of Populations: Methods and Applications, John Wiley & Sons, New York, 1999.

[5] S. K. Thompson, Sampling, John Wiley & Sons, Inc., New York, 1992.

[6] T. M. F. Smith, Sampling in auditing: a statistician's viewpoint, The Statistician 28(4) (1979), 267-280.