3.33pt

Methods in the Practice of Sample Allocation in Stratified Sampling: Review and Perspectives

Dr. Şebnem Er
University of Cape Town, South Africa
Statistical Sciences Department
Sebnem.Er@uct.ac.za

Assoc. Prof. Marcin Kozak
University of Information Technology and Management in Rzeszow, Poland
nyggus@gmail.com

March 24, 2017

Aims and Introduction

Stratified sampling: a heteregeneous population U consisting of N elements is divided into L ($L \geq 2$) homogeneous sub-populations, called strata $(U_1, U_2, \ldots, U_L)$ according to one (univariate) or more characteristics (multivariate) of the population where

Aims and Introduction

Stratified sampling: a heteregeneous population U consisting of N elements is divided into L ($L \geq 2$) homogeneous sub-populations, called strata ($U_1, U_2, \ldots, U_L$) according to one (univariate) or more characteristics (multivariate) of the population where

$$\bigcup_{h=1}^{L} U_h = U$$

$h = 1, 2, \ldots, L$ denotes the stratum.

Aims and Introduction

Stratified sampling: a heteregeneous population U consisting of N elements is divided into L ($L \geq 2$) homogeneous sub-populations, called strata $(U_1, U_2, \ldots, U_L)$ according to one (univariate) or more characteristics (multivariate) of the population where

$$\bigcup_{h=1}^{L} U_h = U$$

$h = 1, 2, \ldots, L$ denotes the stratum.

There are two main problems in stratified sampling that deserve special attention: (1) Constructing the strata and (2) allocating the sample size among strata.

Aims and Introduction

Aim is to consider the problem of sample size allocation among strata in univariate and multivariate stratified sampling from different perspectives such as optimality, construction of the problem and the approach to the solution of the problem.

Sample Size Allocation in Univariate Stratified Sampling

- Equal Allocation:

$$n_h = \frac{n}{L}$$

- Proportional Allocation:

$$n_h = n \cdot \left(\frac{N_h}{N} \right)$$

- Neyman's Optimum Allocation:

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum_{h=1}^{L} N_h \sigma_h}$$

- Optimum allocation with varying sampling costs:

$$n_h = n \cdot \frac{W_h \sigma_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h \sigma_h / \sqrt{c_h}}$$

Neyman's Optimum Allocation

Neyman's (1934) method is very well known in the literature to provide the optimum sample sizes in terms of minimizing the variance of the estimator $Var(\bar{y}_{st})$ with a fixed sample size:

$$min \quad Var(\bar{y}_{st})$$

$$s.t. \quad n \leq n_0$$

Neyman's Optimum Allocation

Neyman's (1934) method is very well known in the literature to provide the optimum sample sizes in terms of minimizing the variance of the estimator $Var(\bar{y}_{st})$ with a fixed sample size:

$$min \quad Var(\bar{y}_{st})$$

$$s.t. \quad n \leq n_0$$

Alternatively, minimizing the total sample size necessary to control the variance of the estimate at a certain level:

$$min \quad n$$

$$s.t. \quad Var(\bar{y}_{st}) \leq Var(\bar{y}_{st})_0$$

Non-optimality in Sample Size Allocation in Univariate Stratified Sampling

When the sample sizes $\mathbf{n} = (n_1, n_2, \ldots, n_L)$ obtained by either of the target functions (given indifferent costs of sampling) satisfy the following conditions, then the sample sizes are optimum under the assumptions of KNOWN variances (for $h = 1, 2, \ldots, L$):

$$2 \leq n_h \leq N_h$$

$$N_h \geq 2$$

Non-optimality in Sample Size Allocation in Univariate Stratified Sampling

However, in practice the sample allocation is nearly always only approximately optimal for the following reasons:

- the sample size of a stratum is found to be greater than the size of the stratum, i.e. $n_h \geq N_h$ or more generally the result obtained does not satisfy the conditions.

Non-optimality in Sample Size Allocation in Univariate Stratified Sampling

However, in practice the sample allocation is nearly always only approximately optimal for the following reasons:

- the sample size of a stratum is found to be greater than the size of the stratum, i.e. $n_h \geq N_h$ or more generally the result obtained does not satisfy the conditions.
- the variances of the strata σ_h^2 are unknown and therefore are estimated by either the sample data or the auxiliary variable which is associated with the survey variable.

Non-optimality in Sample Size Allocation in Univariate Stratified Sampling

However, in practice the sample allocation is nearly always only approximately optimal for the following reasons:

- the sample size of a stratum is found to be greater than the size of the stratum, i.e. $n_h \geq N_h$ or more generally the result obtained does not satisfy the conditions.
- the variances of the strata σ_h^2 are unknown and therefore are estimated by either the sample data or the auxiliary variable which is associated with the survey variable.
- such allocation provides real numbers while integers are needed.

Non-optimality in Sample Size Allocation in Univariate Stratified Sampling

However, in practice the sample allocation is nearly always only approximately optimal for the following reasons:

- the sample size of a stratum is found to be greater than the size of the stratum, i.e. $n_h \geq N_h$ or more generally the result obtained does not satisfy the conditions.
- the variances of the strata σ_h^2 are unknown and therefore are estimated by either the sample data or the auxiliary variable which is associated with the survey variable.
- such allocation provides real numbers while integers are needed.

These issues may arise either all together or one at a time, though the third issue is considered mostly together with the other issues.

Approaches to Violations of the Constraints in Sample Allocation

Kozak and Jankowski (2008) and Keskintürk and Er (2007) examine the violations of the conditions and suggest the following:

1. Do not accept the results provided by the minimization of either of the target functions.
2. When $n_h < 2$, then n_h is set to two $(n_h = 2)$.
3. When $n_h > N_h$, then the sample size is set to the stratum size $(n_h = N_h)$.
4. Under a take-all top stratum, minimize the variance of the estimate using a numerical optimization method:

$$min_{n_1,n_2,\ldots,n_{L-1}} \quad Var(\bar{y}_{st})$$

5. A combination of (2-3) and (4).

Approaches with Integers

However, the results obtained are real numbers. There are two approaches to this problem:

- solve the problem with Lagrange multipliers and then round the real numbers to integers
- use nonlinear optimization techniques for integers

Variances are Known

- Arthanari and Dodge (1981) uses Lagrange multipliers and then round to integers

Variances are Known

- Arthanari and Dodge (1981) uses Lagrange multipliers and then round to integers
- Nordbotten (1956) uses a nonlinear cost function. The argument in assuming a nonlinear cost function lies behind the fact that “a small n_h may perhaps correspond to a higher cost per element than a larger n_h”

Variances are Known

- Arthanari and Dodge (1981) uses Lagrange multipliers and then round to integers
- Nordbotten (1956) uses a nonlinear cost function. The argument in assuming a nonlinear cost function lies behind the fact that "a small n_h may perhaps correspond to a higher cost per element than a larger n_h"
- Bretthauer et al. (1999) examined the same problems with known variances for continuous and integer sample sizes under linear and nonlinear cost function such as a concave sampling cost function.

Variances are Known

- Arthanari and Dodge (1981) uses Lagrange multipliers and then round to integers
- Nordbotten (1956) uses a nonlinear cost function. The argument in assuming a nonlinear cost function lies behind the fact that "a small n_h may perhaps correspond to a higher cost per element than a larger n_h"
- Bretthauer et al. (1999) examined the same problems with known variances for continuous and integer sample sizes under linear and nonlinear cost function such as a concave sampling cost function.
- Rivest et al. (2012) showed that sample allocation based on integers can lead to slightly smaller sample sizes than that based on Neyman allocation.

Variances are Unknown

In practice, the variances of the strata, σ_h^2, are assumed to be known from a recent or preliminary survey, but it is clear that in real life surveys the variances are hardly ever known.

- If the prior distribution of an auxiliary variable (X) related to the study variable (Y) is available, then this can be utilized. This was first discussed by Hanurav (1965), Rao (1968-1979) followed by Prekopa (1995), Uryasev and Pardalos (2001) and Louveaux and Birge (2001).
- Recently, Park et. al (2007) and Diaz-Garcia and Garay-Tapia (2007) provided some solutions for the problem of sample allocation under unknown variances.

Variances are Unknown - Park et.al (2007) Diaz-Garcia and Garay-Tapia (2007)

- Park et. al (2007) assume that the equality of the variances of some strata is known in advance.
- Diaz-Garcia and Garay-Tapia (2007) study the problem of optimum allocation in stratified surveys as problems of stochastic optimization since the population variances are substituted by the sample variances which are then random variables with expected values and variances.

From Univariate Case to Multivariate Case

In the univariate case,

- Neyman's (1934) optimum sample allocation method provides optimum sample sizes.

From Univariate Case to Multivariate Case

In the univariate case,

- Neyman's (1934) optimum sample allocation method provides optimum sample sizes.

In the multivariate case,

- Some compromise must be reached.
- Many compromised criteria have been proposed in the literature.
- They differ in aims and thus usually lead to different results.

Multivariate Stratified Sampling - History

When the survey population under study has “**K**” characteristics, the optimum allocation obtained for one characteristic may not be optimum for other characteristics.

Multivariate Stratified Sampling - History

When the survey population under study has "**K**" characteristics, the optimum allocation obtained for one characteristic may not be optimum for other characteristics.

Some compromise has to be reached. (Cochran (1963) suggest taking the average of two sample sizes (assuming $K = 2$) obtained by optimum allocations.)

Multivariate Stratified Sampling - History

When the survey population under study has "**K**" characteristics, the optimum allocation obtained for one characteristic may not be optimum for other characteristics.

Some compromise has to be reached. (Cochran (1963) suggest taking the average of two sample sizes (assuming $K = 2$) obtained by optimum allocations.)

Many analytical and numerical methods have been proposed in the literature so far.

Multivariate Stratified Sampling - History

When the survey population under study has "**K**" characteristics, the optimum allocation obtained for one characteristic may not be optimum for other characteristics.

Some compromise has to be reached. (Cochran (1963) suggest taking the average of two sample sizes (assuming $K = 2$) obtained by optimum allocations.)

Many analytical and numerical methods have been proposed in the literature so far.

The problem is still popular among today's researchers and there is still a growing amount of research done in this area.

Multivariate Stratified Sampling - History

- Dalenius (1953) used a linear programming approach where there is information available regarding the characteristics used for stratification.
- Dalenius (1957) proposed a nonlinear programming technique.
- Since then, there are several methods proposed in the literature using some type of a mathematical programming method. Some of these methods,
  - minimise the variances (coefficient of variations) of the estimates
  - minimise the total cost
  - try to minimise the covariance of the variance of the estimates
- Neyman's optimum allocation formula is used.

Variance Minimisation Methods - Known Variances

The problem can be formulized as follows:

$$min_{\mathbf{n}} \sum_{j=1}^{K} w_j V(\bar{y}_{st}^{j})$$

$$s.t. \quad \sum_{h=1}^{L} c_h n_h + C_0 = C$$

$$2 \leq n_h \leq N_h, \quad h = 1, 2, \ldots, L, \quad n_h \in \aleph$$

Coefficient of Variation Minimisation Methods - Known Variances

Variances are not unit free!

Squared coefficient of variations (CV) are used instead of variances:

$$min_{\mathbf{n}} = \sum_{j=1}^{K} w_j (CV(\bar{y}_{st}^{j}))^2$$

$$s.t. \quad \sum_{h=1}^{L} c_h n_h + C_0 = C$$

$$2 \leq n_h \leq N_h, \quad h = 1, 2, \ldots, L, \quad n_h \in \aleph$$

Cost Minimisation Methods

Bethel (1985) and Bethel (1989) proposed to use coefficient of variations (CV) which is unit free and the problem is formulized as follows:

$$min_{\mathbf{n}} \sum_{h=1}^{L} c_h n_h + C_0 = C$$

$$s.t. \quad CV(\bar{y}_{st}^{j}) \leq CV_j$$

$$2 \leq n_h \leq N_h, \quad h = 1, 2, \ldots, L, \quad n_h \in \aleph$$

Sample Size Minimisation Methods

Gren (1964) uses Lagrange multipliers method to minimize

$$min_{\mathbf{n}} f(n_1, n_2, \ldots, n_L) = min_{\mathbf{n}} \sum_{h=1}^{L} \sum_{j=1}^{K} (n_{hj} - n_h)^2$$

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

Lagrange multipliers method gives sample sizes, which are simply the average sample sizes for each jth characteristic obtained with Neyman allocation method in the univariate case for the corresponding strata.

Covariance Minimisation Methods

Diaz-Garcia and Ulloa-Cortez (2008) develop a multi-objective optimisation method for optimum allocation and compare the method with nonlinear matrix optimisation of integers under either a cost or sample size constraint.

They approach the problem of optimum allocation under two approaches:

- nonlinear matrix optimisation
- multiobjective optimisation

In nonlinear matrix optimisation approach, their optimisation problem is defined as follows:

$$min_{\mathbf{n}} \hat{Cov}(\bar{y}^{st})$$

$$2 \leq n_h \leq N_h, \quad h = 1, 2, \ldots, L, \quad n_h \in \aleph$$

Papers Dealing with Compromised-Optimum Sample Allocation in the Multivariate Case - Minimise CV

- (Variances known) Valliant and Gentle (1997) used weights and $n_h \in \aleph$
- (Variances known) Khowaja, et.al. (2013) did not use weights and n_h are integers.
- (Variances Unknown) Diaz-Garcia and Cortez (2006) consider the optimum allocation in multivariate stratified sampling as a problem of the multi-objective optimization of integers, under
  - complete (ideal but occurs very rarely)
  - partial
  - zero information
- Khan, Ali and Ahmad (2011) transform the problem to a convex programming problem with several linear transformations of the objective function and the constraints.

Papers Dealing with Compromised-Optimum Sample Allocation in the Multivariate Case - Minimise Cost

- Bethel (1985) applied convex programming to solve the problem. This method can be applied for multivariate stratified surveys using the R package `bethel` written by De Meo (2012).

Papers Dealing with Compromised-Optimum Sample Allocation in the Multivariate Case - Minimise Cost

- Bethel (1985) applied convex programming to solve the problem. This method can be applied for multivariate stratified surveys using the R package `bethel` written by De Meo (2012).
- Huddleston, Claypool and Hocking (1970) use convex programming algorithm that was first given by Hartley and Hocking (1963).

Papers Dealing with Compromised-Optimum Sample Allocation in the Multivariate Case - Minimise Cost

- Bethel (1985) applied convex programming to solve the problem. This method can be applied for multivariate stratified surveys using the R package `bethel` written by De Meo (2012).
- Huddleston, Claypool and Hocking (1970) use convex programming algorithm that was first given by Hartley and Hocking (1963).
- Rahim (1994-1995) uses the weighted squared coefficient of variation values as a constraint.

Papers Dealing with Compromised-Optimum Sample Allocation in the Multivariate Case - Minimise Sample Size

Melaku and Sadasivan (1987) propose two different approaches:

- using quadratic programming problem with an additional constraint of positive sample sizes, with the absolute form of the objective function such that

$$min_{\mathbf{n}} \sum_{h=1}^{L} \sum_{j=1}^{K} |n_{hj} - n_h|$$

- Kozak's (2006a) random search method for optimal sample allocation between strata and domains.

Conclusion - Comparability of Different Approaches

The recent developments in the area of sample allocation in multivariate stratified surveys are growing immensely. The review of the papers shows that:

- many proposed methods are applied to a data set which has few numbers of strata and mostly two characteristics.
- many of the papers lack a simulation study and a comparison with the other methods developed in the literature.
- the data sets used in the newly developed methods are not the data sets that were used before by any other researcher which makes it very difficult to make a comparison of the newly proposed method with the existing methods.

Conclusion - Algorithmic

From the algorithmic point of view, all of the methods use different types of methods to solve the more or less same type of problem:

- Even though the problems seem to be different in their constraints and objective functions, the matter of the use of a different algorithm does not contribute in the solution of the bigger problem.

Thank you for your time...

- Thank you...
- Baie dankie...
- Enkosi...
- Teșekkürler...