Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 177 (2007) 1074–1089
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Stochastics and Statistics

A biobjective method for sample allocation in stratified sampling

Emilio Carrizosa [a], Dolores Romero Morales [b,*]

[a] *Facultad de Matemáticas, Universidad de Sevilla, Tarfia s/n, 41012 Sevilla, Spain*
[b] *Saïd Business School, University of Oxford, Park End Street, Oxford OX1 1HP, United Kingdom*

Received 21 October 2004; accepted 30 November 2005
Available online 17 February 2006

Abstract

The two main and contradicting criteria guiding sampling design are accuracy of estimators and sampling costs. In stratified random sampling, the sample size must be allocated to strata in order to optimize both objectives.

In this note we address, following a biobjective methodology, this allocation problem. A two-phase method is proposed to describe the set of Pareto-optimal solutions of this nonlinear integer biobjective problem. In the first phase, all supported Pareto-optimal solutions are described via a closed formula, which enables quick computation. Moreover, for the common case in which sampling costs are independent of the strata, all Pareto-optimal solutions are shown to be supported. For more general cost structures, the non-supported Pareto-optimal solutions are found by solving a parametric knapsack problem. Bounds on the criteria can also be imposed, directing the search towards implementable sampling plans. Our method provides a deeper insight into the problem than simply solving a scalarized version, whereas the computational burden is reasonable.

Keywords: Integer programming; Stratified random sampling; Sample allocation; Biobjective integer program; Parametric knapsack problem

1. Introduction

The sample allocation problem for stratified simple random sampling is the following: we are given a population of size N divided into n groups (strata), with population sizes $N_1, \ldots, N_n$. Simple random samples without replacement of sizes $x_1, \ldots, x_n$, are to be drawn independently from the different strata. The sampling cost within each stratum is assumed to be linear in its sample size x_i, with unit sampling cost within stratum i equal to a positive integer c_i. The total sampling cost is the sum of the sampling costs within the strata.

* Corresponding author.
E-mail addresses: ecarrizosa@us.es (E. Carrizosa), dolores.romero-morales@sbs.ox.ac.uk (D. Romero Morales).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2005.11.027

The drawn sample is used to estimate some parameter of the variable under study Y. Throughout this paper, we assume that the parameter to be estimated is $\overline{Y}$, the average of the variable Y in the population. Then, the parameter $\overline{Y}$ will be estimated via its Horvitz–Thompson estimator $\widehat{\overline{Y}}$,

$$\widehat{\overline{Y}} = \sum_{i=1}^{n} \frac{N_i}{N} \overline{y}_i, \tag{1}$$

where $\overline{y}_i$ denotes the sample average within stratum i, see e.g. [5] for further statistical details on the problem considered.

Estimator $\widehat{\overline{Y}}$ is unbiased, and its variance $\mathrm{var}(\widehat{\overline{Y}})$ is given by

$$\mathrm{var}(\widehat{\overline{Y}}) = \sum_{i=1}^{n} \left(\frac{N_i}{N}\right)^2 \mathrm{var}(\overline{y}_i) = \sum_{i=1}^{n} \left(\frac{N_i}{N}\right)^2 \left(\frac{1}{x_i} - \frac{1}{N_i}\right) \sigma_{c,i}^2, \tag{2}$$

where $\sigma_{c,i}^2$ is the quasivariance of Y within stratum i.

We assume, as customary in the literature, that the quasivariances $\sigma_{c,i}^2$ are either known from previous similar experiments, or replaced by known upper bounds. For instance, if Y_i, the values of variable Y within stratum i, is a Boolean variable, we can use the upper bound $\frac{N_i}{N_i - 1} \frac{1}{4}$, [5].

The goal is to determine sample sizes $x_1, \ldots, x_n$ minimizing simultaneously

- The total sampling cost.
- The variance of the Horvitz–Thompson estimator $\widehat{\overline{Y}}$.

Two types of constraints are imposed. On the one hand, box constraints are considered on the sample sizes x_i,

$$l_i \leqslant x_i \leqslant u_i \tag{3}$$

for positive integers $l_i \leqslant u_i$, for all $i = 1, 2, \ldots, n$.

Constraints (3) are motivated as follows. First, at least one element must be sampled from each stratum, since, otherwise, the expression (2) is meaningless; moreover, since sampling is without replacement, no more than N_i individuals can be sampled from stratum i.

These trivial bounds $1 \leqslant x_i \leqslant N_i$ may not be sharp enough for practical purposes. Indeed, if we are not only concerned with the variance of the estimator $\widehat{\overline{Y}}$, but also with the variance of the estimators $\overline{y}_i$ within the strata, constraints of the form

$$\mathrm{var}(\overline{y}_i) \leqslant \mu_i, \tag{4}$$

for $\mu_i > 0$ given, may be imposed. Constraint (4) can also be written as

$$x_i \geqslant \left\lceil \frac{\sigma_{c,i}^2 N_i}{N_i \mu_i + \sigma_{c,i}^2} \right\rceil,$$

which, as asserted, yields a constraint of type (3).

On the other hand, the aim of simultaneous minimization of cost and variance may lead to sampling plans in which one of the two objectives attains a low value at the expense of a very high value on the other. To avoid this, we include also in the model target constraints in the form

$$\begin{aligned} &\sum_{i=1}^{n} c_i x_i \leqslant K^*, \\ &\mathrm{var}(\widehat{\overline{Y}}) \leqslant B \end{aligned} \tag{5}$$

for positive K^* and B, allowed also to take the value $+\infty$.