Computational Statistics and Data Analysis 92 (2015) 1–12

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Fast integer-valued algorithms for optimal allocations under constraints in stratified sampling

Ulf Friedrich [a,*], Ralf Münnich [b], Sven de Vries [a], Matthias Wagner [a]

[a] University of Trier, Department of Mathematics, 54286 Trier, Germany

[b] University of Trier, Department of Economics and Social Statistics, 54286 Trier, Germany

ARTICLE INFO

Article history:
Received 21 March 2014
Received in revised form 18 April 2015
Accepted 16 June 2015
Available online 24 June 2015

Keywords:
German Census 2011
Non-linear discrete optimization
Optimal allocation
Greedy algorithm
Polymatroid
Box constraints

ABSTRACT

In stratified random sampling, minimizing the variance of a total estimate leads to the optimal allocation. However, in practice, this original method is scarcely appropriate since in many applications additional constraints have to be considered. Three optimization algorithms are presented that solve the integral allocation problem with upper and lower bounds. All three algorithms exploit the fact that the feasible region is a polymatroid and share the important feature of computing the globally optimal integral solution, which generally differs from a solution obtained by rounding. This is in contrast to recent references which, in general, treat the continuous relaxation of the optimization problem. Two algorithms are of polynomial complexity and all of them are fast enough to be applied to complex problems such as the German Census 2011 allocation problem with almost 20,000 strata.

1. Motivation

Estimation in Official Statistics, in general, is based on survey sampling methods. Hence, the inference has to be drawn with respect to the sampling design. One very important design is stratified random sampling which is widely used in practice, e.g., in the German Census 2011. The universe is split into strata which generally determine certain subgroups of interest such as regions, house size classes or business codes. In stratified random sampling, the total sample size has then to be allocated to the strata by certain conditions.

A finite population U of size N is split into H disjoint strata of size N_h for $h = 1, \ldots, H$. Next, the total sample of size n has to be split into the stratum specific sample sizes n_h which refers to the so-called allocation problem. Within all H strata simple random sampling is applied, either with or without replacement. As an allocation problem, there are several options. In many cases, the total sample size is equally distributed amongst the strata ($n_h = n/H$; equal allocation) or proportionally to the stratum sizes ($n_h \propto N_h$; proportional allocation). In both cases, the allocations are non-integral and, hence, have to be rounded.

In stratified sampling, a well-known unbiased estimator for the total τ_Y of the variable of interest Y in the universe is

$$\hat{t}_y^{StrRS} = \sum_{h=1}^{H} N_h \bar{y}_h,$$

* Corresponding author.
E-mail addresses: friedrich@uni-trier.de (U. Friedrich), muennich@uni-trier.de (R. Münnich), devries@uni-trier.de (S. de Vries), wagnerm@uni-trier.de (M. Wagner).

http://dx.doi.org/10.1016/j.csda.2015.06.003
0167-9473/

with variance

$$V(\hat{\tau}_y^{StrRS}) = \sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) = \sum_{h=1}^{H} \frac{N_h^2 S_h^2}{n_h} - \sum_{h=1}^{H} N_h S_h^2. \tag{1}$$

$\bar{y}_h$ is the sample mean and S_h^2 the variance of variable Y in stratum h. Often, the variances S_h^2 have to be estimated from the sample which can easily be done using the classical inferential variances as plug-in. However, then $n_h \geq 2$ is required as a boundary condition which is not necessarily fulfilled in practice. The expression $(1 - n_h/N_h)$ is the finite population correction for sampling without replacement in each stratum which vanishes in case of sampling with replacement. Note that the last sum in (1) does not depend on the decision variables n_h and, as it is a constant in this sense, can be ignored in the optimization problems hereafter.

In contrast to the equal and proportional allocation mentioned above, variance reduction methods can be applied to derive a more efficient allocation. Neyman (1934) and Tschuprow (1923) minimize the variance (1) of the total estimator with respect to a given total sample size n which leads to the *optimal allocation*. Then, the optimal allocation is formally given by the problem

$$\begin{aligned} \min_{\boldsymbol{n}\in\mathbb{R}_+^H} \quad & V(\hat{\tau}_y^{StrRS}) \\ \text{s.t.} \quad & \sum_{h=1}^{H} n_h = n. \end{aligned}$$

For further details on stratified random sampling we refer to Cochran (1977) or Särndal et al. (2003) who also cover more generalized sampling schemes.

In survey practice, a couple of additional conditions may have to be considered which may lead to difficulties in estimation. First, the optimal allocation may yield solutions where $n_h > N_h$ for certain strata, which by definition is invalid. This may easily occur in so-called open strata, e.g., the largest income class with few incomes but very high variation. In order to avoid this over-allocation problem, upper sampling fractions of 100% may be considered as constraints. Therefore, the constraints $n_h \leq N_h$, $h = 1, \ldots, H$, should be added to the problem formulation. For practical reasons, further constraining of sampling fractions may be applied, e.g., in order to avoid a large spread of response burdens within the population.

Second, the above allocations have to fulfill integrality constraints which, in general, are not met. Rounding may lead to practical solutions in the case of the equal or proportional allocation. However, in the case of the optimal allocation, rounding hardly fulfills the global optimality condition in the class of integral allocations.

Further, since in practice the variances S_h^2 are hardly known a priori, adequate proxies have to be considered, e.g., highly correlated variables or earlier estimates to determine the solution of the optimal allocation. In the case of the German Census 2011, address sizes in the register have been used. The lower fraction may also be fixed in order to achieve minimum sampling fractions or sizes in all relevant areas, e.g., $n_h \geq 2$ for cases when the stratum variances later have to be estimated from the sample. This has to be done when estimating the variance of the total estimator a posteriori or when estimates within the strata may become of interest, the so-called domain estimation problem.

Finally, additional attention may be paid to the variation of the resulting survey weights. According to Gelman's critique on struggles with survey weights (see Gelman, 2007), the variation of survey weights should not be too large (cf. also Burgard and Münnich, 2012) in order to avoid problematic influences of the sampling design on model-based estimates. In stratified random sampling, this variation is defined by the ratio of the highest to the lowest values N_h/n_h of all strata. When using the proportional allocation, this ratio is one due to constant sampling fractions in the strata if the integrality is ignored. The optimal allocation where $n_h \propto N_h \cdot S_h$ can lead to very high ratios, which may be limited by reducing the variation using box constraints.

Altogether, the considerations above lead to setting minimal and maximal sampling fractions such that the variables n_h are bounded according to

$$m_h \leq n_h \leq M_h \quad \forall h = 1, \ldots, H,$$

with $2 \leq m_h < M_h \leq N_h$. Omitting the (for the optimization irrelevant) sum of constant terms in (1) yields the following integral allocation problem of optimal sample sizes in stratified sampling

$$\begin{aligned} \min_{\boldsymbol{n}\in\mathbb{Z}_+^H} \quad & \sum_{h=1}^{H} \frac{d_h^2}{n_h} \\ \text{s.t.} \quad & \sum_{h=1}^{H} n_h \leq n \quad \text{and} \\ & m_h \leq n_h \leq M_h \quad \forall h = 1, \ldots, H, \end{aligned} \tag{2}$$

where $\boldsymbol{n} := (n_1, \ldots, n_H)^T \in \mathbb{Z}_+^H$ defines the sample size in the different strata $h \in \{1, \ldots, H\}$. The total sample size is given by n and d_h^2 is defined as the product of the stratum variance S_h^2 and the squared population size N_h^2 of stratum h, cf. (1).