European Journal of Operational Research 217 (2012) 479–482

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

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

Short Communication

On approximate Monetary Unit Sampling ☆

Emilio Carrizosa *

Facultad de Matemáticas, Universidad de Sevilla, Avda Reina Mercedes s/n, 41012 Sevilla, Spain

ARTICLE INFO

Article history:
Received 20 November 2009
Accepted 13 September 2011
Available online 28 September 2011

Keywords:
Nonlinear programming
Monetary Unit Sampling
Statistical sampling
Karush–Kuhn–Tucker conditions

ABSTRACT

Monetary Unit Sampling (MUS), also known as Dollar-Unit Sampling, is a popular sampling strategy in Auditing, in which all units are to be randomly selected with probabilities proportional to the book value. However, if units sizes have very large variability, no vector of probabilities exists fulfilling the requirement that all probabilities are proportional to the associated book values. In this note we propose a Mathematical Optimization approach to address this issue. An optimization program is posed, structural properties of the optimal solution are analyzed, and an algorithm yielding the optimal solution in time and space linear to the number of population units is given.

1. Monetary Unit Sampling

Consider a finite population $\mathcal{U} = \{u_1, \ldots, u_N\}$, from which a random sample of size n is to be drawn. A popular sampling design in Auditing is the so-called MUS design. In a MUS, n sample units are selected without replacement in such a way that, for each unit u_i, the probability π_i of u_i being part of the sample is proportional to its book value $X_i > 0$. See [6] for an introduction to statistical methods in Auditing and [5,8] for further statistical sampling techniques. In other words, in a MUS design there exists a factor $k > 0$ such that the inclusion probability π_i of sample unit u_i has the form

$$\frac{\pi_i}{X_i} = k, \quad i = 1, 2, \ldots, N. \tag{1}$$

As in any without-replacement design of total sample size n, one has the relations

$$\sum_{i=1}^{N} \pi_i = n, \tag{2}$$

$$0 \leqslant \pi_i \leqslant 1, \quad i = 1, 2, \ldots, N, \tag{3}$$

see [5]. By (1), (2) implies

$$\pi_i = n \frac{X_i}{\sum_{j=1}^{N} X_j}, \quad i = 1, 2, \ldots, N, \tag{4}$$

which essentially provides a recipe for calculating the probabilities π_i.

MUS is the most popular statistical sampling method in Auditing, [7], and it is easily applicable using, for instance, the different procedures implemented in the package `sampling` of `R`, [9]: once the vector of probabilities π is supplied, a random sample m is generated such that the inclusion probability of each u_i, i.e., the probability that $u_i \in m$ is either exactly or approximately π_i. See e.g. [8].

MUS has interesting theoretical statistical properties when sample estimates are to be constructed [5]. Nevertheless, in Auditing MUS is a fundamental tool to select samples to be audited, taking into account the basic principle that units with higher book values deserve a higher probability of being selected.

In spite of its wide use, it is not always possible in practice to implement such a sampling scheme. Indeed, since the probabilities π_i must satisfy (3), it follows that expression (4) is only applicable if book value X_i of any unit u_i satisfies the condition

$$X_i \leqslant \frac{1}{n} \sum_{j=1}^{N} X_j. \tag{5}$$

If condition (5) is not fulfilled for some units, i.e., if strictly speaking MUS cannot be applied, a heuristic strategy is commonly used to obtain a vector of probabilities satisfying approximately (1). To do this, the population $\mathcal{U}$ is split into two groups, so that the r units in the first group, namely, those for which, when applying formula (4), a value strictly greater than 1 is obtained, are sampled with probability 1, whereas a MUS of size $n - r$ is defined on the remaining $N - r$ units. In this new population, a condition analogous to (5) is derived; it may be the case that all units satisfy such a condition, or, contrarily, a new division between large and smaller units is needed, and the process is repeated until condition (5) is satisfied for the remaining units. As a simple illustration, suppose a MUS sample of size $n = 4$ is to be extracted from a population of

☆ Supported by Projects **MTM2009-14039**, Spain and **FQM-329** of Junta de Andalucía.

* Tel./fax: +34 954557943.
E-mail address: ecarrizosa@us.es

0377-2217/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2011.09.037

Table 1
Example of the heuristic MUS.

i	X_i	$n\frac{X_i}{\sum_{j=1}^{N}X_j}$	$n_2\frac{X_i}{\sum_{j=3}^{N}X_j}$	$n_3\frac{X_i}{\sum_{j=4}^{N}X_j}$	π_i (heuristic)	π_i via solving (6)
1	20,000	1.42	–	–	1.00	1.00
2	17,000	1.21	–	–	1.00	1.00
3	9,700	0.69	1.01	–	1.00	1.00
4	6,050	0.43	0.63	0.63	0.63	0.69
5	1,350	0.10	0.15	0.14	0.14	0.12
6	990	0.07	0.11	0.10	0.10	0.09
7	450	0.03	0.06	0.05	0.05	0.04
8	400	0.03	0.05	0.04	0.04	0.03
9	200	0.01	0.03	0.02	0.02	0.02
10	100	0.01	0.01	0.01	0.01	0.01

$N = 10$ units with book values X_i given in the second column of Table 1. The inclusion probabilities, calculated by (4), are given in the third column of Table 1.

Since units u_1, u_2 have a value greater than 1, we set $\pi_1 = \pi_2 = 1$, and we repeat the process with the remaining population and a sample size $n_2 = 4 - 2 = 2$. This way, the fourth column of the table is obtained. Now we set $\pi_3 = 1$, and we repeat the process with the remaining population and sample size $n_3 = 2 - 1 = 1$. All probabilities obtained are smaller than 1, as shown in the fifth column. The process is then stopped, yielding the probabilities in the sixth column of the table. This vector of probabilities is obtained if one uses, for instance, the function `inclusionprobabilities` in the package `sampling` of `R`, [9].

MUS aims to assure that, if one unit u_i has a book value X_i which is k times the value X_j of unit u_j, then the inclusion probability π_i of unit u_i should be k times the inclusion probability π_j of unit u_j. This is fulfilled only approximately if the iterative process illustrated above is followed. Hence, such a process can be seen as a heuristic approach to find a vector π, satisfying the hard constraints (2) and (3) and *hopefully* with a small violation in fulfilling (1). The vector of probabilities obtained with such a heuristic procedure have three important properties:

1. They are always strictly positive.
2. They are monotonic in the book values (if $X_i > X_j$ then $\pi_i \geqslant \pi_j$).
3. They yield a true MUS design, i.e., all probabilities satisfy (1), whenever possible.

Our target is to obtain a vector π of probabilities satisfying the hard constraints (2), (3) and being as close as possible to fulfillment of (1). Since (1) means that one wants the ratios $\frac{\pi_i}{X_i}$ constant, one possible way of measuring closeness to fulfillment of (1) is via the variance $v(\pi)$ of such ratios,

$$v(\pi) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\pi_i}{X_i} - \frac{1}{N}\sum_{j=1}^{N}\frac{\pi_j}{X_j}\right)^2.$$

We propose here to minimize $v(\pi)$ on the feasible region defined by (2) and (3). Minimizing v is equivalent to solving the convex quadratic linearly-constrained optimization problem

$$\begin{aligned}\min\quad & \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\pi_i}{X_i} - k\right)^2 \\ & \sum_{i=1}^{N}\pi_i = n \quad i = 1,2,\ldots,N \\ & 0 \leqslant \pi_i \leqslant 1 \quad i = 1,2,\ldots,N \\ & k \in \mathbb{R},\end{aligned} \tag{6}$$

since, for each π fixed, the optimal value of k is given by $k = \frac{1}{N}\sum_{i=1}^{N}\frac{\pi_i}{X_i}$, and thus $v(\pi)$ coincides with the objective of (6).

The vector π obtained by solving (6) does not necessarily coincide with the solution provided by the heuristic above (e.g. compare the last two columns of Table 1), but, as shown below, such π enjoys the same good properties as the heuristic approach (strictly positive probabilities, monotonic in the book values), guaranteeing also a best-possible fit to (1).

Proposition 1. *Let (π^*, k^*) be optimal to (6). One has:*

1. $k^* = \frac{1}{N}\sum_{i=1}^{N}\frac{\pi_i^*}{X_i} > 0$.
2. *If i is such that $\frac{\pi_i^*}{X_i} < k^*$, then $\pi_i^* = 1$.*
3. $0 < \pi_i^* \leqslant 1$ *for all $i = 1,2,\ldots,N$.*
4. *If i, j are such that $X_i > X_j$, then $\pi_i^* \geqslant \pi_j^*$.*

Proof. Part 1 is obtained by equating to zero the derivative of the objective with respect to k for π fixed at π^*. Part 2 is shown by contradiction: suppose i exists such that $\pi_i^* < 1$ and $\frac{\pi_i^*}{X_i} < k^*$. By part 1, there exists j with $\frac{\pi_j^*}{X_j} > k^*$. Let us define for small positive t the feasible solution $(\pi(t), k(t))$,

$$\pi_r(t) = \begin{cases} \pi_r^*, & \text{if } r \neq i,j, \\ \pi_{i+t}^*, & \text{if } r = i, \\ \pi_{j-t}^*, & \text{if } r = j, \end{cases} \tag{7}$$

$$k(t) = k^* + \frac{t}{N}\left(\frac{1}{X_i} - \frac{1}{X_j}\right).$$

It is easily seen that the derivative $\varphi'(0)$ of the objective for $(\pi(t), k(t))$ at $t = 0$ equals $\frac{2}{N}\left(\left(\frac{\pi_i^*}{X_i} - k^*\right)\frac{1}{X_i} - \left(\frac{\pi_j^*}{X_j} - k^*\right)\frac{1}{X_j}\right)$. Since, by assumption, $\frac{\pi_i^*}{X_i} - k^* < 0 < \frac{\pi_j^*}{X_j} - k^*$, it follows that $\varphi'(0) < 0$, implying that (π^*, k^*) cannot be optimal to (6). Hence, 2 follows.

To show Part 3, suppose by contradiction that some $\pi_i^* = 0$. This would imply $\frac{\pi_i^*}{X_i} < k^*$, and thus, by Part 2, $\pi_i^* = 1$, which is a contradiction.

Part 4 is also shown by contradiction. Suppose i, j exist such that $X_i > X_j$ and $\pi_i^* < \pi_j^*$. In particular, $\pi_i^* < 1$, and thus, by Part 2, $\frac{\pi_i^*}{X_i} \geqslant k^*$, and thus

$$\left(\frac{\pi_j^*}{X_j} - k^*\right) > \left(\frac{\pi_i^*}{X_i} - k^*\right) \geqslant 0.$$

For positive small t, construct the feasible $(\pi(t), k(t))$ as in (7). The derivative of the objective on $(\pi(t), k(t))$ at $t = 0$ equals $\frac{2}{N}\left(\left(\frac{\pi_i^*}{X_i} - k^*\right)\frac{1}{X_i} - \left(\frac{\pi_j^*}{X_j} - k^*\right)\frac{1}{X_j}\right)$, which is negative, implying that (π^*, k^*) cannot be optimal. □

Proposition 1 shows that the vector π^* of probabilities obtained by solving (6) has the desired properties of being positive and monotonic in the book values X_i. Moreover, it turns out that the convex quadratic problem with linear constraints (6) can be solved with a rather simple and quick ad hoc procedure, since one almost gets a closed formula from the Karush–Kuhn–Tucker (KKT) optimality conditions. Indeed, let us assume without loss of generality that units are sorted in nonincreasing values of X, and otherwise perform this $O(N\log(N))$ operation as pre-processing.

KKT conditions can be expressed as

$$
\begin{aligned}
&\left(\frac{\pi_i}{X_i} - k\right)\frac{1}{X_i} + \alpha_i - \beta_i - \vartheta = 0 \quad \forall i = 1,2,\ldots,N,\\
&\frac{1}{N}\sum_{i=1}^{N}\frac{\pi_i}{X_i} - k = 0,\\
&\sum_{i=1}^{N}\pi_i = n,\\
&\pi_i \leqslant 1 \quad \forall i = 1,2,\ldots,N,\\
&\pi_i \geqslant 0 \quad \forall i = 1,2,\ldots,N,\\
&\alpha_i(1-\pi_i) = 0 \quad \forall i = 1,2,\ldots,N,\\
&\beta_i\pi_i = 0 \quad \forall i = 1,2,\ldots,N,\\
&\alpha_i, \beta_i \geqslant 0 \quad \forall i = 1,2,\ldots,N.\\
&k, \vartheta \in \mathbb{R}
\end{aligned}
\tag{8}
$$

By Proposition 1, there exists $s \in \{0,1,\ldots,N-1\}$ such that

$$\pi_1^* = \cdots = \pi_s^* = 1 > \pi_{s+1}^* \geqslant \cdots \geqslant \pi_N^* > 0. \tag{9}$$

Hence $\beta_i = 0$ for all i, and $\alpha_i = 0$ for all $i = s+1,\ldots,N$. Moreover, for any i with $\frac{\pi_i^*}{X_i} \geqslant k^*$, one has

$$\vartheta = \left(\frac{\pi_i^*}{X_i} - k\right)\frac{1}{X_i} + \alpha_i \geqslant 0. \tag{10}$$

With this, (8) yields

$$
\begin{aligned}
&k(s) = \frac{(n-s)\left(\sum_{i=s+1}^{N}X_i\right) + \left(\sum_{i=1}^{s}\frac{1}{X_i}\right)\left(\sum_{i=s+1}^{N}X_i^2\right)}{s\sum_{i=s+1}^{N}X_i^2 + \left(\sum_{i=s+1}^{N}X_i\right)^2},\\
&\vartheta(s) = \frac{sk(s) - \sum_{i=1}^{s}\frac{1}{X_i}}{\sum_{i=s+1}^{N}X_i},\\
&\pi_i(s) = 1 \quad \forall i = 1,2,\ldots s,\\
&\pi_i(s) = k(s)X_i + \vartheta(s)X_i^2 \quad \forall i = s+1,\ldots,N.
\end{aligned}
\tag{11}
$$

Let us rewrite the inequalities in (8). The condition (3) can be rewritten as

$$k(s)X_i + \vartheta(s)X_i^2 \leqslant 1 \quad \forall i > s,$$

which by (10) is equivalent to

$$k(s)X_{s+1} + \vartheta(s)X_{s+1}^2 \leqslant 1. \tag{12}$$

Condition $\pi_i \geqslant 0$ is satisfied since $k(s), \vartheta(s) \geqslant 0$. Complementarity slackness conditions hold by construction. The condition $\alpha_i \geqslant 0$ reads

$$k(s)X_i + \vartheta(s)X_i^2 \geqslant 1 \quad \forall i = 1,2,\ldots,s,$$

which by the non-negativity of $k(s)$ and $\vartheta(s)$ is equivalent to the single condition

$$k(s)X_s + \vartheta(s)X_s^2 \geqslant 1.$$

Hence, together with (11), one must check

$$k(s)X_s + \vartheta(s)X_s^2 \geqslant 1 \geqslant k(s)X_{s+1} + \vartheta(s)X_{s+1}^2. \tag{13}$$

The following strategy can then be used to solve (6): We start by setting $s = 0$, and calculating $k(s)$, $\vartheta(s)$ from (11). If (13) is also fulfilled, we construct $(\pi(s), k(s))$ from (11), which, by construction, satisfies the KKT conditions (8), and thus an optimal solution of (6) has been found. Otherwise, we increase s in one unit, update by (11) $k(s)$, $\vartheta(s)$ and check again (13). The process is repeated until a feasible solution is found (and this always happens, since KKT conditions are necessary and sufficient here). When s^* is found such that (13) is satisfied, the full vector $\pi(s^*)$ is calculated and given as optimal solution to (6).

This argument shows the following.

Proposition 2. *Assuming the units sorted in nonincreasing order of X_i, an optimal solution (π^*, k^*) to (6) is obtained in O(N) time requiring O(N) space.*

2. Discussion

When sample units have a very large range of book values X, a vector of probabilities proportional to the book values may not exist. In this note we have studied the problem of finding the vector of probabilities π minimizing the variance of the ratios $\frac{\pi_i}{X_i}$. This criterion leads to a quadratic optimization problem with a simple form, for which a linear-time algorithm is devised. One interesting question is whether the optimal solution of this quadratic problem strongly differs from the vector of probabilities given by the heuristic approach followed by practitioners. Our empirical findings indicate that the heuristic solution is very close to the minimum-variance one. As illustration, we have simulated N book values from a heavy-tailed distribution with density $f_X(x) = \frac{1}{x^2}(x > 1)$, for three different scenarios for N, namely $N = 200, 300, 500$. For each such N, a small grid of values for the fraction f of units sampled from the population are chosen, and, for the sample size $n = f\,N$, the vectors of probabilities π^* and π obtained respectively by solving (6) with our exact algorithm and applying the heuristic have been obtained. In Fig. 1, for $N = 200, 300, 500$, the correlation $cor(f)$ between π^* and π is plotted as a function of the fraction f of units sampled from the population. It is seen that the correlation is always extremely high. In other words, the heuristic solution systematically yields an almost-optimal solution if the variance of the ratios π_i/X_i is to be minimized.

Problem (6) is not the only way to obtain (approximate) MUS designs. Indeed, the problem addressed in this note is related to the multicriteria decision-making problem in which a set of weights π_i are sought such that their ratio $\frac{\pi_i}{\pi_j}$ are close to given scalars a_{ij}, e.g. [1,4] and the references therein. For instance, one could have replaced the objective function of (6) by a function of the differences between the ratios $\frac{X_i}{X_j}$ and $\frac{\pi_i}{\pi_j}$, yielding the objective $\min_{i,j}\sum_{i,j}\left(\frac{X_i}{X_j} - \frac{\pi_i}{\pi_j}\right)^2$. However, the resulting optimization problem is multimodal, to be addressed by time-consuming Global-Optimization techniques, as those described in [4], which become unfeasible for realworld

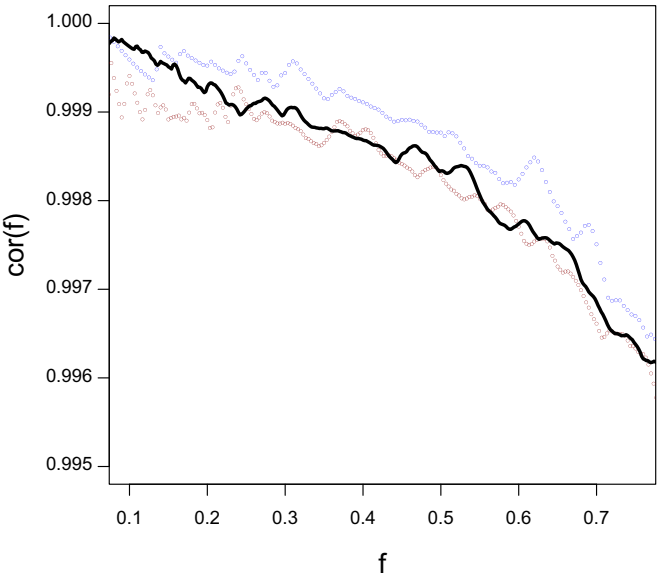

Fig. 1. Correlation between the heuristic and exact solutions.

problems in Auditing. Alternative (tractable) models exist. For instance, instead of minimizing the variance of the ratios π_i/X_i, one could minimize the p-th moment, yielding convex optimization problems. Models incorporating alternative objective functions, accommodating imprecise information on the book values X_i, or prior constraints on the probabilities π_i as e.g. in [2] deserve further study. Designing low-complexity algorithms for such new models is, in our opinion, a promising research line.

A different avenue of research would consist of addressing multiobjective issues, as in [3], by taking into account not only violations in (1), but also alternative criteria, such as the expected coverage $\sum_{i=1}^{N} \pi_i X_i$ of the sample generated. In this case, the resulting problem is convex, and a discrete approximation to the efficient set can be obtained by iteratively solving a variant of problem (6).

This paper has addressed the problem of randomly selecting sample units so that an imposed proportionality principle suffers least-possible deviations. However, this sampling process may also be the first stage for estimating population parameters, and then the main target should be to improve the quality of estimation (by reducing the variance of the resulting estimates). Whether the optimization problem analyzed in this paper or similar ones can help to achieve this goal certainly deserves further attention.

References

[1] R. Blanquero, E. Carrizosa, E. Conde, Inferring efficient weights from pairwise comparison matrices, Mathematical Methods of Operations Research 64 (2006) 271–284.
[2] E. Carrizosa, Deriving weights in multiple-criteria decision making with support vector machines, TOP 14 (2006) 399–424.
[3] E. Carrizosa, Unequal probability sampling from a finite population: A multicriteria approach, European Journal of Operational Research 201 (2010) 500–504.
[4] E. Carrizosa, F. Messine, An exact global optimization method for deriving weights from pairwise comparison matrices, Journal of Global Optimization 38 (2007) 237–247.
[5] W.G. Cochran, Sampling Techniques, Wiley, 1977.
[6] D.M. Guy, D.R. Carmichael, R. Whittington, Audit Sampling. An introduction, Wiley, 2002.
[7] T.W. Hall, J.E. Hunton, B.J. Pierce, Sampling practices of auditors in public accounting, industry, and government, Accounting Horizons 16 (2002) 125–136.
[8] Y. Tillé, Sampling Algorithms, Springer, 2006.
[9] Tillé, Y., A. Matei, Sampling: Survey Sampling. <http://cran.r-project.org/src/contrib/Descriptions/sampling.html>.