# Sample size determination for risk-based tax auditing

Petros Dellaportas

*University College of London, UK,*
*Athens University of Economics and Business, Greece*
*and Alan Turing Institute, UK.*

Evangelos Ioannidis

*Athens University of Economics and Business, Greece.*

Christos Kotsogiannis

*Tax Administration Research Centre (TARC),*
*University of Exeter Business School, UK,*
*and CESifo, Munich, Germany.*

**Summary**. A modern system of Revenue Administration requires an effective and efficient management of compliance which in turn requires a well designed taxpayers audit strategy. The selection of taxpayers to be audited by Revenue Authorities is a non-standard sample size determination problem, involving an initial random sample from the population and, based on the statistical information derived from it, a risk-based auditing scheme whose sole objective is to select for auditing the taxpayers with the highest estimated risk in the population. This paper provides a methodological approach that estimates the initial optimal random sample size such that the Revenue Administration Authority maximises their expected tax revenues. The methodology is illustrated using administrative data from the UK's Revenue Authority, Her Majesty's Revenue and Customs (HMRC).

*Keywords*: Finite population bootstrap; Tax gap; Tax revenue.

## 1. Introduction

An efficient management of tax compliance—and the promotion of voluntary compliance amongst taxpayers—necessitates the development of modern approaches based on risk management. For tax matters, for example, Revenue Authorities (RAs) frequently audit taxpayers in order to assess whether they comply with the tax law and declare their true tax liabilities but also to promote voluntary compliance by increasing the probability of detection perceived (and realised) by non-compliant taxpayers. In practice, of course, not all tax returns are exhaustively examined by RAs. This would not only be infeasible, given budgetary and capacity constraints, but it would also be unnecessary to waste scarce enforcement resources on routinely examining low-risk and compliant taxpayers. Importantly, international experience shows that a small number of large taxpayers (of around 1 percent) are responsible for around 60 percent of domestic tax collections, while a significant number of small taxpayers (of around 30 percent) account for less than 10

*Address for correspondence: Petros Dellaportas Department of Statistical Science, University College , Gower Street , London, WC1E 6BT* E-mail: p.dellaportas@ucl.ac.uk

percent of domestic tax collections (the use of 'taxpayers' is used generically to describe taxable entities, including business). This, too, and for RAs whose objective is to (also) maximise tax revenues, calls for the careful design of the tax audit strategy.

Auditing is not specific, of course, to tax collection matters and Revenue Authorities only, but it applies to much broader themes and issues—including, for example, regulation, management—and across many different economic sectors. While the broader theme of the issue is well recognised the analysis in this paper is primarily motivated by RAs—and the current policy discussions in many RAs involving optimal auditing strategies—and thus the discussion will be centered around the compliance issues pertaining RAs. While the use of random sample audits is common practice in advanced RAs (for example, Her Majesty's Revenue Customs (HMRC), U.S Internal Revenue Services (IRS), Canada Revenue Agency), less advanced RAs have begun considering integrating this in the design of their risk management strategy. The reason for this is not only to estimate the 'tax gap'—defined to be difference between tax voluntarily paid and tax actually owed (the gross compliance gap)—and its component parts, but, also, importantly, to establish the criteria for the selection of taxpayers to be audited by identifying characteristics deemed to be of greatest risk for substantial noncompliance. The selection of taxpayers to be audited, as well as other types of controls, is based on the assessment of risk and the development of risk-based selection techniques and risk-scoring systems. This allows tax audits to be prioritized and enables a more efficient allocation of RA resources; see OECD (2004) for a review and for a description of various approaches and country experiences, Khwaja et al. (2011).

The most extensive analysis based on random audits comes from the IRS studies and the Taxpayer Compliance Measurement Program (TCMP). The studies were conducted to provide the compliance information needed to gauge performance and to inform direction of agency resources through efficient prioritization of audits (see, for example, Rotz et al. (1994)) via the construction of Discriminant Inventory Function scores (Hunter and Nelson, 1996) which quantifies the evasion risk of each tax return in the sense of assigning probability to each return regarding irregularities or evasion; Andreoni et al. (1998) provides discussion on the IRS audit program. This research has identified a considerable compliance gap. More specifically, the 1992 tax gap of federal individual income tax based on TCMP audits in 1988, the last year of that program, was estimated to be in the range of $93 to $95 billion which translates into an individual gross noncompliance rate of about 17 percent. In response to the need for up-to-date measures of taxpayer compliance, and better target of audit resources, the IRS developed the National Research Program (NRP) which is also designed to be less intrusive and burdensome to taxpayers than the TCMP. Under the NRP, a sample of 30,000 returns were chosen for limited in-person audit, and about 2,000 returns for calibration audits, (where each line is examined and supporting documentation is required from a taxpayer). The NRP therefore signifies a substantial reduction in audits from the 54,000 taxpayers who were required to participate in face-to-face audits in the earlier TCMP program.

In the UK, and HMRC, the Random Enquiry Programmes (REPs) involve samples of taxpayers being selected at random and their returns subjected to full enquiries by HMRC officers, Revenue and Customs (2019b). The purpose of these programmes is to identify the proportion of taxpayers under-reporting their tax liabilities and the corresponding

amount of additional tax due. As these audits are randomly selected and constitute a representative sample it can be used for inference for the amount of under-declared tax liability for the whole population. The REPs do not identify all incorrect returns or the full scale of tax gaps, especially where independent information from third parties is not available to verify the data supplied by the taxpayer. The implication of this is that tax gap estimates produced through random enquiries will under-estimate the full extent of the tax gap. To correct for this RAs (such as the IRS and HMRC) use a range of 'multipliers'—supplemented with econometric analysis—to make adjustments for non-detection of under-reported income, (Andreoni et al., 1998). Interestingly, however, RAs have started relying less on REPs as these programs are time consuming, and therefore, costly for both the RAs and the taxpayers. For HMRC, for example, for the three main categories of audits (in terms of tax bases) Self-Assesment, Employer Compliance and Corporate Tax while the sample sizes in 2004-2005 were, respectively, 6,482, 1,649, and 408, in 2015-2016 they were reduced to 2,522, 925, and 362 cases, Revenue and Customs (2019a). From a policy perspective the question then that arises is, what should the optimal size of the random sample be? And this is the aim of the paper, to explore the random sample size determination problem. This is a very practical problem of significant economic and social value for governments as they strive to improve the efficiency and effectiveness of the RA operations.

In the spirit of the HMRC and IRS practices, the analysis will be conducted within the following structure. There is a fixed number of audits that can be conducted by a RA due to budgetary constraints which are split into two Steps. In Step 1, a random sample is drawn from the population and is used to estimate the risk of non-compliance for all population units. In Step 2, the remaining number of audits are chosen to be those that generate the maximal estimated risk of non-compliance in the population. Within this context, we show that optimality of sample size needs to strike a balance: If the RA under-samples in Step 1, it may get hazy estimates of the drivers of non-compliance, and this could lead to poor targeting in Step 2 as to offset the fact that it has not used up many audits in Step 1. If it oversamples at Step 1, it has fewer audits left to make use of, even though its understanding of non-compliance may be better. This, as will be shown later on, results in a non-standard sample size determination problem, which is highly non-linear with a non-trivial analytical solution. We develop a methodological approach which tackles this problem with an innovative numerical algorithm based on bootstrap.

The structure of this paper is as follows. Section 2 sets up the background within which the analysis is conducted. Section 3 presents and analyses the methodological approach devloped. Section 4 presents a numerical illustration based on data provided by the HMRC, which are capable of creating a simulation example mimicking to some extend the population characteristics faced by the HMRC. Finally, Section 5 provides some concluding remarks.

## 2. Preliminaries

### *2.1. Notation*

Indices $i$ and $j$ are used for sampling units and indices 1 and 2 are used for the two Steps of the sampling scheme. When information from previous year(s) audits is used, it is

indexed by *old*. The indicator function is denoted as $I\{.\}$ and understood as $I\{A\} = 1$ if $A$ is true and $I\{A\} = 0$ otherwise. All vectors are taken to be column vectors.

### 2.2. The problem

The RA has resources to select and audit a subset of taxpayer (henceforth 'units') of size $n$ from a population $U$ of size $N$. For simplicity, the cost of an audit is assumed to be the same for all units (an assumption that can be straightforwardly relaxed). The tax gap of the $i$-th unit is denoted by $y_i$ and is defined to be, as noted earlier, the difference between tax actually owed and tax voluntarily paid by that unit. The RA has access to additional information for each $i$ unit, summarised in a vector of covariates $\mathbf{x}_i \in \Re^{p+1}$, which is correlated to $y_i$. Specific risk-propensity characteristics of the taxpayers may be taken into account by including appropriate covariates. Denote $U := \{(y_i, \mathbf{x}_i), i \in J := \{1, ..., N\}\}$, where $J$ is the set of indices in the population. It will be further assumed that the $y_i$'s are independent observations of random variables $Y_i$ with expectation and variance, conditional on the vector of covariates $\mathbf{x}_i$, given via a linear model

$$\mu_i := \mathbb{E}(Y_i|\mathbf{x}_i) = \mathbf{x}_i^T \beta^{(0)}, \quad \mathrm{Var}(Y_i|\mathbf{x}_i) = \sigma^2, i \in J, \tag{1}$$

where $\beta^{(0)}$ is a vector of unknown coefficients. Note that the methodology can be easily extended to cases in which $Y_i$ in (1) is replaced by $\log(Y_i)$ or some other invertible variance-stabilizing transformation taking values in $\Re$. Thus, we assume a super-population model, inducing the correlation between an individual's decision on its tax gap $Y_i$ and the covariates, as is usual in a model-assisted design-based framework; see, for example, Särndal et al. (2003) and Brewer (2013) for a nice discussion on the controversy between model-assisted and model-based approaches. It is worth emphasising that we do not consider a classical model-assisted design-based estimator, as we do not seek to estimate a population total but rather identify individuals with a potentially high tax-gap. We call $\mu_i$ the 'expected tax gap', but it is only in this context that an expectation is taken with respect to the generating super-population model. In what follows our approach will be design-based in that in any other context expectations will be meant with respect to the sample selection at the Step 1. The assumption of model (1) is however necessary, as tax payers' decisions on their tax declaration, and thus on $y_i$, may differ from period to period; this makes $\beta^{(0)}$ a link of the induced populations between periods, which will be required for our methodology. Note that by denoting by $\mathbf{X}_N$ the design matrix having rows $\mathbf{x}_i^T, i = 1, ..., N$, and by $\bar{\beta} := (\mathbf{X}_\mathrm{N}^\mathrm{T}\mathbf{X}_\mathrm{N})^{-\mathbf{1}}\mathbf{X}_\mathrm{N}^\mathrm{T}\mathbf{Y}$ the LS estimator of $\beta^{(0)}$ across all $N$ population units, $\bar{\mu}_i := \mathbf{x}_i^T\bar{\beta}$ is, in the above sense, an unbiased estimator of $\mu_i$.

The RA seeks an approach to utilize (1) to select the subset of $n$ units. It is common practice, as touched upon in the introductory section, to perform this task in two Steps. In Step 1 a random sample of $n_1$ units is drawn from the population. Let $J_1$ denote the set of indices selected in the Step 1 sample. We assume that this sample is drawn without replacement according to first order inclusion probabilities $\pi_i, i = 1, ..., N$, which are all assumed to be positive. Each one unit of the Step 1 sample is then audited and its tax gap is determined. This sample is then used to estimate $\bar{\beta}$, and implicitly $\beta^{(0)}$ in (1), by

$$\hat{\beta} := (\mathbf{X}_1^\mathrm{T}\mathbf{\Pi}^{-\mathbf{1}}\mathbf{X}_1)^{-\mathbf{1}}\mathbf{X}_1^\mathrm{T}\mathbf{\Pi}^{-\mathbf{1}}\mathbf{y}, \tag{2}$$

(Jonrup and Rennermalm, 1976), which yields further estimates $\hat{\mu}_i := \mathbf{x}_i^T \hat{\beta}$. Here $\mathbf{X}_1$ denotes the matrix having rows $\mathbf{x}_i^T, i \in J_1$ and $\mathbf{\Pi} := diag\,(\pi_i, i \in J_1)$. Note that because under sampling without replacement $n_1 = \sum_{i \in U} \pi_i$, the sample inclusion probabilities depend on $n_1$. However, one may still specify them for the entire population because their proportional dependence on $n_1$ cancels out in (2). In Step 2 the RA selects and audits $n_2 = n - n_1$ units, those with a maximal estimated tax gap $\hat{\mu}_i$, after excluding the units already present in the Step 1 sample. Our objective is to provide a way to select $n_1$ in an optimal way that will be precisely defined below.

The objective of the RA is to detect as much non-declared income as possible and therefore minimize the overall tax gap. Given the information obtained from Step 1, the detected tax gap per unit is expected to be higher for the units selected in the Step 2, but it is not optimal for the RA to allocate all of its resources in Step 2 because the Step 2 sample selection relies on the accuracy of the estimate of Step 1: $n_1$ should be sufficiently high so that $\hat{\beta}$ provides adequate information to select the Step 2 units. Section 3 provides an approach that proposes a sample size $n_1$ such that expected tax revenues are maximised.

It is worth emphasizing at this stage that although other sampling designs could be more efficient, the analysis focuses on the above two-Step scheme as they are used widely by RAs. Moreover, the proposed methodology can accommodate alternative to LS estimators when this is deemed necessary. For example, models richer than (1) may capture characteristics such as heteroskedasticity so it might be preferable to use some form of an iterative weighted LS estimator instead of the LS estimators $\bar{\beta}$ and $\hat{\beta}$ to improve the variance in the estimation of $\beta^{(0)}$; see, for example, Särndal et al. (2003).

## 3. Maximizing the expected tax revenue

### *3.1. The target quantity*

We now present a methodological approach to choose $n_1$ that aims at maximising the expected tax revenue. Let the tax revenue of Step 1 be

$$R_1(n_1) := \sum_{i \in J_1} \mu_i, \tag{3}$$

Let us consider a fixed parameter value $\beta$ and define $\mu_i(\beta) := \mathbf{x}_i^T \beta$. For determining the Step 2 sample we proceed as follows. Let

$$\mu_{(1)}(\beta) \leq \mu_{(2)}(\beta) \leq \cdots \leq \mu_{(N-n_1)}(\beta),$$

be the ordered values of $\mu_i(\beta)$ for $i \in J \setminus J_1$ and

$$J_2(\beta) := \left\{ i \in J \setminus J_1 : \mu_i(\beta) \geq \mu_{\left(1 - \frac{n-n_1}{N-n_1}\right)}(\beta) \right\},$$

be the set of indices of the units not present in the Step 1 sample with maximal $\mu_i(\beta)$. If $\beta$ were the RA's 'working' parameter value, the tax revenue of Step 2 would be given by

$$R_2(\beta, n_1) := \sum_{i \in J_2(\beta)} \mu_i. \tag{4}$$

Thus, if the true parameter $\beta^{(0)}$ were known, the revenue from both Steps would be given by

$$R\left(\beta^{(0)}, n_1\right) := R_1(n_1) + R_2\left(\beta^{(0)}, n_1\right).$$

If $\beta^{(0)}$ is not known, but estimated in the Step 1 by $\hat{\beta}$, then the revenue from Step 1 $R_1(n_1)$, remains unaffected, but the revenue from the Step 2 is a random variable $R_2\left(\hat{\beta}, n_1\right)$ and depends on the distribution of $\hat{\beta}$. We propose to choose $n_1$ so as to maximize the expectation of the revenue

$$R\left(\hat{\beta}, n_1\right) := R_1(n_1) + R_2\left(\hat{\beta}, n_1\right).$$

This expectation is taken over the selection of the sample in Step 1 and is, therefore, conditional on the generation of the population $U$. We therefore define:

$$\begin{aligned} \bar{R}(U, n_1) &:= E\left[R\left(\hat{\beta}, n_1\right)|U\right] = E\left[R_1(n_1)|U\right] + E\left[R_2\left(\hat{\beta}, n_1\right)|U\right] \\ &= E\left[R_1(n_1)|U\right] + \sum \mathbb{P}_{n_1}\left(\hat{\beta} = \beta|U\right) E\left[R_2\left(\hat{\beta}, n_1\right)|U, \hat{\beta} = \beta\right], \end{aligned} \quad (5)$$

where the summation in the last expression extends over the possible values $\beta$ of $\hat{\beta}$ for the various outcomes of Step 1 sample and $\mathbb{P}_{n_1}\left(\hat{\beta} = b|U\right)$ denotes the probability of the occurrence of each value, which depends on the population $U$ and $n_1$.

For example, in the case of only one continuous covariate, $\mu_i := \beta_0^{(0)} + \beta_1^{(0)} x_i$ and under $\beta_1^{(0)} > 0$ we would obtain, denoting by $\mu_{(r)}$ the $r$-th order statistic of $\{\mu_i\}_{i=1,\ldots,N}$,

$$\begin{aligned} E\left\{R_2\left(\hat{\beta}, n_1\right)|U\right\} &= \mathbb{P}_{n_1}\left(\hat{\beta}_1 > 0|U\right) E\left\{\sum\nolimits_{i \in J_2\left(\beta^{(\mathbf{0})}\right)} \mu_i | U, \hat{\beta}_1 > 0\right\} \\ &+ \mathbb{P}_{n_1}\left(\hat{\beta}_1 < 0|U\right) E\left\{\sum\nolimits_{i \in J_2'\left(\beta^{(\mathbf{0})}\right)} \mu_i | U, \hat{\beta}_1 < 0\right\} \\ &= \mathbb{P}_{n_1}\left(\hat{\beta}_1 > 0|U\right) E\left\{\sum\nolimits_{i \in J_2\left(\beta^{(\mathbf{0})}\right)} \mu_i | U\right\} \\ &+ \mathbb{P}_{n_1}\left(\hat{\beta}_1 < 0|U\right) E\left\{\sum\nolimits_{i \in J_2'\left(\beta^{(\mathbf{0})}\right)} \mu_i | U\right\}, \end{aligned}$$

where

$$J_2'(\beta) := \left\{i \in J \setminus J_1 : \mu_i(\beta) \leq \mu_{\left(\frac{n-n_1}{N-n_1}\right)}(\beta)\right\}.$$

The reason for the first equality above is that when $\hat{\beta}_1 > 0$ the $n - n_1$ units with largest $\mu_i(\hat{\beta})$ are those with largest $\mu_i$, while when $\hat{\beta}_1 < 0$ the $n - n_1$ units with largest $\mu_i(\hat{\beta})$ are those with lowest $\mu_i$. This also justifies the second equality since $J_2\left(\beta^{(\mathbf{0})}\right)$ does not depend on $\hat{\beta}_1$. Thus, with probability $\mathbb{P}_{n_1}\left(\hat{\beta}_1 < 0|U\right)$ the RA selects in the Step 2 the units with lowest tax gap, instead of those with the highest tax gap. When the number of parameters is greater than one the expressions involved are getting even more complicated and so is the analytical solution to the maximization of (5). We therefore propose the use of a numerical solution based on bootstrap (Efron and Tibshirani, 1994). This approach is explained in the next subsection.

The outcomes of the proposed bootstrap approach will be compared in a simulation study in the last section to two extreme cases. These are (i) The 'Only-Step-1' expected revenue which is achieved by setting $n_1 = n$ and thus devoting all available resources to Step 1 and (ii) The 'Only-Step-2' expected-revenue which is achieved by setting $n_2 = n$ and assuming that all $\mu_i \in J$ are known. Note that (i) is feasible whereas (ii) is an 'oracle' approach in the sense that $\mu_i$ are not available in practice.

### *3.2. The proposed bootstrap approach*

The idea is to approximate the (conditional on $U$) distribution induced by the selection of the Step 1 sample, including the distribution of $\hat{\beta}$, which is involved in $\mathbb{P}_{n_1}\left(\hat{\beta} = \beta | U\right)$ in (5), by a bootstrap analogue based on a past sample of the Step 1. For a survey of finite population bootstrap methods see, for example, Mashreghi et al. (2016). More precisely, let us assume that in a previous period a Step 1 sample $\{(y_i, \mathbf{x}_i), i \in J_{1,old}\}$ of size $n_{1,old}$ was drawn using the same design which will be used in the current period, having first order inclusion probabilities $\{\pi_i, i \in J\}$. We use here a 'Pseudo-population bootstrap method': a pseudo-bootstrap-population $U^* := \{(y_i^*, \mathbf{x}_i^*), i \in J^*\}$ is constructed by repeating $w_i^*$ times each unit from this old sample, a construction which will be made precise further below by specifying $w_i^*$. The goal is to do so in a way that $U^*$ will mimic the population generating mechanism of $U$. Then, for each sample size $n_1$ belonging to a specified grid of values, $B$ bootstrap-samples $\{(y_i^*, \mathbf{x}_i^*), i \in J_{1,b}^*\}, b = 1, ..., B$ of size $n_1$ are drawn from $U^*$ using the same design of interest, re-scaled for the new sample size $n_1$ and the population size, thus inducing first order inclusion probabilities $\{\pi_i^*, i \in J^*\}$, in a way that will also be made precise further below. Note that we drop in our notation the dependence on $n_1$ of various quantities, as e.g. of $J_{1,b}^*$, to avoid further complication of the notation.

For each of the $B$ bootstrap samples we compute the weighted LS estimates $\hat{\beta}_b^*$ as in (2) and subsequently $\mu_i(\hat{\beta}_b^*) := \mathbf{x}_i^T \hat{\beta}_b^*$, from which $J_{2,b}^*(\hat{\beta}_b^*)$ is also determined. Based on these, $R(U, n_1)$ is estimated by $R(U^*, n_1)$ by averaging the sum of the appropriate quantities involved in $R_1(n_1)$, see (3), and $R_2(\hat{\beta}_b^*, n_1)$, see (4), over all $b = 1, ..., B$ bootstrap samples, while, moreover, substituting $\mu_i$ by $\mu_i(\hat{\beta}_{old})$. Thus the following quantity is averaged over all bootstrap samples to obtain $R(U^*, n_1)$

$$R_b^* := \sum_{i \in J_{1,b}^*} \mu_i(\hat{\beta}_{old}) + \sum_{i \in J_{2,b}^*(\hat{\beta}_b^*)} \mu_i(\hat{\beta}_{old}). \tag{6}$$

There are two approximations involved here: the first one is the approximation of $\mu_i$ by $\mu_i(\hat{\beta}_{old})$; the second one concerns mainly picking the units with maximal $\mu_i(\hat{\beta})$ by choosing in the bootstrap world those with maximal $\mu_i(\hat{\beta}^*)$. This actually relies on the approximation of the distribution of $\hat{\beta}$ by the bootstrap-distribution of $\hat{\beta}^*$, which in turn relies in the quality of the approximation of $U$ by $U^*$. What still remains to be specified is the construction of the pseudo-bootstrap-population $U^*$ and of $\{\pi_i^*, i \in J^*\}$. This is done subsequently separately for the case of equal and for the case of unequal selection probabilities, which are here assumed to depend on $\mu_i$.

*3.2.1. Bootstrap for equal probabilities sampling*

Let us first assume simple random sampling without replacement (SRS) so that the design used for selecting Step 1 sample of the previous period had first order inclusion probabilities given by $\pi_i = n_{1,old}/N$, assuming for simplicity that $N_{old} = N$. Following Chao and Lo (1994), the number of times a unit of this sample will be repeated in order to obtain $U^*$ will be given by $w_i^* := trunc(N/n_{1,old})$, where $trunc()$ indicates the largest integer not exceeding the argument. The bootstrap population is then completed to $N$ units, by selecting $N_{var} := N - \sum w_i^*$ units with equal probabilities and with replacement from Step 1 sample, prior to drawing a new bootstrap sample and separately for each one of them. As the size $n_1$ of the bootstrap samples may differ form $n_{1,old}$ we will use $\pi_i^* := n_1/N$. Thus, under SRS:

$$w_i^* := trunc(N/n_{1,old}), N_{var} := N - \sum w_i^*, \pi_i^* := n_1/N. \tag{7}$$

*3.2.2. Bootstrap for unequal probabilities sampling*

If some prior information on $\beta^{(0)}$, and thus on $\mu_i$, is available prior to selection of Step 1 sample, it may be exploited by prioritising, with respect to their selection probabilities, the units with an anticipated expected high tax gap: one might choose $\pi_i$ to be increasing in $\mu_i$. A standard choice here is to choose $\pi_i$ proportional to the size, but as $\mu_i$ may take negative values and may, moreover, contain a number of extreme outliers, we chose a robust alternative which is to allow $\pi_i$ to be proportional to the ranks $R_i^a$, for some constant $a$, of $\mu_i, i \in J$, which is thus here the "size variable". In what follows we set $a = 1$.

In the case of unequal probabilities sampling the pseudo-bootstrap-population is often based on repeating each Step 1 sample unit $w_{HT,i}^* = \pi_i^{-1}$ times, appropriately approximated by an integer. We use the proposal of Barbiero et al. (2015) to calibrate those weights in order to better mimic certain characteristics of $U$. The idea is to define new weights $w_{CAL,i}^*$ to be as close as possible to $\pi_i^{-1}$, while satisfying the constraints that the induced pseudo-population size equals the one of $U$, while the percentage of those anticipated to have a highly ranked tax-gap remains unaffected in the pseudo-bootstrap-population. More precisely, the goal is to minimize

$$\sum_{i \in J_{1,old}} \frac{\left(w_{CAL,i}^* - \pi_i^{-1}\right)^2}{\pi_i^{-1}},$$

under the constraints

$$\sum_{i \in J_{1,old}} w_{CAL,i}^* = N,$$

and further, setting the threshold of 'highly ranked' to some $q \in [0,1]$, under

$$\sum_{i \in J_{1,old}} w_{CAL,i}^* \mathbf{I}\{R_i \geq qN\} = qN,$$

since $\sum_{i \in J} \mathbf{I}\{R_i \geq qN\} = qN$. The solution to this problem is given by

$$\mathbf{w_{CAL}^*} := \mathbf{w_{HT}^*} + \mathbf{\Pi^{-1}} \mathbf{R}_1^{\mathrm{T}} (\mathbf{R}_1^{\mathrm{T}} \mathbf{\Pi^{-1}} \mathbf{R}_1)^{-1} \left(\mathbf{c} - \mathbf{R}_1^{\mathrm{T}} \mathbf{w_{HT}^*}\right),$$

where $\mathbf{w}^*_{\mathbf{CAL}} := \left(w^*_{CAL,i}\right)_{i\in J_{1,old}}$, $\mathbf{w}^*_{\mathbf{HT}} := \left(w^*_{HT,i}\right)_{i\in J_{1,old}}$, $\mathbf{c} := (N, qN)$, $\mathbf{R}_1$ denotes the matrix having for $i \in J_{1,old}$ rows $(1, \mathbf{I}\{R_i \geq qN\})$ and, finally, $\mathbf{\Pi} := diag\,(\pi_i, i \in J_{1,old})$.

Thus, under unequal probability sampling, we use

$$w^*_i := round(w^*_{CAL,i}), N_{var} = 0, \pi^*_i := \pi_i n_1/n_{1,old}. \tag{8}$$

### *3.3. The Algorithm*

The details are given in Algorithm 1.

**Algorithm 1** Calculation of optimum sample size for tax auditing

1: Compute the weights $w^*_i$ and $N_{var}$ using equation (7) for equal probability sampling and (8) for unequal probability sampling.
2: **for all** k of a grid of values $n_{1,k}$, $k = 1, ..., K$ **do**:
3: Generate a bootstrap pseudo-bootstrap-population $U^*_k$ with index set $J^*_k$ by repeating each unit in the random sample of the past audit $w^*_i$ times.
4: Compute $\pi^*_{k,i}, i \in J^*_k$, as specified in (7) for equal probability sampling and in (8) for unequal probability sampling.
5: **for all** $b = 1, ..., B$ **do**:
6: If $N_{var} > 0$ select $N_{var}$ more units from the random sample of the past audit with SRS with replacement to complete the pseudo-bootstrap-population $U^*_k$ to $U^*_{k,b}$. Else, set $U^*_{k,b} = U^*_k$.
7: Draw a sample $\left(y^*_{b,i}, \mathbf{x}^*_{b,i}\right),\ i \in J^*_{1,k,b}$ of size $n_{1,k}$ from $U^*_{k,b}$ using the same design as for the random sample of the past audit but with adjusted first sample inclusion probabilities $\pi^*_{k,i}, i \in J^*_{k,b}$.
8: From $\left(y^*_{b,i}, \mathbf{x}^*_{b,i}\right),\ i \in J^*_{1,k,b}$ compute $\hat{\beta}^*_{k,b}$ as in (2)
9: **for all** $j = 1, ..., N$ **do**:
10: Compute $\mu_j(\hat{\beta}^*_{k,b}) := \mathbf{x}^T_j \hat{\beta}^*_{k,b}$
11: **end for**
12: Determine $J^*_{2,k,b}\left(\hat{\beta}^*_{k,b}\right)$, the index set of the $n_{2,k} = n - n_{1,k}$ largest values of $\{\mu_j(\hat{\beta}^*_{k,b}), j \in J \setminus J^*_{1,k,b}\ \}$.
13: Compute $R^*_{k,b}$ as in (6)
14: **end for**
15: Compute $\widehat{\bar{R}(U, n_{1,k})} = \bar{R}(U^*_k, n_{1,k}) := B^{-1} \sum_{b=1}^{B} R^*_{k,b}$
16: **end for**
17: Set $n_{1,k'}$ as the optimal sample size where $k' = \underset{k=1,...,K}{\arg\min}(\widehat{\bar{R}(U, n_{1,k})})$

## 4. Simulation example

In this Section we conduct a simulation experiment based on real data provided to us by HMRC, in order to check the performance of the bootstrap method proposed in Section 3. The data concern a specific sub-population of taxpayers for which HMRC estimated

a model based on a random sample. For confidentiality reasons HMRC provided us only with the first four moments of the nine covariates, the nature of which was not further specified, and of the residuals, as well as the coefficients $\beta_0, \beta_1, \ldots, \beta_9$ of the estimated regression model. On the basis of these data we generated an artificial population for our simulation study so as to mimic as precisely as possible the true population under study. We generated nine independent vectors of covariates of length $N$ with each one of them drawn from a distribution having the first four moments identical to those of a covariate in the HMRC sample. This was done using an algorithm proposed by Devroye (1986), pp. 690-691, see also Dellaportas and Karlis (2001). We thus obtained $\mathbf{x}_i \in \Re^{10}, i = 1, ..., N$. Then, with the coefficient values ${\beta^{(0)}}^T = (\beta_0, \beta_1, \ldots, \beta_9)$ a 'true' population of tax sample units was generated by setting $y_i := \mathbf{x}_i^T \beta^{(0)} + \epsilon_i$, where $\epsilon_i, i = 1, ..., N$ are errors which were also generated from a distribution having the first four moments identical to those of the residuals in the HMRC sample. We further re-centered the $y_i$ and adjusted the intercept term accordingly in order to mask the true expected tax gap across the population for confidentiality reasons. We were also provided with moments of the 10-percent trimmed data. This had the effect of reducing the standard error of the covariates sometimes by a factor of 100 or even of 1000, while for some of them it still had a value of a few thousands after this reduction. We performed simulations under two scenarios, one using the moments of the raw data (Scenario A) and one using the moments of the trimmed data (Scenario B).

We used $N = 50,000$, although the actual HMRC taxpayer sub-population considered here is larger, in order to keep the necessary computation time limited. This artificial population has been considered to be the 'true' population so it has been kept fixed throughout the simulation exercise. This implies that the simulation variability emanated from the random fluctuations of the estimators $\hat{\beta}_{old}$. We further assumed that the budget constraints restricted the total number of audits to $n = 2000$, a figure which is arbitrarily chosen from us, since the true number of total taxpayer audits is not known to us. The Step 1 sample size of the previous period was set to $n_{1,old} = 100$ for Scenario A and to $n_{1,old} = 200$ for Scenario B, figures which are safely greater than the true optimal values for each scenario. The grid for optimising across $n_1$ which we used varied from 20 to 150 with a step of 5 and from 160 to 500 with a step of 20. The number of bootstrap replications was set to $B = 400$ and the number of the simulation size, i.e. the number of 'old' initial samples for estimating $\beta_{old}$ was set to $M = 1000$. The sampling was done by the 'sample' routine of the R-language, which allows for sampling without replacement and equal-, as well as unequal-, probability sampling with specified weights, which were set to the desired first order selection probabilities.

We present in Figure 1 a separate simulation of the expected revenue as a function of $n_1$ based on 10000 samples for Scenarios A and B. We first remark the very low optimal Step 1 sample sizes $n_1$, particularly under Scenario A. They might be due to the huge variances of the covariates, which are reciprocally related to the variance of the regression coefficient: it seems that even with small sample sizes they allow an estimation of the regression coefficient with sufficient accuracy in order to identify the units in the tail of the distribution. When the variance of the covariates is reduced in scenario B, the optimal $n_1$ climbs from about 50 to about 150. A second striking fact is that sampling with probabilities proportional to the rank of the $\mu_i$ does not increase revenue, at least

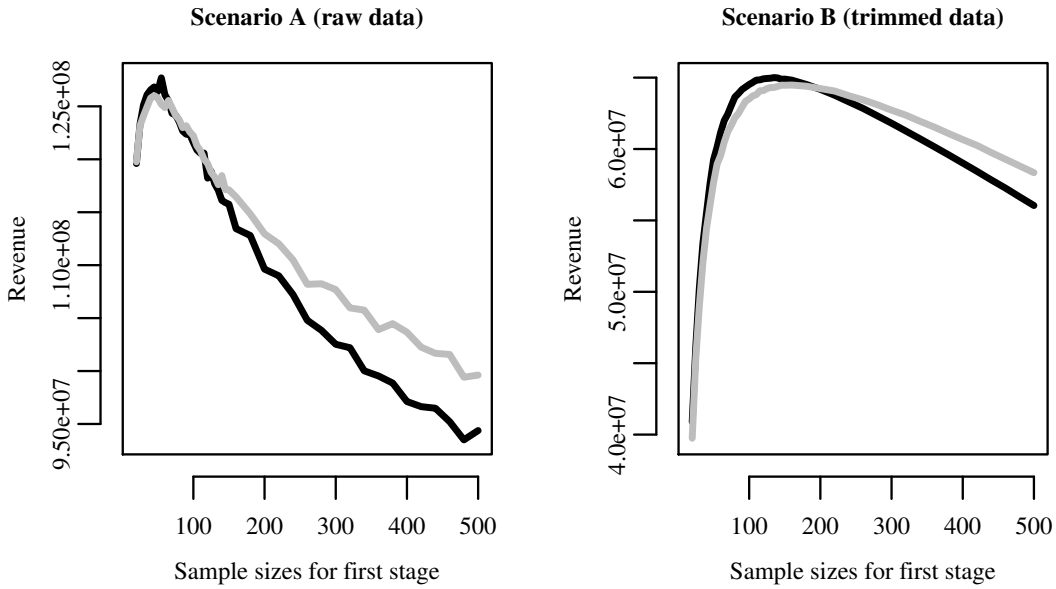



**Fig. 1.** Expected total revenue versus sample size for Step 1. Solid black: Revenue from equal probability sampling. Solid grey: Revenue from unequal probability sampling.

not in the neighborhood of the maximum: while $R_1(n_1)$ is indeed positively affected by unequal probability sampling, especially for values of $n_1$ larger than the optimal ones, it is $E\left[R_2\left(\hat{\beta}, n_1\right)|U\right]$ that dominates the sum of the two components. This latter tends to be slightly lower under unequal probability sampling, especially for small values of $n_1$. This might be due to a possible increase of the variance of the regression coefficient under unequal probability sampling. The above comparisons rely heavily on the assumption that the relation between expected tax gap and the covariates is stable across the population. If, however, this is violated and, for example, different slopes are at work for those with high values of the covariates, then it would be surely advantageous to favour the presence of such values in the sample of Step 1 by assigning higher probabilities to their selection.

We now turn to the simulation of our bootstrap-based method. Our goal is to investigate (i) how the estimated optimal $\hat{n}_1$ which maximises the estimated expected revenue $\widehat{\bar{R}(U, n_1)} = \bar{R}(U_k^*, n_1)$ compares to the true optimal $\bar{n}_1$ which maximises $\bar{R}(U, n_1) := E\left[R\left(\hat{\beta}, n_1\right)|U\right]$ and (ii) how the achieved true expected revenue $\bar{R}(U, \hat{n}_1)$ obtained by the RA when using the estimated optimal $\hat{n}_1$ compares to the maximal true expected revenue $\bar{R}(U, \bar{n}_1)$ which would have hypothetically been obtained by the RA when the (unknown) truly optimal $\bar{n}_1$ that maximises the true expected revenue had been used. We further explore how they both compare to the 'Only-Step-1' and 'Only-Step-2' expected revenues. We run the simulation under two sampling designs, with equal and unequal first order selection probabilities, both without replacement. Unequal selection probabilities were set proportional to the rank of $\mu_i, i \in J$, thus assuming the true $\mu_i, i \in J$

are known. In real life they would be estimated based on some previous Step 1 sample.

In Figure 2 it is shown that the estimated optimal $\hat{n}_1$ is, under Scenario A, strongly skewed with heavy tails on the right. Values of $\hat{n}_1$ in the right tail of the distribution are the result of a shape of $\bar{R}(U_k^*, n_1)$, which was quite skewed to the right as compared to the shape of $\bar{R}(U, n_1)$. This was not the case under Scenario B, where outliers were excluded along with a small percentage of the tails of the distribution of the covariate: Figure 2c shows that the estimation of $\bar{n}_1$ by $\hat{n}_1$ was more successful than under Scenario A (Figures 2a,2b). For the revenue, first note that the maximal true expected revenue indicated by the dashed vertical line, is by definition an upper bound for the achieved true expected revenue, the distribution of which is indicated by the black line. The densities were estimated by kernel estimators, the bandwidth of which were chosen empirically so as to smooth out the effects of evaluating the revenue at a discrete grid of possible values. The distribution of the revenue is under Scenario A also strongly skewed with heavy tails on the left, with values close to the maximum revenue resulting from values of $\hat{n}_1$ close to the mode of their distribution, while values for the revenue further away from the maximum are due to values of $\hat{n}_1$ in the right tail of its distribution. Between the two sampling schemes for Scenario A there is virtually no difference in estimated optimal sample sizes and the achieved revenues, as might be also expected from Figure 1.

Figure 3 reveals that the achieved and maximal expected revenues are relatively close when compared with the two extremes, namely the 'Only-Step-1' expected revenue and the 'Only-Step-2' expected revenue (close to $1.5\ 10^8$ in Scenario A and to $7.1\ 10^7$ in Scenario B). Note that it is natural that the 'Only-Step-1' expected revenue will be lower than the revenue achieved by the risk based method, as it aims at detecting the average tax gap and not the upper tail of the tax gap distribution, as the risk based method does. Moreover, in the SRS case it is close to 0 due to the re-centering of the $\mu_i$ and the $Y_i$ which was applied when generating the population in our simulation. Finally, note that it is also natural that the 'Only-Step-2' expected revenue will be higher than the revenue generated by the method which tries to estimate the individual's expected tax gap: the oracle "knows" exactly which units have the highest expected tax gap, so it does not need to devote resources in spotting them ($n_2 = n$), and moreover, has still a error probability of zero in detecting them.

We conjecture that it is important to control for the presence of outliers in the covariates which will act as leverage points for the LS estimates of the regression coefficients, either by excluding them or by using robust regression estimators. Moewover, unequal probability sampling might be advantageous in case of a departure from the assumption of stability of the relation between expected tax-gap and covariates.

## 5. Concluding remarks

A well functioning economy requires a well-functioning and efficienct RA. Recognising this policymakers across the globe have become increasingly aware of the importance of policies that promote voluntary tax compliance. Inspecting, however, every taxpayer is neither desirable nor feasible, given the availabilitty of resources to RAs. As a consequence RAs are giving considerable attention to the development of risk management practices.

This paper has contributed to this issue by investigating *optimal random sample selec-*

*tion* in tax audits, providing a statistical methodology aiming at maximizing expected tax revenues. The analysis is, of course, limited in several respects. The cost of audit has been assumed uniform (and fixed) across audits. This, clearly, is a simplifying assumption. In reality the probability of detecting income under-reporting is a function of the intensity of audit and the quality of information provided to tax auditors. The assumption that random auditing identifies the tax gap amongst the class of taxpayers who are audited is also quite strong. In reality, auditors do make errors in their assessment and the outcome of audits does rely on tax auditors expertise and experience. Indeed, existing estimates show that there is a considerable heterogeneity in detection rates across examiners for some income items; see, for example, Erard and Feinstein (2010). Incorporating some of these elements in the present analysis is feasible at a small cost of computational effort.

An interesting statistical problem also arises when one combines the issue of model choice and the objective of the RA. Throughout the paper it has been assumed that the 'best' model is fixed and previously estimated by the RA. However, it is clear that a proper model choice procedure may not be based on standard statistical techniques but, rather, on the ultimate objective of the RA. Last but not least, the methodology developed here should be incorporated in a larger audit framework in which audits take place in a stratified sampling fashion with a series of practical constraints. The analysis has provided only the first step of such a large, probabilistically sound approach to deal with this problem.

Although the statistical treatment in the paper is frequentist, one could envisage advantages of a Bayesian treatment. Bayesian updating may be useful in learning the regression parameters by combining past data and data obtained in Step 1. Such an approach can only be based on strong Tax authorities expertise combined with sophisticated Bayesian treatment under model misspecification such as, for example, the one presented in Holmes and Walker (2017). This might be a very challenging exercise if it is combined with a Bayesian model determination approach that assigns different posterior model probabilities in different years.

Our simulations showed that treatment of outliers may be inevitable in reality because their presence affects the estimation of regression parameters. We proposed either the exclusion of outliers or the use of robust regression estimators. Note, however, that exclusion of outliers affects $n_1$ and, therefore, optimising $n_1$ needs to take this into account. On the other hand, the choice of an appropriate robust regression estimator may also have an indirect similar impact. Both approaches have not been dealt in further detail and are left for future investigation.

There remains much scope for the analysis of sample selection and risk management. We hope to have shown that the task is worthwhile and that the conclusions can be instructive.

## Acknowledgements

We thank, without implicating, Anthony Burke, Will England, Sarah Jennings, and Prabhjot Sethi of the HMRC Tax Gap Team for many useful discussions on sampling for tax auditing and for providing us with the data. ESRC funding under ES/S00713X/1 and EP/N510129/1, Alan Turing institute funding under TEDSA2/100056 and Research Cen-

ter of the Athens University of Economics and Business funding under EP-2982-01 are gratefully acknowledged. Extremely constructive comments and guidelines by the associate editor highly improved earlier versions of this manuscript.

## References

Andreoni, J., B. Erard, and J. Feinstein (1998). Tax compliance. *Journal of economic literature 36*(2), 818–860.

Barbiero, A., G. Manzi, and F. Mecatti (2015). Bootstrapping probability-proportional-to-size samples via calibrated empirical population. *Journal of Statistical Computation and Simulation 85*(3), 608–620.

Brewer, K. (2013, 12). Three controversies in the history of survey sampling. *Survey Methodology 39*, 249–262.

Chao, M. T. and S.-H. Lo (1994). Maximum likelihood summary and the bootstrap method in structured finite populations. *Statistica Sinica 4*(2), 389–406.

Dellaportas, P. and D. Karlis (2001). A simulation approach to nonparametric empirical bayes analysis. *International Statistical Rveiew 69*(1), 63–79.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Erard, B. and J. Feinstein (2010). Econometric models for multi-stage audit processes: an application to the irs national research program. *Developing Alternative Frameworks for Explaining Tax Compliance*, 113–137.

Holmes, C. and S. Walker (2017). Assigning a value to a power likelihood in a general bayesian model. *Biometrika 104*(2), 497–503.

Hunter, W. J. and M. A. Nelson (1996). An irs production function. *National Tax Journal*, 105–115.

Jonrup, H. and B. Rennermalm (1976). Regression analysis in samples from finite populations. *Scandinavian Journal of Statistics 3*(1), 33–36.

Khwaja, M. S., R. Awasthi, and J. Loeprick (2011). *Risk-based tax audits: Approach and country experiences*.

Mashreghi, Z., D. Haziza, and C. Leger (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys 10*, 1–52.

OECD (2004). *Compliance risk management: audit case selection systems*. OECD.

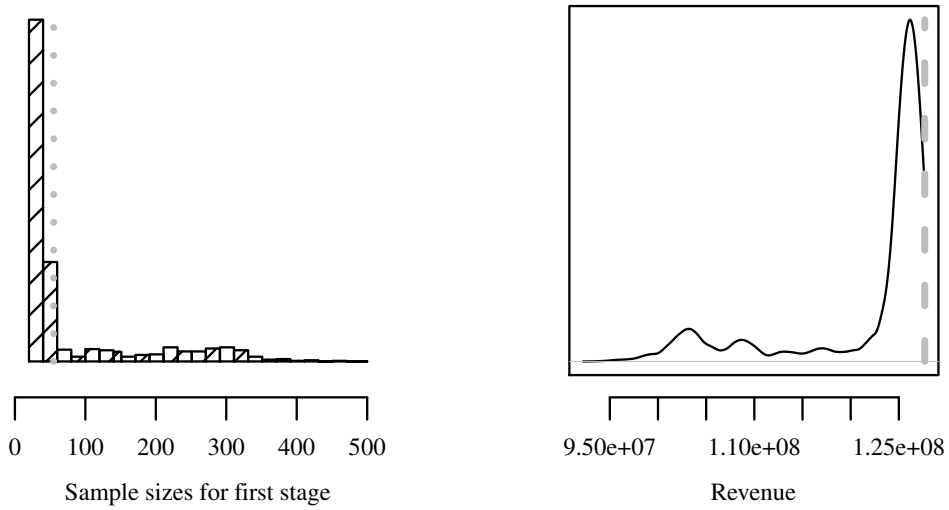
Revenue, H. and Customs (2019a). *HMRC, Measuring tax gaps edition 2019: Methodological annex*.

Revenue, H. and Customs (2019b). *HMRC, Measuring tax gaps edition 2019: Tax gap estimates for 2017-2018*.

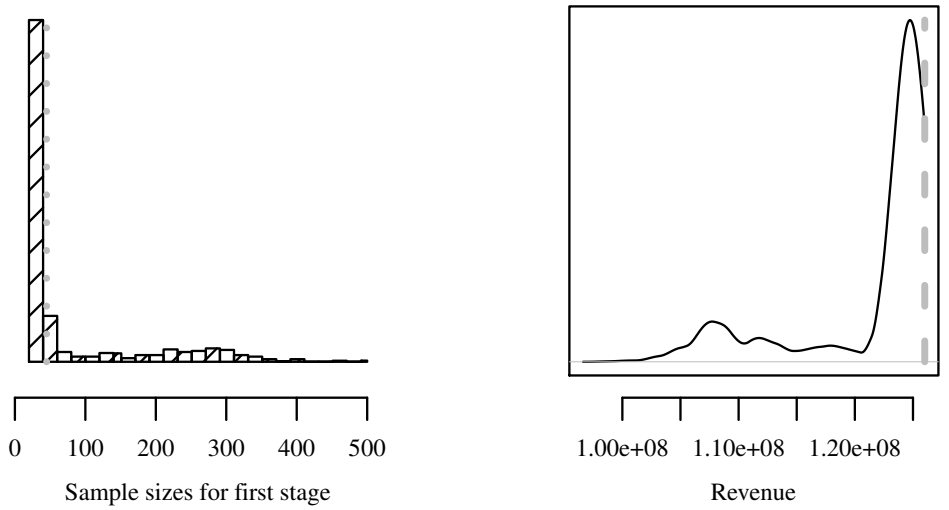
Rotz, W., J. Murlow, and E. Falk (1994). The 1995 taxpayer compliance measurement program (tcmp). sample redesign: A case history. *Turning Administrative System Into Information System. Internal Revenue Service, Washington*, 121.

Särndal, C.-E., B. Swensson, and J. Wretman (2003). *Model Assisted Survey Sampling (Springer Series in Statistics)*. Springer.

a) Equal probability sampling for Scenario A (raw data)




b) Unequal probability sampling for Scenario A (raw data)




c) Equal probability sampling for Scenario B (trimmed data)

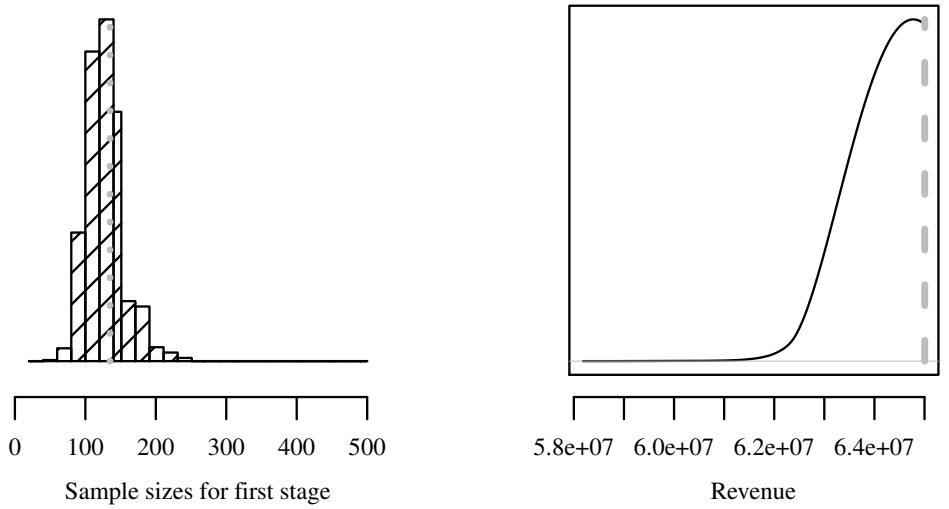



**Fig. 2.** Optimal samples sizes and expected revenue. Left panel: Optimal sample size for Step 1 (vertical line) and distribution of estimated sample size (histogram). Right panel: Maximal true expected revenue (vertical line) and distribution of achieved true expected revenue (kernel estimator).

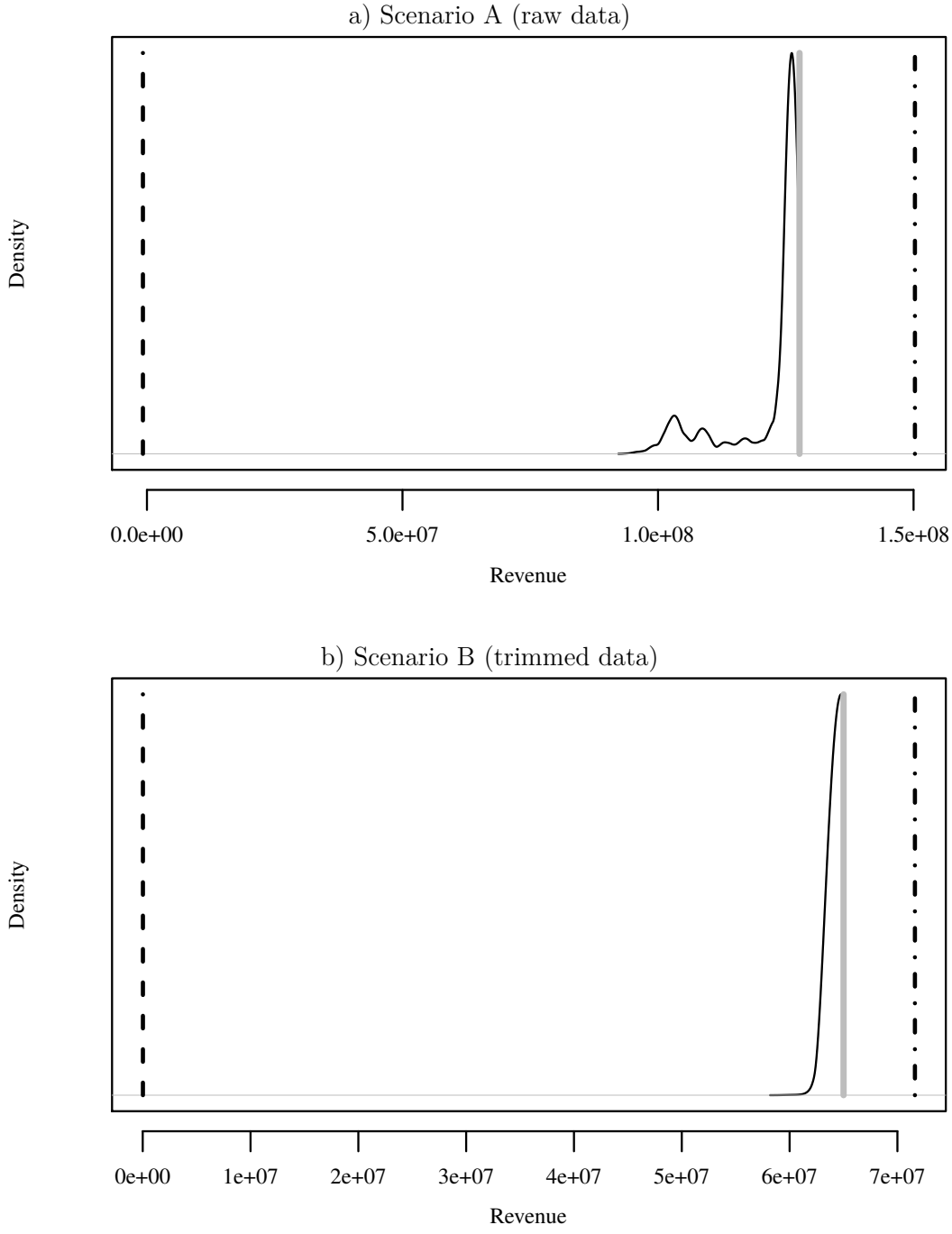



**Fig. 3.** Expected revenue under equal probability sampling. Solid black: Equal probability sampling, density of achieved true expected revenue; Solid grey: Maximal true expected revenue; Dashed black: 'Only-Step-1' expected revenue; Dashed-dotted black: 'Only-Step-2' expected revenue.