
EUROPEAN MASTER IN OFFICIAL STATISTICS
EMOS

# Abstract
# EMOS Master thesis competition 2025

# 'Efficiently Selecting Representative Audit Samples'

Author: **Lotte Mensink, Utrecht University**

Keywords: audit sample, representativeness, efficiency

## 1. INTRODUCTION

Audits are widely used for quality evaluation in many different fields, but they can be costly and time-consuming. When performing an audit, a sample of the population is selected for thorough examination. In performing audits within the context of official statistics, it is not uncommon that part of the population has already been audited. For example, it often occurs that the largest companies are audited by default, due to their substantial influence on the statistic of interest. Given the considerable costs associated with auditing, it would be desirable to re-use the already audited units. However, these units are not randomly chosen, and therefore form a non-probability sample, in which the inclusion probabilities of each unit are either not known or not useful. Traditionally, it is believed that inferences from non-probability samples cannot generalize beyond the sample, because they can contain serious sample selection bias (Rao, 2021).

This puts the researcher in a difficult position. On the one hand, it would be efficient to re-use the available initial audit sample, as it would save considerable resources. On the other hand, the possible non-representativeness of the initial audit sample forms a serious threat for the validity of the conclusions to be drawn from the audit sample.

To tackle this problem, Boeschoten et al. (2019) have developed an approach in which an audit sample can be selected that is representative, in which already audited units can be re-used. In this approach, also referred to as the deviance approach, sample selection viewed as a constrained minimization problem, in which the goal is to obtain the most representative sample possible, under certain audit sample size restrictions. These audit sample size restrictions include a maximum number of additional units to include in the audit sample and a maximum number of previously audited units to exclude from the audit sample.

The implementation of the deviance approach has been investigated using a simulation study, which showed that the deviance approach can be used to efficiently select a representative audit sample (Boeschoten et al., 2019). However, the deviance approach does suffer from a practical limitation. The method is focused on maximizing representativeness, instead of maximizing efficiency. Under the provided sample size restrictions, the deviance approach will find the most representative audit sample, but this does not necessarily mean that this sample is the most efficient in terms of additionally sampled units. In practice, the financial costs associated with auditing additional units are often high, and it might therefore be more desirable to find the most efficient solution, while making sure that representativeness is maintained.

To tackle this practical limitation, this study proposes to restructure the deviance approach in such a way that the number of units to include in and exclude from the sample is minimized, while making sure that representativeness is maintained. We refer to this restructured approach as the sample size approach. We believe the sample size approach fits better with the practical priorities of the researcher.

## 2. OBJECTIVE

First, on a practical level, the aim of this study is to introduce the sample size approach and explain how it may be applied in practice for various situations in which researchers may find themselves.
Second, on an analytical level, the aim of this study is to investigate and evaluate the performance of the sample size approach. Hereby, we intend to provide insight into how the method works, and how particular factors may affect audit sample selection using the sample size approach.
Finally, on a fundamental level, the aim of this study is to enhance insight into the relation between deviance (our measure for representativeness of the audit sample) and bias (in the estimates that the audit sample is intended to provide). The existence of a relation between these concepts forms a fundamental assumption for both the deviance approach as well as the sample size approach. Gaining more insight into the functional form of this relation is crucial for the proper application of the sample size approach.

## 3. THEORETICAL FRAMEWORK

This study assumes that the population that needs to be audited can be divided according to a classification variable of interest. For all units in the population, an error-prone version of the classification variable of interest is observed, denoted by Y. The true value of the classification variable of interest is referred to as W. Eventually, the researcher wants to use the audit sample to estimate the true proportions $P(W = w)$ and the measurement error of the error-prone classification $P(Y = y \mid W = w)$. Furthermore, it is assumed that one or more covariates are available for each unit, denoted by X. Finally, Z denotes the selection indicator of the audit sample, with $Z = 1$ indicating that a unit is included in the audit sample, and $Z = 0$ indicating that a unit is not included in the audit sample.

We rely on the conditional independence model, a model in which the observed classification variable of interest Y is independent of audit inclusion Z, given the covariate(s) X under consideration. If the observed classification and audit inclusion would in fact be conditionally independent, this would mean that our audit sample is representative. Hence, the deviance of the conditional independence model is used to reflect the representativeness of the audit sample. If the deviance is sufficiently small, we assume that the conditional independence model holds. This means the audit sample is sufficiently representative, in the sense that population parameters can be estimated from the audit sample without bias.

The sample size approach is formulated as a constrained minimization problem aimed at selecting a representative audit sample while minimizing the number of additional units to be included. The target function, which is the sum of the units to include in and exclude from the sample, is minimized under the constraint of a user-defined deviance threshold that should not be exceeded to ensure representativeness.

The sample size approach allows the user to choose any deviance threshold. Determining a suitable deviance threshold for an existing data set with an initial audit can be a challenging task. We assume that as deviance decreases, the sample becomes more representative, and bias decreases. However,

we can only guess when it comes to which deviance thresholds yield sufficiently representative audit samples in terms of bias. Moreover, an audit sample that is considered sufficiently representative in one research context, might not be deemed sufficiently representative in another context.

The study describes a detailed approach to find out the maximum amount of bias that can be expected under different deviance threshold. Using this approach, the researcher can model the relation between bias and deviance based on the known characteristics of the data. Combining this relation with the researcher's maximum tolerable bias, the researcher can determine the most appropriate deviance threshold for a given scenario.

## 4. METHODS

To evaluate how the sample size approach performs, a simulation study is conducted. In this simulation study, the sample size approach, including the deviance threshold selection method, iswe applied to several possible situations in which researchers may find themselves. Hereby, we aim to provide insight into how the method works, and how characteristics of a data set may affect different aspects of audit sample selection. Because we want to illustrate the sample size approach for several situations in which researchers may find themselves, the factors that are varied in the simulation study are based on information that is already available to the researcher. The factors to vary are presented in Table 2. For each factor, we will investigate the main effects by varying it while fixing the other factors to the most desirable situation. Furthermore, we will take the least desirable and most desirable situation for every factor and investigate this in a full factorial design. As some of the main effects and full factorial conditions overlap, this yields 12 conditions in total. For every condition, we generate 1000 different distributions of X, Y, W and Z with corresponding data sets. We apply the sample size approach (including the method to determine the most appropriate deviance threshold based on the relation between bias and deviance) to each of these data sets.

Table 2: Conditions simulation study

| Factor | Conditions | Description |
|---|---|---|
| Relation between Y and X | Cramér's V = 0.1<br>Cramér's V = 0.2<br>Cramér's V = 0.4 | The relation between observed classification Y and covariate X |
| Sampling fraction | 0.003<br>0.01<br>0.03<br>0.1 | The ratio of initially audited units to units in the population |
| Maximum measurement error | 50%<br>33%<br>25% | The maximum measurement error that the researcher expects |

The goal of the simulation is to see whether the proposed sample size approach works as expected. Using the described deviance threshold selection method, we can obtain audit samples for any given tolerable amount of bias. The question is, however, whether these audit samples in fact contain no more bias than the tolerable amount. We can answer this question by determining the bias in the final audit samples that result from applying the sample size approach.

## 5. RESULTS

In the majority of the conditions, the average bias after the sample size approach was applied is below the maximum tolerable bias. When the sampling fraction is smallest, and when the maximum tolerable bias is low, average bias is sometimes higher than maximum tolerable bias. In many of the other conditions, the average bias falls well below the maximum tolerable bias. We are also

interested in seeing in how many data sets (out of the 1000 data sets per condition) the observed bias exceeds the maximum tolerable bias based on which the deviance threshold was chosen. The proportions of data sets in which maximum tolerable bias was exceeded are shown in Table 7.

Table 7: The proportion of final audit samples in which observed bias in $P(W = w)$ or $P(Y = y|W = w)$ exceeds maximum tolerable bias

| Sampling fraction | Measurement error | YX relation | Bias 1% | Bias 2% | Bias 5% |
|---|---|---|---|---|---|
| 0.003 | Max 50% | Weak | 0.338 | 0.220 | 0.061 |
| 0.003 | Max 50% | Strong | 0.368 | 0.272 | 0.114 |
| 0.003 | Max 25% | Weak | 0.304 | 0.190 | 0.057 |
| 0.003 | Max 25% | Strong | 0.336 | 0.230 | 0.086 |
| 0.01 | Max 25% | Strong | 0.230 | 0.110 | 0.030 |
| 0.03 | Max 25% | Strong | 0.118 | 0.045 | 0.012 |
| 0.1 | Max 50% | Weak | 0.056 | 0.022 | 0.004 |
| 0.1 | Max 50% | Strong | 0.059 | 0.021 | 0.004 |
| 0.1 | Max 33% | Strong | 0.067 | 0.023 | 0.002 |
| 0.1 | Max 25% | Weak | 0.039 | 0.015 | 0.005 |
| 0.1 | Max 25% | Medium | 0.035 | 0.011 | 0.005 |
| 0.1 | Max 25% | Strong | 0.043 | 0.014 | 0.000 |

These results show that there are some cases in which the sample size approach works well, and the proportion is (very close to) zero, but there are also cases in which the performance is quite bad. First, we can conclude that the method works better as we tolerate more bias. Overall, the performance is best in the conditions where the maximum tolerable bias is 5%, and worst in the conditions where the maximum tolerable bias is 1%. Furthermore, we again observe a strong effect of the sampling fraction. The method appears to work better when the sampling fraction is larger. As the final audit sample sizes were also a lot larger in these situations, this conclusion is not surprising.
Although this study mainly focused on performance in terms of bias, efficiency is another key aspect of performance. This aspect was studied more thoroughly in a pilot simulation study, which was also part of the thesis project (Mensink et al., 2023). The results of this simulation study showed that the sample size approach made it possible to obtain a representative audit sample through sampling fewer additional units than the deviance approach. Furthermore, the sample size approach outperforms the deviance approach by recognizing samples that are already representative to begin with. For representative initial audit samples, the deviance approach would suggest including more units in the sample to improve further upon representativeness, whereas the sample size approach recognizes that no additional units are needed because the sample is not selective to begin.

All in all, we conclude that the sample size approach could be a valuable addition to the auditing practice. Especially in situations when efficiency in terms of additional units to sample is prioritized, the sample size approach can provide good solutions. Further research is necessary to explore the full potential and investigate implementation in various auditing contexts.

## 6. CONTRIBUTION

In the context of official statistics, auditing the data used to produce a statistic is crucial to ensuring its' credibility. In many cases, part of the data has already been audited for other purposes. This study tackles the challenge of obtaining a representative sample while at the same time making optimal use of the audit sample that is already available. On a practical level, this study contributes by introducing the sample size approach, providing researchers with a valuable tool that allows them to optimize their efficiency in auditing.

Furthermore, on an analytical level, this study contributes by evaluating the performance of the sample size approach through extensive simulation studies, offering insights into how various factors affect the application of the sample size approach.

Finally, on a fundamental level, this study contributes by gaining more insight into the theoretical foundation of the sample size approach. The study thoroughly investigates the relation between bias and deviance. It is shown that the relation exists, and guidance is provided how the researcher can take this into account in determining a deviance threshold to be able to use the sample size approach.

## 7. REFERENCES

[1] Boeschoten, L., S. Scholtus, and A. van Delden (2021). A note on efficient audit sample selection. https://arxiv.org/abs/2105.10737.

[2] Mensink, L., S. Scholtus, and L. Boeschoten (2023). A comparison between two approaches for efficient audit sample selection. https://github.com/lottemensink/pilot_simulation_research_archive/tree/main.

[3] Rao, J. (2021). On making valid inferences by integrating data from surveys and other sources. Sankhya B 83 (1), 242–272.