Research Report

A Comparison Between Two Approaches for Efficient Audit Sample Selection

Lotte Mensink
9585842

Supervisors: Laura Boeschoten & Sander Scholtus

Word Count: 2498
FETC Approved 22-1861

Methodology and Statistics for the Behavioral, Biomedical and Social Sciences
Utrecht University
January 2023

Introduction

Official statistics are often presented for sub-populations, that are defined by categorical variables (Boeschoten et al., 2021). For example, statistics on turnover of establishments are grouped by economic activity, where establishments are classified according to the NACE rev 2. codes (Eurostat, 2008). It can be challenging to accurately determine the right NACE rev 2. classification (Burger et al., 2015), which results in classification error. A study by van Delden et al. (2016) looked at how classification error in economic activity affect turnover estimates. The results showed that classification error can cause up to 10% bias in published turnover estimates. These large biases can occur because classification errors do not average out to zero (Schwartz, 1985). To illustrate, the economic activity 'sale and repair of passenger cars and light motor vehicles' (code 45112) has a relatively high probability to be wrongly classified as the economic activity 'specialised repair of motor vehicles' (code 45200) (van Delden et al., 2016). Because the sector 45112 is very large, and because the average turnover in sector 45112 is much larger than in sector 45200, there are many establishments with relatively large turnover wrongly classified in sector 45200, resulting in an overestimation of turnover for sector 45200. Because the impact of classification error on the statistic of interest can be high, it is vital to check whether registers contain classification errors. This can be done by performing an audit.

Audits are widely used for quality improvement, applied in fields like official statistics, clinical research, finance and machine learning (see for example Chataway et al. (2004), Derks et al. (2021) and Hernández et al. (2014)). When performing an audit, a subset of the population is sampled in an audit sample. The audit sample is thoroughly evaluated in order to infer the true value with respect to a classification variable. With this, we make the implicit assumption that this true value can in fact be inferred. All units in the population should have a chance to be selected in the audit sample, in order to provide a reasonable basis to draw conclusions about the population (The Financial Reporting Council, 2016). At national statistical institutes such as Statistics Netherlands, the population often consists of error-prone classifications that are observed for every unit. By drawing an audit sample, the quality of the observed classifications can be evaluated, because the true classifications from the audit sample can be compared to the error-prone classifications from the population. Throughout this paper, we will mainly consider audit samples in the context of official statistics. However, it is worthwhile to note that this method can be relevant to other fields in which audits are performed.

In official statistics, a subset of units that are audited for other purposes is often already available (Boeschoten et al., 2021). For example, it often occurs that the largest companies are audited by default, due to their substantial influence on the statistic of interest. In a sense, a subset of already audited units

forms a non-probability sample, in which the inclusion probabilities of each unit are either not known or not useful. The initial audit can contain serious sample selection bias (Rao, 2021), which would cause the sample to be unrepresentative. As auditing comes at considerable costs, it is desired to re-use as many already audited units as possible, and minimize the number of additional units that need to be sampled. However, this poses a problem, because it becomes difficult to draw valid conclusions from an audit sample when we are unsure about the representativeness of the already audited cases (Langer and Krosnick, 2018).

To tackle this problem, Boeschoten et al. (2021) have developed a framework in which an audit sample can be selected that is representative with respect to the classification variable, while re-using as many earlier audited units as possible. In this framework, the true value of the classification variable of interest is referred to as W. For all units in the population, an error-prone version of the classification variable of interest is measured, denoted by Y. Furthermore, it is assumed that one or more covariates are available for each unit, denoted by X. Finally, we let Z denote 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. Audit sample selection is then considered to be a constrained minimization problem, in which the goal is to obtain the most representative sample possible, under certain audit sample size restrictions. 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 sample.

The framework relies on the conditional independence (CI) 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 are in fact conditionally independent, this would mean that our audit sample is representative. Hence, the deviance of the CI model is used to reflect the representativeness of the audit sample. If the deviance is sufficiently small, the CI model cannot be rejected, which means that we assume that conditional independence holds, and that the audit sample is sufficiently representative. To decide whether an obtained deviance value is sufficiently small, Boeschoten et al. (2021) use a chi-square distribution with an α of 0.05 to provide a threshold (Agresti, 2013).

The framework proposed by Boeschoten et al. (2021) suffers from two limitations:

1. The method compares the deviance against a threshold to decide whether a sample is sufficiently representative. However, insight is lacking when it comes to the amount of bias that corresponds to a certain deviance threshold. More insight into the exact meaning of "sufficiently representative" is desired.

2. The method minimizes the deviance, instead of minimizing the number of additionally sampled units. It will find the optimal solution in terms of representativeness, but it will not find the most efficient solution in terms of additionally sampled units. In practice, the costs associated with auditing additional units are often high, and it might therefore be more desirable to find the most efficient solution, while maintaining representativeness.

The aim of the current study is to address both limitations of the existing framework, which will be referred to as the *deviance approach* henceforth. We address the limitations of the deviance approach by restructuring in such a way that the number of additionally sampled units is minimized, while maintaining a fixed level of representativeness. We will refer to this approach as the *sample size approach*. The sample size approach allows us to (1) investigate how different thresholds for deviance relate to bias in the statistics of interest and (2) to find more efficient solutions in terms of additionally sampled units. The research question that will be investigated throughout this project is: how do different thresholds for deviance affect the bias in the statistics of interest? The aim is to eventually summarize the results of this study in a practical guideline for efficiently drawing audit samples using the sample size approach.

In this report, we aim to introduce the sample size approach for audit sample selection. The mathematical formulation will be presented, and a pilot simulation study will be performed in order to compare the sample size approach to the deviance approach.

Theoretical Framework

In this section, we present the mathematical formulation of the sample size approach. We are interested in obtaining an audit sample that is representative with regard to the classification variable of interest. We use the deviance of the conditional independence (CI) model as a criterion for representativeness. This means we analyze the joint distribution (X, Y, Z) in the observed target population using the non-saturated log-linear CI model $(XY)(XZ)$. This model contains two direct relations, one between the covariate(s) and the observed (error-prone) classification of interest, and one between the covariate(s) and the audit selection indicator. This model is compared against a saturated log-linear model, that also includes the direct relation between the observed classification of interest Y and audit sample inclusion Z and the three-way interaction (XYZ). This comparison results in the deviance, which gives an indication of the fit of the CI model. Analogously to Boeschoten et al. (2021), we let n_{ijk} denote the number of observations in each cell $(X = i, Y = j, Z = k)$. The formula for the deviance D can now be expressed as follows:

$$D = C + 2 \sum_{i,j,k} n_{ijk} \, log \, n_{ijk} - 2 \sum_{i,k} n_{i+k} \, log \, n_{i+k},$$

where

$$C = 2\sum_{i} n_{i++}\, log\, n_{i++} - 2\sum_{i,j} n_{ij+}\, log\, n_{ij+}.$$

This term is constant, because it only depends on the distribution of X and Y, and will therefore be the same for every possible audit sample (Bishop et al., 1975). Here, $n_{i+k} = \sum_j n_{ijk}$, $n_{i++} = \sum_{i,k} n_{ijk}$ and $n_{ij+} = \sum_k n_{ijk}$. For a more elaborate motivation for using this model, including the entire mathematical derivation, see Boeschoten et al. (2021).

We consider the general situation in which additional units can be selected for auditing (moved from $Z = 0$ to $Z = 1$), and previously audited units may be excluded from the audit sample (moved from $Z = 1$ to $Z = 0$). After applying the method, the adjusted number of units in cell $(X = i, Y = j, Z = k)$ is denoted as m_{ijk}:

$$m_{ij1} = n_{ij1} + \delta^{+}_{ij} - \delta^{-}_{ij};$$

$$m_{ij0} = n_{ij0} - \delta^{+}_{ij} + \delta^{-}_{ij}.$$

Here, δ^{+}_{ij} represents the number of additional units with $X = i$ and $Y = j$ to include in the sample, and δ^{-}_{ij} represents the number of initially audited units with $X = i$ and $Y = j$ to remove from the sample.

The goal is to minimize the number of additionally sampled units, while reusing as many cases as possible. Hence, the target function of the minimization problem can be written as:

$$\delta = \sum_{i,j} \delta^{+}_{ij} + \sum_{i,j} \delta^{-}_{ij}.$$

When minimizing δ^{+}_{ij} and δ^{-}_{ij}, a number of constraints apply. First of all, it is important that the sample is sufficiently representative. Under the CI model, the deviance asymptotically follows a chi-square distribution with $I(J-1)$ degrees of freedom, where I and J represent the number of categories in X and Y respectively (Agresti, 2013). The user can specify a maximum for the deviance via α. The higher α is set, the more strict we are in accepting the CI model that assumes representativeness. The user-specified deviance threshold leads to the following constraint:

$$D \leq \chi^2_{I(J-1)}(1-\alpha).$$

Second, for each combination $(X = i, Y = j)$, there exist bounds on δ^{+}_{ij} and δ^{-}_{ij} based on the initial counts n_{ijk}, since no more units can be moved from $Z = 0$ to $Z = 1$ or vice versa than are initially available:

$$0 \leq \delta^{+}_{ij} \leq n_{ij0};$$

$$0 \leq \delta^{-}_{ij} \leq n_{ij1}.$$

Third, the user might desire to set a minimum audit sample size M, that is required for doing the inferences needed from the audit. The user-specified minimum for the audit sample leads to the following constraint:

$$\sum_{i,j} m_{ij1} \geq M.$$

The minimization procedure can now be written as follows:

$$min\{\sum_{i,j} \delta_{ij}^{+} + \sum_{i,j} \delta_{ij}^{-}\},$$

under constraints:

$$\begin{aligned} m_{ij1} &= n_{ij1} + \delta_{ij}^{+} - \delta_{ij}^{-}; \\ m_{ij0} &= n_{ij0} - \delta_{ij}^{+} + \delta_{ij}^{-}; \\ D &\leq \chi^2_{I(J-1)}(1-\alpha); \\ 0 &\leq \delta_{ij}^{+} \leq n_{ij0}; \\ 0 &\leq \delta_{ij}^{-} \leq n_{ij1}; \\ \sum_{i,j} m_{ij1} &\geq M. \end{aligned}$$

This is a linear minimization problem with both linear and nonlinear constraints, which can be solved using the 'nloptr' package (Johnson, 2022) in R (R Core Team, 2022).

Pilot Simulation Study

Aims

To investigate the sample size approach, it is compared with the deviance approach in a simulation study. The performance of both approaches is assessed in various situations.

Data-generating mechanisms

In this simulation study, we make use of a subset of the conditions investigated in Boeschoten et al. (2021). In the original study, conditions were created based on varying relationships between X and W, Y and W and Y and Z. For each of the three relationships, we take the (1) most desirable and the (2) least desirable condition, and we investigate them in a full factorial design, yielding 8 conditions in total. In all conditions, a joint probability density is generated based on the bivariate relationships specified in Table 1. For each of the conditions, 1000 population data sets of size $N = 10000$ are sampled from the generated joint probability density. The size of the initial audit sample is more or less 300 in all conditions, as follows from the probabilities in Table 1.

Selectivity initial audit (relation YZ)			(1) No selectivity Y			(1) Strong selectivity Y		
			1	2	3	1	2	3
	Z	0	.323	.323	.323	.323	.323	.323
		1	.010	.010	.010	.018	.010	.002

Measurement error (relation WY)			(1) Perfect measurement W			(2) Imperfect measurement W		
			1	2	3	1	2	3
		1	.333	.000	.000	.300	.017	.017
	Y	2	.000	.333	.000	.033	.267	.033
		3	.000	.000	.333	.050	.050	.233

Explanatory power of the covariate (relation WX)			(1) Strong relationship W			(2) Weak relationship W		
			1	2	3	1	2	3
		1	.267	.033	.033	.267	.017	.017
	X	2	.033	.267	.033	.067	.200	.033
		3	.033	.033	.267	.100	.100	.133

Table 1: Overview of the bivariate probability distributions to generate the data for the simulation study.

Estimands

For this study, there is one target parameter. We consider the true proportion of units g in the population with $W_g = w$ for each category w:

$$P_w^W = \frac{1}{N} \sum_{g=1}^{N} I(W_g = w).$$

Under the assumption that the CI model holds, P_w^W can be estimated without bias from the audit sample by

$$\hat{P}_w^W = \sum_y P_x^X p_{w|x}^{W|X},$$

where P_x^X is the proportion of units in the population with $X_g = x$, and $p_{w|x}^{W|X}$ is the observed proportion of cases with $X_g = x$ in the audit sample that also have $W_g = w$ (Boeschoten et al., 2021).

Methods

To allow for a direct comparison, both the deviance approach and the sample size approach are applied to each simulated data set. Subsequently, a random audit sample is selected based on the proposed solution of the constrained minimization problem. The selected audit sample is used to calculate the target parameter $\hat{P}_w^W$.

Performance Measures

To evaluate the performance of the sample size approach, we will focus on comparing the number of added units and the deviance of the final audit sample to the deviance approach. Furthermore, we will evaluate the bias in the target parameter $\hat{P}_w^W$. The bias will be calculated by subtracting the estimated proportions from the audit sample from the true proportions that were used to generate the data.

Results

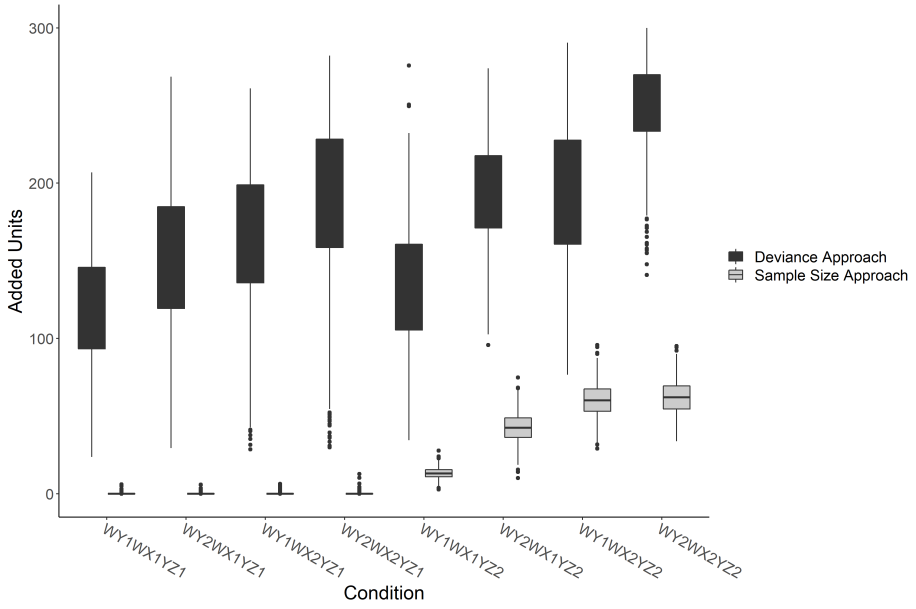

Figure 1: Comparison of the number of added units to the final audit sample using the deviance approach and the sample size approach. The boxplots demonstrate the spread of 1000 solutions obtained for every condition. WY1 = no measurement error, WY2 = strong measurement error, WX1 = strong relation between W and X, WX2 = weak relation between W and X, YZ1 = no selectivity, and YZ2 = strong selectivity.

Figure 1 shows the comparison between the sample size approach and the deviance approach in terms of added units to the final audit sample. In Boeschoten et al. (2021), the maximum number of additional units to sample was set to 300. From the boxplots of the added units, it can be observed that the number is generally much lower for the sample size approach as compared to the deviance approach.

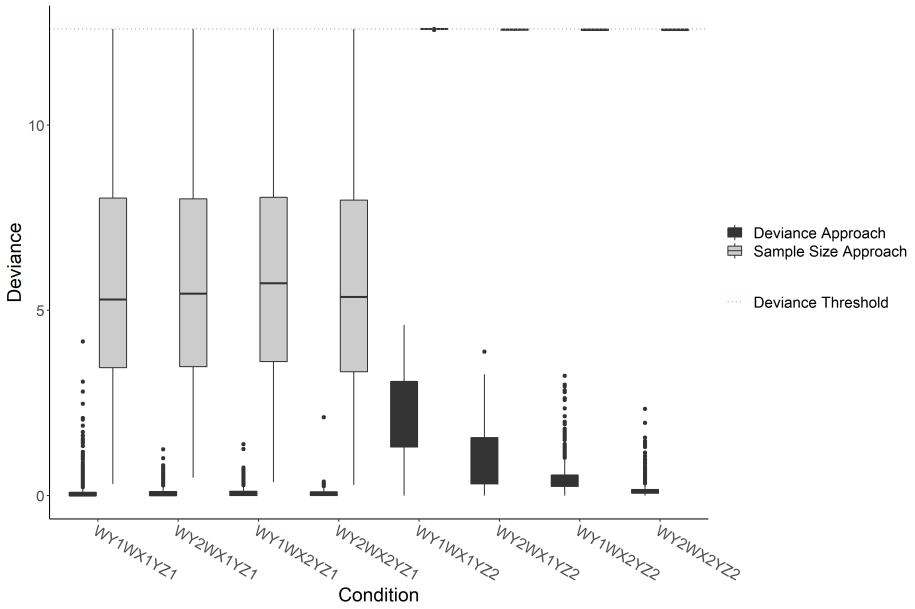

Figure 2: Comparison of the deviance of the final audit sample using the deviance approach and the sample size approach. The boxplots demonstrate the spread of the deviance from 1000 audit samples obtained for every simulation condition. WY1 = no measurement error, WY2 = strong measurement error, WX1 = strong relation between W and X, WX2 = weak relation between W and X, YZ1 = no selectivity, and YZ2 = strong selectivity.

Figure 2 shows the comparison between the sample size approach and the deviance approach in terms of the deviance of the final audit sample. In the original study, an audit sample was considered sufficiently representative when the deviance was smaller than or equal to the 95^{th} percentile of a Chi-square distribution with 6 degrees of freedom. This corresponds to a deviance threshold of 12.59. For an accurate comparison, this deviance threshold was used in the current study as well. From the boxplots, we can observe that when deviance is minimized, as is done in the deviance approach, it falls well below the threshold that is considered sufficiently representative. Furthermore, we can observe that using the sample size approach, deviance conforms to the threshold in the last four conditions, in which there is selectivity in the initial audit.

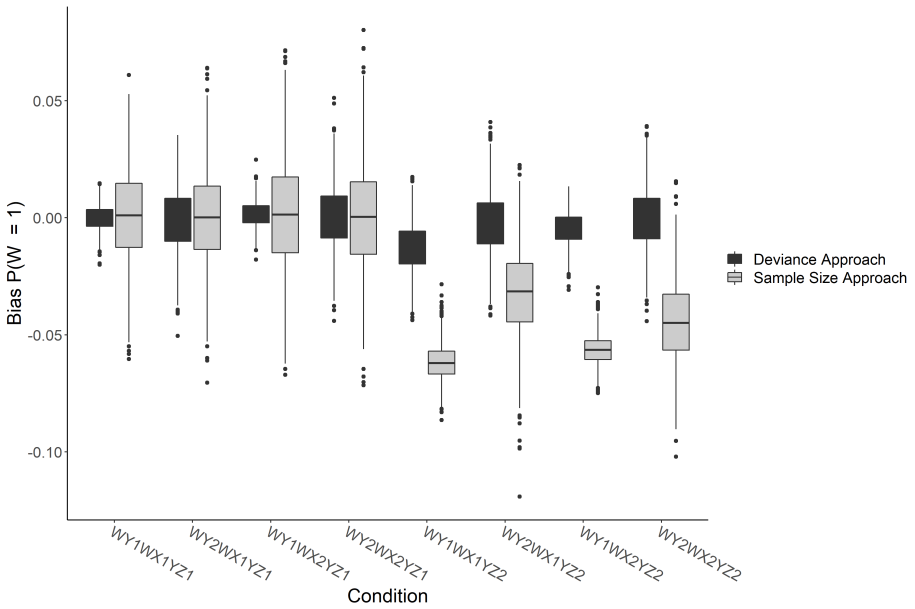

Figure 3: Comparison of bias in $P(W = 1)$ in the final audit sample using the deviance approach and the sample size approach. The boxplots demonstrate the spread of bias in 1000 audit samples obtained for every simulation condition. WY1 = no measurement error, WY2 = strong measurement error, WX1 = strong relation between W and X, WX2 = weak relation between W and X, YZ1 = no selectivity, and YZ2 = strong selectivity.

Figure 3 shows that average bias is around zero in the first four conditions. In the last four conditions, bias becomes more substantial when using the sample size approach. This difference in bias between the sample size approach and the deviance approach might be explained by the differences in the deviance values of the audit samples from both approaches.

Discussion

In this report, we aimed to introduce the sample size approach, and compare it against the deviance approach. The results indicate that the sample size approach works as expected, in the sense that fewer additional units need to be sampled as compared to the deviance approach. The first four conditions were generated with no selectivity in the initial audit sample. Hence, these samples should have been representative in the first place. The sample size approach confirms this, as we can observe that the additionally sampled units for these conditions are either 0 or close to 0. Furthermore, the deviance follows the expected distribution using the sample size approach, except that it is cut off at the deviance threshold. If the deviance approach is used in these conditions, the fact that the sample is already representative is ignored, and additional units

are sampled in order to lower the deviance even further. This is something that might be undesirable in practice, due to the substantial costs that are associated with auditing additional units.

Finally, it is vital to consider the bias in $\hat{P}_w^W$. Ideally, bias in $\hat{P}_w^W$ is close to zero. The results show however, that especially with the sample size approach, the bias is quite substantial in more challenging conditions. One possible reason for this, is that the deviance threshold that was considered acceptable in this study is too high. Over the course of the upcoming months, we will investigate how different deviance thresholds relate to bias. The goal is to obtain a deeper insight into the relationship between deviance thresholds and bias in the audit sample.

References

Agresti, A. (2013). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley.

Bishop, Y. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete multivariate analysis: theory and practice*. Springer Science & Business Media.

Boeschoten, L., S. Scholtus, and A. van Delden (2021). A note on efficient audit sample selection.

Burger, J., A. van Delden, and S. Scholtus (2015). Sensitivity of mixed-source statistics to classification errors. *Journal of official statistics 31*, 489–506.

Chataway, J., N. Davies, S. Farmer, R. Howard, E. Thompson, and K. Ward (2004, 06). Herpes simplex encephalitis: an audit of the use of laboratory diagnostic tests. *QJM: An International Journal of Medicine 97*(6), 325–330.

Derks, K., J. de Swart, E.-J. Wagenmakers, J. Wille, and R. Wetzels (2021). Jasp for audit: Bayesian tools for the auditing practice. *Journal of Open Source Software 6*(68), 2733.

Eurostat (2008). Nace rev. 2: Statistical classification of economic activities in the european community.

Hernández, B., A. Parnell, and S. R. Pennington (2014). Why have so few proteomic biomarkers "survived" validation?(sample size and independent validation considerations). *Proteomics 14*(13-14), 1587–1592.

Johnson, S. G. (2022). *The NLopt nonlinear-optimization package*. CRAN. R package version 2.0.3.

Langer, G. and J. A. Krosnick (2018). *The Importance of Probability-Based Sampling Methods for Drawing Valid Inferences*, pp. 7–12. Cham: Springer International Publishing.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Schwartz, J. E. (1985). The neglected problem of measurement error in categorical data. *Sociological Methods & Research 13*(4), 435–466.

The Financial Reporting Council (2016). *International Standard on Auditing (UK) 530 Audit Sampling*. The Financial Reporting Council.

van Delden, A., S. Scholtus, and J. Burger (2016). Accuracy of mixed-source statistics as affected by classification errors. *Journal of Official Statistics 32*(3), 619–642.