Noname manuscript No.
(will be inserted by the editor)

Bounds for monetary-unit sampling in auditing: an adjusted empirical likelihood approach

Yves G. Berger · Paola M. Chiodini · Mariangela Zenga

Received: date / Accepted: date

Abstract It is common practice for auditors to verify only a sample of recorded values to estimate the total error amount. Monetary-unit sampling is often used to over-sample large valued items which may be overstated. The aim is to compute an upper confidence bound for the total errors amount. Naïve bounds based on the central limit theorem are not suitable, because the distribution of errors are often very skewed. Auditors frequently use the Stringer bound which known to be too conservative. We propose to use weighted empirical likelihood bounds for Monetary-unit sampling. The approach proposed is different from mainstream empirical likelihood. A Monte-Carlo simulation study highlights the advantage of the proposed approach over the Stringer bound.

Keywords coverages · external audit · nominal level · Stringer bound · tolerable error amount · unequal probability sampling

Mathematics Subject Classification (2010) 62D05 · 62G15 · 62G20

1 Introduction

In practice, it is natural to audit only a sample of accounting records to establish the correctness of the entire financial reporting process. Audit techniques are divided into two main areas: the so-called "*internal audit*" which is carried

Yves G. Berger
Economic, Social and Political Sciences
University of Southampton, SO17 1BJ, UK
E-mail: y.g.berger@soton.ac.uk
url: http://yvesberger.co.uk
ORCID: 0000-0002-9128-5384

Paola M. Chiodini & Mariangela Zenga
Department of Statistics and Quantitative Methods
University Milano-Bicocca, Italy

out internally to monitor the accounting process, and "*external audit*" carried out by accounting experts who certify the correctness of the accounting recording process. We shall focus on the latter. In general, auditing aims to verify whether there are material errors in a set of N accounting records or items. The inferential problem facing the auditor is to decide, on the basis of sample information, whether the errors found on the accounting records are attributable only by random material errors or by fraudulent actions. Each item in the sample provides the auditor with two types of information: the recorded amount (or book amount) and the audited amount (or corrected amount). The difference between these two amounts is called the error which is used to estimate the overall unknown error amount.

Auditors want to verify if the total error falls below a pre-assigned "*tolerable error amount*" denoted $\mathcal{A}$ hereafter. This can be achieved by calculating an upper confidence bound for the total error. If this bound is lower than $\mathcal{A}$, the auditor concludes that no misstatement has been made. On the contrary, if this bound is larger than $\mathcal{A}$, then the auditor may decide to verify all the recorded amounts. Alternatively, a p-value calculated at $\mathcal{A}$ can be used instead.

The primary focus is on the upper bound of the confidence interval rather than point estimation. The fraction of incorrect items present in a sample can also be very variable leading to unreliable estimates and confidence bounds. In fact, two scenarios are possible. We may have a relatively high number of small errors which results in a high overall error rate. The second scenario is when we have a small amount of larger errors and a small overall error rate. Thus, the distribution of errors may be very skewed with many null errors. As a consequence, the upper limits of the confidence intervals based upon variance estimates and the central limit theorem are no longer adequate (e.g. Cox and Snell, 1979). The actual coverage of these intervals is frequently lower than the chosen nominal level (Kaplan, 1973; Neter and Loebbecke, 1975; Beck, 1980). In practice, auditors tend to use unconventional confidence interval limits (e.g. Horgan, 1996), such as Stringer's (1963) bounds. This approach, however, tends to give conservative limits with coverages larger than the nominal level.

Audit sample are often selected with "*probability proportional to size*" sampling without replacement, also called "*monetary-unit sampling*" (MUS) (e.g. Arens and Loebbecke, 1981; Higgins and Nandram, 2009). Large valued items containing the greatest potential of large overstatement, have more chance of being sampled.

Chen et al. (2003) proposed an empirical likelihood bound for population containing many zero values. This approach is limited to simple random sampling, and cannot be directly used with MUS. A analogous parametric likelihood-based approach based on mixture models was proposed by Kvanli et al. (1998). However, non-parametric are preferable because it avoids making assumption about the distribution of the errors. We propose to use Berger and Torres's (2016) non-parametric weighted empirical likelihood approach, which takes into account of the unequal selection probabilities inherent with MUS. Empirical likelihood providing confidence bounds driven by the distribution of the data (Owen, 2001); that is, it tends to give large upper bounds with

skewed data. This makes it particularly suitable for MUS. Bootstrap is another well-known non-parametric approach for confidence bounds. However, it may perform poorly with data containing many zero errors. In this paper, we compare numerically the empirical likelihood bound proposed by Berger and Torres's (2016) with the Stringer's (1963) bound.

This paper is organized as follows. §2 describes MUS and the point estimator of the total overstatement error. In §3, we describe the empirical bound proposed, the Stringer's (1963) bound and other alternative bounds. The results of the simulation study are presented in §4.

2 Statistical sampling method in auditing

An accounting population consists of N line items with recorded (or book) values, $\{z_i : i = 1, \ldots, N\}$, where $z_i > 0$. The audited (correct) amount of the N line items in the population is denoted by $\{x_i : i = 1, \ldots, N\}$. The values x_i are unknown before sampling, whereas z_i are known.

The error in item i, is $y_i := z_i - x_i$. When $y_i > 0$, the *i-th* item is overstated and when $y_i < 0$, it is understated. We have 100% overstatement if $y_i = z_i$. When $y_i = 0$, the account is error free. A large fraction of the items in the population are error free while the non-zero errors are usually highly skewed to the right (Johnson et al., 1981; Neter et al., 1985). The total error amount is defined as

$$Y_{_N} := \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} t_i \, z_i, \tag{1}$$

where

$$t_i := \frac{y_i}{z_i}$$

is called the fractional error or "*taint*" that is the fraction of error within z_i.

The purpose is to estimate, on sample basis, the total error amount $Y_{_N}$. More precisely, the auditors is mostly interested in obtaining an upper bound of a confidence interval derived from an estimate of $Y_{_N}$. If the upper bound exceeds a "*tolerable error amount*" $\mathcal{A}$, we conclude that there are significant material errors in the book values, or on the contrary there are only minor errors.

Generally, the audit processes consists in selecting samples with "*monetary-unit sampling*" (MUS) also called "*dollar unit sampling*" (Arens and Loebbecke, 1981). According to this approach, an accounting balance can be considered as a group of monetary units that can be either correct or incorrect. If the selected monetary-unit falls within the i-th item then a taint is observed. In practice a systematic random sample S of size n with unequal probabilities proportional to z_i is often selected (e.g. Madow, 1949; Tillé, 2006, §7.2). However, the approach proposed is not limited to systematic sampling.

Under MUS, an audit amounts x_i is selected with probability $\pi_i := nz_i Z_N^{-1}$; where

$$Z_N := \sum_{i=1}^{N} z_i$$

denotes the known total book amount. We usually have $nz_i Z_N^{-1} < 1$. However, with small population or right-skewed z_i, we may have $nz_i Z_N^{-1} > 1$ for some units. In this case, we need to adjust the π_i with the usual scaling method that can be found in Tillé (2006, §2.10); that is, $\pi_i = 1$ if $nz_i Z_N^{-1} > 1$ and the remaining π_i are adjusted so that $\sum_{i=1}^{N} \pi_i = n$. The Horvitz and Thompson (1952) estimator of Y_N is given by

$$\widehat{Y}_n := \sum_{i \in S} w_i \, y_i, \tag{2}$$

with $w_i := \pi_i^{-1}$. If $nz_i Z_N^{-1} \leqslant 1$ for all i, scaling is not needed and (2) reduces to the mean per-unit $\widehat{Y}_n = Z_N \, \overline{t}$, where $\overline{t} := n^{-1} \sum_{i \in S} t_i$ is the sample mean of the taints.

3 Confidence bound for the total error amount

The interest of the auditors usually focuses on obtaining an upper confidence bound for Y_N, at a specified confidence level $1 - \alpha \in [0.5, 1)$, e.g. $\alpha = 0.05$ or 0.01. If this upper bound exceeds a tolerable error amount $\mathcal{A}$, then there is statistical evidence of a possible material error. When this bound is less than $\mathcal{A}$, we conclude that the recorded values are a fair reflection of the accounts.

It is important to compute confidence intervals whose limits are reliable. The presence of low error rates means that y_i usually have a strongly positive asymmetric distribution, because small y_i are much more frequent than large y_i. As a result, the upper limits of the naïve confidence intervals based on variance estimation and the central limit theorem (see (13) below) can be problematic as their coverage is generally below the confidence level $1 - \alpha$ (Kaplan, 1973; Neter and Loebbecke, 1975; Beck, 1980), because the sampling distribution is usually not normal (Stringer, 1963; Kaplan, 1973; Neter and Loebbecke, 1975, 1977). In addition, a negative correlation between $\widehat{Y}_n$ and standard error estimates can increase the probability of type II error and reduces the probability of type I error (Kaplan, 1973). The lack of normality is the main reason for not using classical statistical inference, based on the central limit theorem (Ramage et al., 1979; Johnson et al., 1981; Neter et al., 1985; Ham et al., 1985).

Non-traditional heuristic estimation methods have been developed to overcome the above problems (e.g. Horgan, 1996). These methods are known as "*Combined Attribute and Variable*" (CAV) (Goodfellow et al., 1974a,b) some of which will be described in §§ 3.4 and 3.3. The Stringer's (1963) bound, described in §3.3, is widely used by auditors. Swinamer et al.'s (2004) simulation

study show that the upper bound is too conservative, with confidence level frequently greater than $1-\alpha$.

3.1 Weighted empirical likelihood's bounds proposed for MUS

Berger and Torres (2016) developed an empirical likelihood approach for unequal probability sampling. We show how this method can be used to derive a confidence bound for the total error Y_N.

The "*maximum empirical likelihood estimator*" is defined by

$$\widehat{Y}_{EL} = \arg\max \ell(Y),$$

where $\ell(Y)$ is the following "*weighted empirical likelihood function*".

$$\ell(Y) := \max_{p_i : i \in S} \left\{ \sum_{i \in S} \log(np_i) : p_i > 0, \sum_{i \in S} p_i = 1, \sum_{i \in S} p_i w_i \Big(y_i - \frac{Y \pi_i}{n} \Big) = 0 \right\}, \quad (3)$$

where $w_i := \pi_i^{-1}$ are weights. The function (3) is different from Owen's (1988) and Chen et al.'s (2003) empirical likelihood functions, because the constraint within (3) contains the adjustments w_i which take into account of the fact that the y_i are selected with unequal probabilities, under MUS. The function (3) can also be adjusted to accommodate stratification (see Berger and Torres, 2016, for more details).

Using Lagrangian multipliers, we have that the set of p_i that maximises $\sum_{i \in S} \log(np_i)$ for a given Y is given by

$$p_i(Y) := \frac{1}{n} \Big\{ 1 + w_i \boldsymbol{c}_i(Y)^\top \boldsymbol{\eta} \Big\}^{-1},$$

where $\boldsymbol{c}_i(Y)$ is the 2×1 vector function

$$\boldsymbol{c}_i(Y) := \Big\{ \pi_i, \Big(y_i - \frac{Y \pi_i}{n} \Big) \Big\}^\top \quad (4)$$

and $\boldsymbol{\eta}$ is the Lagrangian vector which is such that the constraint

$$n \sum_{i \in S} p_i(Y)\, w_i\, \boldsymbol{c}_i(Y) = (n, 0)^\top \quad (5)$$

holds. Thus, (3) reduces to

$$\ell(Y) = \sum_{i \in S} \log\big\{ np_i(Y) \big\} = - \sum_{i \in S} \log\big\{ 1 + w_i \boldsymbol{c}_i(Y)^\top \boldsymbol{\eta} \big\}. \quad (6)$$

This function can be calculated numerically from the observed y_i ($i \in S$) and a given value Y.

In practice, the function (3) is not needed for point estimation, because it can be shown that $\widehat{Y}_{EL} = \widehat{Y}_n$ given by (2). This function is used to derive an

upper confidence bound. The $(1-\alpha)$ "*empirical likelihood confidence bound*" b_α is the largest root of

$$\chi^2_{1,1-2\alpha} - 2\ell(b_\alpha) = 0, \tag{7}$$

where $\chi^2_{1,1-2\alpha}$ denotes the upper $(1-2\alpha)$-th quantile of a χ^2-distribution with one degree of freedom. A root-finding algorithm, such that the Brent (1973) and Dekker's (1969) method, can be used to find b_α.

The quantity b_α is an upper confidence bound, because the convexity of $-2\ell(Y)$ implies that the equation $\chi^2_{1,1-2\alpha} - 2\ell(Y) = 0$ has two roots $b_{\scriptscriptstyle L}$ and b_α, such that $b_{\scriptscriptstyle L} < b_\alpha$. Berger and Torres (2016) showed that

$$-2\ell(Y_{\scriptscriptstyle N}) \xrightarrow{d} \chi^2_1, \tag{8}$$

where χ^2_1 denotes the χ^2-distribution with one degree of freedom. Hence $Pr\{-2\ell(Y_{\scriptscriptstyle N}) \leqslant \chi^2_{1,1-2\alpha}\} \to 1-2\alpha$ or $Pr\{b_{\scriptscriptstyle L} \leqslant Y_{\scriptscriptstyle N} \leqslant b_\alpha\} \to 1-2\alpha$, by using the convexity of $-2\ell(Y)$. Thus, $[b_{\scriptscriptstyle L}, b_\alpha]$ is a two-sided $(1-2\alpha)$. Hence, b_α is indeed an upper confidence bound.

The computation of b_α involves a root-finding algorithm. A simpler and less computationally intensive approach based on a "*p-value*" of a one-side test can be used to check if $b_\alpha \leqslant \mathcal{A}$ at a given level α, where $\mathcal{A}$ denotes the "*tolerable error amount*". A p-value less than α means that b_α is likely to be below $\mathcal{A}$. In other words, $b_\alpha \leqslant \mathcal{A}$ if p-value $\leqslant \alpha$, and $b_\alpha > \mathcal{A}$ otherwise. This p-value is given by

$$\text{p-value} := \frac{1}{2}\Big[1 + (-1)^{\delta\{\widehat{Y}_n \leqslant \mathcal{A}\}} F\{-2\ell(\mathcal{A})\}\Big] \tag{9}$$

is the p-value of a one-side test. Here, $F\{\cdot\}$ is the cumulative distribution of a χ^2-distribution with one degree of freedom. Here, $\delta\{\widehat{Y}_n \leqslant \mathcal{A}\} = 1$ if $\widehat{Y}_n \leqslant \mathcal{A}$ and $\delta\{\widehat{Y}_n \leqslant \mathcal{A}\} = 0$ otherwise. The value of $F\{-2\ell(\mathcal{A})\}$ can be found from the usual statistics table of χ^2-distributions. Note that $\mathcal{A} < \widehat{Y}_n$ implies p-value $\geqslant 0.5$, because $F\{-2\ell(\mathcal{A})\} \geqslant 0$. It can be shown that p-value $\leqslant \alpha$ implies $2\ell(\mathcal{A}) \geqslant \chi^2_{1,1-2\alpha}$ and $\widehat{Y}_n \leqslant \mathcal{A}$. The strict concavity of (6) implies that $\mathcal{A}$ is larger than the largest root b_α of (7). Hence $b_\alpha \leqslant \mathcal{A}$. The trivial case $\mathcal{A} < \widehat{Y}_n$ always implies $b_\alpha > \mathcal{A}$. In this case, we always have that p-value $\geqslant 0.5$.

Berger and Torres (2016) showed that (8) holds conditionally on $\{y_i, \pi_i : i = 1, \ldots, N\}$. Property (8) relies on regularity conditions, such as the existence of fourth moments of $\widehat{Y}_n$, $n/N \to 0$ and that the central limit theorem holds for $\widehat{Y}_n$. In fact, $\widehat{Y}_n$ may not be normaly distributed, because of the skewness of the distribution of y_i. It turn out that for moderate n, simulation studies have shown that the distribution of $-2\ell(Y_{\scriptscriptstyle N})$ is still well approximated by a χ^2-distribution, even with skewed y_i (Owen, 1988; Berger and Torres, 2016). Since empirical likelihood is a data driven approach, the bound b_α should capture the skewness of the y_i.

3.2 Extension for large sampling fraction or strong correlation between selection probabilities and the errors

The approaches described so far rely on $n/N \to 0$, because (8) hold under this assumption. Berger and Torres (2016) proposed an empirical likelihood for non-negligible n/N or when the π_i are strongly correlated with the y_i. We described briefly this approach. The technical details can be found in Berger and Torres (2016) and Berger (2018). This approach is based on a "*penalised empirical likelihood function*" defined by

$$\widetilde{\ell}(Y) := \max_{p_i: i\in S}\Big\{\sum_{i\in S}\log(np_i) - n\sum_{i\in S}p_i + n : p_i > 0,$$
$$\sum_{i\in S}\Big(np_iq_i - \frac{q_i-1}{n}\Big) = 1,\ n\sum_{i\in S}\Big(p_iq_i - \frac{q_i-1}{n}\Big)w_i\Big(y_i - \frac{Y\pi_i}{n}\Big) = 0\Big\},$$

where $p_i = n^{-1}$, if $q_i = 0$. Here, $q_i = (1-\pi_i)^{1/2}$ are Hájek's (1964) finite population correction. Note that $n/N \to 0$ implies $\pi_i \to 0$ and $q_i \to 1$. If we replace q_i by 1, we have that (10) reduces to (3). Berger and Torres (2016) showed that $-2\widetilde{\ell}(Y_N) \xrightarrow{d} \chi^2_1$ for non-negligible n/N. Thus, the $(1-\alpha)$ "*penalised empirical likelihood confidence bound*" $\widetilde{b}_\alpha$ is the largest quantity which is the solution to

$$\chi^2_{1,1-2\alpha} - 2\widetilde{\ell}(\widetilde{b}_\alpha) = 0\cdot \tag{10}$$

The function $\widetilde{\ell}(Y)$ can be calculated by using the Lagrangian method as in (6). We expect $\widetilde{b}_\alpha$ to be smaller than b_α with non-negligible n/N. The p-value of the tolerable amount $\mathcal{A}$ is p-value $:= 0.5[1 + (-1)^{\delta\{\widehat{Y}_n \leqslant \mathcal{A}\}}F\{-2\widetilde{\ell}(\mathcal{A})\}]$.

Our simulation study in §4 also show that b_α given by (7) can be too conservative, when y_i is strongly correlated with π_i. This could be the case when the errors are mainly within the tail of z_i or when the π_i are strongly correlated with the y_i. In these situations, $\widetilde{b}_\alpha$ is less conservative and have better coverages, even when n/N is negligible. The bound $\widetilde{b}_\alpha$ should be lower than b_α, when the number of units with $\pi_i = 1$ is large, because the variance of the sampling distribution is smaller (see Berger and Torres, 2016, for more details).

With accounting populations with a very low error rates, we may have a "*zero-error sample*"; that is $y_i = 0$ for all the sampled items. In this case, $\widehat{Y}_n = 0$ and the auditor evaluates the book amount as free of error. In this case, it is not be possible to obtain empirical likelihood bounds, because the functions $\ell(Y)$ and $\widetilde{\ell}(Y)$ cannot be computed when $y_i = 0$ for all $i \in S$. The bound $\widetilde{b}_\alpha$ cannot be computed when $q_iy_i = 0$ for all $i \in S$; that is, when $y_i = 0$ for all $i \in S$ such that $\pi_i < 1$. In this case, it may not be possible to find p_i that satisfies the constraint within $\widetilde{\ell}(Y)$, for a given Y. The more conservative bound b_α can still be computed in this situation, as long as $y_i \neq 0$ for some units with $\pi_i = 1$.

3.3 The Stringer bound

Suppose that we are interested in cases of overstatement, i.e. $x_i = z_i$ if $x_i \leqslant z_i$. Let us also assume that the value of each overstatement does not exceed the declared value such that $0 \leqslant t_i \leqslant 1$. Let $T_1, \ldots, T_n$ be independent random variables that describe the taints, such that $Pr(0 \leqslant T_i \leqslant 1) = 1$. Here, t_i is an observation of T_i. Let $0 \leqslant t_{(1)} \leqslant t_{(2)} \leqslant \ldots \leqslant t_{(n)} \leqslant 1$ be the ordered statistics of $\{T_1, \ldots, T_n\}$. Let u_i be the $(1-\alpha)$ upper confidence limit for the binomial parameter when i errors are observed in a sample of size n. The quantity u_i is the unique solution to

$$\sum_{k=0}^{i} \binom{n}{k} u_i^k (1-u_i)^{n-k} = \alpha, \qquad \text{for } i = 0, 1, ..., n-1; \tag{11}$$

with $u_n = 1$. The u_i can sometimes be calculated using the Poisson approximation instead of a binomal within (11). An upper bound for the total overstatement error can be obtained by combining the upper limits of the sample errors with the observed taints. The Stringer bound, at the significance level α is defined as (e.g. Pap and van Zuijlen, 1995).

$$b_\alpha^{(S)} := Z_N \Big\{ u_0 + \sum_{i=1}^{n} (u_i - u_{i-1}) t_{(n-i+1)} \Big\}.$$

The bound relies on $0 \leqslant t_i \leqslant 1$, which is not necessary for the empirical limit b_α and $\widetilde{b}_\alpha$. When some taints t_i are negative, we can use the "*Stringer offset bound*" (e.g. Clayton and McMullen, 2007) given by

$$b_\alpha^{(SO)} := Z_N \Big\{ u_0 + \sum_{i=1}^{n} (u_i - u_{i-1}) \max(0, t_{(n-i+1)}) + \frac{1}{n} \sum_{i=1}^{n} \min(t_i, 0) \Big\}.$$

We have that $b_\alpha^{(SO)} = b_\alpha^{(S)}$, when $0 \leqslant t_i \leqslant 1$.

The Stringer bound has been extensively studied in literature and many empirical studies confirm that the coverage level is at least equal to its nominal level. However, this bound is very conservative (Leitch et al., 1982; Reneau, 1978; Anderson and Teitlebaum, 1973; Wurst et al., 1989; Higgins and Nandram, 2009) and is usually much larger than the total error (1). This is also confirmed by the simulation study in §4. The direct consequence is that auditors may reject an acceptable accounting populations (Leitch et al., 1982). Bickel (1992) studied the asymptotic behaviour of the Stringer bound and showed that in case of large samples the confidence level is frequently higher than its nominal level. Pap and van Zuijlen (1996) showed that the Stringer bound is asymptotically conservative. In §4, we how that The Stringer bound is more conservative than the empirical likelihood bound. Indeed, b_α and $\widetilde{b}_\alpha$ are usually smaller than $b_\alpha^{(S)}$ and have confidence levels close to α. The "*Stringer offset bound*" $b_\alpha^{(SO)}$ can be less conservative, when some t_i are negative.

It is not possible to compute the empirical likelihood bounds (b_α or $\widetilde{b}_\alpha$) with zero-error sample. However, the Stringer approach has the advantage of providing a bound in this situation. Indeed, when $y_i = 0$ for all the sampled items, $t_i = 0$, $\widehat{Y}_n = 0$ and $b_\alpha^{(S)} = Z_N u_0 = Z_N(1-\alpha^{1/n})$. Since usually, $\widehat{Y}_n = Z_N\,\overline{t}$, we can view $(1-\alpha^{1/n})$ as an upper bound for the average taints. This upper bound decreases with n, reflecting the fact that with a large zero-error sample, the average taints has more change of being small.

3.4 Other confidence bounds

Fienberg et al. (1977) introduced a less conservative bound based on a multinomial distribution derived from MUS. The method is rather complex because it is necessary to maximize over a joint confidence region. Leslie et al. (1979) proposed a "*cell bound*" which can be much greater than the actual error amount when we have a low error rate (Plante et al., 1985). Dworin and Grimlund (1984, 1986) introduced the so-called "*moment bound*" which is obtained by approximating the sampling distribution with a three-parameters gamma distribution. The method of moments is used to estimate these parameters. Simulation studies shows that the moment bound gives coverage close to the nominal level, and is less conservative than the Stringer bound.

Fishman (1991) showed that Hoeffding's inequality can be used to derive a confidence bound, which can be more conservative than the Stringer bound. Howard (1994) proposed a bound based on bootstrap and Hoeffding's inequality. This bound is not uniformly better than the Stringer (1963) bound, when the accounts are characterized by low error rates.

When the non-zero accounts values can be described by a suitable parametric model, Kvanli et al. (1998) showed that it is possible to use a parametric likelihood ratio statistics to define a two-sided confidence interval for the mean error. The nominal value is achieved when this parametric model holds. However, the bound depends entirely on the parametric model. Assuming a model that does not follow the distribution of the account may affect the coverage. The method introduced in paragraph §3.1 is very similar, but it has the advantage of being non-parametric, because it is not necessary assume a model for the errors.

4 Simulation studies

In this §, we compare the numerical performance of the empirical likelihood bounds proposed with the Stringer bound. The recorded values z_i are simulated from a skewed log-normal distribution,

$$\log(Nz_i) \sim \mathcal{N}(1, \sigma^2 = 1.44), \qquad i = 1, \ldots, N\cdot \tag{12}$$

We use this distribution, because z_i are monetary values which usually follow a right-skewed distribution. Furthermore, the main reason for using MUS is

the skewness of z_i. The resulting π_i, proportional to z_i, are right-skewed, with some $\pi_i = 1$, depending on the values of N and n. The distribution of the taints t_i is crucial, because it drives the sampling distribution of $\widehat{Y}_n$ and the upper bounds. Indeed, when $\pi_i < 1$, we have that $\widehat{Y}_n = Z_N\, \overline{t}$ and the sample mean $\overline{t}$ of the taints drives the sampling distribution of $\widehat{Y}_n$. We shall consider uniform, and skewed distributions, with positive and negative t_i, and large fractions of $t_i = 0$ and 1. In the different simulation setup considered, we shall vary n as well as the distribution of t_i. The values z_i generated are fixed and the same, for a given N. This isolates the effects of the distribution of the taints t_i.

Consider $N = 10\,000,\ 1000$ and 700. The error in item i, is $y_i = t_i\, z_i$, with $(100 - r)\%$ of t_i are equal zero and the remaining $r\%$ positive taints are generated randomly from uniform distributions $\mathrm{Un}(t_L, t_U)$. Here, r denotes the error rate. We shall consider $r = 2\%,\ 5\%$ and 10%. Several ranges of positive taints are considered: $[t_L, t_U] = [0.1, 0.3]$, $[0.2, 0.7]$ and $[0.5, 0.7]$. The values generated $y_i,\ t_i,\ z_i$ are treated as fixed. The MUS sample is based on a systematic procedure with random ordering of line items, selected with probability proportional to z_i. The sample sizes considered are $n = 100,\ 200$ and 500. We consider 1000 replications. Consider a nominal coverage of $1-\alpha = 0.95$. The results are given in Table 1 and 2.

We shall compare $b_\alpha^{(S)}$ with b_α, $\widetilde{b}_\alpha$, $b_\alpha^{(S)}$ and $b_\alpha^{(\mathcal{N})}$, where $b_\alpha^{(\mathcal{N})}$ is the following naïve bound based on the normal approximation.

$$b_\alpha^{(\mathcal{N})} := \widehat{Y}_n + \Phi^{-1}(1-\alpha)\ \widehat{v}(\widehat{Y}_n)^{\frac{1}{2}}, \tag{13}$$

where $\Phi(\cdot)$ is the cumulative function of a standardised normal distribution and $\Phi^{-1}(1-\alpha)$ is its $1-\alpha$ quantile. Here, $\widehat{v}(\widehat{Y}_n)$ is Hartley and Rao's (1962) consistent variance estimator for systematic sampling. It is well known that $b_\alpha^{(\mathcal{N})}$ tends to be too small. Here, $b_\alpha^{(\mathcal{N})}$ is used as a benchmark.

Several indicators are computed to assess the accuracy of the bounds. The coverage probability of a specific bound is the proportion of replications for which a bound is greater than or equal to the true population error amount. A bound is considered unreliable if its coverage is significantly different from $1 - \alpha = 0.95$. The observed mean of a bound b is denoted by $\mathrm{Mean}(b)$, with $b = b_\alpha^{(\mathcal{N})}$, b_α, $\widetilde{b}_\alpha$ or $b_\alpha^{(S)}$. In the tables, we report the value of $\mathrm{Mean}(b)/Y_{0.95}$, where $Y_{0.95}$ denotes the 95% quantile of the observed distribution of $\widehat{Y}_n$. The quantities $\mathrm{Mean}(b)/Y_{0.95}$ gives unit free values which are usually close to 1. The uncertainty of the bound is measured by the observed standard deviation ($s.d.$) of the bounds. In the tables, we have the relative efficiencies $s.d.(b_\alpha)/s.d.(b_\alpha^{(S)})$ and $s.d.(\widetilde{b}_\alpha)/s.d.(b_\alpha^{(S)})$. A relative efficiency larger (smaller) than 1 indicates that b_α is less (more) stable than $b_\alpha^{(S)}$. We also compute the decile ranges of $b_\alpha/b_\alpha^{(S)}$ and $\widetilde{b}_\alpha/b_\alpha^{(S)}$ which assesses the variation of b_α and $\widetilde{b}_\alpha$ with respect to $b_\alpha^{(S)}$. It will reveal that the empirical likelihood bound are often lower and approximately proportional $b_\alpha^{(S)}$.

Table 1 Positive taints $\sim U(t_L, t_U)$. $N = 10\,000$ and n is the sample size. Coverages (%). Nominal level $= 1 - \alpha = 95\%$. Mean(b) denotes the observed mean, with $b = b_\alpha^{(\mathcal{N})}$, b_α or $b_\alpha^{(S)}$. The quantity $Y_{0.95}$ is the 95% quantile of the observed distribution of $\widehat{Y}_n$. The function $s.d.(\cdot)$ gives the observed standard deviation. r is the error rate. 1000 replicates.

			Coverages (%)			Mean(b)/$Y_{0.95}$			$\frac{s.d.(b_\alpha)}{s.d.(b_\alpha^{(S)})}$	Decile range of $b_\alpha/b_\alpha^{(S)}$
$[t_L\,;\,t_U]$	r	n	$b_\alpha^{(\mathcal{N})}$	b_α	$b_\alpha^{(S)}$	$b_\alpha^{(\mathcal{N})}$	b_α	$b_\alpha^{(S)}$		
[0.1 ; 0.3]	2%	100	89.7†	98.3†	100†	0.99	1.20	3.89	1.13	[0.18 ; 0.41]
		200	89.0†	95.3	100†	0.96	1.11	2.77	1.05	[0.29 ; 0.49]
		500	91.9†	96.6†	100†	0.99	1.10	1.93	1.00	[0.49 ; 0.63]
	5%	100	86.4†	93.2†	100†	0.96	1.08	2.58	1.05	[0.30 ; 0.51]
		200	90.1†	94.2	100†	0.96	1.03	1.88	1.01	[0.45 ; 0.62]
		500	92.9†	96.3	100†	0.98	1.03	1.45	0.99	[0.67 ; 0.75]
	10%	100	90.4†	93.3†	100†	0.99	1.05	1.91	1.01	[0.46 ; 0.62]
		200	92.5†	95.8	100†	0.99	1.04	1.53	1.00	[0.62 ; 0.72]
		500	93.4†	96.6†	100†	1.00	1.03	1.26	0.99	[0.79 ; 0.83]
[0.2 ; 0.7]	2%	100	88.6†	97.4†	100†	0.97	1.18	2.14	1.17	[0.33 ; 0.68]
		200	86.3†	93.6†	100†	0.96	1.11	1.71	1.06	[0.50 ; 0.74]
		500	89.8†	95.6	99.8†	0.98	1.08	1.38	1.00	[0.71 ; 0.82]
	5%	100	88.1†	94.2	100†	0.96	1.08	1.56	1.05	[0.56 ; 0.77]
		200	90.0†	94.0	100†	0.97	1.05	1.33	1.01	[0.71 ; 0.83]
		500	91.6†	95.0	99.7†	0.98	1.04	1.19	0.99	[0.85 ; 0.89]
	10%	100	88.3†	93.1†	99.9†	0.96	1.02	1.32	1.01	[0.69 ; 0.82]
		200	93.7	96.1	100†	1.00	1.04	1.21	1.00	[0.82 ; 0.88]
		500	93.4†	96.6†	99.8†	1.00	1.03	1.12	0.99	[0.90 ; 0.92]
[0.5 ; 0.7]	2%	100	100†	100†	100†	1.12	1.35	2.02	1.05	[0.54 ; 0.75]
		200	87.7†	93.3†	100†	0.98	1.12	1.51	1.02	[0.63 ; 0.81]
		500	91.9†	96.6†	99.7†	0.99	1.08	1.27	0.99	[0.81 ; 0.88]
	5%	100	87.6†	95.8	100†	0.99	1.10	1.44	1.02	[0.65 ; 0.82]
		200	91.9†	95.4	99.4†	1.00	1.07	1.27	1.00	[0.79 ; 0.88]
		500	92.9†	96.5†	99.3†	0.98	1.03	1.13	0.99	[0.90 ; 0.92]
	10%	100	91.0†	93.8	99.3†	0.99	1.04	1.23	1.01	[0.79 ; 0.88]
		200	92.7†	95.6	99.0†	1.00	1.04	1.15	1.00	[0.88 ; 0.92]
		500	93.7	96.5†	98.8†	0.99	1.02	1.08	0.99	[0.94 ; 0.95]

† significantly different from 95% (p-value $\leqslant 0.05$)

Table 2 $t_i \sim U(t_L, t_U)$. $N = 700$ or 1000. Sample size $n = 200$. Coverages (%). Nominal level $= 1 - \alpha = 95\%$. Mean(b) denotes the observed mean, with $b = b_\alpha$, $\widetilde{b}_\alpha$ or $b_\alpha^{(S)}$. The quantity $Y_{0.95}$ is the 95% quantile of the observed distribution of $\widehat{Y}_n$. The function $s.d.(\cdot)$ gives the observed standard deviation. r is the error rate. 1000 replicates.

			Coverages (%)				Mean(b)/$Y_{0.95}$			$\frac{s.d.(\widetilde{b}_\alpha)}{s.d.(b_\alpha^{(S)})}$	Decile range of $\widetilde{b}_\alpha/b_\alpha^{(S)}$
N	$[t_L\,;\,t_U]$	r	$b_\alpha^{(\mathcal{N})}$	b_α	$\widetilde{b}_\alpha$	$b_\alpha^{(S)}$	b_α	$\widetilde{b}_\alpha$	$b_\alpha^{(S)}$		
700	[0.1 ; 0.3]	2%	88.7†	100†	96.1	100†	1.00	1.06	2.45	0.85	[0.38 ; 0.47]
		5%	92.4†	100†	95.2	100†	1.00	1.03	1.56	0.86	[0.63 ; 0.68]
		10%	92.9†	100†	95.3	100†	1.00	1.02	1.41	0.85	[0.71 ; 0.74]
	[0.2 ; 0.7]	2%	87.9†	100†	91.4†	100†	0.98	1.04	1.33	0.99	[0.74 ; 0.82]
		5%	92.0†	100†	93.7	98.5†	1.00	1.02	1.14	0.87	[0.88 ; 0.91]
		10%	92.4†	100†	94.6	99.8†	1.00	1.02	1.23	0.85	[0.82 ; 0.84]
	[0.5 ; 0.7]	2%	87.4†	100†	94.0	98.5†	0.99	1.04	1.16	0.90	[0.87 ; 0.92]
		5%	90.8†	100†	93.0†	94.4	0.99	1.02	1.06	0.87	[0.94 ; 0.98]
		10%	93.4†	100†	94.6	97.5†	1.00	1.02	1.09	0.84	[0.93 ; 0.96]
1000	[0.1 ; 0.3]	2%	87.6†	100†	92.7†	100†	0.99	1.07	2.28	0.99	[0.39 ; 0.52]
		5%	89.0†	100†	94.3	100†	0.97	1.02	1.88	0.93	[0.48 ; 0.59]
		10%	91.9†	100†	94.0	100†	0.99	1.01	1.45	0.94	[0.67 ; 0.72]
	[0.2 ; 0.7]	2%	85.6†	100†	90.7†	100†	0.98	1.05	1.36	1.05	[0.69 ; 0.82]
		5%	89.9†	100†	94.7	99.9†	0.99	1.04	1.27	0.94	[0.79 ; 0.84]
		10%	90.8†	100†	93.3†	99.3†	1.00	1.02	1.15	0.93	[0.87 ; 0.89]
	[0.5 ; 0.7]	2%	87.8†	100†	93.6†	99.6†	0.96	1.03	1.26	0.97	[0.77 ; 0.85]
		5%	89.9†	100†	95.0	99.3†	1.00	1.05	1.20	0.93	[0.85 ; 0.88]
		10%	93.7	100†	96.1	99.5†	1.00	1.02	1.13	0.93	[0.90 ; 0.91]

† significantly different from 95% (p-value $\leqslant 0.05$)

In Table 1, the sampling fraction n/N is small. The coverage of $b_\alpha^{(\mathcal{N})}$ is significantly smaller than 95% and $b_\alpha^{(S)}$ gives large coverages. On average b_α is slightly larger than $b_\alpha^{(\mathcal{N})}$ and smaller than $b_\alpha^{(S)}$. The bounds b_α and $b_\alpha^{(S)}$ have similar standard deviations. With $n = 100$, the standard deviations b_α is slightly larger. Some coverages of b_α may be significantly different from 95%. With $n = 100$, and $r = 2\%$, about 13% of the samples contains only zero values for y_i. Those samples have been ignored when computing the coverages. With $n = 500$ and $r = 10\%$, we observe large coverages significantly different from 95%, because in this case $n/N = 0.05$ is small but not negligible enough, leading to a more conservative bound b_α. With $n/N = 0.05$, the bound $\widetilde{b}_\alpha$ is more suitable and should have better coverage.

In Table 2, we consider $n = 200$ with $N = 1000$ or 700; that is, n/N is not negligible. Usually, the bound $b_\alpha^{(\mathcal{N})}$ has a low coverage and $b_\alpha^{(S)}$ has a large coverage. The bound b_α is too conservative with 100% coverage, but b_α is usually smaller than $b_\alpha^{(S)}$, because $\text{Mean}(b_\alpha) \leqslant \text{Mean}(b_\alpha^{(S)})$. The coverages of $\widetilde{b}_\alpha$ are closer to the nominal value, because the effect of the Hájek's (1964) corrections q_i are more pronounced than with $N = 10\,000$, in Table 1. However, some coverages are still significantly different from 95%. The bound $\widetilde{b}_\alpha$ is smaller than b_α. The bound $\widetilde{b}_\alpha$ is mostly smaller than $b_\alpha^{(S)}$ because the upper deciles are less than 1. The bound $\widetilde{b}_\alpha$ is more stable than $b_\alpha^{(S)}$, because we observe a smaller s.d. for $\widetilde{b}_\alpha$. With $N = 1000$, the bound b_α is only slightly more stable than $b_\alpha^{(S)}$.

For the next series of simulation, we consider the situation when the errors are only in the right tail, which could be the case with fraudulent behaviour. Let Z_{1-r} denote the $1-r$ quantile of z_i generated from (12), where r denotes the error rate. We generate t_i randomly from uniform distributions $\text{Un}(t_L, t_U)$, when $z_i > Z_{1-r}$. If $z_i \leqslant Z_{1-r}$, we set $t_i = 0$. We consider $r = 2\%$, 5% and 10%. The ranges are $[t_L, t_U] = [0.1, 0.3]$, $[0.2, 0.7]$ and $[0.5, 0.7]$. Since the errors are in the right tail, we expect a strong correlation between π_i and y_i. The results are given in Table 3. The coverages of b_α are larger than with $\widetilde{b}_\alpha$. Usually, we observe coverages closer to 95% with $\widetilde{b}_\alpha$. The Stringer bound $b_\alpha^{(S)}$ has a very large coverages and is usually larger than $\widetilde{b}_\alpha$. The coverage of $b_\alpha^{(\mathcal{N})}$ is smaller than $\widetilde{b}_\alpha$, when $n = 100$. Most coverages of $b_\alpha^{(\mathcal{N})}$ are not significantly different from 95%, when $n > 100$. The bound b_α is more conservative than $\widetilde{b}_\alpha$, even when n/N is negligible, because of the following reasons. Some π_i can be large even when n/N is negligible; thus some q_i can be very different from 1 for the units that are more likely to be selected. Furthermore, the correlation between y_i and π_i makes $\widetilde{b}_\alpha$ less conservative, because the self-normalising property of $\widetilde{\ell}(Y)$ implies that $\widetilde{\ell}(Y)$ can be approximated by a quadratic form (Berger and Torres, 2016) involving a small variance because of q_i and the correlation between y_i and π_i (e.g. Rao, 1966). The bound $\widetilde{b}_\alpha$ seems to be the most appropriate.

Table 3 $t_i \sim U(t_L, t_U)$ if $z_i > Z_{1-r}$ and $t_i = 0$ otherwise. $N = 10\,000$ and n is the sample size. Coverages (%). Nominal level $= 1 - \alpha = 95\%$. Mean(b) denotes the observed mean, with $b = b_\alpha^{(\mathcal{N})}$, b_α, $\widetilde{b}_\alpha$ or $b_\alpha^{(S)}$. The quantity $Y_{0.95}$ is the 95% quantile of the observed distribution of $\widehat{Y}_n$. The function $s.d.(\cdot)$ gives the observed standard deviation. r is the error rate. 1000 replicates.

			Coverages (%)				Mean(b)/$Y_{0.95}$			$\frac{s.d.(\widetilde{b}_\alpha)}{s.d.(b_\alpha^{(S)})}$	Decile range of $\widetilde{b}_\alpha/b_\alpha^{(S)}$
$[t_L\,;\,t_U]$	r	n	$b_\alpha^{(\mathcal{N})}$	b_α	$\widetilde{b}_\alpha$	$b_\alpha^{(S)}$	b_α	$\widetilde{b}_\alpha$	$b_\alpha^{(S)}$		
[0.1 ; 0.3]	2%	100	92.7†	95.4	94.7	100†	1.01	1.04	1.57	1.00	[0.61 ; 0.70]
		200	94.7	96.8†	95.3	100†	1.01	1.02	1.33	0.99	[0.74 ; 0.79]
		500	94.2	98.9†	94.6	100†	1.00	1.01	1.16	0.98	[0.86 ; 0.87]
	5%	100	92.4†	94.2	93.9	100†	0.99	1.00	1.33	1.00	[0.72 ; 0.78]
		200	94.1	95.7	94.7	100†	1.00	1.01	1.20	1.00	[0.82 ; 0.85]
		500	93.4†	97.9†	93.4†	100†	1.00	1.00	1.09	0.99	[0.91 ; 0.92]
	10%	100	94.2	95.3	94.9	100†	1.00	1.01	1.25	1.00	[0.78 ; 0.82]
		200	94.5	96.4†	95.2	100†	1.00	1.00	1.14	1.00	[0.87 ; 0.89]
		500	93.9	96.8†	94.0	99.9†	1.00	1.00	1.06	0.99	[0.94 ; 0.94]
[0.2 ; 0.7]	2%	100	90.8†	94.1	93.2†	99.5†	0.99	1.01	1.21	1.00	[0.80 ; 0.86]
		200	94.0	97.3†	95.7	99.7†	1.00	1.01	1.15	0.99	[0.87 ; 0.90]
		500	93.2†	98.6†	93.5†	99.7†	1.00	1.00	1.07	0.98	[0.93 ; 0.94]
	5%	100	93.1†	95.2	94.4	99.6†	1.00	1.01	1.13	1.00	[0.88 ; 0.91]
		200	94.2	96.8†	95.2	99.2†	1.00	1.01	1.09	1.00	[0.92 ; 0.93]
		500	95.2	98.4†	95.3	99.5†	1.00	1.00	1.04	0.99	[0.96 ; 0.96]
	10%	100	94.5	95.9	95.1	99.6†	1.00	1.01	1.10	1.00	[0.90 ; 0.92]
		200	94.8	95.6	95.2	99.2†	1.00	1.00	1.06	1.00	[0.94 ; 0.95]
		500	95.4	97.4†	95.4	99.0†	1.00	1.00	1.03	0.99	[0.97 ; 0.97]
[0.5 ; 0.7]	2%	100	93.3†	96.1	95.1	99.5†	0.98	1.00	1.12	1.00	[0.86 ; 0.91]
		200	94.0	96.1	94.7	98.3†	1.00	1.01	1.10	1.00	[0.91 ; 0.93]
		500	95.1	99.2†	95.4	99.3†	1.00	1.00	1.05	0.99	[0.95 ; 0.96]
	5%	100	95.1	96.8†	96.4†	99.2†	1.00	1.00	1.08	1.00	[0.92 ; 0.94]
		200	94.1	95.7	94.7	98.7†	1.00	1.00	1.05	1.00	[0.95 ; 0.96]
		500	95.1	97.4†	95.2	98.0†	1.01	1.01	1.03	1.00	[0.97 ; 0.97]
	10%	100	94.6	95.2	95.1	98.6†	1.00	1.00	1.06	1.00	[0.94 ; 0.95]
		200	95.3	96.1	95.6	98.5†	1.00	1.00	1.03	1.00	[0.96 ; 0.97]
		500	94.6	96.4†	94.2	97.9†	1.00	1.00	1.01	1.00	[0.98 ; 0.98]

† significantly different from 95% (p-value $\leqslant 0.05$)

Table 4 $(1-\gamma)r\%$ taints generated from a Beta$(2,5)$ distribution and $\gamma r\%$ taints are equal to 1. $N = 10\,000$ and n is the sample size. Coverages (%). Nominal level $= 1-\alpha = 95\%$. Mean(b) denotes the observed mean, with $b = b_\alpha^{(\mathcal{N})}$, b_α or $b_\alpha^{(S)}$. The quantity $Y_{0.95}$ is the 95% quantile of the observed distribution of $\widehat{Y}_n$. The function $s.d.(\cdot)$ gives the observed standard deviation. r is the error rate. 1000 replicates.

			Coverages (%)			Mean$(b)/Y_{0.95}$			$\frac{s.d.(b_\alpha)}{s.d.(b_\alpha^{(S)})}$	Decile range of $b_\alpha/b_\alpha^{(S)}$
r	n	γ	$b_\alpha^{(\mathcal{N})}$	b_α	$b_\alpha^{(S)}$	$b_\alpha^{(\mathcal{N})}$	b_α	$b_\alpha^{(S)}$		
2%	100	10%	77.9†	85.7†	100†	0.89	1.14	1.97	1.49	[0.18 ; 0.78]
		20%	81.6†	88.0†	100†	0.99	1.25	1.97	1.43	[0.27 ; 0.84]
		40%	81.2†	85.9†	100†	0.95	1.19	1.63	1.32	[0.28 ; 0.85]
	200	10%	81.1†	87.2†	100†	0.94	1.17	1.66	1.28	[0.39 ; 0.84]
		20%	81.5†	87.2†	100†	0.90	1.10	1.47	1.20	[0.45 ; 0.84]
		40%	86.3†	88.6†	100†	0.95	1.14	1.41	1.15	[0.53 ; 0.89]
	500	10%	88.5†	96.6†	99.7†	0.98	1.18	1.41	1.05	[0.77 ; 0.88]
		20%	88.3†	97.2†	99.9†	0.97	1.13	1.32	1.02	[0.79 ; 0.90]
		40%	88.2†	95.9	99.1†	0.99	1.13	1.27	1.02	[0.84 ; 0.92]
5%	100	10%	82.7†	89.4†	100†	0.92	1.08	1.56	1.23	[0.44 ; 0.84]
		20%	82.4†	86.6†	100†	0.89	1.04	1.37	1.19	[0.44 ; 0.87]
		40%	87.4†	89.9†	100†	0.95	1.10	1.33	1.14	[0.58 ; 0.90]
	200	10%	88.7†	92.3†	99.9†	0.98	1.11	1.39	1.09	[0.65 ; 0.87]
		20%	89.3†	95.3	99.8†	0.96	1.07	1.27	1.05	[0.77 ; 0.89]
		40%	89.5†	95.3	98.9†	0.98	1.09	1.22	1.04	[0.84 ; 0.93]
	500	10%	91.5†	96.8†	99.9†	0.99	1.09	1.24	1.00	[0.84 ; 0.90]
		20%	91.3†	96.0	99.4†	1.01	1.10	1.21	1.00	[0.88 ; 0.92]
		40%	93.1†	96.9†	99.1†	0.99	1.06	1.13	1.00	[0.91 ; 0.95]
10%	100	10%	86.5†	91.2†	99.9†	0.98	1.09	1.39	1.12	[0.62 ; 0.86]
		20%	85.4†	91.3†	99.3†	0.95	1.06	1.28	1.09	[0.70 ; 0.89]
		40%	88.1†	94.0	98.6†	0.95	1.04	1.17	1.04	[0.84 ; 0.93]
	200	10%	91.4†	94.9	99.4†	0.99	1.06	1.24	1.04	[0.81 ; 0.89]
		20%	90.1†	95.1	99.3†	0.98	1.05	1.18	1.02	[0.85 ; 0.92]
		40%	93.0†	95.8	98.8†	1.00	1.06	1.14	1.02	[0.90 ; 0.95]
	500	10%	90.8†	95.6	99.7†	0.97	1.02	1.12	1.00	[0.89 ; 0.93]
		20%	92.4†	95.9	99.0†	1.00	1.04	1.12	1.00	[0.92 ; 0.94]
		40%	93.5†	96.7†	98.6†	1.00	1.04	1.09	1.00	[0.95 ; 0.96]

† significantly different from 95% (p-value $\leqslant 0.05$)

Now, we consider the situation when we have 100% overstatement for some items; that is, we allow $t_i = 1$, for some i. We also consider a right-skewed beta-distribution for $0 < t_i < 1$. Consider $(100 - r)\%$ of t_i are equal zero and $(1 - \gamma)r\%$ taints generated randomly from a Beta$(2, 5)$ distribution. The remaining $\gamma r\%$ taints are equal to one. We consider $N = 10\,000$, with $n = 100,\ 200$ and 500. The error rates are $r = 2\%$, 5% and 10%. The fraction γ of taints equal to one among the $t_i > 0$ is $\gamma = 10\%$, 20% or 40%. Systematic sampling is used, with probability proportional to z_i. The results are given in Table 4. We also observe low coverages for $b_\alpha^{(\mathcal{N})}$ and a large coverage for $b_\alpha^{(S)}$.

By comparing Table 4 with Table 1, we see that we have a lower coverage for b_α with $n = 100$, when $r = 2\%$ or 5%. In these situations, the bound b_α has larger $s.d.$ than $b_\alpha^{(S)}$. We have $\text{Mean}(b_\alpha^{(S)})/Y_{0.95} < 2$. In Table 1, this ratio can be larger than 2 for $r = 2\%$. With $r = 10\%$, the coverage of b_α is the closest to 95%. The fraction γ of $t_i = 1$ does not seem to affect the precision and the coverage of b_α.

For the last series of simulation, we consider understatements; that is negative taints. We follow approximately Clayton and McMullen's (2007) simulation setup. Now, r denotes the fraction of $t_i > 0$, and ν represents the fraction of $t_i < 0$, which are given by $t_i = -a_i$, with a_i generated randomly from a Beta$(2, 5)$ distribution. The fraction of $t_i = 0$ is given by $(100 - r - \nu)\%$. We have $(1-\gamma)r\%$ taints between 0 and 1, following a Beta$(2, 5)$ distribution. The fraction of $t_i = 1$ is $\gamma r\%$. We consider $\gamma = 20\%$, $N = 10\,000$ and $n = 200$. The fraction of $t_i > 0$ is $r = 2\%$, 5% or 10%. The fraction ν of $t_i < 0$ is $\nu = 2\%$, 5% or 10%. Systematic sampling is used, with probability proportional to z_i. The results are given in Table 5.

For Tables 4 and 5, the positive taints are generated the same way with $\gamma = 20\%$. The differences observed between Tables 5 and 4 can be just due to the negative taints. We notice that the coverage of $b_\alpha^{(SO)}$ can be lower than 95% and decreases with ν, because the Stringer offset bound $b_\alpha^{(SO)}$ is used. The offset reduces the bound and is more pronounced with large ν. We observe large coverages for $b_\alpha^{(\mathcal{N})}$. The coverages of b_α are the closest to 95%, in all cases. We observed lower coverages in Table 4 for $n = 100$, because the distribution of the taints is more skewed than in Table 5. Note that we have smaller $s.d$ for b_α compared to $b_\alpha^{(SO)}$. The large values of $\text{Mean}(b)/Y_{0.95}$ observed in Table 5 are due to the fact that $Y_{0.95}$ can be close to zero.

In Table 2, the number of units with $\pi_i = 1$ is 46 with $N = 700$ and 33 with $N = 1000$. With $N = 10\,000$ and $n = 500$, we only have 5 units with $\pi_i = 1$ (Table 1, 3, 4 and 5). For $n = 100$, 200 and $N = 10\,000$, we have $\pi_i < 1$ for all i. These numbers are the same for different distribution of taints, because we use the same z_i generated by (12), for a given N. We expect $\widetilde{b}_\alpha$ to be noticeably lower than b_α, when the number of units with $\pi_i = 1$ is large. This is what we observe in Table 2.

Table 5 r denotes the % of $t_i > 0$, and ν represents the % of $t_i < 0$ generated from a Beta(2,) distribution. The % of $t_i = 0$ is given by $(100-r-\nu)\%$. We have $(1-\gamma)r\%$ taints between 0 and 1, following a Beta(2, 5) distribution. The fraction of $t_i = 1$ is $\gamma = 20\%$. $N = 10\,000$ and n is the sample size. Coverages (%). Nominal level $= 1-\alpha = 95\%$. Mean(b) denotes the observed mean, with $b = b_\alpha^{(\mathcal{N})}$, b_α or $b_\alpha^{(S)}$. The quantity $Y_{0.95}$ is the 95% quantile of the observed distribution of $\widehat{Y}_n$. The function $s.d.(\cdot)$ gives the observed standard deviation. r is the error rate. 1000 replicates. The Stringer offset bound $b_\alpha^{(SO)}$ is used

			Coverages (%)			Mean(b)/$Y_{0.95}$			$\frac{s.d.(b_\alpha)}{s.d.(b_\alpha^{(SO)})}$	Decile range of $b_\alpha/b_\alpha^{(SO)}$	Number of $\pi_i = 1$
r	n	ν	$b_\alpha^{(\mathcal{N})}$	b_α	$b_\alpha^{(SO)}$	$b_\alpha^{(\mathcal{N})}$	b_α	$b_\alpha^{(SO)}$			
2%	100	2%	98.6†	93.9	90.4†	1.18	0.78	1.18	0.77	[-0.15 ; 1.22]	0
		5%	98.8†	95.2	83.9†	0.05	5.75	10.02	0.69	[-0.06 ; 1.16]	0
		10%	98.1†	95.9	76.4†	0.96	1.20	1.79	0.72	[0.55 ; 0.73]	0
	200	2%	98.6†	95.0	85.9†	1.32	0.67	0.19	0.75	[0.01 ; 1.52]	0
		5%	98.0†	96.2	78.5†	1.06	1.44	2.36	0.73	[0.35 ; 0.77]	0
		10%	97.2†	95.6	72.1†	0.97	1.04	1.43	0.79	[0.68 ; 0.75]	0
	500	2%	98.1†	97.1†	86.0†	0.66	0.85	2.27	0.77	[-0.28 ; 1.50]	5
		5%	96.9†	96.3	76.3†	0.96	1.00	1.40	0.79	[0.65 ; 0.75]	5
		10%	96.1	96.3	68.4†	0.99	0.99	1.23	0.85	[0.78 ; 0.82]	5
5%	100	2%	98.8†	92.3†	94.4	1.02	1.00	1.23	0.90	[0.37 ; 1.05]	0
		5%	98.2†	95.7	84.7†	1.11	0.29	−0.57	0.74	[-0.05 ; 1.43]	0
		10%	97.6†	95.5	77.7†	0.90	1.14	1.79	0.75	[0.42 ; 0.73]	0
	200	2%	98.2†	94.8	92.2†	1.06	1.03	1.08	0.83	[0.70 ; 1.24]	0
		5%	97.4†	95.7	81.7†	0.73	1.30	2.86	0.77	[-0.23 ; 1.36]	0
		10%	96.9†	95.5	75.4†	0.96	1.03	1.45	0.80	[0.63 ; 0.75]	0
	500	2%	96.6†	95.6	91.5†	1.07	1.06	0.95	0.83	[0.92 ; 1.66]	5
		5%	96.5†	95.9	78.9†	0.99	1.03	1.57	0.82	[0.40 ; 0.73]	5
		10%	96.3	96.3	71.7†	0.99	0.99	1.23	0.86	[0.78 ; 0.82]	5
10%	100	2%	98.3†	94.8	94.7	1.01	1.02	1.13	0.87	[0.76 ; 1.07]	0
		5%	98.3†	95.6	86.3†	1.05	0.79	0.42	0.77	[-0.27 ; 1.98]	0
		10%	97.6†	95.7	80.9†	0.86	1.35	2.68	0.76	[0.04 ; 1.14]	0
	200	2%	97.9†	96.5†	92.8†	1.05	1.04	1.06	0.85	[0.89 ; 1.21]	0
		5%	97.1†	95.6	84.7†	1.16	0.75	−0.70	0.80	[-0.40 ; 2.13]	0
		10%	97.0†	95.4	76.8†	1.02	1.12	1.72	0.81	[0.40 ; 0.73]	0
	500	2%	97.2†	96.9†	93.0†	1.00	1.02	0.97	0.86	[0.97 ; 1.23]	5
		5%	96.3	96.3	83.4†	0.99	0.94	2.23	0.84	[-0.27 ; 1.63]	5
		10%	97.2†	97.6†	78.0†	0.99	0.98	1.28	0.87	[0.71 ; 0.79]	5

† significantly different from 95% (p-value $\leqslant 0.05$)

5 Conclusions

Our simulation study confirms that the naïve bound based on the central limit theorem can be too small, with an observed coverage significantly lower than the nominal level. On the other hand, the Stringer bound is too conservative, with a coverage close to 100%, unless we have understatements. The empirical bounds proposed have coverages closed to the nominal level and usually lies between the naïve bound and the Stringer bound. The penalised empirical likelihood bound described in §3.2 seems to be the most appropriate, because it takes into account of the sampling fraction and possible correlation between the error and the selection probabilities. For example, when the errors are mainly within the tail of the recorded values, better bounds are obtained with the penalised empirical likelihood approach, even with small n/N. We recommend using the penalised empirical bound, because it has better observed coverages and may be more stable than the stringer bound.

Both empirical likelihood bounds have the advantage of respecting the confidence level and of being less conservative than the Stringer bounds. However, they are more numerically intensive than the Stringer bound, because they rely on a Lagrangian parameter. The approach based on p-values is a simpler alternative to check if total error exceed a tolerable error amount. We need to compute $\ell(\mathcal{A})$ which requires solving (5) with a root-search method, to obtain the value of $\boldsymbol{\eta}$ for $Y = \mathcal{A}$. Once $\boldsymbol{\eta}$ is known, $\ell(\mathcal{A})$ can be computed from (6). The analogue to (5) for the penalised version of Section 3.2 can be easily derived. Computing the bound b_α (or $\widetilde{b}_\alpha$) is more numerically intensive than the approach based on p-values, because it involves $\boldsymbol{\eta}$ for different values of Y, in order to solve (7) (or (10)).

Like the Stringer bound, both empirical likelihood bounds may not be suitable with very small sample sizes. It cannot provide a bound for samples containing no errors. It can be unstable with sample containing an tiny amount of errors. Recent empirical likelihood approaches tackle the former (e.g. Chen et al., 2003; Chen et al., 2008; Jing et al., 2017) but the are not designed to handle unequal probability sampling, used in MUS. It would be useful to investigate how these extensions can be used under MUS.

Acknowledgements We wish to thank the European Union's Seventhes Programme for Research, Technological Development and Demonstration (Grant Agreement No 312691 - InGRID), for supporting the visits of Paola M. Chiodini and Mariangela Zenga to the University of Southampton.

Conflict of interest

The authors have no conflict of interest.

References

Anderson R, Teitlebaum A (1973) Dollar-unit sampling. Canadian Chartered Accountant pp 30–39

Arens A, Loebbecke J (1981) Applications of Statistical sampling to Auditing. Prentice-Hall, Prentice-Hall

Beck P (1980) A critical analysis of the regression estimator in audit sampling. Journal of accounting research 18:16–37

Berger YG (2018) Empirical likelihood approaches in survey sampling. The Survey Statistician 78:22–31

Berger YG, Torres ODLR (2016) An empirical likelihood approach for inference under complex sampling design. Journal of the Royal Statistical Society Series B, doi: 101111/rssb12115 78(2):319–341

Bickel PJ (1992) Inference and auditing: The stringer bound. International Statistical Review 60(2):197–209

Brent RP (1973) Algorithms for Minimization without Derivatives. Prentice-Hall ISBN 0-13-022335-2, New-Jersey

Chen J, Chen SR, Rao JNK (2003) Empirical likelihood confidence intervals for the mean of a population containing many zero values. The Canadian J of Stat 31(1):53–68

Chen J, Variyath AM, Abraham B (2008) Adjusted empirical likelihood and its properties. Journal of Computational and Graphical Statistics 17(2):426–443

Clayton H, McMullen P (2007) Combining approaches for evaluating auditing populations: A simulation study. European Journal of Operational Research 178(3):907 – 917, DOI https://doi.org/10.1016/j.ejor.2006.01.043

Cox DR, Snell EJ (1979) On sampling and the estimation of rare errors. Biometrika 66:125–132

Dekker TJ (1969) Finding a zero by means of successive linear interpolation. In: Dejon B, Henrici P (eds) Constructive Aspects of the Fundamental Theorem of Algebra, Handbook of Statistics, Wiley-Interscience, London, pp 37–489

Dworin L, Grimlund RA (1984) Dollar unit sampling for accounts receivables and inventory. The Accounting Review 59:218–241

Dworin L, Grimlund RA (1986) Dollar-unit sampling: A comparison of the quasi-bayesian and moment bounds. The Accounting Review 61(1):236–58

Fienberg S, Neter J, Leitch R (1977) Estimating the total overstatement error in accounting population. Journal of American Statistical Association 358(72):295–302

Fishman G (1991) Confidence intervals for the mean in the bounded case. Statistics and Probability Letters 12:223–227

Goodfellow J, Loebbecke J, Neter J (1974a) Some perspectives on CAV sampling plans. In: Part I, October, CA Magazine, pp 23–30

Goodfellow J, Loebbecke J, Neter J (1974b) Some perspectives on CAV sampling plans. In: Part II, November, CA Magazine, pp 46–53

Hájek J (1964) Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics 35(4):1491–1523

Ham J, Losell D, Smieliauskas W (1985) An empirical study of error characteristics in accounting populations. The Accounting Review 60(3):387–406

Hartley HO, Rao JNK (1962) Sampling with unequal probabilities without replacement. The Annals of Mathematical Statistics 33:350–374

Higgins HN, Nandram B (2009) Monetary unit sampling: Improving estimation of the total audit error. Advances in Accounting 25(2):174–182

Horgan J (1996) he moment bound with unrestricted random, cell and sieve sampling of monetary units. Journal of Accounting and Business Research 26(3):215–223

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Statist Assoc 47:663–685

Howard RC (1994) A combined bound for errors in auditing based on hoeffding's inequality and bootstrap. Journal of Business and Economic Statistics 12(2):437–448

Jing BY, Tsao M, Zhou W (2017) Transforming the empirical likelihood towards better accuracy. The Canadian Journal of Statistics 45(3):340–352

Johnson J, Leitch R, Neter J (1981) Characteristics of errors in accounts receivables and inventory audits. Accounting Review 58:270–293

Kaplan R (1973) Statistical sampling in auditing with auxiliary information estimators. Journal of Accounting Research 11(2):238–258

Kvanli AH, Shen YK, Deng LY (1998) Construction of confidence intervals for the mean of a population containing many zero values. Journal of Business and Economics Statistics 16:362–368

Leitch R, Neter J, Plante R, Sinha P (1982) Modified multinomial bounds for larger number of errors in audits. The Accounting Review 57(2):384–400

Leslie D, Teitlebaum A, Anderson R (1979) Dollar-Unit Sampling - A Practical Guide for Auditors. Pitman, London

Madow WG (1949) On the theory of systematic sampling, ii. The Annals of Mathematical Statistics pp 333–354

Neter J, Loebbecke J (1975) Behaviour of major statistical estimators in sampling accounting population-an empirical study. AICP, New York

Neter J, Loebbecke J (1977) On the behavior of statistical estimators when sampling accounting populations. Journal of the American Statistical Association 359(72):501–507

Neter J, Johnson J, Leitch R (1985) Characteristics of dollar-unit taints and error rates in accounts receivables and inventory. The Accounting Review 60:488–499

Owen AB (1988) Empirical likelihood ratio confidence intervals for a single functional. Biometrika, doi: 101093/biomet/752237 75(2):237–249

Owen AB (2001) Empirical Likelihood. Chapman & Hall, New York

Pap G, van Zuijlen M (1995) The stringer bound in cased of uniform taintings. Computer Math Applic 29(10):51–59

Pap G, van Zuijlen M (1996) On the asymptotic behaviour of the stringer bound. Statistica Neerlandica 50(3):367–389

Plante R, Neter J, Leitch RA (1985) Comparative performance of the multinomial, cell and stringer bound. Auditing: A Journal of Practice and Theory 5:40–56

Ramage J, Kreieger A, LL S (1979) An empirical study of error characteristics in audit populations (supplement). Journal of Accounting Research 17:72–102

Rao JNK (1966) Alternative estimators in pps sampling for multiple characteristics. Sankhyā A28:47–60

Reneau J (1978) CAV bounds in dollar unit sampling: Some simulation results. The Accounting Review 53(3):669–680

Stringer KW (1963) Practical aspects of statistical sampling in auditing. In: Proceedings of the Business and Economics Statistics Section, ASA, Dublin, pp 405–411

Swinamer K, Lesperance ML, Will H (2004) Optimal bounds used in dollar unit sampling: A comparison of reliability and efficiency. Communications in Statistics - Simulation and Computation 33(1):109–143

Tillé Y (2006) Sampling Algorithms. Springer Series in Statistics, Springer, New York

Wurst J, Neter J, Godfrey J (1989) Comparison of sieve sampling with random and cell sampling of monetary units. The Statistician 11(2):235–249