DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NIJMEGEN The Netherlands

CONFIDENCE BOUNDS FOR THE MEAN IN NONPARAMETRIC MULTISAMPLE PROBLEMS

V. Bentkus, M. Kalosha, M. van Zuijlen

Report No. 0406 (May 2004)

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NIJMEGEN
Toernooiveld
6525 ED Nijmegen
The Netherlands

Confidence bounds for the mean in nonparametric multisample problems

Vidmantas Bentkus
Institute of Mathematics and Informatics
Akademijos 4, 2600 Vilnius, Lithuania

Mikalai Kalosha and Martien C. A. van Zuijlen
Department of Mathematics, University of Nijmegen
Toernooiveld 1, 6525 ED Nijmegen, The Netherlands

Abstract

In auditing practice it often occurs that a statement regarding the accounting error in a population consisting of several subpopulations has to be made. Since the relative proportion of errors can differ dramatically across these subpopulations, it is desirable to take independent fixed size dollar-unit samples from each of them, as this often leads to lower variability compared to dollar-unit sampling from the whole population. It also occurs that the results of the separate investigations of, e.g., different branches of one company need to be combined to make a statement on the bookkeeping quality in general.

The problem of estimating the total accounting error is thus related to the problem of estimating linear combinations of the means corresponding to several families of identically distributed independent random variables.

In this article, we propose several confidence upper bounds for such linear combinations based on Hoeffding type inequalities and show how they can be applied in the actual auditing problems. Simulation results comparing these modifications to the Hoeffding-based bounds for the one-sample case are also provided. It must be emphasized that the technique that we propose in this paper is fully justified from a mathematical point of view.

Although the simulations show the proposed bounds to be highly conservative, they still present great interest, since we are not aware of any other method for estimation of the total accounting error in the multisample setting. Moreover, it is shown that significant improvements are hardly possible given the present conditions.

Keywords— bookkeeping quality, lower and upper confidence bounds and intervals, conservativeness, auditing, Hoeffding inequalities, finite populations, multisample problems, computer simulations, Stringer bound.

This research was supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Ministry of Economic Affairs.

1 Introduction

An important problem in the world of statistical auditing is the so-called multi-sample problem. In auditing practice it often occurs that the auditor has to make a statement concerning the total error in a population of dollars, which consists of several individually sampled subpopulations. For example, it is often needed to obtain a combined confidence upper bound from the sample results in the individual subpopulations (without using, for instance, a Bonferroni type of inequality, as the latter technique would generally lead to bad confidence upper bounds). It also happens that a company is comprised of several branches, and the results of quantitative analysis indicate that there is a significant difference in the bookkeeping quality between them. In this case, random sampling without accounting for this difference is going to result in the unwanted increase of the variability in the sample.

In this section we are going to discuss this combination problem in the two-sample setting.

Suppose that a population $\mathbf{A}$ of $N = K + L$ monetary items is a union $\mathbf{A} = \mathbf{B} \cup \mathbf{C}$ of two subpopulations $\mathbf{B}$ and $\mathbf{C}$ that contain respectively K and L items with book values

$$B_1, B_2, \ldots, B_K, \quad \text{and} \quad C_1, C_2, \ldots C_L.$$

Let the errors which are hidden in these book values be respectively

$$E_1^B, E_2^B, \ldots, E_K^B \quad \text{and} \quad E_1^C, E_2^C, \ldots, E_L^C.$$

We assume that only overstatement errors are present, that is

$$0 \le E_i^B \le B_i \quad \text{and} \quad 0 \le E_j^C \le C_j$$

for all possible values of i and j.

Then the total book values of the subpopulations are respectively

$$B = B_1 + B_2 + \cdots + B_K \quad \text{and} \quad C = C_1 + C_2 + \cdots + C_L,$$

and the total errors hidden in B and C are given by

$$E^B = E_1^B + E_2^B + \cdots + E_K^B \quad \text{and} \quad E^C = E_1^C + E_2^C + \cdots + E_L^C.$$

Therefore, the book value and the combined error in the total population $\mathbf{A}$ can be written as

$$A = B + C \quad \textit{and} \quad E = E^B + E^C.$$

As usual, the taintings associated with the accounts in the subpopulations B and C are defined as

$$t_i^B = \frac{E_i^B}{B_i} \quad \text{and} \quad t_j^C = \frac{E_j^C}{C_j}$$

for all possible values of i and j.

Finally, the relative book values of the subpopulation and the relative errors are

$$\alpha_1 = \frac{B}{A}, \quad \alpha_2 = \frac{C}{A}$$

and

$$\mu_1 = \frac{E^B}{B}, \quad \mu_2 = \frac{E^C}{C}$$

respectively. Then for the proportion of errors in the whole population defined as

$$\mu = \frac{E}{A} \tag{1.1}$$

we have

$$\mu := \alpha_1 \mu_1 + \alpha_2 \mu_2.$$

In order to derive a confidence upper bound for the total error E or equivalently for μ, we take independent samples from the subpopulations **B** and **C**. Using the dollar-unit sampling technique as described in Bickel (1992) we draw (with replacement) k independent taintings (relative errors) from the population **B** and l taintings from the population **C**. This leads us to the total of $n = k + l$ independent samples

$$U_1, U_2, \ldots, U_k \quad \text{and} \quad V_1, V_2, \ldots, V_l,$$

where

$$\mathbb{P}\left(U_s = t_i^B\right) = \frac{B_i}{B} \quad \text{and} \quad \mathbb{P}\left(V_t = t_j^C\right) = \frac{C_j}{C}$$

for all possible values of s, t, i and j.

The sample means will be written as

$$\overline{U} = \frac{U_1 + U_2 + \cdots + U_k}{k} \quad \text{and} \quad \overline{V} = \frac{V_1 + V_2 + \cdots + V_l}{l}.$$

For all $s = 1, 2, \ldots, k$ and $t = 1, 2, \ldots, l$ we clearly have

$$\mathbb{E}U_s = \frac{E^B}{B} = \mu_1 \quad \text{and} \quad \mathbb{E}V_t = \frac{E^C}{C} = \mu_2.$$

An important special case is the situation where the sample sizes are proportional to the population sizes, i.e.

$$\frac{k}{n} = \alpha_1 \quad \text{and} \quad \frac{l}{n} = \alpha_2. \tag{1.2}$$

Our results are based on the relationship between upper confidence bounds for a parameter and upper bounds for the tail probabilities depending on this parameter, as described in the appendix. In the nonparametric one-sample setting, the best known estimates for the tail probabilities come from Hoeffding inequalities. A direct adaptation of these results to the two-sample auditing problem is possible, and is described in section 2. However, the efficiency of this approach leaves much to be

desired, requiring extensions of Hoeffding's theorems. Despite considerable analytical difficulties, such an extension has been proved. It is presented in section 3, followed by a construction of a more efficient bound incorporating prior information.

In the end of sections 2 and 3 the applications of our results to the auditing problem are clarified, and computer simulation results are presented in section 4.

The appendix also contains the proof of the inequality leading to the improvement of Hoeffding's results mentioned above.

2 A confidence upper bound for the total auditing error in two populations

Let

$$S_1, S_2..., S_k \quad \text{and} \quad T_1, T_2..., T_l$$

be independent random variables, where k and l are positive integers, $k + l = n$. We further assume that the random variables $S_1, S_2..., S_k$ are identically distributed according to the distribution function F_1^*, and that the same holds for $T_1, T_2..., T_l$ with the corresponding distribution function F_2^*.

In the auditing context we can assume without loss of generality that the values of S_i and T_j are in the interval $[0, 1]$, hence

$$\mathbb{P}(0 \leq S_i \leq 1) = 1 \quad \text{and} \quad \mathbb{P}(0 \leq T_i \leq 1) = 1.$$

Denote the expectations and the variances of these random variables by

$$\lambda_1 = \mathbb{E}S_i \quad \text{and} \quad \eta_1^2 = \mathrm{Var}(S_i); \quad \lambda_2 = \mathbb{E}T_j \quad \text{and} \quad \eta_2^2 = \mathrm{Var}(T_j)$$

respectively. Consider the sample means

$$\overline{S} = \frac{S_1 + ... + S_k}{k}, \quad \overline{T} = \frac{T_1 + ... + T_l}{l},$$

and

$$\overline{S \cup T} = \frac{S_1 + ... + S_k + T_1 + ... + T_l}{n}.$$

Clearly, we have

$$\mathbb{E}\overline{S} = \lambda_1, \quad \mathbb{E}\overline{T} = \lambda_2 \quad \text{and} \quad \mathbb{E}\overline{S \cup T} = \lambda,$$

where

$$\lambda = \frac{k}{n}\lambda_1 + \frac{l}{n}\lambda_2.$$

Let $H(a; \nu)$, $0 \leqslant \nu \leqslant 1$, be the Hoeffding function, which is defined as

$$H(a; \nu) = \left(\frac{1-\nu}{1-a}\right)^{1-a} \left(\frac{\nu}{a}\right)^{a} \qquad \text{for} \quad \nu < a \leqslant 1$$

and $H(a;\nu) = 1$ for $a \leqslant \nu$; $H(a;\nu) = 0$ for $a > 1$. The simplest Hoeffding inequality can be formulated as follows.

Lemma 1. *Assuming that random variables S_i and T_j satisfy the conditions given above, the following inequality holds for $x \in [0,1]$*

$$\mathbb{P}\left(\overline{S \cup T} \leq x\right) \leq \mathbb{P}(1 - \overline{S \cup T} \geq 1 - x) \leq H^n(1-x, 1-\lambda). \tag{2.1}$$

Proof of Lemma 1. Lemma 1 is a simple corollary of Th.1 in Hoeffding (1963). □

Let us construct a confidence upper bound for λ based on this inequality. For $\alpha \in (0,1)$, $x \in [0,1]$ and $n \in \mathbb{N}$ write

$$b_1(x) = \max\left\{\mu : 0 \leq \mu \leq 1 \text{ and } H(1-x, 1-\mu) \geq \alpha^{1/n}\right\}. \tag{2.2}$$

Statement 1. *The statistic $b_1(\overline{S \cup T})$ is a $(1-\alpha)$-confidence upper bound for λ, i.e.*

$$\mathbb{P}\left(\lambda \leq b_1(\overline{S \cup T})\right) \geq 1 - \alpha.$$

Proof of Statement 1. This statement can be easily derived from Lemma 1 using the methods described in Bentkus and van Zuijlen (2003). A more general description of this methodology can be found in Bentkus et al. (2001) and Finkelstein et al. (2000). Note that this bound is, in fact, the bound b_1 by Bentkus and van Zuijlen (2003), which has been rewritten for the two-sample case.

□

Going back to the auditing setup presented in the introduction, consider the random variables

$$U_1, U_2, \ldots, U_k, \quad \text{and} \quad V_1, V_2, \ldots, V_l.$$

It is easy to see that in the case of proportional sampling as defined by (1.2) we have

$$\frac{k}{n}\mu_1 + \frac{l}{n}\mu_2 = \mu,$$

and hence the statistic

$$Ab_1(\overline{U \cup V})$$

is a $(1-\alpha)$-confidence upper bound for E.

Statement 1 can be generalized for the non-proportional case in the following way. Write

$$c_1 = \frac{n}{k}\alpha_1, \;\; c_2 = \frac{n}{l}\alpha_2, \;\; c = \max\{c_1, c_2\}. \tag{2.3}$$

Random variables defined as

$$U_i' = \frac{c_1}{c}U_i, \; i = 1, 2, \ldots, k, \;\; V_j' = \frac{c_2}{c}V_j, \; j = 1, 2, \ldots, l$$

satisfy the probability inequalities $\mathbb{P}(0 \le U_i' \le 1) = 1$ and $\mathbb{P}(0 \le V_j' \le 1) = 1$. According to Statement 1, the statistic $b_1(\overline{U' \cup V'})$ is a $(1-\alpha)$-confidence upper bound for the value

$$\mu' = E(\overline{U' \cup V'}) = \frac{k}{n}\frac{c_1}{c}\mu_1 + \frac{l}{n}\frac{c_2}{c}\mu_2 = \frac{\alpha_1\mu_1 + \alpha_2\mu_2}{c} = \frac{\mu}{c}.$$

Corollary 1. *Let $U_1, \dots, U_k$ and $V_1, \dots, V_l$ be dollar-unit tainting samples taken from populations B and C respectively. Then the statistic*

$$A \cdot c \cdot b_1\left(\frac{c_1\overline{U} + c_2\overline{V}}{c}\right)$$

is a $(1-\alpha)$-confidence upper bound for the total error E.

Proof of Corollary 1. By the argument above, the statistic

$$b_1(\overline{U' \cup V'}) = b_1\left(\frac{c_1\overline{U} + c_2\overline{V}}{c}\right)$$

is a $(1-\alpha)$-confidence upper bound for μ/c. This, together with (1.1), proves the corollary.

□

It is clear that, in general, larger values of c will correspond to more conservative upper bounds for E. Therefore, for practical applications it is advised to choose k and l in such a way that c is minimal.

3 A confidence upper bound for the total auditing error in two populations based on a modification of the second Hoeffding inequality

Under the assumptions made in Section 2, let us consider the case where some apriori information is available in the form $\lambda_1 \le \lambda_1^{(0)}$, $\lambda_2 \le \lambda_2^{(0)}$, $\eta_1 \le \eta_1^{(0)}$, $\eta_2 \le \eta_2^{(0)}$.

As in the previous section, we define $c_1 = \alpha_1(n/k)$, $c_2 = \alpha_2(n/l)$. Write

$$X_i = c_1 S_i, \;\; Y_i = c_2 T_i. \tag{3.1}$$

The mean values and the variances of X_i and Y_i are respectively

$$\mu_1 = \lambda_1 c_1, \quad \mu_2 = \lambda_2 c_2, \quad \sigma_1^2 = \eta_1^2 c_1^2, \quad \sigma_2^2 = \eta_2^2 c_2^2.$$

We also have $\mathbb{P}(0 \le X_i \le c_1) = 1$, $\mathbb{P}(0 \le Y_i \le c_2) = 1$.

Let μ be defined by

$$\mu = \frac{k}{n}\mu_1 + \frac{l}{n}\mu_2.$$

The bounds below are given for the case where $\mu < 1/2$. Since the proportion of errors in the populations investigated by auditors is typically quite low, this assumption hardly leads to any loss of generality in the auditing context.

For the statistic

$$\overline{X \cup Y} = \frac{X_1 + ... + X_k + Y_1 + ... + Y_l}{n}$$

we have $E(\overline{X \cup Y}) = \frac{k}{n}\mu_1 + \frac{l}{n}\mu_2 = \mu$. Note that this is essentially a reduction of the problem to the case of proportional sampling as defined by (1.2).

In order to prove an analogue of the second Hoeffding inequality for the two-sample case, we are going to need the following result.

Lemma 2. *Let s be a real number, and let*

$$f(b, \sigma^2, h) = \ln\left(\frac{b^2}{b^2+\sigma^2}e^{-\sigma^2/bh} + \frac{\sigma^2}{b^2+\sigma^2}e^{bh}\right).$$

Then for all positive real numbers b_1, b_2, σ_1, σ_2 and h the inequality

$$kf(b_1, \sigma_1^2, h) + lf(b_2, \sigma_2^2, h) \le nf\left(\frac{kb_1 + lb_2}{n}, \frac{k}{n}\sigma_1^2 + \frac{l}{n}\sigma_2^2 + \frac{s+1}{2}\frac{k}{n}\frac{l}{n}(b_1 - b_2)^2, h\right) \tag{3.2}$$

holds for $s \ge 0$, and for any real $s < 0$ there exists a combination of parameters b_1, b_2, σ_1, σ_2, h such that the opposite inequality holds.

Remark. The heuristic considerations presented later in this article quickly led to the formulation of (3.2) with $s = 1$ as a conjecture, which was subsequently confirmed by computer simulations. However, an analytical proof, which can be found in the appendix, was not at all easy to obtain due to the complex form of the underlying function. Also, it must be noted that Lemmas 2 and 3, as well as the corresponding upper confidence bound, allow for a natural extension to the general multisample case.

Lemma 3. *Let random variables X_i and Y_j be defined by (3.1). The following inequality holds for $t > 0$:*

$$\mathbb{P}(\overline{X \cup Y} \le \mu + t) \le \left\{\left(1 + \frac{bt}{\sigma^2}\right)^{-(1+bt/\sigma^2)\sigma^2/(b^2+\sigma^2)}\left(1 - \frac{t}{b}\right)^{-(1-t/b)b^2/(b^2+\sigma^2)}\right\}^n, \tag{3.3}$$

where

$$b_1 = c_1 - \mu_1, \;\; b_2 = c_2 - \mu_2, \;\; b = (k/n)(c_1 - \mu_1) + (l/n)(c_2 - \mu_2)$$

and

$$\sigma^2 = \frac{k}{n}\sigma_1^2 + \frac{l}{n}\sigma_2^2 + \frac{k}{n}\frac{l}{n}(b_1 - b_2)^2. \tag{3.4}$$

Proof of Lemma 3. Consider the random variables $X_i - \mu_1$ and $Y_j - \mu_2$. Their expectations are equal to 0, and they are bounded from above by b_1 and b_2 respectively. By applying (1.7), (1.8) and Lemma 2 from Hoeffding (1963), we obtain for $h > 0$

$$\mathbb{P}(\overline{X \cup Y} \leq \mu + t) \leq e^{-hnt + kf(b_1, \sigma_1^2, h) + mf(b_2, \sigma_2^2, h)} \tag{3.5}$$

Then, by Lemma 2,

$$\mathbb{P}(\overline{X \cup Y} \leq \mu + t) \leq e^{-hnt + nf\left(\frac{kb_1 + mb_2}{n}, \frac{k}{n}\sigma_1^2 + \frac{l}{n}\sigma_2^2 + \frac{k}{n}\frac{l}{n}(b_1 - b_2)^2, h\right)} \tag{3.6}$$

The right-hand side of (3.6) attains its minimum at

$$h = \frac{b}{b^2 + \sigma^2} \ln\left(\frac{1 + tb/\sigma^2}{1 - t/b}\right).$$

By inserting this value in (3.6) we obtain (3.3). □

Remark. The modifications in the right hand side of (3.2) compared to inequality (2.8) in Hoeffding (1963) originate from the following intuitive argument.
Let $\widetilde{X}_i$, $i = 1, 2, \ldots, n$ and $\widetilde{Y}_j$, $j = 1, 2, \ldots, n$ be independent random variables such that all $\widetilde{X}_i$ and $\widetilde{Y}_j$ are distributed identically to S_1 and T_1 respectively. Let ξ_k, $k = 1, 2, \ldots, n$, be i.i.d. random Bernoulli variables such that $\mathbb{P}(\xi_k = 1) = \alpha_1$, $\mathbb{P}(\xi_k = 0) = \alpha_2 = 1 - \alpha_1$ and let the random variables in the set

$$\{\widetilde{X}_i, i = 1, 2, \ldots, n\} \cup \{\widetilde{Y}_j, j = 1, 2, \ldots, n\} \cup \{\xi_k, k = 1, 2, \ldots, n\}$$

be independent.

Write $Z_i = \xi_i \widetilde{X}_i + (1 - \xi_i)\widetilde{Y}_i$, $i = 1, 2, \ldots, n$. These Z_i are i.i.d. random variables, with the expected value $E(Z_i) = \mu$ and the variance equal to $\mathrm{Var}(Z_i) = \alpha_1\sigma_1^2 + \alpha_2\sigma_2^2 + \alpha_1\alpha_2(\mu_1 - \mu_2)^2$. This corresponds to a sampling procedure from $\widetilde{X}$ and $\widetilde{Y}$ where a random number of items from each population is included in the sample, following a binomial distribution with parameters n, α_1 and α_2. Let $HB_1(\mu)$ be the Hoeffding-based upper estimate for the corresponding tail probability, i.e.

$$\mathbb{P}\left(\frac{Z_1 + \ldots + Z_n}{n} > \mu + t\right) \leq HB_1(\mu),$$

where t is greater than zero. For $t > 0$ one would expect $HB_1(\mu)$ to be also an upper estimate for the tail probability in the case of fixed size sampling, i.e.

$$\mathbb{P}\left(\frac{\widetilde{X}_1 + \ldots \widetilde{X}_k + \widetilde{Y}_1 + \ldots \widetilde{Y}_l}{n} > \mu + t\right) \leq HB_1(\mu),$$

since this corresponds to a sampling procedure which is intuitively less random and also gives a lower variance for the relevant statistic when $\mu_1 \neq \mu_2$.

Lemma 2 is a generalization of theorem 3 in Hoeffding (1963). However, it doesn't improve the corresponding inequalities for all possible combinations of the parameters b_1, b_2, σ_1, σ_2 and t. Instead, it is necessary to take the minimum over all valid combinations of b_1 and b_2 as given below.

Corollary 2. *Let random variables X_i and Y_j be defined by* (3.1). *Let $b_1 = c_1 - \mu_1$, $b_2 = c_2 - \mu_2$. Without loss of generality we can assume that $b_1 \leq b_2$. The following inequality holds for $t > 0$:*

$$\mathbb{P}(\overline{X \cup Y} \leq \mu + t) \leq \min_{b_1 \leq b_1' \leq b_2} \left\{ \left(1 + \frac{bt}{\sigma^2}\right)^{\frac{-(1+bt/\sigma^2)\sigma^2}{b^2+\sigma^2}} \left(1 - \frac{t}{b}\right)^{\frac{-(1-t/b)b^2}{b^2+\sigma^2}} \right\}^n, \quad (3.7)$$

where $b = (k/n)b_1' + (l/n)b_2$ and

$$\sigma^2 = \frac{k}{n}\sigma_1^2 + \frac{l}{n}\sigma_2^2 + \frac{k}{n}\frac{l}{n}(b_1' - b_2)^2.$$

Proof of Corollary 2. Taking $c_1' = b_1' + \mu_1'$ we have that $c_1' \geq c_1$ and therefore

$$\mathbb{P}(0 \leq X_i \leq c_1') = 1.$$

The inequality (3.7) follows directly from (3.3) with c_1 substituted by c_1'. □

Now, by using the methods described in Bentkus and van Zuijlen (2003) it is possible to introduce a confidence upper bound for μ based on the inequality (3.7) that utilizes the available apriori information and also depends on the sample variances.

Let $b_1 = c_1 - \mu_1$, $b_2 = c_2 - \mu_2$. Let $\hat{\sigma}_1^2$ be the sample variance of $X_1, ..., X_k$. Given below are auxiliary functions needed to construct a certain confidence upper bound for a variance. For a detailed description of this bound refer to Bentkus and van Zuijlen (2003). Let

$$t_3^B(\mu_1) = \begin{cases} \min\left\{t: \ H(\tau(\mu_1) + 2t; \tau(\mu_1)) \leqslant \beta_1^{1/[k/2]}\right\}, \text{ if } (\tau(\mu_1)) \leqslant \beta_1^{1/[k/2]}; \\ \mu_1 - \mu_1^2, \ \text{otherwise}, \end{cases}$$

$$t_2^B = t_3^B\left(\lambda_1^{(0)}\right);$$

$$t_4^B = \begin{cases} \min\left\{t: \ H\left(\varkappa\left(\eta_1^{(0)}\right) + 2t; \varkappa\left(\eta_1^{(0)}\right)^2\right) \leqslant \beta_1^{1/[k/2]}\right\}, \text{ if } \varkappa\left(\eta_1^{(0)}\right) \leqslant \beta_1^{1/[k/2]}; \\ \left(\eta_1^{(0)}\right)^2, \text{ otherwise}, \end{cases}$$

where $\tau(x) = 1 - 2x + 2x^2$, $\varkappa(x) = 1 - 2x^2$ and β_1 lies in the interval $(0, 1)$.

Finally, write

$$t_5^B(\lambda_1) = \min\left\{\left(\eta_1^{(0)}\right)^2, \ \mu_1 - \mu_1^2, \ \hat{\sigma}_1^2 + \min\left\{t_2^B, t_3^B(\mu_1), t_4^B\right\}\right\}.$$

It is shown in Bentkus and van Zuijlen, (2003) that the bound $t_5^B(\lambda_1)$ is a confidence upper bound for the variance η_1^2 with risk β_1 given that $\eta_1 \le \eta_1^{(0)}$. It is possible to define a bound t_5^C for the variance η_2^2 with risk β_2 in the same way.

Write

$$t(\lambda_1, \lambda_2) = \left(\frac{k}{n} c_1^2 t_5^B(\lambda_1) + \frac{l}{n} c_2^2 t_5^C(\lambda_2) \right).$$

It is a confidence upper bound for the value $\frac{k}{n}\sigma_1^2 + \frac{l}{n}\sigma_2^2$ with risk $\beta_1 + \beta_2$.

For a given function $z : [0, 1/2) \times [0, 1/2) \to \mathbb{R}$, consider the functional

$$B(x, z(\cdot)) = \max \left\{ \rho = \frac{k}{n} c_1 \lambda_1 + \frac{l}{n} c_2 \lambda_2 : \right.$$
$$\left. H_2 \left(1 - \rho', x - \rho', z(\lambda_1, \lambda_2) + \frac{1}{2} \frac{k}{n} \frac{l}{n} \big(c_1(1 - \lambda_1') - c_2(1 - \lambda_2') \big)^2 \right) \ge (\alpha - \beta_1 - \beta_2)^{1/n} \right\}, \tag{3.8}$$

where

$$H_2(b, t, \sigma) = \left(1 + \frac{bt}{\sigma^2} \right)^{-(1+bt/\sigma^2)\sigma^2/(b^2+\sigma^2)} \left(1 - \frac{t}{b} \right)^{-(1-t/b)b^2/(b^2+\sigma^2)},$$

and $\lambda_1 \le \lambda_1' \le \max\{\lambda_1, \lambda_2\}$, $\lambda_2 \le \lambda_2' \le \max\{\lambda_1, \lambda_2\}$, $\rho' = \frac{k}{n} c_1 \lambda_1 + \frac{l}{n} c_2 \lambda_2$.

Statement 3. *The statistic $b_2^*(\overline{X \cup Y})$ defined by*

$$b_2^*(\overline{X \cup Y}) = B(\overline{X \cup Y}, t(\cdot)) \tag{3.9}$$

is a confidence upper bound for μ with risk α.

Proof of Statement 3. The proof can be easily obtained by using corollary 2 and applying the methods described in Bentkus and van Zuijlen (2003) for the bound b_6. □

Returning to the original auditing problem, we can prove the following corollary.

Corollary 3. *Assume that the auditor has been able to translate his judgement of the bookkeeping quality (which is based on his previous experience) into apriori upper estimates for the proportion of errors in each population, namely $\mu_1^{(0)}$ and $\mu_1^{(0)}$, as well as for the variances of the taintings, namely $\left(\eta_1^{(0)}\right)^2$ and $\left(\eta_2^{(0)}\right)^2$. If no information of that kind is available, take $\mu_1^{(0)} = \mu_2^{(0)} = 1/2$ and $\left(\eta_1^{(0)}\right)^2 = \left(\eta_1^{(0)}\right)^2 = 1/4$. Let $U_1, \ldots, U_k$ and $V_1, \ldots, V_l$ be the dollar-unit tainting samples taken from populations B and C respectively.*

Then the statistic

$$A \cdot b_2^*(c_1\overline{U} + c_2\overline{V})$$

is a $(1-\alpha)$*-confidence upper bound for the total error* E*.*

Proof of Corollary 3. This follows directly from (1.1) and statement 3. □

As in the previous section, we can say that in general, bigger values of $c = \max(c_1, c_2)$ will result in more conservative bounds for E. Therefore, it is advised to choose k and l so that c is minimal.

4 Simulation results

In order to evaluate the performance of the new bounds presented in the sections 2 and 3 in auditing applications, computer simulations were used. The simulation design followed Grimlund and Felix (1987).

First, two test population **B** and **C** were generated. Since the proposed bounds only depend on the taintings, and not on the actual account sizes and error values, the simulated population consisted of the taintings corresponding to the individual dollars. Thus, instead of the accounts $B_1, B_2, \ldots, B_K$ and the corresponding errors $E_1^B, E_2^B, \ldots, E_K^B$ we consider the taintings $T^{(1)} = \{t_1^B, t_2^B, ..., t_B^B\}$, where, assuming that the i-th dollar was taken from the j-th account,

$$t_i^B = \frac{E_j^B}{B_j}.$$

It is easy to see that sampling with replacement from $T^{(1)}$ results in the same distribution of samples as dollar-unit sampling from the population **B**. The same is true for **C** and the combined population **A**. The monetary values of both populations were taken to be equal to 10000 dollars, i.e. $B = C = 10000$.

The distribution of taintings in each population was derived from the error model described in Grimlund and Felix (1987), which is commonly used in the simulations related to auditing problems. This model uses four parameters: r – the proportion of accounts that are in error relative to all accounts, p_1 – proportion of non-100% overstatement errors among all erroneous items, $p_2 = 1 - p_1$ – proportion of 100% overstatement errors and ν_1 – mean value of the chi-squared distribution used to generate non-100% overstatement errors. The actual values of the parameters used in our computations, as well as the mean value μ and the standard deviation σ of the taintings in each population as well as the union $\mathbf{B} \cup \mathbf{C}$ are presented in Table 1.

Table 1. Generating parameters and descriptive statistics of the two test populations.

	B	**C**	**B** $\cup$ **C**
r	0.05	0.3	–
p_1	0.95	0.8	–
p_2	0.05	0.2	–
ν_1	0.05	0.4	–
μ	0.00502	0.11680	0.06091
σ	0.05337	0.27378	0.20501

Thus, each test population consisted of $(1-r)\cdot 10000$ zero taintings, $r\cdot p_2\cdot 10000$ taintings equal to 1 and $r\cdot p_1\cdot 10000$ randomly generated taintings that followed chi-squared distribution with 1 d.f. scaled by ν_1 and truncated at 1.

The proposed upper confidence bound b_2^* has been compared with the bound b_6 from Bentkus and van Zuijlen (2003). Confidence level was taken to be 95% ($\alpha = 0.05$). It must be noted that a direct comparison with the well-known Stringer bound, as described in Bickel (1992), is not possible, since this bound does not allow an extension to the two-sample case.

The sample sizes $n = 30$, 60, 120 and 240 were chosen for the simulation, and each round was performed in the following way.

First, the desired number of samples was drawn randomly from the test populations. For the bound b_2^*, $n/2$ taintings were randomly drawn from each populations (since the monetary values of the populations are the same, $\alpha_1 = \alpha_2 = 1/2$ and taking $k = l = n/2$ means sampling proportionally to the monetary value). For the bound b_6, n taintings were drawn from the union **B** $\cup$ **C**. The corresponding bounds were calculated and compared with the actual mean value of the items in the test population. Since the samples evaluated in each case were different and random, in order to obtain reliable results this procedure was repeated 10000 times for each sample size. Finally, average values, coverages and variabilities for the analyzed bounds were calculated. Here coverage is the percentage of rounds where the calculated bound was greater than or equal to the actual mean value, and the variability is the variance of the calculated bounds. In order to guarantee the absence of correlation in the data, a cryptographic pseudo-random number generator by Kelsey et al. (2000) was used to create the test populations and to draw the samples.

The following table presents the simulation results.

Table 2. Simulation results.

Sample size	$\overline{b_2^*}$	$\overline{b_6}$	Cov_2^*	Cov_6	Var_2^*	Var_6	Var_2^*/Var_6
30	0.21981	0.21915	100	100	0.003437	0.003710	0.926
60	0.16374	0.163287	99.97	99.86	0.001440	0.001542	0.934
120	0.12881	0.12874	99.90	99.80	0.000600	0.000665	0.901
240	0.10537	0.10541	99.76	99.81	0.000268	0.000290	0.922

Here $\overline{b_2^*}$ and $\overline{b_6}$ are the mean values; Cov_2^* and Cov_6 are the coverage statistics; Var_2^* and Var_6 are the variabilities for the bounds b_2^* and b_6 respectively. In order to measure the relative advantage of sampling proportionally and using the bound b_2^* compared to random sampling with b_6, the ratio Var_2^*/Var_6 was calculated.

Simulation results can be summarized as follows.

- The proposed bounds with proportional subsample sizes offer lower variability compared to the case of random sampling from the combined population $\mathbf{B} \cup \mathbf{C}$.
- The coverage is approximately the same in both cases.
- The tightness of both bounds used in the simulations is also approximately the same.

Therefore, proportional sampling and the bound b_2^* are more effective than random sampling with the bound b_6.

It must be emphasized that the relation between confidence bounds for a parameter and bounds for tail probabilities (which is established by Statement 4 in the appendix) implies that, although very conservative, the presented bounds are in a certain sense the best possible, since they originate in the most accurate provable bounds for tail probabilities in a nonparametric setting. It is natural to expect that significant improvements in the tightness of these bounds can only be obtained by certain restrictions on the class of the possible tainting distributions.

5 Appendix A: Proof of Lemma 2

Lemma 2 follows directly from a more general statement that uses positive real numbers α and β, $\alpha + \beta = 1$, in place of k/n and k/n.

We are going to use the following notation. Let ε be a Bernoulli random variable (that is, a random variable which assumes at most two different values) such that

$$\mathbb{P}\{\varepsilon = a\} = p, \quad \mathbb{P}\{\varepsilon = b\} = q, \quad \mathbb{E}\varepsilon = 0, \quad \mathbb{E}\varepsilon^2 = \sigma^2,$$

for some $b > 0$, $a < 0$, $\sigma^2 > 0$ and $0 \leqslant p, q \leqslant 1$. Using the conditions

$$p + q = 1, \quad \mathbb{E}\varepsilon = ap + bq = 0, \quad \mathbb{E}\varepsilon^2 = a^2 p + b^2 q = \sigma^2$$

we can write

$$p = \frac{b^2}{b^2 + \sigma^2}, \quad q = 1 - p = \frac{\sigma^2}{b^2 + \sigma^2}, \quad a = -\sigma^2/b.$$

For $h \geqslant 0$, introduce the generating function

$$A = A(x) = \mathbb{E}\,\mathrm{e}^{h\varepsilon} = p\,\mathrm{e}^{ha} + q\,\mathrm{e}^{hb},$$

where for brevity we write $x = (b, \sigma^2)$. Write

$$f(x) = \ln A(x).$$

We shall prove that, for any $h \geqslant 0$, the function f satisfies the following convexity type inequality.

Lemma 4. *Let* $0 \leqslant \alpha \leqslant 1$ *and* $\beta = 1 - \alpha$. *Write*

$$x = (b, \sigma) \ \textit{with} \ b = \beta b_0 + \alpha b_1 \ \textit{and} \ \sigma^2 = \beta\sigma_0^2 + \alpha\sigma_1^2 + \frac{(s+1)}{2}\alpha\beta(b_0 - b_1)^2,$$

and

$$x_0 = x\Big|_{\alpha=0} = (b_0, \sigma_0^2), \quad x_1 = x\Big|_{\alpha=1} = (b_1, \sigma_1^2), \tag{5.1}$$

with some

$$b_0, b_1, \sigma_0^2, \sigma_1^2, h \geqslant 0.$$

Then the smallest constant s *such that the inequality*

$$\beta f(x_0) + \alpha f(x_1) \leqslant f(x)$$

holds for all x_0, x_1 *and* $0 \leqslant \alpha \leqslant 1$, *is equal to* 0. *Moreover,* (5.1) *holds for all* $s \geqslant 0$.

Inequality (5.1) extends Hoeffding's (1963) Lemma 3 which is a particular case of (5.1) with $b_0 = b_1$.

Proof of Lemma 4. We start by showing that it suffices to prove (or disprove) (5.1) for $h = 1$. Indeed, if $h = 0$ then (5.1) becomes the equality $0 = 0$. If $h > 0$, we can replace

$$hb_0, hb_1, h\sigma_0, h\sigma_1$$

by new variables

$$b_0, b_1, \sigma_0, \sigma_1.$$

Therefore we assume henceforth that $h = 1$.

Consider the function

$$g(\alpha) = \ln A(x) - \beta \ln A(x_0) - \alpha \ln A(x_1), \quad 0 \leqslant \alpha \leqslant 1. \tag{5.2}$$

This function satisfies

$$g(0) = g(1) = 0.$$

Therefore, in order to prove (5.1) it suffices to show that g is a concave function of α, that is, that

$$g''(\alpha) \leqslant 0. \tag{5.3}$$

In order to show that (5.1) does not hold for $s < 0$, it suffices for such s to find $b_0, b_1, \sigma_0^2, \sigma_1^2 \geqslant 0$ such that g is a strictly convex function of α, that is, that

$$g''(\alpha) > 0, \quad \text{for all} \quad 0 \leqslant \alpha \leqslant 1. \tag{5.4}$$

To simplify formulas, we introduce additional notation. Write

$$B = p\,\mathrm{e}^{-\vartheta} + q, \quad \vartheta = b - a, \quad g_0(\alpha) = \ln B.$$

Then $A = \mathrm{e}^b B$ and

$$g(\alpha) = C + g_0(\alpha), \quad C = b - \beta \ln A(x_0) - \alpha \ln A(x_1).$$

Since C is a linear function of α, we have $C'' = 0$. Therefore g'' and g_0'' have the same sign. Furthermore, $B^2 g_0'' = B''B - B'^2$. Hence, instead of (5.3) (respectively (5.4)) it suffices to check that

$$B'^2 - B''B \geqslant 0, \tag{5.5}$$

and respectively

$$B'^2 - B''B < 0. \tag{5.6}$$

We have

$$B' = (p' - p\vartheta')\,\mathrm{e}^{-\vartheta} + q'$$

and

$$B'' = (p'' - 2p'\vartheta' - p\vartheta'' + p\vartheta'^2)\,\mathrm{e}^{-\vartheta} + q''.$$

Furthermore

$$\begin{aligned} B'^2 &= (p'^2 - 2pp'\vartheta' + p^2\vartheta'^2)\,\mathrm{e}^{-2\vartheta} + (2p'q' - 2pq'\vartheta')\,\mathrm{e}^{-\vartheta} + q'^2, \\ B''B &= (pp'' - 2pp'\vartheta' - p^2\vartheta'' + p^2\vartheta'^2)\,\mathrm{e}^{-2\vartheta} \\ &\quad + (p''q - 2p'q\vartheta' - pq\vartheta'' + pq\vartheta'^2 + pq'')\,\mathrm{e}^{-\vartheta} + qq''. \end{aligned}$$

Hence

$$B'^2 - B''B = m_1\,\mathrm{e}^{-2\vartheta} + m_2\,\mathrm{e}^{-\vartheta} + m_3$$

with

$$\begin{aligned} m_1 &= p'^2 - pp'' + p^2\vartheta'', \\ m_2 &= 2p'q' - pq'' - qp'' - 2pq'\vartheta' + 2p'q\vartheta' + pq\vartheta'' - pq\vartheta'^2, \\ m_3 &= q'^2 - qq''. \end{aligned}$$

We have $q' = -p'$ and $q'' = -p''$, since $p + q = 1$. Replacing q' and q'' by $-p'$ and $-p''$ respectively, we obtain

$$\begin{aligned} m_1 &= p'^2 - pp'' + p^2\vartheta'', \\ m_2 &= -2p'^2 + pp'' - qp'' + 2p'\vartheta' + pq\vartheta'' - pq\vartheta'^2, \\ m_3 &= p'^2 + qp''. \end{aligned} \tag{5.7}$$

Introducing

$$u(\vartheta) = m_1 + m_2\,\mathrm{e}^{\vartheta} + m_3\,\mathrm{e}^{2\vartheta},$$

inequality (5.5) (respectively (5.6)) is equivalent to

$$u(\vartheta) \geqslant 0 \tag{5.8}$$

and respectively

$$u(\vartheta) < 0. \tag{5.9}$$

To simplify forthcoming calculations, we introduce a new parametrization. Write

$$t = \frac{\sigma^2}{b^2}.$$

Then

$$p = \frac{1}{1+t}, \quad q = \frac{t}{1+t}, \quad p' = -p^2 t', \quad p'' = 2p^3 t'^2 - p^2 t''. \tag{5.10}$$

We introduce a new variable

$$\delta = b_1 - b_0 > 0.$$

The assumption that $\delta > 0$ does not restrict generality. Indeed, if $\delta = 0$ then the inequality to prove reduces to the Hoeffding's (1963) Lemma 3. Or one can derive the result by continuity arguments using the inequality for $\delta \neq 0$. 0 If $\delta < 0$, then we can exchange roles of b_0 and b_1. Introduce new variables

$$\hat{\sigma}_0, \hat{\sigma}_1, \hat{b}_0$$

by

$$\sigma_0 = \delta\hat{\sigma}_0, \quad \sigma_1 = \delta\hat{\sigma}_1, \quad b_0 = \delta\hat{b}_0.$$

Note that the variable

$$t = \frac{\sigma_0^2 + \alpha\varrho + (s+1)\alpha\beta\delta^2/2}{(b_0 + \alpha\delta)^2} = \frac{\hat{\sigma}_0^2 + \alpha\hat{\varrho} + (s+1)\alpha\beta/2}{(\hat{b}_0 + \alpha)^2}$$

with $\varrho = \sigma_1^2 - \sigma_0^2$ and $\hat{\varrho} = \hat{\sigma}_1^2 - \hat{\sigma}_0^2$ is independent of δ. Writing

$$c = \hat{b}_0 + \alpha, \quad \gamma = \frac{c}{p},$$

we have

$$t = \frac{\hat{\sigma}_0^2 + \alpha\hat{\varrho} + (s+1)\alpha\beta/2}{c^2}, \quad \vartheta = \delta\gamma$$

since $\vartheta = b + \sigma^2/b = b/p = c\delta/p$. Furthermore, we have

$$\vartheta' = \delta\gamma' = \frac{\vartheta\gamma'}{\gamma}, \quad \vartheta'' = \delta\gamma'' = \frac{\vartheta\gamma''}{\gamma}.$$

Using the notation, we can rewrite (5.7) as

$$\begin{aligned}
m_1 &= p'^2 - pp'' + \frac{p^2\gamma''}{\gamma}\vartheta, \\
m_2 &= -2p'^2 + pp'' - qp'' + \frac{2p'\gamma'}{\gamma}\vartheta + \frac{pq\gamma''}{\gamma}\vartheta - pq\frac{\gamma'^2}{\gamma^2}\vartheta^2, \\
m_3 &= p'^2 + qp''.
\end{aligned}$$

Hence

$$\begin{aligned}\gamma^2 m_1 &= c_0 + c_1\vartheta,\\ \gamma^2 m_2 &= d_0 + d_1\vartheta + d_2\vartheta^2,\\ \gamma^2 m_3 &= e_0.\end{aligned} \tag{5.11}$$

with some c_0,c_1,d_0,d_1,d_2,e_0 independent of ϑ such that

$$\begin{aligned}c_0 &= \gamma^2(p'^2 - pp''),\\ c_1 &= \gamma p^2\gamma'',\\ d_0 &= \gamma^2(-2p'^2 + pp'' - qp''),\\ d_1 &= 2p'\gamma\gamma' + pq\gamma\gamma'',\\ d_2 &= -pq\gamma'^2,\\ e_0 &= \gamma^2(p'^2 + qp'').\end{aligned} \tag{5.12}$$

We introduce the notation

$$y = \gamma'.$$

Let us show that

$$pct' = py - 1, \tag{5.13}$$

$$pc^2t'' = 2 - 4py + (1-s)p, \tag{5.14}$$

$$p' = \frac{1-py}{\gamma}, \tag{5.15}$$

$$p'' = \frac{2py^2 - 1 + s}{\gamma^2}, \tag{5.16}$$

$$\gamma'' = \frac{1 - s - 2y}{p\gamma}. \tag{5.17}$$

Let us prove (5.13). Since $\gamma = c/p$ and $c' = 1$, we have

$$y = \gamma' = \frac{1}{p} - \frac{cp'}{p^2}.$$

Using $p' = -p^2t'$ (cf. (5.10)), we derive (5.13).

Let us prove (5.14). We have

$$t = \frac{u}{c^2}, \quad u = \hat{\sigma}_0 + \alpha\hat{\varrho} + (s+1)\alpha\beta/2.$$

Hence

$$t' = \frac{u'}{c^2} - \frac{2u}{c^3}, \quad ct' = \frac{u'}{c} - 2t$$

and, using $u'' = -s - 1$,

$$t'' = \frac{u''}{c^2} - \frac{2u'}{c^3} - \frac{2u'}{c^3} + \frac{6u}{c^4} = -\frac{s+1}{c^2} - \frac{4u'}{c^3} + \frac{6t}{c^2}.$$

Inserting $\frac{u'}{c} = ct' + 2t$ and using $pct' = py - 1$ (cf. (5.13)), we have

$$c^2 t'' = -s - 1 - \frac{4u'}{c} + 6t = -s - 1 - 4ct' - 2t = 1 - s - 2(1+t) - 4y + 4/p.$$

Multiplying by $p = 1/(1+t)$, we derive (5.14).

Let us prove (5.15). We have $p' = -p^2 t'$ (cf. (5.10)), and an application of (5.13) together with $\gamma = c/p$ yields (5.15).

Let us prove (5.16). By (5.10) we have $p'' = 2p^3 t'^2 - p^2 t''$. Now an application of (5.13), (5.14) and $\gamma = c/p$ yields (5.16).

To prove (5.17) we note that $p\gamma = c$. Differentiating twice, using $c'' = 0$, $\gamma' = y$ and (5.15), (5.16), we easily derive (5.17).

Inserting in (5.12) $\gamma' = y$ and the values (5.13)–(5.17), we obtain

$$\begin{aligned}
c_0 &= 1 - 2py - p^2 y^2 + p - ps, \\
c_1 &= -2py + p - ps, \\
d_0 &= 4py - 2qpy^2 - 2 + q - p + ps - qs, \\
d_1 &= 2py - 2py^2 + q - qs, \\
d_2 &= -qpy^2, \\
e_0 &= (1 - py)^2 + 2qpy^2 - q + qs.
\end{aligned} \tag{5.18}$$

Now we can start the proof of the positive part of the statement of the lemma or, equivalently, of (5.8). Since $p\gamma^2 = c^2/p > 0$, inequality (5.8) is equivalent to (cf. (5.11))

$$w(\vartheta) \geqslant 0, \quad \text{for } \vartheta \geqslant 0, \tag{5.19}$$

with

$$w(\vartheta) = n_1 + n_2 \, \mathrm{e}^{\vartheta} + n_3 \, \mathrm{e}^{2\vartheta},$$

where

$$\begin{aligned}
n_1 &= c_0 + c_1 \vartheta, \\
n_2 &= d_0 + d_1 \vartheta + d_2 \vartheta^2, \\
n_3 &= e_0.
\end{aligned} \tag{5.20}$$

It suffices to prove (5.8) for $s = 0$. It is clear (cf. (5.18) and (5.20)) that the function $s \mapsto w(s)$ is a linear increasing function of s, that is, $\partial_s w \geqslant 0$ (here ∂_s stands for the partial derivative with respect to s). Indeed, using (5.18) and (5.20), we have

$$\partial_s w = -p - p\vartheta + (p - q - q\vartheta) \, \mathrm{e}^{\vartheta} + q \, \mathrm{e}^{2\vartheta} \, .$$

Using $\mathrm{e}^{2\vartheta} \geqslant \mathrm{e}^{\vartheta}(1 + \vartheta)$ and $\mathrm{e}^{\vartheta} \geqslant 1 + \vartheta$ we get

$$\partial_s w \geqslant -p - p\vartheta + p \, \mathrm{e}^{\vartheta} \geqslant 0.$$

Letting $s = 0$ in (5.18) yields

$$\begin{aligned} c_0 &= 1 - 2py - p^2y^2 + p, \\ c_1 &= -2py + p, \\ d_0 &= 4py - 2qpy^2 - 2 + q - p, \\ d_1 &= 2py - 2py^2 + q, \\ d_2 &= -qpy^2, \\ e_0 &= (1-py)^2 + 2qpy^2 - q. \end{aligned} \tag{5.21}$$

Let us show that the function $p \mapsto w$ is a concave function of $0 \leqslant p \leqslant 1$, that is, that $\partial_p^2 w \leqslant 0$. We have

$$h(\vartheta) = \partial_p^2 w = -2y^2 + (4y^2 + 2y^2\vartheta^2)\,\mathrm{e}^{\vartheta} - 2y^2\,\mathrm{e}^{2\vartheta}$$

and

$$\partial_\vartheta h(\vartheta) = (4y^2 + 4y^2\vartheta + 2y^2\vartheta^2)\,\mathrm{e}^{\vartheta} - 4y^2\,\mathrm{e}^{2\vartheta}\,.$$

Using $\mathrm{e}^{2\vartheta} \geqslant \mathrm{e}^{\vartheta}(1+\vartheta+\vartheta^2/2)$, we obtain $\partial_\vartheta h(\vartheta) \leqslant 0$. Therefore the function $\vartheta \mapsto h(\vartheta)$ is decreasing and

$$\partial_p^2 w = h(\vartheta) \leqslant h(0) = 0, \quad \text{for } \vartheta \geqslant 0.$$

Since the function $p \mapsto w$ is a concave function of $0 \leqslant p \leqslant 1$, in order to prove $w \geqslant 0$ it suffices to check that

$$w\Big|_{p=0} \geqslant 0 \quad \text{and} \quad w\Big|_{p=1} \geqslant 0. \tag{5.22}$$

Using (5.21) we can rewrite (5.22) as

$$h_0(\vartheta) = 1 + (\vartheta - 1)\,\mathrm{e}^{\vartheta} \geqslant 0 \tag{5.23}$$

and

$$h_1(\vartheta) = 2 - 2y - y^2 + (1-2y)\vartheta + (4y - 3 + (2y - 2y^2)\vartheta)\,\mathrm{e}^{\vartheta} + (y-1)^2\,\mathrm{e}^{2\vartheta} \geqslant 0. \tag{5.24}$$

To prove (5.23) it suffices to note that $h_0(0) = 0$ and $\partial_\vartheta h_0(\vartheta) = \vartheta\,\mathrm{e}^{\vartheta} \geqslant 0$, for $\vartheta \geqslant 0$.

To conclude the proof of (5.8) we have to check (5.24). It suffices to show that

$$h_1(0) = \partial_\vartheta h_1(0) = 0, \quad \partial_\vartheta^2 h_1(\vartheta) \geqslant 0, \quad \text{for } \vartheta \geqslant 0.$$

We have

$$\partial_\vartheta h_1(\vartheta) = 1 - 2y + (6y - 2y^2 - 3 + (2y - 2y^2)\vartheta)\,\mathrm{e}^{\vartheta} + 2(y-1)^2\,\mathrm{e}^{2\vartheta} \tag{5.25}$$

and

$$\partial_\vartheta^2 h_1(\vartheta) = (8y - 4y^2 - 3 + (2y - 2y^2)\vartheta)\,\mathrm{e}^{\vartheta} + 4(y-1)^2\,\mathrm{e}^{2\vartheta}\,. \tag{5.26}$$

Now (5.25) and (5.26) clearly imply $h_1(0) = h_1'(0) = 0$. Hence it remains to show that $\partial_\vartheta^2 h(\vartheta) \geqslant 0$, which is equivalent to

$$8y - 4y^2 - 3 + (2y - 2y^2)\vartheta + 4(y-1)^2\,\mathrm{e}^{\vartheta} \geqslant 0. \tag{5.27}$$

We apply $\mathrm{e}^{\vartheta} \geqslant 1 + \vartheta + \vartheta^2/2$. Then (5.27) is implied by

$$g(y,\vartheta) = 1 + 2(y-1)(y-2)\vartheta + 2(y-1)^2\vartheta^2 \geqslant 0. \tag{5.28}$$

If $(y-1)(y-2) \geqslant 0$ then (5.28) obviously holds. If $(y-1)(y-2) < 0$, that is, if $1 < y < 2$, then a unique minimizer of the quadratic function $\vartheta \mapsto g(y,\vartheta)$ in (5.28) is $\vartheta_0 = (2-y)/(2(y-1))$ and we get

$$\inf_{1<y<2} g(y,\vartheta) \geqslant \inf_{1<y<2} g(y,\vartheta_0) = \inf_{1<y<2} \Big(1 - \frac{(2-y)^2}{2}\Big) = \frac{1}{2} > 0,$$

which concludes the proof of (5.28) and of (5.8).

Let us prove the negative part of the statement of the lemma, that is, inequality (5.9). We assume that $s < 0$. Below C_k, $k = 0, 1, 2, \ldots$ stand for positive constants which can depend only on s. We have to choose $b_0, b_1, \sigma_0^2, \sigma_1^2 > 0$ such that (5.9) holds. Instead of these parameters we can choose $\hat b_0, \delta, \hat\sigma_0^2, \hat\sigma_1^2 > 0$. Let n be a sufficiently large natural number. Choose

$$\hat b_0 = 1, \quad \hat\sigma_0^2 = n, \quad \hat\sigma_1^2 = 2n.$$

Then, writing $d = \frac{1+s}{2}\alpha\beta$, we have

$$c = 1 + \alpha, \quad \hat\sigma^2 = n + \alpha n + d, \quad t = \frac{n + \alpha n + d}{(1+\alpha)^2}.$$

It is clear that $t \to \infty$, $p \to 0$, $q \to 1$, as $n \to \infty$. Using $p' = -p^2 t'$ (cf. (5.10)) and $1/p = 1 + t$, for $y = \gamma' = (c/p)'$ we have

$$y = \frac{1}{p} - \frac{cp'}{p^2} = 1 + t + ct'. \tag{5.29}$$

It is clear that $ct' = -2t + (n + d')/c$ and $t = (n + \alpha n + d)/c^2$. Hence, using $c = 1 + \alpha$, relation (5.29) yields

$$y = 1 - t + (n + d')/c = 1 + (cd' - d)/c^2.$$

Therefore $|y| \leqslant C_0$. It is clear that

$$\frac{s+1}{2}\alpha\beta \geqslant -C_1.$$

We have

$$p \leqslant 1/t \leqslant 4/(n - C_1) \leqslant 8/n, \quad \text{for } n \geqslant 2C_1.$$

Using $p \leqslant 1$ and $|y| \leqslant C_0$, for the coefficients given by (5.18) we have

$$|c_0| + |c_1| + |d_0| + |d_1| + |d_2| \leqslant C_2. \tag{5.30}$$

In the case of e_0 we obtain

$$e_0 = p(1-y)^2 + pqy^2 + qs \leqslant pC_3 + s/2 \leqslant s/4 \tag{5.31}$$

provided that n is so large that $pC_3 \leqslant -s/4$ and $q \geqslant 1/2$ (recall that $p \to 0$, as $n \to \infty$). We have

$$\gamma^2 u(\vartheta) \leqslant \mathrm{e}^{\vartheta}(|m_1| + |m_2| + m_3\,\mathrm{e}^{\vartheta}).$$

Hence, it suffices to show that $|m_1| + |m_2| + m_3\,\mathrm{e}^{\vartheta} < 0$. Using (5.30), (5.31) and $m_3 = e_0$, we derive

$$|m_1| + |m_2| + m_3\,\mathrm{e}^{\vartheta} \leqslant C_3(1+\vartheta^2) + \frac{s}{4}\,\mathrm{e}^{\vartheta} < 0 \tag{5.32}$$

provided that we choose a sufficiently large $\vartheta > C_5$ (or, equivalently, $\delta > C_6$). Of course, while proving (5.32) we used that $s < 0$.

□

6 Appendix B: The relationship between confidence bounds for a parameter and bounds for tail probabilities

In order to prove the validity of our confidence bounds we have used the relation between upper bounds for the tail probabilities and confidence upper bounds for a parameter as described in Bentkus et al (2001). Let us establish the inverse relation, that is, show how to translate confidence upper bounds for a certain parameter into bounds for tail probabilities.

Let $\Theta \subset \mathbb{R}$ be a parameter space, and consider $\mathcal{T} = \bigcup_{\theta\in\Theta} \mathcal{T}_\theta$, where $\mathcal{T}_\theta$ is a family of random variables with a characteristic θ, where $\theta = \theta(T)$ depends only on the distribution of T. Suppose that there exists a $(1-\alpha)$-confidence upper bound for the parameter θ, i.e. a function

$$b : \mathbb{R} \times (0,1) \to \mathbb{R},$$

which is increasing in the first argument, decreasing in the second argument and

$$\inf_{T\in\mathcal{T}_\theta} \mathbb{P}\left(\theta < b(T,\alpha)\right) \geq 1-\alpha, \quad \forall \alpha \in (0,1), \ \forall \theta \in \Theta. \tag{6.1}$$

Furthermore, we assume that this function is right-continuous with respect to α.

Statement 4. Let $V : \mathbb{R} \times \Theta \to [0,1]$ be defined as

$$V(x,\theta) = \inf\left\{\alpha \in (0,1) \,\middle|\, b(x,\alpha) < \theta\right\}, \tag{6.2}$$

where x is real and $\theta \in \Theta$. Then $V(x,\theta)$ is increasing in x, decreasing in θ and

$$\sup_{T \in \mathcal{T}_\theta} \mathbb{P}\left(T \leq x\right) \leq V(x,\theta), \quad \forall x \in \mathbb{R}, \ \forall \theta \in \Theta. \tag{6.3}$$

Proof of Statement 4. Monotonicity properties follow from the monotonicity of b and the definition of $V(x,\theta)$. In order to prove (6.3) let us substitute α in (6.1) by $V(x,\theta)$. If it is necessary, substitute $b(x,1)$ by $\inf\{\theta \in \Theta\}$ and $b(x,0)$ by $\sup\{\theta \in \Theta\} + \delta$ with an arbitrarily small $\delta > 0$. We have

$$\mathbb{P}\left(\theta < b(T, V(x,\theta))\right) \geq 1 - V(x,\theta),$$

or equivalently

$$\mathbb{P}\left(\theta \geq b(T, V(x,\theta))\right) \leq V(x,\theta). \tag{6.4}$$

From (6.2) and the monotonicity of b in the second argument it follows that for any $\varepsilon > 0$ we have $b(x, V(x,\theta) + \varepsilon) < \theta$. Since the function b is also right-continuous, we can write that

$$b(x, V(x,\theta)) = \lim_{\varepsilon \to +0} b(x, V(x,\theta) + \varepsilon),$$

and therefore

$$b(x, V(x,\theta)) \leq \theta. \tag{6.5}$$

Looking at the left-hand side of (6.4), we can write

$$\mathbb{P}\left(\theta \geq b(T, V(x,\theta))\right) = \mathbb{P}\left(T \in \{t \,|\, b(t, V(x,\theta)) \leq \theta\}\right).$$

Since b is increasing in its first argument, from (6.5) we can conclude that

$$\{t \,|\, t \leq x\} \subset \{t \,|\, b(t, V(x,\theta)) \leq \theta\},$$

and therefore

$$\mathbb{P}\left(T \leq x\right) \leq \mathbb{P}\left(\theta \geq b(T, V(x,\theta))\right) \leq V(x,\theta).$$

□

References

[1] Bentkus, V. and van Zuijlen, M.C.A. (2003). *On conservative confidence intervals.* Lietuvos Matematikos Rinkinys (Lithuanian Mathematical Journal). Vol. **43**, No. 2, 169–193.

[2] Bentkus, V., Pap, G. and van Zuijlen, M.C.A. (2001). *Confidence bounds for a parameter.* Report No. 0125, October 2001. Department of Mathematics, University Nijmegen.

[3] Bickel, P.J. (1992). *Inference and auditing: The Stringer bound.* International statistical review. Vol. **60**, Issue 2, 197–209.

[4] Finkelstein, M., Tucker, H.G. and Veeh, J.A. (2000). *Conservative confidence interval for a single parameter.* Communications in Statistics: Theory and Methods. Vol. **29**, No. 8, 1911–1928.

[5] Grimlund, R. A. and Felix, W. L. (1987). *Simulation Evidence and Analysis of Alternative Methods of Evaluating Dollar-Unit Samples.* The Accounting Review, No. **62**, 455–479.

[6] Hoeffding, W. (1963) *Probability inequalities for sums of bounded random variables.* Journal of The American Statistical Association. Vol. **58**, 13–30.

[7] Kelsey, J., Scheider, B. and Ferguson, N. (2000). *Yarrow-160: Notes on the Design and Analysis of the Yarrow Pseudorandom Number Generator.* Sixth Annual Workshop on Selected Areas in Cryptography. Springer Verlag, 13–33.