Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 178 (2007) 907–917
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Decision Support

Combining approaches for evaluating auditing populations: A simulation study

Howard R. Clayton [a,*], Patrick R. McMullen [b,1]

[a] Department of Management, Auburn University, 415 W. Magnolia Ave., Auburn, AL 36849-5241, USA
[b] Wake Forest University, Babcock Graduate School of Management, Winston-Salem, NC 27109, USA

Received 10 December 2004; accepted 30 January 2006
Available online 2 May 2006

Abstract

This research explores ways of combining four distinct bounds for the mean error in an auditing population. Two competing objectives for a bound are to be close to the true mean being estimated and to be reliable: not less than the true mean in more than 5% of estimations. The optimal combination should provide the best balance of these competing objectives. Estimating the mean error by a single approach is difficult because typically most accounts have no error and the distribution of the errors among those that do is discontinuous and highly skewed. This study reveals that the weights in the optimal combination are not constant but depend on the characteristic of the population being estimated. The optimally combined bound is only 7% smaller overall than the best of the constituents. However, while the best of the constituents fails in 50% of most challenging populations, the optimal combination never fails.

Keywords: Combinatorial optimization; Simulation; Heuristics; Auditing

1. Introduction

The purpose of the research in this paper is to explore ways of combining four distinct approaches to estimating the mean error in an auditing population. The problem of estimating the mean error in an auditing population, in which most items have no error, has been the subject of much research over the last 40 years. The problem which has been referred to as the ''zeros problem'' has been made even more difficult to solve because of the highly skewed distribution of the errors among those line items that do indeed contain an error. The ''zeros problem'' occurs in areas other than auditing. Applied statisticians encounter a similar estimation dilemma in medicine, environmental science, and meteorology, for example. For the remainder of this introductory section and the section on motivation for the research, terms and references specific to auditing and accounting will be made. Detailed explanations of these terms along with illustrative examples will be dealt with in Section 3 where we believe they most naturally fit into the narrative.

To determine the accuracy of a firm's financial statements, auditors typically use the information

* Corresponding author. Tel.: +1 334 844 6512.
E-mail addresses: claythr@auburn.edu (H.R. Clayton), patrick.mcmullen@mba.wfu.edu (P.R. McMullen).
[1] Tel.: +1 336 758 4574.

0377-2217/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2006.01.043

from a sample of account balances to construct an upper confidence bound for the total error amount in the population. Early attempts to form these upper confidence bounds employed classical methods based on simple Normal theory. The bounds were constructed from simple random sampling and stratified random sampling of individual accounts. But the bounds performed poorly because of the peculiar distribution of errors found in accounting populations.

Subsequent proposals to construct the upper confidence bounds were based on a more efficient sampling method, called dollar unit sampling (DUS), in which the accounting population is viewed as a collection of dollar units as opposed to individual accounts. Methods that use DUS include the Stringer bound (1963), Grimlund and Felix's Modified Moment bound (1987), Bickel's Bootstrap bound (1992), and Rohrbach's Variance Augmented bound (1993). The research in this paper will focus on these four methods.

Other noteworthy bounds, based on DUS, but not included in this research are the Load and Spread (Leslie et al., 1976), the Multinomial (Neter et al., 1978), the Bayesian Normal (Menzefricke and Smieliauskas, 1984), the Modified Cox and Snell (Godfrey and Neter, 1984; Neter and Godfrey, 1985), and the Multinomial Dirichlet (Tsui et al., 1985).

2. Motivation for the research

Although some methods, such as the Modified Moment bound and the Variance Augmented bound, have proven to be superior to others in limited groups of populations, none of them is entirely satisfactory. A bound that combines the leading contenders, however, might provide a solution to the "zeroes" problem. The challenge will be to determine the best way of combining the methods, i.e. the relative weights.

Combining existing methods to provide a bound that is superior to the constituents is based on a tried and proven concept. Dworin and Grimlund (1984) employed this strategy in their development of the moment bound. Clayton (1994) also used it in his study on the Hoeffding and bootstrap bounds. The combination idea has extended into other areas of business research such as forecasting. Examples can be seen in the works of Clemen and Winkler (1986), Winkler and Makridakis (1983), Makridakis and Winkler (1983), Newbold and Granger (1974), and others. Even in a completely different field such as medicine, the best known treatment of life threatening illnesses rely on a "cocktail" of a variety of drugs.

3. Description of the bounds

3.1. Dollar unit sampling

We first present a detailed example of dollar unit sampling and an explanation of the various kinds of errors to provide a contextual framework for the description of the bounds. The example is taken from Clayton (1994). Imagine there exists a company with a total of six clients, to whom the company extends credit for services rendered. To track the amount each client owes, the company maintains a list, as shown in Table 1, of each client's balance called the book value. The list is referred to as accounts receivable and each balance, an account receivable or line item. When the company hires an auditor to verify the balances and to make corrections whenever errors are found, the auditor reports an audit value for each client's balance representing the true amount owed. Examples of errors that might be found are listed in the last column of Table 1. The table also shows the audit values as well as cumulative (running) total of the accounts. As will be shown shortly, the cumulative total is important in the sampling process.

Table 1
Illustration of dollar unit sampling

Line item	Item book value	Cumulative book value	Random numbers	Item audit value	Item error (book–audit)
Apex	\$70	\$70	25	\$63	\$7 (overstated)
Bonds	\$30	\$100		\$30	\$0 (no error)
Cars	\$40	\$140	103	\$45	−\$5 (understated)
Duds	\$100	\$240	141, 240	\$0	\$100 (overstated)
Everlast	\$50	\$290		\$50	\$0 (no error)
Foxtrot	\$10	\$300		\$10	\$0 (no error)

The sample unit in dollar unit sampling is an individual dollar rather than an account balance or line item. Once an auditor selects a sample dollar, he or she will identify the line item to which the sample dollar belongs and include that account receivable in the audit. To obtain a random sample of, say, four dollar units from the population of 300 dollar units (given by the cumulative book value), the auditor generates four random numbers between 1 and 300. Suppose the first random number is 25. Then the corresponding sampled dollar will be the 25th in order of the list and the Apex account will be audited since it contains the 25th dollar. If the second random number is 103, then the Cars account will also be audited since it contains the 103rd dollar. Finally, the Duds account will be included in the audit if the third and fourth random numbers happen to be 141 and 240, respectively. Notice here that two sample dollars come from the same line item. When this occurs the information gleaned from the two sample dollars is treated in the same way as if it were obtained from separate line items.

Having two or more sample dollars from the same line item is common in sampling with probability proportional to size of which dollar unit sampling is a variant. The phenomenon is appealing to auditors because the larger line items, containing more information than smaller ones, are more likely to be "hooked" into the auditing process.

Now suppose upon auditing the Apex account the auditor determines that its correct (audit) value is \$63. Then there would exist a \$7 overstatement (OS) error in the book value since \$70 − \$63 = \$7. The \$7 error would be prorated to each dollar in the Apex account to give what is called a taint of 7/70 or \$.10. Thus, every dollar in this account (including the one sampled) would be regarded to possess an OS taint of \$.10. If the auditor finds the audit value of the Cars account to be \$45, then there would be, on the other hand, an understatement (US) error of \$5 (\$40 − \$45) and such US taint associated with every dollar in the Cars account would be −5/40 or −\$.125.

In special situations, most commonly caused by failure to delete an account receivable that is fully paid up, the OS error amount equals the book value amount. This leads to an OS taint of 100% for each dollar unit in that account. This is the situation exemplified in Table 1 by the Duds account whose audit value is \$0. For each dollar in this listed book value the taint is 100/100 = \$1. Thus, the random sample of four dollar units taken from the population of six line item accounts yields the taints \$.10, −\$.125, \$1, and \$1. If one makes the reasonable assumption that no book or audit value is negative, then OS taints will lie in the closed interval between 0 and 1. In contrast, US taints may have absolute values greater than 1.

An auditor can obtain the total error amount in a population of accounts by multiplying the population mean taint (say, μ) by the known population book amount (say, B). Thus, *an upper bound* on the population error is obtained from the product $B * \underline{\mu}$ where $\underline{\mu}$ is an upper bound for the population mean taint.

3.2. Performance measures

A bound that is correct should be greater than the population mean it is attempting to estimate. However, the bound's value should be as close as possible to the true population mean. This property is referred to as its tightness. A bound that far exceeds the true population mean is considered to be conservative and is undesirable because it may lead to costly over-auditing. We measure a bound's tightness with its mean value over repeated applications.

A competing criterion for a desirable bound is its reliability as measured by the coverage. This is the percentage of times the bound is greater than the population mean. Bounds typically are designed to have 95% coverage, referred to as the nominal level. There is a natural trade-off between reliability and tightness. Some bounds, such as Bickel's (1992) bootstrap, often achieve tightness at the expense of reliability; their coverage is typically below the nominal level. Other bounds, such as the Stringer (1963), achieve reliability at the expense of tightness. Their coverages will be typically above nominal but they are often too conservative.

A third criterion is a bound's stability over repeated applications. This is measured by the standard deviation over repeated replications. A desirable bound is one that provides coverage close to the nominal (95%) level, while possessing tightness (i.e. the bound mean should be close to the population mean, and the standard deviation should be small).

3.3. Variance augmented bound

Of the four bounds included in this study, the Variance Augmented bound (AVE), developed by Rohrbach (1993), performs the best under

varied conditions. A brief description of the bound follows.

If a random sample of n taints is obtained from the population of dollar unit taints, the $100(1-\alpha)\%$ large-sample upper bound for the mean taint is given by

$$\underline{\mu}_1 = \underline{t} + [z(1-\alpha)/\sqrt{n}]s, \tag{1}$$

where $\underline{t}$ is the sample mean taint, s, the sample standard deviation of the taints, and $z(1-\alpha)$, the $100(1-\alpha)$ percentile of the standard normal distribution. To derive an expression for the augmented variance, Rohrbach (1993) preferred to express the bound in terms of the complement of the taints, w_i, which also represents the ratio of the audit to book value of the dollars in the ith line item. Thus, we have

$$\underline{\mu}_1 = 1 - \underline{w} + [z(1-\alpha)/\sqrt{n}]s, \tag{2}$$

where $s^2 = [1/(n-1)]\sum^n (w_i - \underline{w})^2$ and $\underline{w} = (1/n)\sum^n w_i$. Rohrbach (1993) showed that the variance expression for s^2 is equivalent to

$$s^2 = (1/n)\sum^n w_i^2 - [2/(n(n-1))]\sum^n \sum^n w_i w_j. \tag{3}$$

By inserting an augmentation factor, $2.7/n$, determined experimentally, this variance becomes

$$s^2 = (1/n)\sum^n w_i^2 - [(2-2.7/n)/(n(n-1))] \times \sum^n \sum^n w_i w_j. \tag{4}$$

Thus, the AVE bound is computed by (2) using the variance expression in (4).

3.4. Modified moment (MM) bound

The MM bound is much more complicated than the AVE bound and will require extensive explanations that are best obtained from the references (Dworin and Grimlund, 1984, 1986). A very brief outline is given here. The first step in computing the MM bound is to determine the mean *nonzero* sample taint $\underline{t}$ and a hypothetical taint $t^* = 0.81[1 - 0.667\tanh(10\,\mathrm{abs}(\underline{t}))][1 + 0.667\tanh(n/10)]$. The hypothetical taint is a heuristic needed to introduce a degree of conservatism into the bound. These two sample statistics are then used to estimate the mean, variance and third central moment of the sampling distribution of mean error tainting which is approximated by a gamma distribution. Estimation of the three moments of the gamma distribution involves the introduction of a number of auxiliary distributions and their central and noncentral moments. Denoting the three central moments of the gamma distribution by UC_1, UC_2, and UC_3, the parameters of the gamma are given by $A = 4(UC_2)^3/(UC_3)^2$; $B = 0.5[(UC_3)/(UC_2)]$; and $G = UC_1 - 2(UC_2)^2/(UC_3)$. Then MM is computed by

$$\underline{\mu}_2 = G + AB[1 + z(1-\alpha)/\sqrt{(9A)} - 1/(9A)]^3. \tag{5}$$

3.5. Bickel Bootstrap bound

Bickel's bound is computed by a modified bootstrap procedure. Assume that an initial sample (size n) of taints is taken from the population of dollar units. This sample may contain m nonzero taints with the remaining $n - m$ being zero taints. The first step in the usual procedure would be to repeatedly draw random samples, also of size n, *with replacement*, from the initial sample of taints to obtain 1000 or more new samples. These new samples obtained in this way, treating the initial sample as if it were a population, are called bootstrap samples. Such bootstrap samples would contain zero and nonzero taints whose distribution would be expected to approximate, to high degree of accuracy, the real population tainting distribution.

The modification proposed by Bickel (1992) is to first consider the m^*, the number of nonzero taints in a bootstrap sample of size n, to have a binomial distribution with parameters n and $p(m, 1-\alpha)$ rather than parameters n and m/n as would be dictated by the regular bootstrap. Here, the expression $p(m, 1-\alpha)$ represents the $(1-\alpha)$ upper bound on m/n $(=\pi)$, the proportion of nonzero taints in the initial sample. The second part of the modification is to then generate the bootstrap samples (size m^*) of nonzero taints by resampling, *with replacement*, only the nonzero taints in the initial sample.

Since only the nonzero taints contribute to the sum of taints in a sample, one can approximate the probability

$$P\left[\sum^n t_k \geqslant n\mu - nt\right] \quad \text{by} \quad P^*\left[\sum^{m^*} v_k^* \geqslant np(m, 1-\alpha)\underline{v} - nt\right],$$

where t_k, with mean $\underline{t}$, represents a taint in the initial sample; v_k, with mean $\underline{v}$, represents a nonzero taint in the initial sample; and v_k^*, represents a nonzero taint in a bootstrap sample (size m^*). From the

bootstrap process one may solve $P*[\sum^{m^*} v_k^* \geqslant np(m,1-\alpha)\underline{v} - nt] = 1-\alpha$, with the α quantile, t_α, to obtain the $1-\alpha$ upper bound for the mean taint given by

$$\underline{\mu} = \underline{t} + t_\alpha.$$

However, a reasonable and less computer-intensive approach to obtain a bound when $m > 1$, is to use the normal approximation to the bootstrap distribution of $\sum^{m^*} v_k^*$. Noting the expected value by

$$E^* \sum^{m^*} v_k^* = np(m,1-\alpha)\underline{v},$$

and the variance by

$$\mathrm{Var}^* \sum^{m^*} v_k^* = n\{p(m,1-\alpha)s^2 + p(m,1-\alpha) \times [1-p(m,1-\alpha)]\underline{v}^2\},$$

where $s^2 = [1/(m-1)]\sum^m (v_k - \underline{v})^2$, one obtains the following for the approximate bound:

$$\underline{\mu}_3 = \underline{t} + [z(1-\alpha)/\sqrt{n}]\{p(m,1-\alpha) \times [s^2 + (1-p(m,1-\alpha))\underline{v}^2]\}^{1/2}. \quad (6)$$

This bound is asymptotically correct because as $n \to \infty$, $p(m,1-\alpha) \to \pi$, $s^2 \to \mathrm{Var}(V)$, and $\underline{v} \to E(V)$ where V is a random variable denoting the nonzero taints in the population.

3.6. Stringer bound

Computation of the Stringer (1963) bound starts out by assuming conservatively that all the m observed errors in a dollar unit sample are 100% overstatements (i.e. all taints are equal to +1). If this were the case then the bound would be given by $p(1-\alpha;n,m)$ which, a slight modification of the notation in Section 3.5, represents the $1-\alpha$ upper confidence bound for the population proportion, π, based on the binomial distribution, when m errors are found in a random sample of n. The next step is to adjust this bound for taints that are less than 1. If the m observed nonzero taints are $t_1 \geqslant t_2 \geqslant \cdots \geqslant t_m$ then the Stringer bound, for OS errors only, can be expressed as

$$\underline{\mu}_4 = p(1-\alpha;n,m) - \sum_{(k=1)}^{m} [p(1-\alpha;n,k) - p(1-\alpha;n,k-1)](1-t_k).$$

The bound however is usually expressed in the following equivalent form:

$$\underline{\mu}_4 = p(1-\alpha;n,0) + \sum_{(k=1)}^{m} [p(1-\alpha;n,k) - p(1-\alpha;n,k-1)]t_k. \quad (7)$$

When dealing with a sample containing negative as well as positive taints, a procedure known as the Stringer offset method is employed. The Stringer offset method reduces the initial Stringer bound for OS errors only, by subtracting the projected US errors inferred from the observed sample. For example, if, in addition to the m positive taints there are w negative taints with absolute values, $g_1 \geqslant g_2 \geqslant \cdots g_w$, then the Stringer offset bound (SO) is given by $\mathrm{SO} = \underline{\mu}_4 - \sum^w g_k/n$.

3.7. Comparison among the bounds

The MM and AVE bounds have been studied and compared by various authors including Rohrbach (1993) and Clayton (1995). Both bounds have been shown to perform equally well in regular and extreme audit situations, providing a tight bound that is reliable most times. However, they do not always give the tightest bound or the most reliable one in all settings. The Stringer bound is very reliable, always giving coverage above the nominal level, but achieves this reliability at the expense of being too conservative. At the other extreme we have the Bickel bound, which is the tightest among the four bounds but has shown, in simulation studies, to be unreliable with coverages well below nominal in many settings.

The complementary property of the four bounds is the reason for their inclusion in this study. The desired outcome is that the optimal combination among these bounds will provide a useful tool for evaluating populations with skewed distributions.

4. Methodology

4.1. Tainting distributions

Based on the description of the empirical distribution of dollar unit taints in accounts receivable and inventory populations reported by Neter et al. (1985), the tainting distribution model for the dollar units was taken as a mixture of four components. These were (a) a continuous distribution for absolute US taints, (b) a mass at zero, (c) a separate continuous distribution truncated at 1.0 for OS taints under 100%, and (d) a mass at 1.0 for 100% OS

taints. Fig. 1 illustrates an example of the tainting distribution model. The highly skewed continuous distributions for absolute US and non-100% OS taints were aptly described by χ^2 distributions (Dworin and Grimlund, 1984). A variety of tainting distributions were generated by manipulating the total dollar unit error rate, the conditional proportions of taints that were OS, US, and 100% OS, and the means of the χ^2 distributions.

4.2. Simulation design

For this investigation 48 study populations were simulated. To generate these study populations, first the unconditional total error rate (π) was varied between 5% and 10%. This represented a low and a mid-range value for rates that typically vary between 3% and 30%. Next the conditional proportion of 100% OS errors was varied among 0%, 10%, 20%, and 40%. For US errors, the conditional proportion was varied among 0%, 10%, and 20%. In any given study population, the conditional proportion of pure OS errors was taken as the remainder after the 100% OS and US error rates were determined. For example, with 10% of the errors being US and 20% of them being 100% OS, the proportion of pure OS would be 70%. Finally, two models for the means of the distributions for OS and US were employed. The first model took the means to be 0.3 and 0.1, respectively, and the second model set the mean values at 0.1 and 0.1, respectively. Thus, the 48 study populations were a result of $2 \times 4 \times 3 \times 2$ combinations.

For each of the 48 study populations, two sample sizes, 50 of 100, were investigated. From each sample the four bounds at the 95% confidence level were computed 1000 times and the values stored for the combination phase in the simulation.

4.3. Combining the bounds

Enumeration of the possible weighting combinations of the bounds was accomplished in the following manner: Step 1, specify a fixed total number (e.g. 30) of sub-units to be made available for splitting up among the four bounds. Step 2, allocate a certain number of sub-units to each of the bounds so that the sum of the allocated sub-units always equals the pre-specified total. For example, $5 + 6 + 3 + 16 = 30$. Actual allocation of sub-units was done systematically as described below in Section 4.4. Step 3, determine the weight associated with each bound by dividing the allocated number of sub-units by the pre-specified total. For example, $wt_1 = 5/30 = 0.167$, $wt_2 = 6/30 = 0.200$, $wt_3 = 3/30 = 0.100$ and $wt_4 = 16/30 = 0.533$. For every possible combination of the four bounds, the weights must sum to unity. For example, $0.167 + 0.200 + 0.100 + 0.533 = 1$.

In deciding on the total number of sub-units we had to be aware that the larger the number, the greater will be the precision in allocation of weights to the bounds. But the price paid for increased precision is an increase in CPU requirements. So a compromise between precision and CPU demands had to be reached in specifying the total number

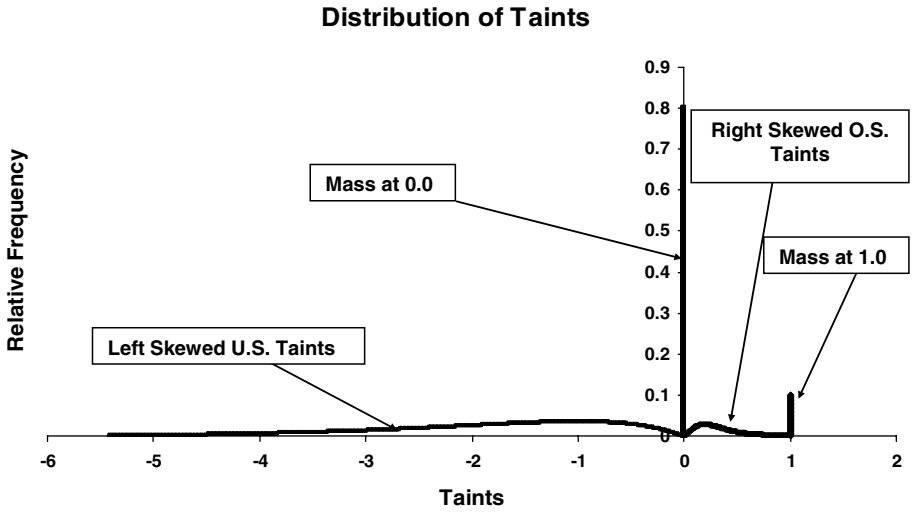

Fig. 1. Distribution of taints.

of sub-units. We chose 30 sub-units in this study because it yielded an acceptable level of precision without an overwhelming burden on our CPU.

4.4. Data gathering

The choice of a weighting scheme with 30 sub-units for combining the four bounds yielded a total of 5456 possible combinations. The total can be derived in the following way: Allocating 30 sub-units to an arbitrary first bound, there is only 1 way for allocating the remaining (zero) sub-units. Then allocating 29 sub-units to this arbitrary first bound, there are three ways of allocating the remaining sub-unit. Next, allocating 28 sub-units to this arbitrary first bound, there are six ways of allocating the remaining sub-units. Continuing in this fashion the number of combinations follows a "triangle" series (1, 3, 6, 10, 15,...,496) with 31 terms. The sum of this triangle series is given by the binomial coefficient $\binom{33}{30}$ which equals $(33\times 32\times 31)/(3\times 2)$ and results in the total 5456.

For each of the possible 5456 combinations we computed the mean, standard deviation, and coverage of the combined bound over 1000 replications. This was done for each of the 48 study populations and for each sample size. Through this process, by having zero weights on three bounds and unit weight on one bound, the performance measures of the four constituent bounds in every study population were automatically generated. This made it possible to compare the performance among various combinations as well as the constituents.

Because of the trade-off between the tightness of a bound, as measured by its mean, and its reliability, as measured by its coverage, it is instructive to visualize an "efficient frontier" over the range encompassed by all possible combinations of the bounds (see left curved edge in Fig. 2). Points on the efficient frontier will represent combinations for which no other combinations can be identified to provide a tighter bound for a given coverage level. The optimal combination, under our chosen weighting scheme, will logically correspond to a point on the efficient frontier. It will be the combination with the smallest mean at exactly 95% coverage. We determined the optimal combination for each of the 48 study populations and two sample sizes.

5. Simulation results

5.1. Optimal bound weights

Table 2 shows the percentage of the 48 study populations, investigated with sample sizes 50 and 100, for which each bound achieved a certain weighting in the optimal combination. One can see that the Stringer bound and Moment bound achieved a weighting less than 0.10 in nearly 90% of the populations while never appearing in any optimal combination with above 50% of the weight. In contrast, the Bickel and AVE bounds achieve low weights under 0.30 in about 40% of the study populations and high weights above 0.70 in over 30% of the study populations. Clearly the optimal combination in a wide variety of settings is dominated by the Bickel and AVE bounds.

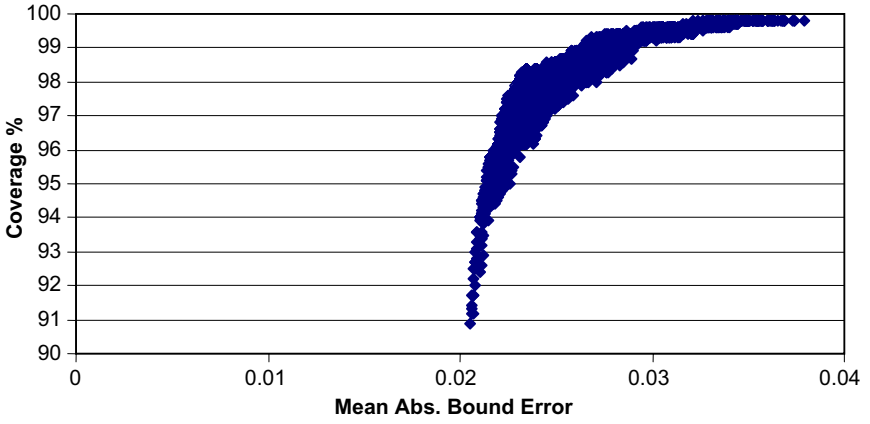

Fig. 2. Coverage vs. tightness of bound for a population with 10% error rate (80% of conditional errors are overstatements and 20% of conditional errors are understatements).

Table 2
Percentage of study populations with varying levels of assigned weights of constituent bounds in the optimal combination

Wt. class	Percentage of the study populations			
	Str	Bkl	Mom	Ave
<0.10	88.5	38.5	89.6	27.1
0.10–0.30	9.4	1.0	9.4	16.7
0.30–0.50	2.1	9.4	1.0	14.6
0.50–0.70	0.0	17.7	0.0	5.2
0.70–0.90	0.0	19.8	0.0	14.6
0.90–1.00	0.0	13.5	0.0	21.9
Total %	100.0	100.0	100.0	100.0

To illustrate the particular type of accounting population that is favored by the dominant bounds in the optimal combination, Table 3 shows the distribution of mean bound weights among populations having only OS errors, high and low incidents of US errors, as well as low and high incidents of 100% OS errors. We observe that for populations having only regular OS errors, the optimal combination is practically a Bickel bound capturing over 80% of the weights. For populations with low or high incidents of US errors the Bickel and AVE bounds are roughly equally weighted when the small sample size is used but the AVE gains prominence when the larger sample size is used. An interesting reversal of weighting occurs in the presence of 100% OS errors. At low levels of 100% OS errors, the Bickel dominates with between 76% and 67% weighting (for sample size 50 and 100, respectively). But at higher levels of 100% OS, the AVE dominates with mean weights between 72% and 79%, respectively.

Table 3
Mean weights in the optimal combinations for various types of populations

Pop. types	Mean weights			
	Str	Bkl	Mom	Ave
n = 50				
All OS	0.000	0.875	0.017	0.108
US = 20%	0.021	0.500	0.069	0.410
US ⩽ 10%	0.020	0.454	0.058	0.468
100 OS ⩽ 10%	0.000	0.763	0.115	0.122
100 OS ⩾ 20%	0.037	0.226	0.018	0.718
Averages	0.016	0.564	0.055	0.365
n = 100				
All OS	0.000	0.817	0.017	0.167
US = 20%	0.081	0.381	0.021	0.517
US ⩽ 10%	0.039	0.350	0.015	0.595
100 OS ⩽ 10%	0.005	0.667	0.032	0.297
100 OS ⩾ 20%	0.096	0.107	0.006	0.792
Averages	0.044	0.464	0.018	0.473

5.2. Shortcomings of optimal weights

Despite the dominance of the Bickel and AVE bounds in the optimal weightings, it is clear from Table 3 that the optimal weighting was far from being constant over the various types of accounting populations under investigation. This lack of constancy presents a problem in the usefulness of the optimal bound. The weights are dependent on characteristics of the population being audited and one cannot reasonably assume that a practitioner will always have full knowledge of the characteristics of the accounting population he or she is working with. Having a fixed optimal weighting would have been ideal. With a fixed set of weights for the constituent bounds it is easy to program a computer to estimate the mean population error. Regardless of the nature of the errors in the sample, the ultimate bound would be computed in the same way. At the same time, the overarching goal is for the resulting combined bound to always achieve the nominal level of reliability (e.g. 95%) while being as close as possible to the true population mean error. This goal is attained by the optimal bound.

One possible solution to having constant weights would be to use the average values of the mean weights for the bounds over all the study populations. Unfortunately the combined bound generated by these average weights proved to be quite unreliable achieving as low as 90% coverage for some populations. In addition, for most of the populations in which this combined bound proved unreliable, its mean value exceeded that of the AVE bound. For these reasons, the combined bound generated by the average of the mean optimal weights could not be considered a viable solution.

5.3. Sub-optimal compromise

Faced with the dilemma of either having an optimal bound whose weights depended on population characteristics or a combined bound with coverage and tightness problems, we decided on a compromise solution. We manually searched for the unique combination that was not necessarily optimal for any study population but which gave a combined bound with the smallest mean value and which

could attain at least 95% coverage in *every* study population. This bound, though not globally optimal, represents the best combination that can guarantee nominal reliability without the practitioner having to apply knowledge about the population. In every setting the bound is calculated with predetermined weights. We named this combined bound appropriately the Global95. In Tables 4 and 5, the properties of the Global95 are compared with the properties of the optimal combination and the AVE bound which is the best of the constituent bounds. The purpose of the comparison is to illustrate how much of the optimal properties are given up in the compromise while still improving on the performance of the best constituent.

Table 4 compares the coverage performance of the Global95 with that of the optimal combination and the AVE bound. We see that, with coverages averaging above 99% for sample size 50 and over 98% for sample size 100, the Global95 is more conservative than both the optimal and AVE. Note how much closer to the nominal 95% is the optimal

Table 4
Coverage performance, in various types of populations, for the optimal, AVE, and Global95 bounds

	Mean coverage values			Percentage of coverages below nominal 95% level		
	Optimal	Ave	Global95	Optimal	Ave	Global95
n = 50						
All OS	95.5	100.0	100.0	0	0	0
US = 20%	96.4	99.1	99.1	0	12.5	0
US ⩽ 10%	96.3	99.1	99.4	0	14.3	0
100 OS ⩽ 10%	95.7	100.0	99.9	0	0	0
100 OS ⩾ 20%	96.9	98.2	98.7	0	25.0	0
Averages	96.3	99.1	99.3			
n = 100						
All OS	96.2	99.6	99.9	0	0	0
US = 20%	95.7	96.9	98.2	0	37.5	0
US ⩽ 10%	95.8	97.5	98.6	0	25.0	0
100 OS ⩽ 10%	95.7	99.1	99.4	0	5.0	0
100 OS ⩾ 20%	95.8	95.5	97.5	0	50.0	0
Averages	95.8	97.3	98.5			

Table 5
Means and standard deviations, in various types of populations, for the optimal, AVE, and Global95 bounds

	Mean			Standard deviation		
	Optimal	Ave	Global95	Optimal	Ave	Global95
n = 50						
All OS	0.0471	0.0604	0.0680	0.0225	0.0153	0.0184
US = 20%	0.0650	0.0698	0.0796	0.0304	0.0246	0.0283
US ⩽ 10%	0.0697	0.0747	0.0858	0.0305	0.0252	0.0287
100 OS ⩽ 10%	0.0530	0.0621	0.0718	0.0276	0.0194	0.0234
100 OS ⩾ 20%	0.0843	0.0843	0.0963	0.0341	0.0313	0.0346
Averages	0.0682	0.0731	0.0837	0.0305	0.0250	0.0286
n = 100						
All OS	0.0357	0.0409	0.0463	0.0133	0.0107	0.0121
US = 20%	0.0494	0.0498	0.0562	0.0185	0.0171	0.0185
US ⩽ 10%	0.0534	0.0545	0.0621	0.0187	0.0172	0.0187
100 OS ⩽ 10%	0.0397	0.0425	0.0495	0.0160	0.0135	0.0153
100 OS ⩾ 20%	0.0651	0.0637	0.0713	0.0218	0.0213	0.0225
Averages	0.0521	0.0530	0.0601	0.0187	0.0172	0.0186

coverage with average values of 96.3% and 95.8%, respectively, for sample sizes 50 and 100. On the other hand, as a result of its definition, the Global95 does not exhibit the reliability problems that the AVE clearly has. In none of the study populations is the coverage of the Global95 compromise bound below the nominal 95%. In contrast, using sample size 50, the AVE bound shows below nominal coverage in 25% of the populations with high incidents of 100% OS errors. When the larger sample size is employed, the AVE's coverage is below nominal in as many as 37% of the populations with high US errors and 50% of populations with high 100% OS errors.

Table 5 compares the means and standard deviations of the Global95 compromise with those of the optimal and AVE bounds. As a reminder, it is desirable to have small values for the mean and standard deviation without the bound becoming unreliable by having below nominal coverage. As one would expect, the optimal bound is the smallest in nearly every setting. The exception is for high incidents of 100% OS errors when the AVE is smaller than the optimal (but becomes unreliable to achieve this). On the other hand, the Global95 is clearly less tight than the AVE in every setting in order to maintain its reliability. The consistency of the Global95, supported by its relatively smaller standard deviation, is better than that of the optimal bound in most types of populations. The AVE with the smallest standard deviation, being a "pure" bound instead of a combination, is the most consistent among the three bounds.

6. Summary and conclusions

This study has shown modest achievement in combining the two leading methods, modified moment (MM) and variance augmented (AVE), for evaluating audit populations, with the conservative Stringer bound and the aggressive, but unreliable, Bickel bound. The optimal weights, though not constant over the wide cross-section of study populations, gave a combined bound that dominates the best among the constituent bounds. Compared to the variance augmented bound, the combined bound is only 7% smaller, on average, over all the study populations. However, the major advantage of the optimal combined bound lies in its reliability. Whereas the variance augmented bound gives below nominal coverage in as many as half of the study populations that possess a high proportion of 100% overstatement (OS) errors, the optimal combined bound always gives coverage that is above, yet close to, the nominal level.

One short-coming of the optimal weights developed in this study is that they depend on population characteristics which will not always be known by the practitioner. Nevertheless, in some cases, the practitioner will be familiar with the population characteristics if he or she has past experience with auditing accounts from a given client. Also, some general knowledge of the error distribution based on the industry from which the accounts are taken, and on the accounting category, is available in the literature (Neter et al., 1985; Ham et al., 1985). An attempt to determine a reliable combined bound whose weights are fixed, and thus independent of population characteristics, resulted in a bound (the "Global95") that was not tight. The overall mean of this compromise bound was as much as 23% higher than the optimal using samples of size 50 and 16% higher using samples of size 100. The trade-off for having constant weights and guaranteed reliability appears rather expensive. In a future investigation, it would be useful to continue the search of weights that still produce a reliable bound but which rely entirely on sample information.

References

Bickel, P.J., 1992. Inference and auditing. International Statistical Review 60, 197–209.

Clayton, H.R.A., 1994. Combined bound for errors in auditing based on Hoeffding's inequality and the bootstrap. Journal of Business and Economic Statistics 12 (4), 437–448.

Clayton, H.R.A., 1995. Simulation study to compare two leading methods for evaluating audit populations using dollar unit sampling. In: Proceedings of the 26th Annual National Decision Sciences Institute, Boston November.

Clemen, R.T., Winkler, R.L., 1986. Combining economic forecasts. Journal of Business and Economic Statistics 4, 39–46.

Dworin, L., Grimlund, R.A., 1984. Dollar unit sampling for accounts receivable and inventory. The Accounting Review 59, 218–241.

Dworin, L., Grimlund, R.A., 1986. Dollar unit sampling: A comparison of the quasi-bayesian and moment bounds. The Accounting Review 61, 36–57.

Godfrey, J., Neter, J., 1984. Bayesian bounds for monetary unit sampling in accounting and auditing. Journal of Accounting Research 22, 497–525.

Grimlund, R.A., Felix Jr., W.L., 1987. Simulation evidence and analysis of alternative methods of evaluating dollar-unit samples. The Accounting Review 62, 455–479.

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

Leslie, D.A., Teitlebaum, A.D., Anderson, R.J., 1976. Dollar-unit Sampling: A Practical Guide for Auditors. Copp, Clark, & Pitman, Toronto.

Makridakis, S., Winkler, R.L., 1983. Averages of forecasts: Some empirical results. Management Science 29, 987–996.

Menzefricke, U., Smieliauskas, W., 1984. A simulation study of the performance of parametric dollar unit sampling statistical procedures. Journal of Accounting Research 22, 588–603.

Neter, J., Godfrey, J., 1985. Robust Bayesian bounds for monetary unit sampling in auditing. Applied Statistics 34, 157–168.

Neter, J., Leitch, R.A., Fienberg, S.E., 1978. Dollar unit sampling: Multinomial bounds for total overstatement and understatement errors. The Accounting Review 53, 77–93.

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

Newbold, P., Granger, C.W.J., 1974. Experience with forecasting univariate time series and the combination of forecasts. Journal of the Royal Statistical Society Series A 137, 131–164.

Rohrbach, K.J., 1993. Variance augmentation to achieve nominal coverage probability in sampling from audit populations. Auditing: A Journal of Practice and Theory 12, 79–97.

Stringer, K.W., 1963. Practical aspects of statistical sampling in auditing. In: Proceedings of the Business and Economic Statistics Section. American Statistical Association, pp. 405–411.

Tsui, K.W., Matsumura, E.M., Tsui, K.L., 1985. Multinomial-Dirichlet bounds for dollar-unit sampling in auditing. The Accounting Review 60, 76–96.

Winkler, R.L., Makridakis, S., 1983. The combination of forecasts. Journal of the Royal Statistical Society Series A 146, 150–157.