Speeding Up the Asymptotics When Constructing One-sided Coverage Intervals with Survey Data

Phillip S. Kott[1] and Yan K. Liu[2]

[1] RTI International, 6110 Executive Blvd., Rockville, MD 20852.
[2] Statistics of Income Division, IRS, P.O. Box 2608, Washington, DC 20013

Abstract

Coverage intervals for a parameter estimate are frequently derived from a survey sample by assuming that the randomization-based parameter estimate is asymptotically normal and that the associated measure of the estimator's variance is roughly chi-squared. In many situations, however, the size of the sample and the nature of the parameter being estimated render the conventional Wald technique dubious, especially when a one-sided coverage interval is needed. We will propose a method of coverage-interval construction that "speeds up the asymptotics" so that the resulting one-sided intervals can have much better coverage properties than corresponding Wald intervals. For the important case of a mean computed from a stratified, simple random sample with ignorably small sampling fractions, no model need be assumed. Moreover, whether or not a model is invoked, our intervals are asymptotically equivalent to Wald intervals. As a result, they share the same large-sample, randomization-based properties. A simulation demonstrates the usefulness of our intervals.

Key Words: Audit data, randomization-based, skewness, stratified sample design, Wald interval.

1. Introduction

Suppose $\hat{t}$ is a nearly unbiased estimator for a finite-population or model parameter t based on a survey sample. A one-sided Wald coverage interval for t is

$$t \le \hat{t} + \Phi^{-1}(\alpha)\sqrt{v} \quad \text{or} \quad t \ge \hat{t} - \Phi^{-1}(\alpha)\sqrt{v}, \tag{1.1}$$

where v is an estimator for V the variance of $\hat{t}$, and $\Phi(.)$ is the cumulative distribution function (*cdf*) of a standard normal distribution. It is well known that if the sample size is large enough, then both inequalities hold for roughly α-percent of the samples of that size (when the sampling mechanism under consideration produces a sample of random size, the modifier "expected" needs to be added to "sample size" and "size" in the last sentence).

In practice, however, the sample size may not be nearly large enough for a one-sided Wald interval to contain ("cover") t with the frequency suggested by the asymptotic theory. This may be because $\hat{t}$ has a skewed distribution, as in the case of an estimated proportion not near enough to 1/2, or because the relevant sample size is the number of sampled primary sampling units not the number of enumerated elements, as when v is computed from a stratified multistage sample using probability-sampling (randomization-based) principles. Ineither of those situations, we propose the following one-sided intervals:

$$t \le \hat{t} + \delta + \sqrt{z^2 v + \delta^2} \quad \text{or} \quad t \ge \hat{t} + \delta - \sqrt{z^2 v + \delta^2}, \tag{1.2}$$

where $\delta = \frac{1}{6}(1 - z^2)\frac{m_3}{v} + \frac{z^2}{2} b,$ (1.3)

$z = \Phi^{-1}(\alpha)$, and m_3 is a nearly unbiased estimator for the third central moment of $\hat{t}$ and b is a nearly unbiased estimator for

$$B = \frac{E[v(\hat{t} - t)]}{V}. \tag{1.4}$$

Often, $b = m_3/v$, and δ collapses to

$$\delta = \left(\frac{1}{6} + \frac{z^2}{3}\right)\frac{m_3}{v}. \tag{1.5}$$

We are interested here in one-sided coverage intervals with good randomization-based properties. In particular, if the sample sizes were large enough (whatever that means), then the Wald intervals in equation (1.1) with v computed using probability-sampling principles should obtain. With that in mind, observe that if $|\delta|$ is of a smaller asymptotic order than v, then equation (1.2) has the same large-sample randomization-based properties as (1.1). What equation (1.2) does, if anything, is “speed up the asymptotics” required by equation (1.1).

Section 2 derives equation (1.2) by closely following arguments in Kott and Liu (2009) for the coverage interval of a proportion under a stratified simple random sample with large sampling fractions within all strata. Section 3 focuses on two special cases: population means and totals based on stratified simple random samples with at least three sampled units per stratum, and parameters based on stratified multistage samples when the first-stage sampling fractions can be ignored. Section 4 looks at potential applications at the Internal Revenue Service, where stratified simple random sampling is combined with three different estimators. Section 5 describes an empirical investigation based on those applications. Section 6 offers some concluding remarks.

2. Deriving the Intervals

We begin with an Edgeworth expansion for $\hat{t}$:

$$\Pr\left(\frac{\hat{t} - t}{\sqrt{V}} \le z\right) = \Phi(z) + (1/6)(1 - z^2)\varphi(z)\tau + O(1/n), \tag{2.1}$$

where $\varphi(z)$ is the probability density function (*pdf*) of the standard normal distribution, and

$$\tau = \frac{E[(\hat{t}-t)^3]}{E[(\hat{t}-t)^2]\ ^{3/2}} = \frac{M_3}{V^{3/2}}$$

is the skewness of $\hat{t}$. Under mild conditions for the sampling design and the underlying population, which will we assume to hold, V/t^2 is O(1/*n*), while M_3/t^3 is O(1/n^2). Thus, τ is O(1/$n^{1/2}$).

Strictly speaking, Edgeworth expansions only apply for continuous distributions, while $\hat{t}$ may have a discrete distribution. We are ignoring an "oscillatory term" of the probability function of $\hat{t}$ which differentiates it from a continuously distributed approximation, call it $\hat{t}$ sharing its first two moments. As a result of ignoring the oscillatory term, one of our α-percent intervals will (at best) cover *t* in α-percent of all possible samples on average across small ranges of potential values for *t* rather than covering *t* in *at least* α-percent of all possible samples for *each t*. This is why we use the term "coverage interval" to describe the intervals in equation (1.2) rather than the more common "confidence interval."

Letting $a = (1/6)(1-z^2)M_3/V^{3/2}$, and employing the Taylor-series expansion, $\Phi(z-a) = \Phi(z) - a\varphi(z) + O(1/n)$, we can replace *z* in equation (2.1) with *z* – *a* and write:

$$\Pr\left(\frac{\hat{t}-t}{\sqrt{V}} \le z - (1/6)(1-z^2)\frac{M_3}{V^{3/2}} \right) = \Phi(z) + O(1/n)$$

or equivalently

$$\Pr\left(\frac{\hat{t}-t}{\sqrt{V}} \le z - (1/6)(1-z^2)\frac{M_3}{V^{3/2}} + O(1/n) \right) = \Phi(z).$$

Dropping the O(1/*n*) term, this implies

$$\Pr\left(\hat{t}-t \le z\sqrt{V} - (1/6)(1-z^2)\frac{M_3}{V} \right) \approx \Phi(z),$$

$$\Pr\left(\left[\hat{t}-t + (1/6)(1-z^2)\frac{M_3}{V} \right]^2 \le z^2V \right) \approx \Phi(z),$$

and

$$\Pr\left(\; t-\hat{t} \;\; ^{2} - (1/3)(1-z^2)\frac{M_3}{V} \;\; t-\hat{t} \;\; \le z^2V \right) \approx \Phi(z).$$

To use this last equation in constructing coverage intervals, we will need (among other things) to estimate the unknown *V*. Here we follow Andersson and Nerman (2000) and replace *V* on the right-hand side, not with *v* as one might expect, but with the much more efficient *idealized variance estimator*:

$$\tilde{v} = v - B\ \hat{t} - t\ \ , \tag{2.2}$$

where $B = \dfrac{Cov\ \ v, \hat{t}}{V}$.

Unfortunately, $\tilde{v}$, although having the minimum variance of an estimator for *V* in the form $v - \lambda\ \hat{t} - t$, can still possess an error that should not, strictly speaking, be ignored in this context (formally, $(\tilde{v} - V)/t^2$ is usually $O_P(1/n^{1/2})$). Nevertheless, we will ignore it for the time being.

Substituting z^2V by $z^2\tilde{v}$ and rearranging brings us to

$$\Pr\left(\ t - \hat{t}\ \ ^2 - \left[\tfrac{1}{3}(1-z^2)\frac{M_3}{V} + z^2B\right]\ t - \hat{t}\ \ - z^2v \le 0\right) \approx \Phi(z).$$

By solving the quadratic equation, we then have

$$\Pr\left(t - \hat{t} \le \frac{\tfrac{1}{3}(1-z^2)\dfrac{M_3}{V} + z^2B + \sqrt{\left[\tfrac{1}{3}(1-z^2)\dfrac{M_3}{V} + z^2B\right]^2 + 4z^2v}}{2}\right) \approx \Phi(z)$$

or

$$\Pr\left(t - \hat{t} \ge \frac{\tfrac{1}{3}(1-z^2)\dfrac{M_3}{V} + z^2B - \sqrt{\left[\tfrac{1}{3}(1-z^2)\dfrac{M_3}{V} + z^2B\right]^2 + 4z^2v}}{2}\right) \approx \Phi(z).$$

Estimating $\Delta = (1/6)(1-z^2)(M_3/V) + (z^2/2)B$ by

$$\delta = (1/6)(1-z^2)(m_3/V) + (z^2/2)b, \tag{2.3}$$

and the intervals in equation (1.2) result. Under mild conditions, Δ is $O(1/n^{1/2})$, while the bias of δ an estimator for Δ is usually $O(1/n)$.

3. Two Special Cases

3.1 Means and Totals under a Stratified Simple Random Sample

Let the population be divided into *H* mutually exclusive strata, U_h denote the population of stratum *h*, $h = 1, \ldots, H$, each containing N_h units, and $U = \sum_{h=1}^{H} U_h$. Let $\mathbb{S}_h$ be the simple random sample of n_h units selected without replacement from U_h, and $\mathbb{S} = \sum_{h=1}^{H} \mathbb{S}_h$. Let *t* be the finite-population total and $\hat{t}$ be the estimate from the sample. Suppose we are interested in constructing one-sided coverage intervals for a finite-population total or mean based on a stratified simple random sample. The former can be expressed as $t = \sum_{h=1}^{H} \sum_{k \in U_h} y_k$, where *y* is the variable of interest. The corresponding population mean is $\bar{Y} = t/N = \sum_{h=1}^{H} W_h \bar{Y}_h$, where $N = \sum_H N_h$, $W_h = N_h / N$ and $\bar{Y}_h = N_h^{-1} \sum_{U_h} y_k$.

An unbiased estimator for the finite-population total *T* using probability-sampling principles is

$$\hat{t} = N\bar{y}\ , \tag{3.1}$$

where $\bar{y} = \sum_{h=1}^{H} W_h \bar{y}_h$ and $\bar{y}_h = n_h^{-1} \sum_{\mathbb{S}_h} y_k$.

When every n_h is at least 3, it is a simple matter to construct one-sided coverage intervals for $\bar{y}$ based entirely on probability-sampling principles using equation (1.2). One sets

$$v = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sum_{k \in \mathbb{S}_h} (y_k - \bar{y}_h)^2}{n_h (n_h - 1)},$$

$$m_3 = \sum_{h=1}^{H} W_h^3 \left(1 - \frac{n_h}{N_h}\right)\left(1 - \frac{2n_h}{N_h}\right) \frac{\sum_{k \in \mathbb{S}_h} (y_k - \bar{y}_h)^3}{n_h (n_h - 1)(n_h - 2)}, \text{ and} \tag{3.2}$$

$$\text{b} = \frac{\sum_{h=1}^{H} W_h^3 \left(1 - \frac{n_h}{N_h}\right) \frac{\sum_{k \in \mathbb{S}_h} (y_k - \bar{y}_h)^3}{n_h (n_h - 1)(n_h - 2)}}{v}.$$

The first two are unbiased estimators for the second and third central moments of $\bar{y}$, while *b* is a consistent estimator for *B* under mild conditions.

One-sided coverage intervals for t are constructed conformably. They are simply N times the analogous intervals for $\bar{y}$. It is easy to see that only when all the $2n_h/N_h$ are small enough to be ignored does equation (1.3) collapse into equation (1.5) for all intents and purposes.

3.2 A Parameter under a Stratified Multistage Sample

In this subsection, we will consider a coverage interval for a parameter t based on stratified multistage sample when a nearly unbiased estimator for that parameter can be put in the form:

$$\hat{t} = \sum_{h=1}^{H} \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{t}_{hi},$$

where there are n_h primary sampling units (PSU's) in stratum h, and each $\hat{t}_{hi}$ is a nearly unbiased estimator for the same value. The parameter t may be a model parameter or a finite-population parameter. In the latter case, we are often assuming that the PSU's were selected using probability-sampling principles but with replacement.

As an example of what we mean, consider the ratio estimator,

$$\hat{t} = \sum^{H} \sum^{n_h} \sum_{k \in \mathbb{S}_{hi}} w_k y_k \Big/ \sum^{H} \sum^{n_h} \sum_{k \in \mathbb{S}_{hi}} w_k x_k,$$

where w_k is the survey weight associated with element k, and $\mathbb{S}_{hi}$ is the set of sampled elements in PSU i of stratum h. Here,

$$\hat{t}_{hi} = n_h \sum_{k \in \mathbb{S}_{hi}} w_k y_k \Big/ \sum^{H} \sum^{n_g} \sum_{k \in \mathbb{S}_{gi}} w_k x_k.$$

A univariate component of an estimated regression coefficient can also be put into this form.

When all $n_h \geq 3$, the following linearization estimators can be used in equations (1.2) and (1.3):

$$v = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{(e_{hi} - \bar{e}_h)^2}{n_h(n_h - 1)}, \quad m_3 = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \frac{(e_{hi} - \bar{e}_h)^3}{n_h(n_h - 1)(n_h - 2)}, \quad \text{and } b = \frac{m_3}{v},$$

where the e_{hi} are linearized sample residuals such that $e_{hi} \approx u_{hi}$, where $\hat{t} - t \approx \sum\sum u_{hi}$, the u_{hi} – and thus the e_{hi} – are nearly independent random variables with a common mean within each stratum, and $\bar{e}_h = n_h^{-1} \sum_{i=1}^{n_h} e_{hi}$. For the ratio estimator,

$$u_{hi} = n_h \sum_{k \in \mathbb{S}_{hi}} w_k (y_k - t x_k) \Big/ \sum_{g=1}^{H} \sum_{j=1}^{n_g} \sum_{k \in \mathbb{S}_{gi}} w_k x_k,$$

and

$$e_{hi} = \sum\nolimits_{k \in \mathbb{S}_{hi}} w_k (y_k - \hat{t}x_k) \Big/ n_h \sum\nolimits^{H} \sum\nolimits^{n_g} \sum\nolimits_{k \in \mathbb{S}_{gi}} w_k x_k .$$

When one or more first-stage strata have less than three sampled PSU's, some collapsing of strata will be necessary to compute m_3 and b. Putting asterisks on the collapsed strata and counts (which affects the definition of the e_{hi}*) we have these replacements to use in equations (1.2) and (1.3):

$$v* = \sum_{h=1}^{H*} \sum_{i=1}^{n_h*} \frac{(e_{hi} * - \overline{e}_h *)^2}{n_h * (n_h * -1)}, \quad m_3 * = \sum_{h=1}^{H*} \sum_{i=1}^{n_h*} \frac{(e_{hi} * - \overline{e}_h *)^3}{n_h * (n_h * -1)(n_h * -2)}, \text{ and } \quad \mathrm{b}* = \frac{m_3 *}{v*}.$$

Recall that using these, by themselves, does not affect the large-sample randomization-based properties of the coverage intervals.

4. Applications

In this section, we discuss potential uses of our improved one-sided intervals at the Internal Revenue Service (IRS), contrasting our intervals with the conventional Wald intervals.

The *Meals and Entertainment (M&E) Study* is an example in tax situation where coverage intervals are needed. Many companies incur significant meals and entertainment (M&E) expense as part of their normal cost of doing business. Examples of M&E expenses include costs like travel reimbursements, client entertainments, and employee group meetings or events. Some of these expenses are 100% tax deductible. For tax purpose, companies would like to know the total amount of expenses that are 100% tax deductible. The universe of these expenses may include so many items that it would be too costly to review each of them. To facilitate accounting of fully deductible M&E, the IRS issued revenue procedure 2004-29 to allow the use of statistical sampling to account for such expenses (Batcher, 2004). An M&E universe is a company's list of expenses, and the sample is typically a stratified simple random sample with expense amount as the stratifier. The sampled expenses are reviewed, and the 100% deductible expenses are identified. If x is the original expense amount, and y is the 100% tax deductible amount, then the value of y is either x (if the expense is classified as 100% deductible) or 0 (if the expense is not qualified as 100% deductible). Overall, 30% - 60% of expenses are qualified for the 100% deductible. Based on the sample results, the total amount of the 100% deductible expenses is estimated, and the lower bound of a one-sided 95% Wald interval is constructed (which is the same as the lower bound of a 90% two-sided Wald interval). The distribution of x is skewed, however, so that the distribution of y is highly skewed. The use of the Wald interval in this context may be dubious.

The *IRS Individual Income Tax Underreporting Study* is an example where the qualified amount is part of the expense amount. The IRS created the National Research Program (NRP) in 2000 to champion the agency's efforts to measure taxpayer compliance. The first NRP reporting compliance study sampled more than 46,000 tax returns from tax year 2001 Forms 1040. The sample design for the tax year 2001 study is a stratified random sample design that includes 30 strata based on examination class, adjusted gross income, business receipts, and other measures. A new individual underreporting study covering tax years 2006 to 2008 with approximately 13,000 tax returns each year is in progress.

The new study has a stratified random sample design that includes 58 strata based on examination class, filing status, the presence of the form 2106 and other measures. The audit results of sample returns are used to make population estimates. IRS uses a number of different statistics to measure taxpayer compliance. One of the compliance measures is the Net Misreported Amount (NMA). This is the total difference between the amount reported and the amount that should have been reported for deduction items. Let x be the reported amount and y be the deduction amount that should have been reported for an item, then $d = x - y$ is the misreported or error amount. Here, y is less than x, that is, $0 \leq y \leq x$. In addition to the estimate of the total error amount, the upper coverage bound of the total error amount is of interest to the IRS individual income tax underreporting study.

The Research and Development (R&D) Study estimates the dollar amount that qualifies as research and development for tax purposes. The sampling units can be employees, projects or locations. For example, if the sample selections are conducted at the individual employee level, then the sample is often stratified by *Tier* and W-2 wages. The Tiers are employee groups based on the expected qualification rate for their job title. Within each Tier, employees are grouped according to their W-2 wages, which can aid in improving precision by controlling the variability of the qualifying dollars. The randomly selected individuals are interviewed to determine the amount of qualifying activities the individual had for the research and experimental expenditures under Section 174 of the Internal Revenue Code. Because the population tends to be highly variable, a large sample is needed to achieve acceptable precision of the point estimate. Unfortunately, a large sample can be very costly. Consequently, the lower bound of the qualifying amount at 90% coverage level is often used instead. For each sampled unit, the qualifying percent varies from 0% to 100%. If x is the original expense amount, then the qualifying amount y is somewhere between 0 and x with a large number of 0's. As a result, the distribution of y is highly skewed.

The IRS permits the use of three randomization-based estimators given a stratified simple random sample: the expansion, (separate) ratio, and difference estimators. The notations of lower and upper bounds (intervals) are given in Table 1. The subscripts denote the estimation methods, and the superscript '*a*' stands for the adjusted method using our proposed approach. Some additional notations are defined as follows.

Table 1: Notations of Coverage Bounds

Estimation Method	Wald Approach	Proposed Approach
Expansion	L_E and U_E	L_E^a and U_E^a
Ratio (Separate)	L_R and U_R	L_R^a and U_R^a
Difference	L_D and U_D	L_D^a and U_D^a

The Wald coverage intervals for the expansion estimator are

$$L_E = \hat{t} - z\sqrt{v} \quad or \quad U_E = \hat{t} + z\sqrt{v} \ ,$$

where $\hat{t} = N\sum_{h=1}^{H} W_h \bar{y}_h$, $v = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$ and $s_h^2 = \frac{\sum_{k\in\mathbb{S}_h}(y_k - \bar{y}_h)^2}{n_h - 1}$.

Our improved intervals have the form:

$$L_E^a = \hat{t} + \delta - \sqrt{z^2 v + \delta^2} \quad or \quad U_E^a = \hat{t} + \delta + \sqrt{z^2 v + \delta^2} \quad ,$$

where $\delta = \frac{1}{v}\sum_{h=1}^{H} W_h^3 \left\{ \frac{1-z^2}{6}\left(1 - \frac{2n_h}{N_h}\right) + \frac{z^2}{2}\right\}\left(1 - \frac{n_h}{N_h}\right)\frac{s_h^3}{n_h^2} g_h$, and $g_h = \frac{n_h \sum_{k\in\mathbb{S}_h}(y_k - \bar{y}_h)^3}{(n_h - 1)(n_h - 2)s_h^3}$.

Here δ is based on equations (2.3) and (3.2), and g_h is the estimated skewness of y for stratum h.

The Wald coverage intervals for the ratio estimator can be expressed as

$$L_R = \hat{t}_R - z\sqrt{v_R} \quad or \quad U_R = \hat{t}_R + z\sqrt{v_R} \ ,$$

where $\hat{t}_R = \sum_{h=1}^{H} \hat{R}_h t_{xh}$, $\hat{R}_h = \frac{\bar{y}_h}{\bar{x}_h}$, and t_{xh} is the total of x in stratum h,

$v_R = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_{Rh}^2}{n_h}$, $s_{Rh}^2 = c_h^2 s_{Rh0}^2$, $c_h = \frac{t_{xh} / N_h}{\bar{x}_h}$, and $s_{Rh0}^2 = \frac{\sum_{k\in\mathbb{S}_h}(y_k - \hat{R}_h x_k)^2}{n_h - 1}$.

We are using the weighted-residual-variance estimator for v_R (Särndal et al., 1989) because it has been shown to have better coverage properties than the more standard alternative where c_h is replaced by 1. Our improved intervals are

$$L_R^a = \hat{t}_R + \delta_R - \sqrt{z^2 v_R + {\delta_R}^2} \quad or \quad U_R^a = \hat{t}_R + \delta_R + \sqrt{z^2 v_R + {\delta_R}^2} \quad ,$$

where $\delta_R = \frac{1}{v_R}\sum_{h=1}^{H} W_h^3 \left\{ \frac{1-z^2}{6}\left(1 - \frac{2n_h}{N_h}\right) + \frac{z^2}{2}\right\}\left(1 - \frac{n_h}{N_h}\right)\frac{s_{Rh}^3}{n_h^2} g_{Rh}$,

and $g_{Rh} = {c_h}^3 \frac{n_h \sum_{k\in\mathbb{S}_h}(y_k - \hat{R}_h x_k)^3}{(n_h - 1)(n_h - 2)s_{Rh}^3}$.

Note that we also apply the correction term c_h in estimating the skewness in stratum h.

The Wald coverage intervals for the difference estimator are

$$L_D = \hat{t}_D - z\sqrt{v_D} \quad or \quad U_D = \hat{t}_D + z\sqrt{v_D} \;,$$

where $\hat{t}_D = t_x + \sum_{h=1}^{H} N_h \bar{d}_h$, $\bar{d}_h = n_h^{-1} \sum_{\mathbb{S}_h} d_k$, $d = y - x$, $v_D = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{Dh}^2}{n_h}$ and $s_{Dh}^2 = \frac{\sum_{k \in \mathbb{S}_h} (d_k - \bar{d}_h)^2}{n_h - 1}$. Here, $\bar{d}_h$ is the sample mean of variable d in stratum h. Our alternatives are

$$L_D^a = \hat{t}_D + \delta_D - \sqrt{z^2 v_D + \delta_D{}^2} \quad or \quad U_D^a = \hat{t}_D + \delta_D + \sqrt{z^2 v_D + \delta_D{}^2} \;,$$

where $\delta_D = \frac{1}{v_D} \sum_{h=1}^{H} W_h^3 \left\{ \frac{1 - z^2}{6} \left(1 - \frac{2n_h}{N_h}\right) + \frac{z^2}{2} \right\} \left(1 - \frac{n_h}{N_h}\right) \frac{s_{Dh}^3}{n_h^2} g_{Dh}$

and $g_{Dh} = \frac{n_h \sum_{k \in \mathbb{S}_h} (d_k - \bar{d}_h)^3}{(n_h - 1)(n_h - 2) s_{Dh}^3}$.

5. Simulations

In this section, we compare the coverage intervals described in Section 4. A population of 10,020 values of x is first generated from the highly skewed lognormal distribution with parameters μ=8, σ=1, shown in Figure 1 (these are the parameters of the normal distribution from which the lognormal is derived). The largest 20 values are dropped from the subsequent analysis because these values would likely be treated as certainties in practice. The remaining population is divided into five strata using equal dollar amount of x . Table 2 summarizes the population.

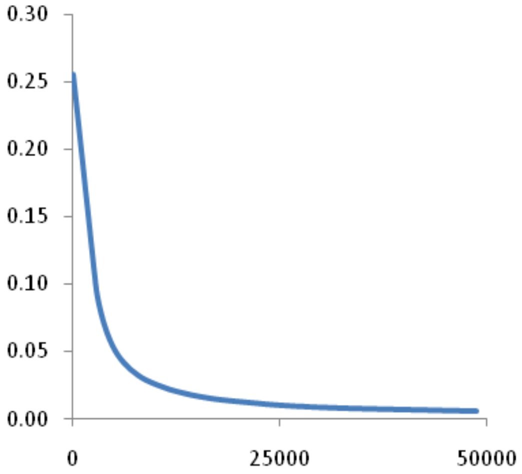

Figure 1. Population Distribution of *x*

Table 2: Summary of the Simulation Population

Stratum h	**Range of x**		**Population Size** N_h	**Total of** x T_x
	Minimum	**Maximum**		
1	38.2	3,448.8	5,563	9,690,057
2	3,449.3	6,139.6	2,085	9,692,312
3	6,141.5	10,164.0	1,244	9,694,484
4	10,170.0	17,914.8	733	9,675,960
5	17,925.2	48,669.2	375	9,710,327
Total			**10,000**	**48,463,140**

We create eight setting of values for a stratum-specific parameter p_h shown in Table 3. This parameter corresponds to the "probability" that the variable of interest (to be labeled "*y*") is zero. Settings 1 – 4 represent rare events in all strata with some variations in the p_h values. Settings 6 and 7 represent not rare events overall and have wide variations in p_h values among strata. Setting 7 and 8 have high values of p_h in all strata.

Table 3: p_h Settings for the Simulation

p_h **Setting**	**Stratum Qualifying Proportions** $(p_1,\ p_2,\ p_3,\ p_4,\ p_5)$	**Po1pulation total of** y**,** T	**Ratio,** $\frac{T}{T_x}$
1	0.10, 0.08, 0.05, 0.03, 0.02	2,706,178	5.6%
2	0.02, 0.03, 0.05, 0.08, 0.10	2,738,314	5.7%
3	0.20, 0.15, 0.10, 0.10, 0.05	5,812,281	12.0%
4	0.05, 0.10, 0.10, 0.15, 0.20	5,891,010	12.2%
5	0.10, 0.30, 0.50, 0.70, 0.90	24,314,679	50.2%
6	0.90, 0.70, 0.50, 0.30, 0.10	24,300,641	50.1%
7	0.90, 0.92, 0.95, 0.97, 0.98	45,785,287	94.5%
8	0.98, 0.97, 0.95, 0.92, 0.90	45,775,904	94.5%

We set n_h = 30 in every stratum. We then draw 1,000 stratified random. Let x_{hi} be the i-th unit in stratum h from the randomly generated values of x. We generate y values using two models:

Model One:
$$y_{hi} = \begin{cases} x_{hi}, & \text{if } i \le N_h p_h \\ 0, & \text{otherwise} \end{cases} , \tag{5.1}$$

and

Model Two:
$$y_{hi} = \begin{cases} u_{hi} x_{hi}, & \text{if } i \le N_h p_h \\ 0, & \text{otherwise} \end{cases} , \tag{5.2}$$

where $0 \le u_{hi} \le 1$ is the random variable from the uniform(0, 1) distribution. Model One of equation (5.1) roughly mimics the data for the M&E study where a unit is either not qualified ($y = 0$) or fully qualified ($y = x$), and Model Two roughly mimics the data in the other two applications where a unit can be partially qualified ($0 \le y \le x$).

We calculate the coverage bound of total amount of y for each of 1000 samples. A constraint is put on the calculation such that the lower bound should not be smaller than the sample total $\sum_{k\in\mathbb{S}} y_k$, and the upper bound should not be larger than $T_x - (\sum_{k\in\mathbb{S}} x_k - \sum_{k\in\mathbb{S}} y_k)$. We use the *coverage rate* (CR) and *average distance* (AD) to compare different methods. The coverage rate is the fraction of samples whose intervals cover the true population value. The average distance is a measure of how close the boundary of the coverage interval is to the true value. It is defined as the mean of the absolute distance of the lower (or upper) bound from the true population value divided by the true value:

$$AD = \frac{1}{T}\frac{1}{1000}\sum_{j=1}^{1000} | B_j - T | , \qquad (5.3)$$

where B_j is the coverage bound calculated from the sample B_j, $j = 1, 2, \cdots, 1000$.

Figures 2 compare the coverage rates (CR) and average distance (AD) of the lower bound for the eight settings of different p_h values and different estimation methods. The nominal level is 95%. The coverage plots show that our proposed improvement brings the Wald coverage closer to the nominal level in almost all situations. When the qualifying rates are small (settings 1 – 4), the Wald lower bound for the expansion and ratio estimators tend to over-cover, while our proposed improvement adjusts coverage levels down to be close to the 95% nominal level. In addition, our method has shorter average distances. When coupled with the difference estimator, the Wald interval performs well in settings 1 – 4. Its properties are very similar to those of our proposed alternative. In these settings, however, average distances are much longer than those produced by the expansion and separate estimators. When the qualifying rates are not extreme (settings 5 and 6), our proposed intervals are not very different from the Wald no matter the estimator. When the qualifying rates are large (settings 7 and 8), our improved intervals have coverage rates closer to the nominal level for the ratio and difference estimators.

Figure 3 displays the same comparisons for upper bounds. Note that the scale of AD in Figure 3 is larger than that in Figure 2 in order to fit the much larger average distance of upper bounds using difference estimators. The conclusions drawn from lower bounds tend to hold again except that the skewness is in the opposite direction. The upper bounds have under-coverage in settings where the lower bounds have over-coverage. The relative sizes of the average distances are also reversed.

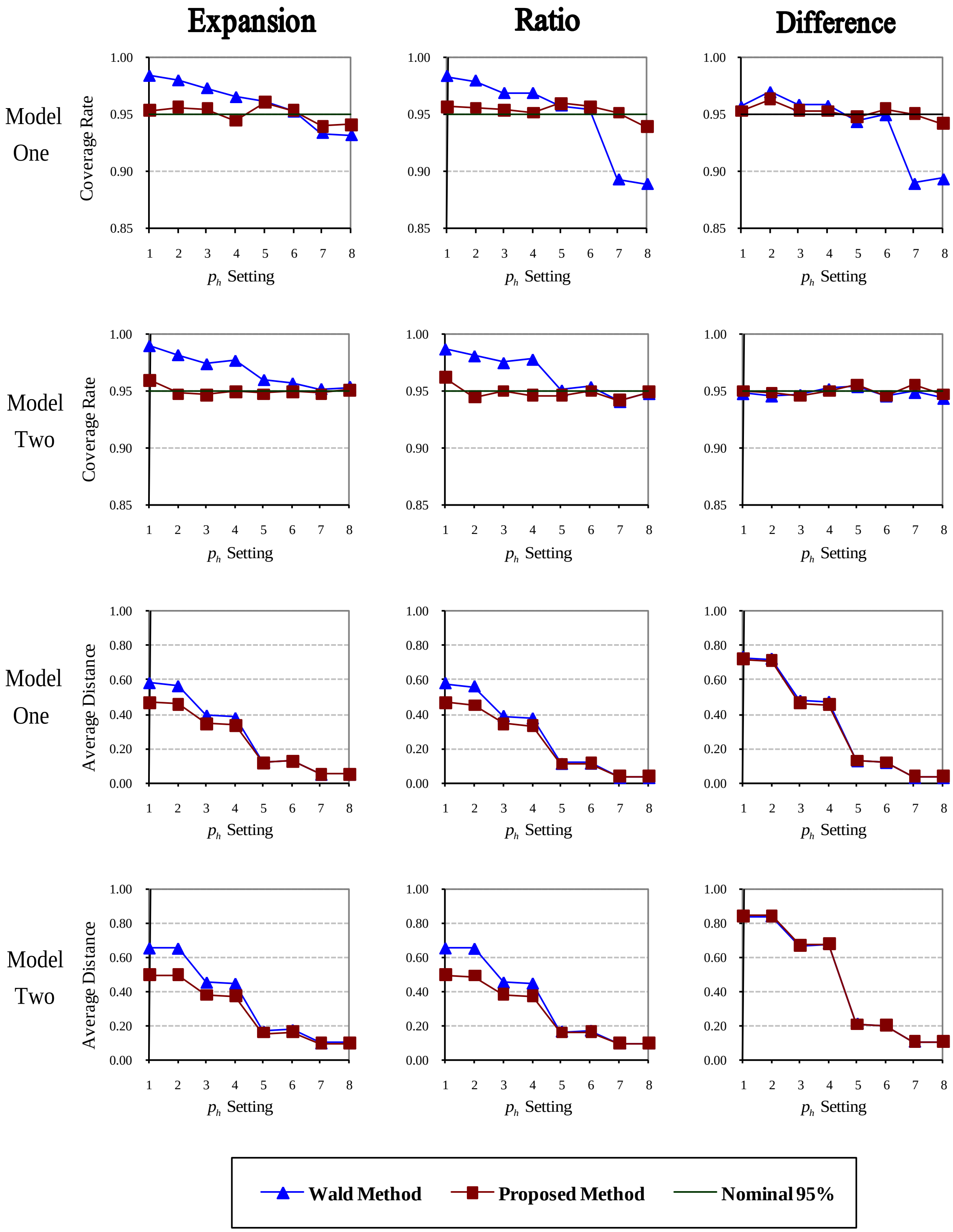

Figure 2: Coverage Rate and Average Distance of Lower Bound at 95%

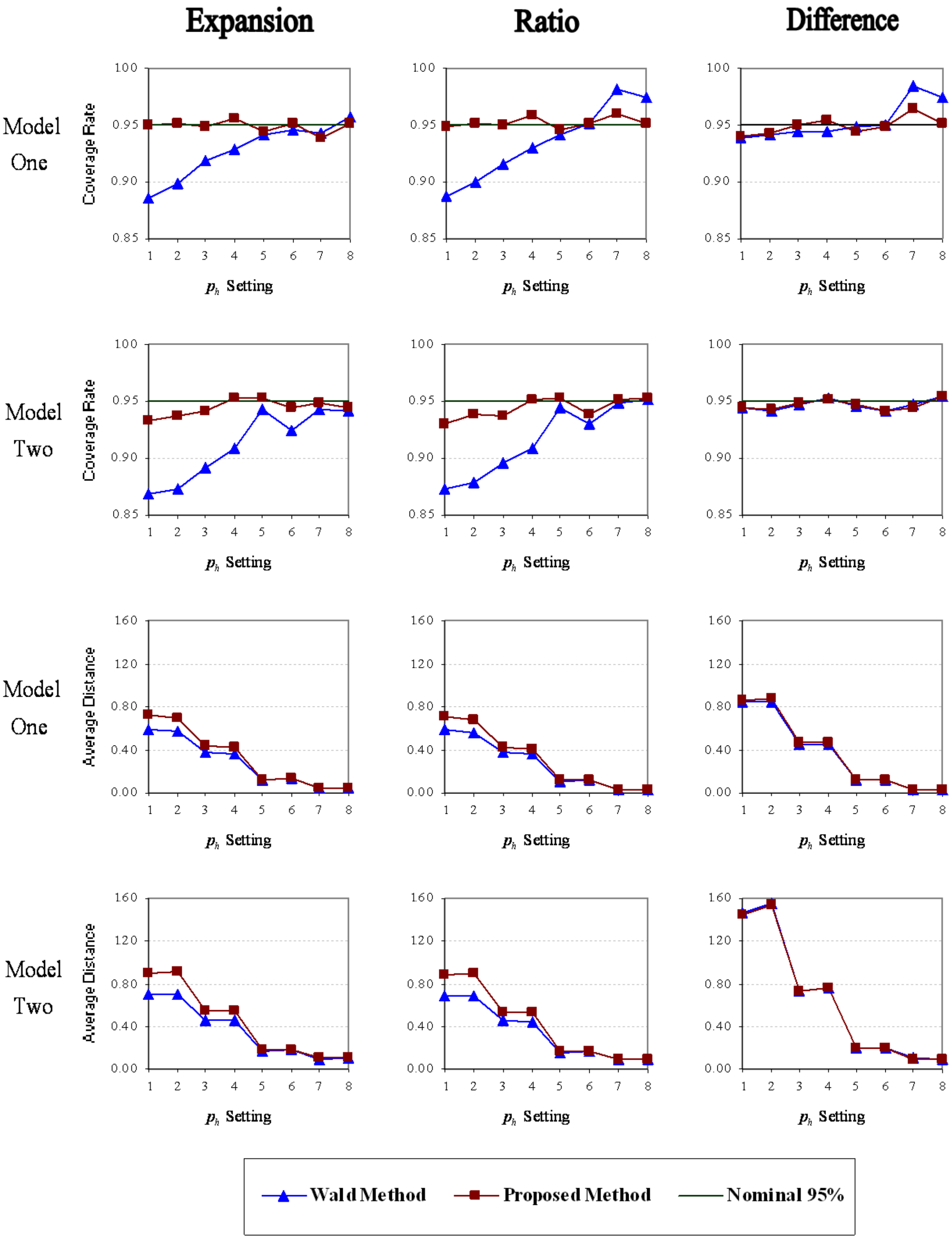

Figure 3: Coverage Rate and Average Distance of Upper Bound at 95%

6. Conclusions

We have proposed a relatively simple improvement on one-sided Wald coverage intervals for a parameter estimated with survey data. We have shown how our improvement can be applied broadly, demonstrating it empirically in the important special case of a stratified simple random auditing sample. For alternative, and more complex, treatments of auditing samples, see Bimpeh (2008) and the Panel on Nonstandard Mixtures of Distributions (1989).

There are limitations to our methodology. It does not produce *confidence* intervals, because it does not protect against the oscillations in discrete distributions. Our intervals, at best, cover the true parameter at the designated level (e.g. 95% of the time) on average rather than (almost) always.

The methodology simply speeds up the asymptotics in coverage interval construction. The key statistic is the skewness (τ) of the parameter estimate. The skewness converges to zero asymptotically, but we live in a finite world. When this parameter is nonzero, our intervals will improve on the Wald. Nevertheless, when the skewness is too large − say, greater than ½ in absolute value − our intervals may not be very reliable (because we assumed τ^2 was ignorably small in our Edgeworth approximation, *i.e*, equation (2.1)).

Another limitation of our method is that, although it increases the efficiency of the implicitly estimated variance of the parameter estimate, it does not remove all random noise from variance estimation, even asymptotically. How much of a practical problem this is has yet to be seriously explored.

Acknowledgements

The authors express a special note of thanks to Mary Batcher and Eric Falk at Ernst & Young LLP for the useful discussion on applications in the tax situation.

References

Anderson, P.G. and Nerman, O. (2000). A Balanced Adjusted Confidence Interval Procedure Applied to Finite Population Sampling. *Presented at the International Conference of Establishment Surveys, Buffalo, NY.*

Batcher, M. (2004). Statistical Sampling in Tax Filings: New Confirmation from the IRS. *The Tax Executive*. May 1, 2004.

Bimpeh, Y. (2008). Statistical Modeling and Inference for Financial Auditing. *PhD thesis, Dublin City University.*

Kott, P. S. and Liu, Y. K. (2009). One-sided Coverage Intervals for a Proportion Estimated from a Stratified Simple Random Sample. *Internal Statistical Review* 77, 2, 251–265.

Panel on Nonstandard mixtures of Distributions (1989). Statistical Models and Analysis in Auditing. *Statistical Science*, Vol.4, No.1, pp 2-33.

Särndal, C- E, B. Swensson, and J. Wretman (1989). The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of a Finite Population total. *Biometrika* 76, 527-537.