

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

Phillip S. Kott  
RTI International

<http://www.nass.usda.gov/research/OD5.htm>

Yan K. Liu  
Internal Revenue Service

# Outline

*One-sided Wald **Coverage** Intervals*

*Improved Intervals*

*use an Edgeworth expansion*

*and a better implicit variance estimator*

*An Empirical Exploration*

*Concluding Remarks*

# *One-Sided Wald Intervals*

Suppose  $\hat{t}$  is a nearly unbiased estimator for  $t$ .

A one-sided 95%-percent Wald coverage interval for  $t$  is

$$t \leq \hat{t} + 1.645\sqrt{v} \quad (\text{the upper bound})$$

or

$$t \geq \hat{t} - 1.645\sqrt{v} \quad (\text{the lower bound}),$$

where  $v$  is an estimator for  $V$  the variance of  $\hat{t}$ .

# *One-Sided Wald Intervals*

When the sample size is *large enough*, both inequalities hold for roughly 95% percent of the samples.

In practice, however, the sample size will often not be 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 or because  $v$  is not a very stable estimator of variance.

# *Improved One-Sided Intervals*

In either of those situations, we propose the following one-sided intervals:

$$t \leq \hat{t} + \delta + \sqrt{z^2 v + \delta^2}$$

or

$$t \geq \hat{t} + \delta - \sqrt{z^2 v + \delta^2},$$

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

$z = 1.645$  (or more generally,  $\Phi^{-1}(\alpha)$  for the  $\alpha^{\text{th}} \times 100\%$  interval),

# *Improved One-Sided Intervals*

$m_3$  is a nearly unbiased estimator for  $\hat{t}$ 's third central moment, (this part comes from the Edgeworth expansion)

$b$  is a nearly unbiased estimator for  $B = \frac{\text{Cov}(v, \hat{t})}{V}$   
(this part come from regressing  $v$  on  $\hat{t} - t$ )

Often,  $b = m_3/v$ , and  $\delta$  collapses to  $\delta = \left( \frac{1}{6} + \frac{z^2}{3} \right) \frac{m_3}{v}$ .

# Look ma, no models! (almost)

The parameter estimates can be randomization-based (except when there are less than three PSUs in a first-stage stratum), but they need not be.

We are modeling  $\hat{t}$  as a continuous random variable.

As the sample grows in size,  $\delta$  goes to zero faster than  $\sqrt{v}$ , so the usual asymptotic randomization properties hold.

# Empirical Exploration

We will look at stratified simple random samples with 30 draws in each of five large strata.

The population will be generated using audit-sample models.

We will observe the empirical coverages in 1000 samples of the expansion, separate-ratio, and difference estimators of a total.

# *One-sided Intervals for the Expansion Estimator*

$$L = \hat{t} + \delta - \sqrt{z^2 v + \delta^2} \qquad U = \hat{t} + \delta + \sqrt{z^2 v + \delta^2},$$

$$\text{where } \hat{t} = \sum_{h=1}^H N_h \bar{y}_h, \quad v = \sum_{h=1}^H N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h}, \quad s_h^2 = \frac{\sum_{k \in S_h} (y_k - \bar{y}_h)^2}{n_h - 1},$$

$$\delta = \frac{1}{v} \sum_{h=1}^H N_h^3 \left\{ \frac{(1 - z^2)}{6} \left(1 - \frac{2n_h}{N_h}\right) + \frac{z^2}{2} \right\} \left(1 - \frac{n_h}{N_h}\right) \frac{\sum_{k \in S_h} (y_k - \bar{y}_h)^3}{n_h (n_h - 1)(n_h - 2)}$$

## One-sided Intervals for the Difference Estimator:

Replace  $y_k$  in equations with  $y_k - x_k$ .

## for the Separate-Ratio Estimator:

When  $k$  is in stratum  $h$ , replace  $y_k$

with  $\frac{t_{xh}}{\hat{t}_{xh}} y_k$  for computing  $\hat{t}$ ,

and with  $\frac{t_{xh}}{\hat{t}_{xh}} \left( y_k - \frac{\bar{y}_h}{\bar{x}_h} x_k \right)$  for  $s_h^2$  and  $\delta$ .

# The Simulated Population: from a truncated lognormal ( $\mu = 8, \sigma = 1$ )

Largest 20 were removed (treated as a certainty stratum)

Stratum $h$	Range of $x$	$N_h$	$t_{xh}$
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
		10,000	48,463,140

# The $y$ -values were generated from two models

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

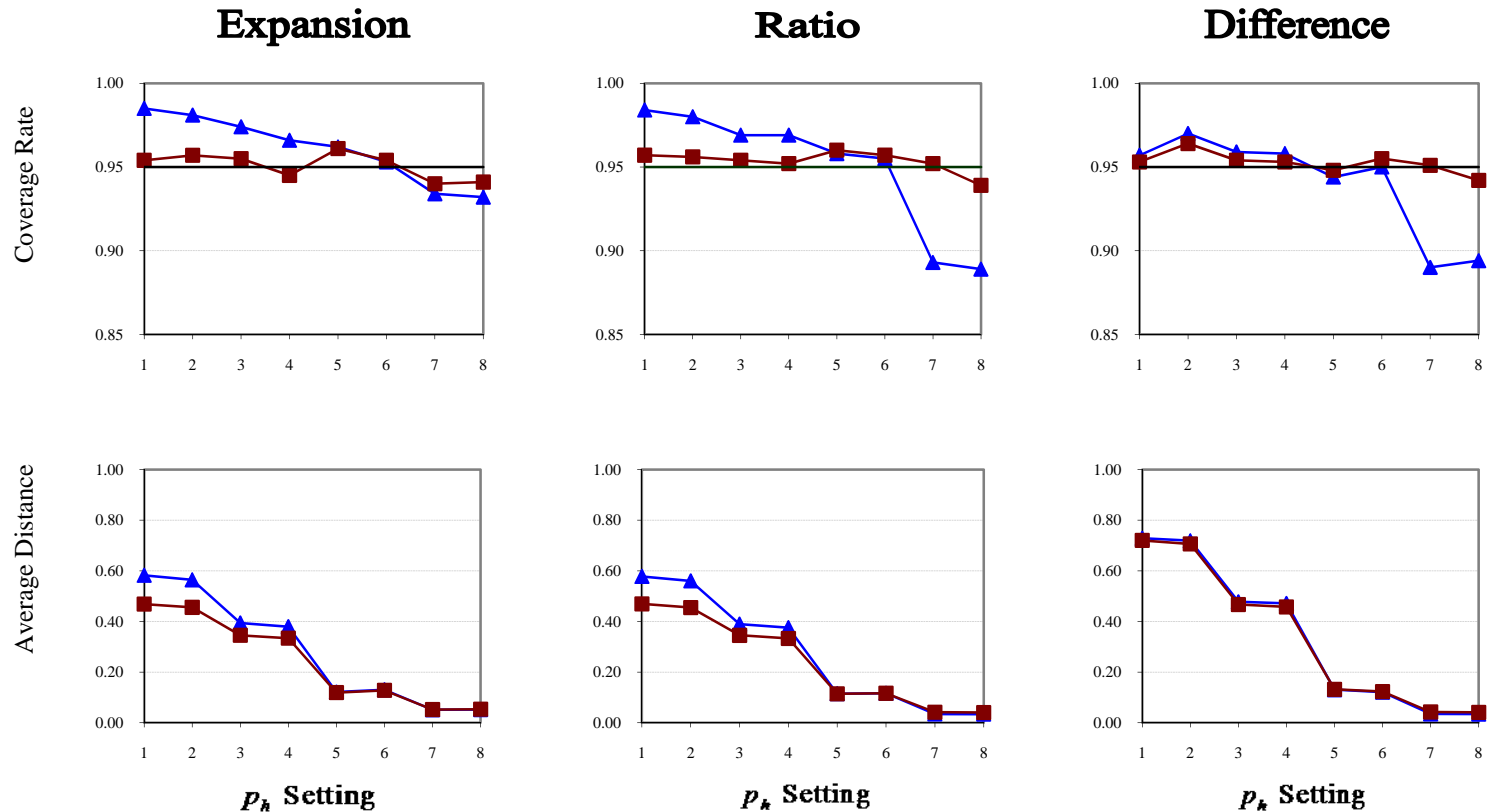
Model Two: 
$$y_{hi} = \begin{cases} u_{hi} x_{hi}, & \text{if } i \leq N_h p_h \\ 0, & \text{otherwise} \end{cases}$$

where each  $u_{hi}$  was generated from a uniform distribution, and the  $p_h$  (probabilities of being positive) have the following settings:

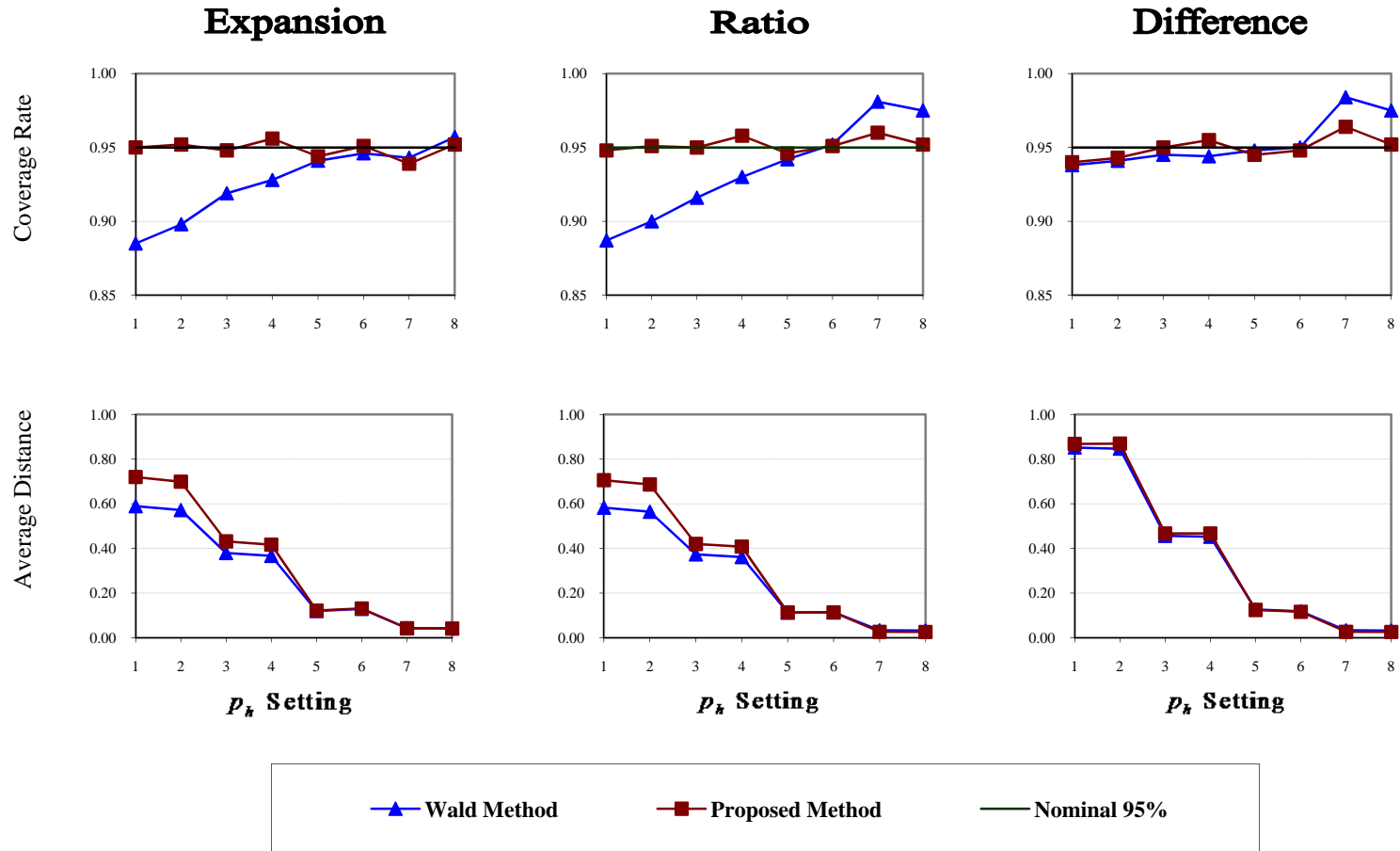
# Settings of Stratum Proportions

Setting	Stratum Positive Proportions ( $p_1, p_2, p_3, p_4, p_5$ )	$t_y$
1	0.10, 0.08, 0.05, 0.03, 0.02	2,706,178
2	0.02, 0.03, 0.05, 0.08, 0.10	2,738,314
3	0.20, 0.15, 0.10, 0.10, 0.05	5,812,281
4	0.05, 0.10, 0.10, 0.15, 0.20	5,891,010
5	0.10, 0.30, 0.50, 0.70, 0.90	24,314,679
6	0.90, 0.70, 0.50, 0.30, 0.10	24,300,641
7	0.90, 0.92, 0.95, 0.97, 0.98	45,785,287
8	0.98, 0.97, 0.95, 0.92, 0.90	45,775,904

# Lower Bound at 95%, Model One



# Upper Bound at 95%, Model One



# Concluding Remarks

Coverage intervals vs. confidence intervals  
(covering on average rather than with certainty)

The impact of finite population adjustment

Even when a model is needed (e.g., because there is only two PSUs per stratum), our intervals collapse asymptotically to Wald intervals.

# Concluding Remarks

## *Limitation 1*

Our method requires the skewness of  $\hat{t}$  (i.e.,  $M_3/V^{3/2}$ ) to be smallish in absolute value.

## *Limitation 2*

Using the residual of the randomization-based estimated variance ( $v$ ) regressed on the error ( $\hat{t} - t$ ) as the estimated variance in the pivotal  $((\hat{t} - t)/\tilde{v}^{1/2})$  reduces but does not remove the noise from the estimated variance.

# More Information

Visit RTI International in booth 513/612

NASS and IRS also have booths.