Some Effects of Measurement Error in the Auxiliary Variate on the Design-based Properties of Ratio Estimators of the Population Total

T. G. Gregoire   C. Salas

School of Forestry and Environmental Studies
Yale University
New Haven, CT, USA

SNS, Egilsstadir, Iceland, August 2008

Purpose

This study aimed to discern the effect of measurement error (ME) on the statistical properties of the ratio estimator of the population total,

$$\tau_y = \sum_{k=1}^{N} y_k$$

for some attribute, Y, in a population of discrete units.

With ratio estimation, use is made of an auxiliary variate, X.

See any book on statistical sampling of past 60 years for background.

How is ratio estimation better than HT?

If X is well and positively correlated with Y, more precise estimation is possible than with $\widehat{\tau}_{y\pi} = \sum_{k \in \mathcal{S}} \frac{y_k}{\pi_k}$.

To be feasible, X must be comparatively inexpensive to acquire for all elements of the population (well, at least τ_x). Hence the possibility of ME.

The effect of ME in X on the performance of the ratio estimator seems never to have been studied.

Study extent

We have looked at SRSwoR as the sampling design, no other design yet.

We examined additive ME only: if x_k denotes the value of the auxiliary variate for the kth unit of the population the error contaminated value is $x_k^* = x_k + \delta_k$.

It is x_k^* rather than x_k that is used in estimation of τ_y.

Ratio estimators of interest

There is no unique ratio estimator. We looked at three, which may be expressed in the absence of ME as

- $\widehat{\tau}_{y1} = \tau_x \widehat{R}$, where $\widehat{R} = \overline{y}/\overline{x}$;

Ratio estimators of interest

There is no unique ratio estimator. We looked at three, which may be expressed in the absence of ME as

- $\widehat{\tau}_{y1} = \tau_x \widehat{R}$, where $\widehat{R} = \overline{y}/\overline{x}$;
- $\widehat{\tau}_{y2} = \overline{r}\tau_x$ where $\overline{r}$ is the average ratio $r_k = y_k/x_k$ of the n units in the sample;

Ratio estimators of interest

There is no unique ratio estimator. We looked at three, which may be expressed in the absence of ME as

- $\widehat{\tau}_{y1} = \tau_x \widehat{R}$, where $\widehat{R} = \overline{y}/\overline{x}$;
- $\widehat{\tau}_{y2} = \overline{r}\tau_x$ where $\overline{r}$ is the average ratio $r_k = y_k/x_k$ of the n units in the sample;
- $\widehat{\tau}_{y3} = \widehat{\tau}_{y2} + \left(\frac{N-1}{N}\right)\left(\frac{n}{n-1}\right)\left(\widehat{\tau}_{y\pi} - \overline{r}\,\widehat{\tau}_{x\pi}\right)$,

  in which $\widehat{\tau}_{y\pi}$ and $\widehat{\tau}_{x\pi}$ are the HT estimators of τ_y and τ_x, respectively.

Properties in the absence of ME I

Both $\widehat{\tau}_{y1}$ (ratio of means) and $\widehat{\tau}_{y2}$ (mean of ratios) admit a design bias,

whereas $\widehat{\tau}_{y3}$ (introduced by Hartley and Ross in 1954 in *Nature*) adds a correction for the bias in $\widehat{\tau}_{y2}$ to yield a design-unbiased estimator of τ_y.

Ratio estimators in the presence of ME

Recall that δ_k is ME attached to kth unit, so that the measured value of the auxiliary variate is $x_k^* = x_k + \delta_k$.

The analog to $\widehat{\tau}_{y1}$ with x_k^* in place of x_k is

$$\widehat{\tau}_{y1}^* = \widehat{R}^* \tau_x^* = \widehat{\tau}_{y1} \left(1 + \frac{\mu_\delta}{\mu_x}\right) \Big/ \left(1 + \frac{\overline{\delta}}{\overline{x}}\right),$$

where $\widehat{R}^* = \overline{y}/\overline{x}^* = \widehat{R}/\left(1 + \frac{\overline{\delta}}{\overline{x}}\right)$, and where μ_δ and $\overline{\delta}$ are the average ME in the population and the sample, respectively.

Bias ratio

In upcoming *Biometrics* article we derive the ratio of the the ratio of the bias of $\widehat{\tau}^*_{y1}$ to that of $\widehat{\tau}_{y1}$, which is

$$\frac{B\left[\widehat{\tau}^*_{y1} : \tau_y\right]}{B\left[\widehat{\tau}_{y1} : \tau_y\right]} = \left(\frac{\mu_y \sigma_x^{*2} - C(x^*, y)(\mu_x + \mu_\delta)}{\mu_y \sigma_x^2 - C(x, y)\mu_x}\right) \Big/ \left(1 + \frac{\mu_\delta}{\mu_x}\right)^2 .$$

The utility of this expression is that it is independent of sample size, *n*.

It shows, also, that even when the population mean error, μ_δ, is identically zero, the bias of $\widehat{\tau}^*_{y1}$ is affected by the variability of measurement error.

Variance approximations

With no ME, the usual approximation to the variance of $\widehat{\tau}_{y1}$ is

$$V\left[\widehat{\tau}_{y1}\right] = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2_{rm},$$

where $\sigma^2_{rm} = \frac{1}{N-1}\sum_{\mathcal{P}}(y_k - Rx_k)^2$.

With ME, this becomes

$$V\left[\widehat{\tau}^*_{y1}\right] = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\sigma^{*2}_{rm},$$

where $\sigma^{*2}_{rm} = \frac{1}{N-1}\sum_{\mathcal{P}}\left(y_k - R\left(x_k + \delta_k\right)/\left(1 + \frac{\mu_\delta}{\mu_x}\right)\right)^2$.

In the article we derive expressions for the bias, bias ratio with:without ME, and variance with and without ME for $\widehat{\tau}_{y2}$ and for the variance of $\widehat{\tau}_{y3}$ with and without ME, inasmuch as it remains unbiased with additive ME.

Design-based considerations

In keeping with precepts of design-based inference, we assume further that δ_k is fixed in the sense that repeated measurements of the kth unit of $\mathcal{P}$ would result in the same value x_k^*.

Fixed error is reasonable, for example, in the case where a measure of length is rounded to the nearest cm: we presume that repeated measurements of the same length would result an identical measurement, the error of which would be unknown but have constant magnitude among the repeated measurements.

Types of ME considered (continued)

In a remote sensing context LiDAR readings of height will contain error, the magnitude of which will vary among pixels, yet for a given scene it will be fixed for each pixel in the scene.

As a further example, measurement error of a constant magnitude may result from faulty instrumentation, thereby leading to the same magnitude of measurement error among all elements of $\mathcal{P}$. Cochran (1977, §13.9) terms the latter “constant bias over all units,” yet he discusses the case where such biased measurement only affects y, not x.

Types of ME considered (continued)

We considered both the case where the magnitude of δ_k may vary among units, and when it is constant for all units.

For variable ME among units, we distributed the ME according to a uniform, Gaussian, and beta distribution, all centered at $\mu_\delta = 0$ but with a range of dispersions.

Constant ME Results: 1

In one sense, the analytical results presented in equations (5) through (12) are all that are needed. They answer the question of how the average size and and dispersion of ME affect the bias and variance of the ratio estimators of τ_y.

They do not provide much of an intuitive interpretation or understanding.

Therefore we evaluated these expressions for a specific "population" of leaves.

Aggregate leaf area was τ_y and the auxiliary variate was the rectangular area of each leaf (product of leaf width and length).

Constant ME Results: 2

Under fixed ME, $\delta_k = \mu_k, \forall k$, hence $\sigma_\delta = 0$.

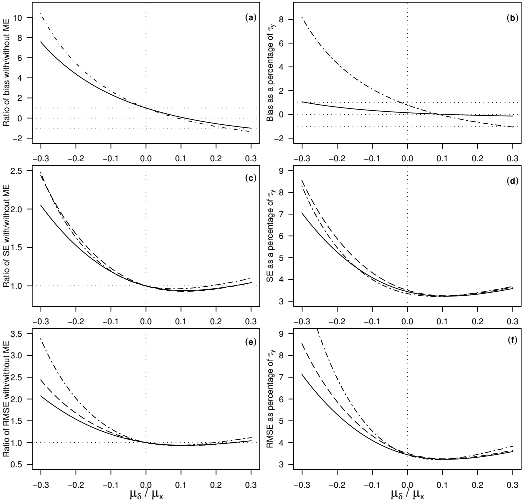
The graphical results of the next page show the bias, standard error, and RMSE of the three estimators under fixed ME.

The horizontal axis shows the magnitude of the ME as a proportion of μ_x. Therefore the zero point corresponds to the absence of ME.

The vertical axis on the left side of panel show ratios of bias, SE, and RMSE with:without ME.

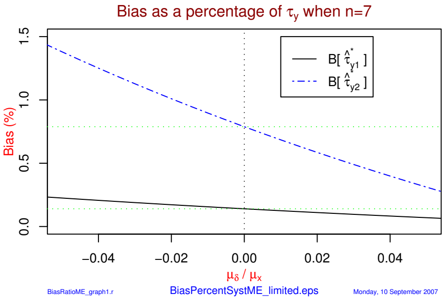
The vertical axis on the right side of panel show bias, SE, and RMSE as a percentage of τ_y.

Constant ME Results: 3



Interpreting Constant ME results

When $\mu_\delta < 0$, bias is increased; bias is decreased when $\mu_\delta > 0$; figure below show bias % when $\mid \mu_\delta \mid < .05\mu_x$



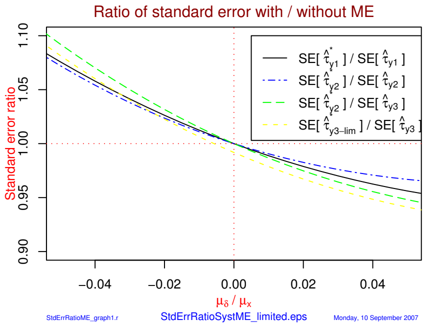
Within this range of tolerable ME, bias is not affected very much, especially for $\hat{\tau}_{y1}$.

Interpreting Constant ME results (continued)

Standard error decreases with increasing ME up to a point, and then increases again.

For these data, ME is capable of increasing precision of estimation, which is a surprise.

Within the range of $|\mu_\delta| < .05\mu_x$, standard error monotonically decreases with increasing ME.



Variable ME Results: 1

Ratio Estimation with Measurement Error

Gregoire & Salas

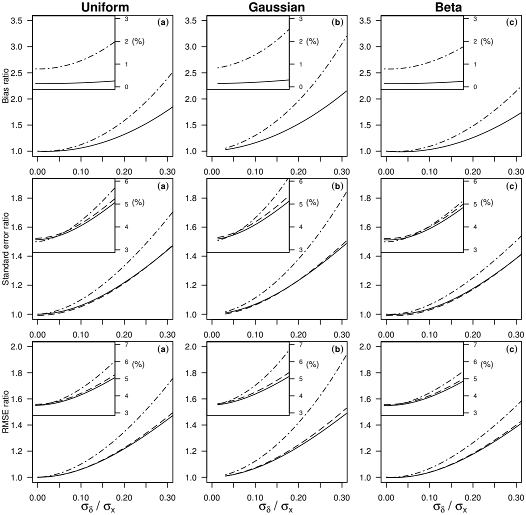
Ratio Estimation

Overview

Specific estimators examined

Types of ME

Results



Horizontal axis is scaled by increasing dispersion of the ME

Variable ME results: 2

- Bias of $\widehat{\tau}_{y1}$ and $\widehat{\tau}_{y2}$ increases with increasing σ_δ, with $\widehat{\tau}_{y2}$ being more affected than $\widehat{\tau}_{y1}$;
- Standard error directly increasing with increasing σ_δ, with $\widehat{\tau}_{y2}$ being more affected than $\widehat{\tau}_{y1}$ and $\widehat{\tau}_{y3}$;
- on the basis of RMSE, $\widehat{\tau}_{y1}$ is best, followed closely by $\widehat{\tau}_{y3}$.

Other stuff

- We also looked at the performance of the usual approximations to the variance of these estimators based upon simulated sampling from the leaf population; see *Biometrics* for details;
- We also looked at the performance of the usual estimators of variance;
- Many additional avenues of inquiry remain: interval estimation; regression estimator; alternative sampling designs; model-based analysis.