Estimation of a difference of two ratios under the infinite population approach to NFI sampling

Radim Adolt

October 13, 2017

Let's define two ratio estimators $\hat{R}_{12}$ and $\hat{R}_{34}$

$$\hat{R}_{12} = \frac{\hat{T}_{y_1}}{\hat{T}_{y_2}} \qquad \hat{R}_{34} = \frac{\hat{T}_{y_3}}{\hat{T}_{y_4}}, \tag{1}$$

where $\hat{T}_{y_1}$, $\hat{T}_{y_2}$, $\hat{T}_{y_3}$ and $\hat{T}_{y_4}$ are (one-phase) estimators of totals T_{y_1}, T_{y_2}, T_{y_3} and T_{y_4} of functions $y_1(x)$, $y_2(x)$, $y_3(x)$ and $y_4(x)$ values of which are observable at any point x of an infinite population of points (a continuum).

By definition,

$$T_{y_k} = \int_D y(x)dx, \tag{2}$$

where T_{y_k} is a total of an arbitrary function $y_k(x)$ of local density corresponding to particular variable denoted by lower index k. The ratio estimators $\hat{R}_{12}$ and $\hat{R}_{34}$ can be equivalently defined by (spatial) means instead of totals.

In case n_D sample points are selected uniformly and independently in a geographical domain D (URS, Uniform Random Sampling) variance of each of the two ratio estimators can be obtained by an approximate formula

$$\hat{\mathbb{V}}_{URS}(\hat{R}_{kl}) = \frac{n_D^2 \left[\widehat{\overline{Z^2_{kl}}} - \hat{\bar{Z}}^2_{kl}\right]}{(n_D - 1)\left[\sum_{\substack{x\in s\\ x\in D}} y_l(x)\right]^2}, \tag{3}$$

where $\widehat{\overline{Z^2_{kl}}}$ and $\hat{\bar{Z}}^2_{kl}$ are the arithmetic mean of a squared residual variable z_{kl} and the square of the residual variable itself. The residual variable is defined by

$$z_{kl}(x) = y_k(x) - y_l(x)\hat{R}_{kl}. \tag{4}$$

The mean of squares and square of the mean are evaluated using all (n_D) sample points in the study area D.

It can be easily shown that the arithmetic mean $\hat{\bar{Z}}_{kl}$ is by definition zero, so Eq. (3) is simplified to

$$\hat{\mathbb{V}}_{URS}(\hat{R}_{kl}) = \frac{n_D \sum_{\substack{x\in s\\ x\in D}} z^2_{kl}(x)}{(n_D - 1)\left[\sum_{\substack{x\in s\\ x\in D}} y_l(x)\right]^2}. \tag{5}$$

Now, imagine that our parameter to estimate is R_δ i.e. the difference of two ratios defined below.

$$R_\delta = R_{12} - R_{34} \tag{6}$$

Trivially an approximate point estimator $\hat{R}_\delta$ of R_δ can be obtained by

$$\hat{R}_\delta = \hat{R}_{12} - \hat{R}_{34}, \tag{7}$$

which is a mere difference of two ratio estimators. Obviously, the key question here is how to estimate the variance of $\hat{R}_\delta$.

Variance of a sum of two random quantities Y_1 and Y_2 is defined as

$$\mathbb{V}(Y_1 + Y_2) = \mathbb{V}(Y_1) + \mathbb{V}(Y_2) + 2\mathbb{C}(Y_1, Y_2), \tag{8}$$

where $\mathbb{C}(Y_1, Y_2)$ is the covariance between Y_1 and Y_2. In a direct analogy to Eq. (8) the variance of a difference can be derived using an equality $Y_1 - Y_2 = Y_1 + (-Y_2)$. Following relationships hold the variances and covariances

$$\mathbb{V}(Y_2) = \mathbb{V}(-Y_2) \tag{9}$$

$$\mathbb{C}(Y_1, -Y_2) = -\mathbb{C}(Y_1, Y_2) \tag{10}$$

Consequently the variance of a difference of Y_1 and Y_2 can be written as

$$\mathbb{V}(Y_1 - Y_2) = \mathbb{V}(Y_1) + \mathbb{V}(Y_2) - 2\mathbb{C}(Y_1, Y_2) \tag{11}$$

and estimated by

$$\hat{\mathbb{V}}(Y_1 - Y_2) = \hat{\mathbb{V}}(Y_1) + \hat{\mathbb{V}}(Y_2) - 2\hat{\mathbb{C}}(Y_1, Y_2). \tag{12}$$

The variance of $\hat{R}_\delta$ can be estimated using

$$\hat{\mathbb{V}}\big(\hat{R}_\delta\big) = \hat{\mathbb{V}}\big(\hat{R}_{12}\big) + \hat{\mathbb{V}}\big(\hat{R}_{34}\big) - 2\hat{\mathbb{C}}\big(\hat{R}_{12}, \hat{R}_{12}\big). \tag{13}$$

For the URS design both $\hat{\mathbb{V}}\big(\hat{R}_{12}\big)$ and also $\hat{\mathbb{V}}\big(\hat{R}_{34}\big)$ are given by Eq. (5) and the covariance $\hat{\mathbb{C}}\big(\hat{R}_{12}, \hat{R}_{12}\big)$ can be obtained by

$$\hat{\mathbb{C}}_{URS}\big(\hat{R}_{12}, \hat{R}_{34}\big) = \frac{n_D \sum_{\substack{x\in s\\ x\in D}} z_{12}(x) z_{34}(x)}{(n_D - 1) \sum_{\substack{x\in s\\ x\in D}} y_2(x) \sum_{\substack{x\in s\\ x\in D}} y_4(x)}. \tag{14}$$

Putting all pieces together the estimator of variance of $\hat{R}_\delta$ under the URS sampling design is expressed as follows

$$\begin{aligned}\hat{\mathbb{V}}_{URS}\big(\hat{R}_\delta\big) &= \frac{n_D}{n_D - 1} \times \\ &\times \left\{ \frac{\sum_{\substack{x\in s\\ x\in D}} z_{12}^2(x)}{\Big[\sum_{\substack{x\in s\\ x\in D}} y_2(x)\Big]^2} + \frac{\sum_{\substack{x\in s\\ x\in D}} z_{34}^2(x)}{\Big[\sum_{\substack{x\in s\\ x\in D}} y_4(x)\Big]^2} - \frac{2\sum_{\substack{x\in s\\ x\in D}} z_{12}(x) z_{34}(x)}{\sum_{\substack{x\in s\\ x\in D}} y_2(x) \sum_{\substack{x\in s\\ x\in D}} y_4(x)} \right\}\end{aligned} \tag{15}$$

Remarks

1. If tracts[1] are used and local densities have not been modified to compensate for the edge effect due to tracts, calculations should be performed in an extended region D^+. This corresponds to an area where tract origins were generated (to guarantee that each point of D can be selected by each type of plot considering the given tract geometry and fixed or random

[1] Clusters of sample plots with predefined geometry.

tract rotation). In such a case local densities on plots outside D are set to zero and the sample size n_D has to be replaced by n_{D^+}.

2. Alternatively the calculations can be performed using a reduced set of tracts by leaving out those with local densities $y_1(x)$, $y_2(x)$, $y_3(x)$ and $y_4(x)$ simultaneously zero. Using this approach the set of all samples as well as sampling design are defined over a (bounded) geographical domain A^+ with generally unknown map and total area. Sample size n_{A^+} is used instead of n_D or n_{D^+} respectively.