One-phase estimation techniques

Based on Horwitz-Thompson theorem for continuous populations

Radim Adolt

ÚHÚL Brandýs nad Labem, Czech Republic

USEWOOD WG2, Training school in Dublin, 16.-19. September 2013

Contents

Essential terminology I

One-phase estimators

Sampling is conducted in one-phase only. This type of estimators uses **only one source of data - typically NFI ground data**. This only data source is considered **error-free**. No auxiliary data are involved in the estimation.

Two-phase estimators

Two-phase estimators use:

1. **first-phase data - the auxiliary dataset**, typically in a form of GIS layer(s), this data does not have to be unbiased nor homogeneous over the study area
2. **second-phase data - typically NFI ground data considered error-free**

Besides various forms of first-phase data **the process of estimation itself generates several variants of two-phase estimators**.

Essential terminology II

Direct versus indirect estimators

- **direct estimators use data from the estimation domain** D only
- **indirect estimators borrow strength** from areas outside of the estimation domain D

All estimators shown during the sessions are direct (in the strict sense).

Design-based versus model based inference

Uncertainty matters:

- **design-based - uncertainty due to sampling**, population under study is fixed
- **model-based inference - the uncertainty is due to a random process that generated the population** which is just one out of many (infinite) possible realizations of a stochastic process

Essential terminology III

Geographical domain D

For the whole lesson, if not stated otherwise, D is a study area or an area of interest. **Our task is to estimate the value of a target parameter in D. The exact area of D is known** and an exact map of D is available.

Estimation of population's total using HTC [Cordy, 1993]

Let Y be the total of a resource in geographical domain D made up by an infinite set of points x (D is a continuous population).

$$Y = \int_D Y(x)dx = \lambda(D)\bar{Y} \tag{1}$$

An unbiased estimator $\hat{Y}$ of the total Y is given by

$$\hat{Y} = \sum_{x \in s} \frac{Y(x)}{\pi(x)}, \tag{2}$$

where the symbol $Y(x)$ stands for local density found on the point x of the sample s with fixed size n_D, $\pi(x)$ is the **inclusion density function** on the point x (local measure of the expected number of sample points per unit area).
Estimator (2) is unbiased if the function $Y(x)$ is positive or bounded in D and if $\pi(x) > 0 \ \forall x \in D$, i. e. any point $x \in D$ can be selected.

Variance of total estimator [Cordy, 1993]

If the local density $Y(x)$ is bounded, conditions $\pi(x) > 0 \ \forall x \in D$ and $\int_D \frac{1}{\pi(x)} dx < \infty$ hold, the **variance of total estimator** $\hat{Y}$ is given:

$$\mathbb{V}_{HTC}(\hat{Y}) = \int_D \frac{Y^2(x)}{\pi(x)} dx + \\ + \int_D \int_D Y(x_i) Y(x_j) \left[\frac{\pi(x_i, x_j) - \pi(x_i)\pi(x_j)}{\pi(x_i)\pi(x_j)} \right] dx_i dx_j \qquad (3)$$

Estimation of the variance of total estimator [Cordy, 1993]

Provided condition $\pi(x_i, x_j) > 0 \ \forall x_i, x_j \in D$ is satisfied, an **unbiased estimator of variance** $\mathbb{V}(\hat{Y})$ is given by

$$\hat{\mathbb{V}}_{HTC}(\hat{Y}) = \sum_{x_i \in s} \left[\frac{Y(x_i)}{\pi(x_i)} \right]^2 + \sum_{x_i \in s} \sum_{\substack{x_j \in s \\ x_i \neq x_j}} Y(x_i) Y(x_j) \left[\frac{\pi(x_i, x_j) - \pi(x_i)\pi(x_j)}{\pi(x_i, x_j)\pi(x_i)\pi(x_j)} \right], \tag{4}$$

and equivalently by

$$\hat{\mathbb{V}}_{HTC}(\hat{Y}) = \sum_{x_i \in s} \left[\frac{Y(x_i)}{\pi(x_i)} \right]^2 + \sum_{x_i \in s} \sum_{\substack{x_j \in s \\ x_i \neq x_j}} \frac{Y(x_i) Y(x_j)}{\pi(x_i)\pi(x_j)} - \sum_{x_i \in s} \sum_{\substack{x_j \in s \\ x_i \neq x_j}} \frac{Y(x_i) Y(x_j)}{\pi(x_i, x_j)}. \tag{5}$$

Total and variance estimators for URS

In case of **Uniform Random Sampling (URS)** with a **fixed number of n_D sample points selected in geographical domain** D the general HTC estimators (2) and (4), (5) take the form of (6), (7), (8):

$$\hat{Y} = \sum_{x \in s} \frac{Y(x)}{\pi(x)} = \lambda(D) \frac{\sum_{x \in s} Y(x)}{n_D} = \lambda(D) \hat{\bar{Y}} \tag{6}$$

$$\hat{\mathbb{V}}_{URS}(\hat{Y}) = \frac{\lambda^2(D)}{n_D(n_D - 1)} \sum_{x \in s} \left[Y(x) - \hat{\bar{Y}} \right]^2 \tag{7}$$

$$\hat{\mathbb{V}}_{URS}(\hat{Y}) = \frac{\lambda^2(D)}{n_D - 1} \sum_{x \in s} \left[\widehat{\bar{Y}^2} - \hat{\bar{Y}}^2 \right] \tag{8}$$

Variance of the estimator of mean density over D

Sometimes we are interested in **mean density of a resource**[1] **in** D which is defined as

$$\hat{\bar{Y}} = \frac{\hat{Y}}{\lambda(D)}, \tag{9}$$

where, as usual, the symbol $\hat{Y}$ stands for total resource estimator.
An **estimator of variance of the mean estimator** $\hat{\bar{Y}}$ is given by

$$\hat{\mathbb{V}}(\hat{\bar{Y}}) = \frac{\hat{\mathbb{V}}(\hat{Y})}{\lambda^2(D)}, \tag{10}$$

see Anděl [1978, theorem 2 on page 14] for instance.

[1] Let's mention proportion of forest land in D as a typical example.

Definition of a ratio of two totals or means in D

Many NFI target parameters are or can be defined as **a ratio $R_{1,2}$ of two totals Y_1 and Y_2**, or equivalently **a ratio of two means densities $\bar{Y}_1$ and $\bar{Y}_1$** of a resource in geographical domain D

$$R_{1,2} = \frac{Y_1}{Y_2} = \frac{\bar{Y}_1}{\bar{Y}_2}. \tag{11}$$

A typical example is the mean per hectare growing stock (or basal area) i. e. a ratio of the total growing stock (or basal area) in D and the total forest-stand area (one of the subclasses of the forest land-use category).

Estimator of a ratio of two totals or means in D

An approximately unbiased estimator $\hat{R}_{1,2}$ of a ratio is calculated as **a ratio of two total estimators $\hat{Y}_1$ and $\hat{Y}_2$** or equivalently as **a ratio of two mean estimators $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_1$ in D**.

$$\hat{R}_{1,2} = \frac{\hat{Y}_1}{\hat{Y}_2} = \frac{\hat{\bar{Y}}_1}{\hat{\bar{Y}}_2} \tag{12}$$

Approximate variance estimation of the ratio estimator

An approximate variance estimator $\hat{\mathbb{V}}(\hat{R}_{1,2})$ of a ratio estimator can be generally expressed as

$$\hat{\mathbb{V}}(\hat{R}_{1,2}) \approx \frac{1}{\hat{Y}_2^2}\hat{\mathbb{V}}(\hat{Z}_\circ), \tag{13}$$

where

$$\hat{Z}_\circ = \sum_{x \in \boldsymbol{s}} \frac{Z_\circ(x)}{\pi(x)} = \lambda(D)\hat{\bar{Z}}_\circ, \tag{14}$$

and

$$Z_\circ(x) = Y_1(x) - \hat{R}_{1,2}Y_2(x). \tag{15}$$

The term $\hat{Z}_\circ$ represents a total estimator of **residual variable** $Z_\circ(x)$ **in** D. Details on ratio estimation using Taylor approximation (the delta method) can be found in Särndal et al. [2003, sections 5.5 and 5.6 starting on page 172].

Yet another formula to estimate variance of ratio

Equivalently to (13) the variance of a ratio estimator can be estimated by

$$\hat{\mathbb{V}}\left(\hat{R}_{1,2}\right) \approx \frac{1}{\hat{Y}_2^2}\left[\hat{\mathbb{V}}\left(\hat{Y}_1\right) + \hat{R}_{1,2}^2\hat{\mathbb{V}}\left(\hat{Y}_2\right) - 2\hat{R}_{1,2}^2\hat{\mathbb{C}}\left(\hat{Y}_1, \hat{Y}_2\right)\right], \tag{16}$$

where $\hat{\mathbb{V}}\left(\hat{Y}_1\right)$, $\hat{\mathbb{V}}\left(\hat{Y}_2\right)$ are the variance estimators of of one-phase totals in the numerator $\left(\hat{Y}_1\right)$ and the denominator $\left(\hat{Y}_2\right)$ of the ratio $\hat{R}_{1,2}$.
The term $\hat{\mathbb{C}}\left(\hat{Y}_1, \hat{Y}_2\right)$ is an estimator of covariance of given one-phase estimators see Särndal et al. [2003, section 5.6 starting on page 176] for details.

In case of a positive and high enough covariance $\hat{\mathbb{C}}\left(\hat{Y}_1, \hat{Y}_2\right)$ **there is a potential for a higher precision of a ratio estimator** in comparison to total estimators included in its numerator and denominator.

Estimation of a ratio, URS design

For URS (fixed sample size in D), **the estimator of a ratio given by** (12) **remains unchanged**.

Its **variance estimator** takes the form of

$$\hat{\mathbb{V}}_{URS}\left(\hat{R}_{1,2}\right) \approx \frac{\lambda^2(D)}{n_D(n_D-1)\hat{Y}_2^2} \sum_{x \in s} Z_\circ^2(x), \tag{17}$$

or equivalently the form of

$$\hat{\mathbb{V}}_{URS}\left(\hat{R}_{1,2}\right) \approx \frac{\sum_{x \in s}\left[Y_1(x) - \hat{R}_{1,2}\, Y_2(x)\right]^2}{n_D(n_D-1)\hat{\bar{Y}}_2^2}. \tag{18}$$

Problems with NFI variance estimation I

NFI sampling design properties influencing variance estimation

- most NFIs use a variant of **systematic** or **spatially stratified design**
- **condition** $\pi(x_i, x_j) > 0 \ \forall x_i, x_j \in D$ **doesn't hold** for systematic sampling and for stratified sampling with fixed origin

There are several workarounds how to approximate and estimate the design-based variance of systematic samples see Wolter M. [1985], Cordy and Thompson [1995], Heikkinen [2006, section 10 starting on page 155] and Cooper [2006] for instance.

Problems with NFI variance estimation II

Variance estimators based on URS approximation

- for designs for which $\pi(x_i, x_j) > 0 \ \forall x_i, x_j \in D$ doesn't hold, **an approximation of $\pi(x_i, x_j) = n_D(n_D - 1)/\lambda^2(D)$ corresponding to URS with fixed sample size in D** is used
- **variance estimators are approximated** by (7), (8), (10) (variance of totals and means) and (17), (18) (variance of a ratio) respectively
- under the condition of **positive spatial correlation** and sufficient sample size **these approximations of variance estimator are mostly but not necessarily[a] conservative** [Heikkinen, 2006, section 10 starting on page 155] and [Mandallaz, 2007, section 4.1 starting on page 53]

[a] Sampling grid resolution might coincide with periodicity of the local density function.

Problems with NFI variance estimation III

Variance estimators based on first order contrasts

For systematic and spatially stratified designs **an usually less conservative variance estimator** (ST estimator) is defined as

$$\hat{\mathbb{V}}_{ST}\left(\hat{\theta}\right) = \frac{\lambda^2(D)}{2kn_D} \sum_{\substack{x \in s \\ x \in D}} \sum_{\substack{\dot{x} \in s \\ \dot{x} \in D}} \left[Y(x) - Y(\dot{x}) \right]^2. \tag{19}$$

Summands represent **differences of local density between sample point** x **and the neighboring sample point** $\dot{x}$. Nearest neighbors in north, east, south and west direction are considered. The term k in the denominator stands for total number of differences (pairs of nearest neighbors) that could be formed in D.

Properties of (19) for various sampling designs were analyzed by Cordy and Thompson [1995]. The authors reported very good results in terms of true coverage (true confidence level) of classically constructed confidence intervals (using normal approximation).

Non-constant sample size in D I

Reasons for and implications of non-constant sample size

- most NFIs use square, less frequently rectangular, triangular or hexagonal sampling grids with random origin
- sampling grid divide the **the support**[a] $\mathcal{S} \supseteq D$ into congruent, non-overlapping cells of equal size and shape - **inventory blocks**
- D **can't be tessellated by a whole number of inventory blocks**
- because of the random origin of the grid and for some designs also due to random location of sample points in inventory blocks, **the number of sample points in** D **varies from sample to sample**

---

[a] An area used for the implementation of sampling algorithm - typically a rectangle covering D.

Non-constant sample size in D II

Reaching unbiasedness of estimators under non-constant sample size

One option is **to implement the estimation on $\mathcal{S}$ instead of D** level[a] and to redefine local density to zero in areas outside D. This technique **works well for designs not using the URS approximation for variance estimation**, otherwise it further increases the conservativeness of variance estimator.

Cordy [1993] describes two solutions:

1. replace inclusion density and pairwise inclusion density functions on inventory points by their expected values taken over all possible sample sizes, estimators and their variances are then **unbiased unconditionally on the realized sample size in D, total estimators are additive**
2. express the inclusion density and pairwise inclusion density functions with respect to realized sample size in D, estimators are **unbiased conditionally on the realized sample size in D, total estimators are no longer additive**

[a] Specific topology for $\mathcal{S}$ [Stevens, 1997, section 3.1 starting on page 172], [Mandallaz, 2007, section 5.6 starting on page 92] the number of sample points in $\mathcal{S}$ is kept constant.

Geographical additivity of total estimators

Geographical additivity of total estimator $\hat{Y}$ is defined by

$$\hat{Y}_{\bigcup_{i=1}^{k} D_i} = \sum_{i=1}^{k} \hat{Y}_{D_i}. \tag{20}$$

A geographically additive total estimator evaluated for arbitrary domain being an union $\bigcup_{i=1}^{k} D_i$ of k sub-domains $D_1, D_2 \ldots D_k$ equals to the sum of k total estimators of the same kind calculated individually for each sub-domain.

Unconditional versus conditional unbiasedness of estimators

Under **unconditional unbiasedness** the expected value of an estimator $\hat{\theta}$, **when evaluated over all possible sample sizes**, equals the true value of the parameter of given population

$$\mathbb{E}\left[\hat{\theta}\right] = \theta. \tag{21}$$

Under **conditional unbiasedness** (conditioning on sample size) the expected value of an estimator $\hat{\theta}$, **when evaluated over all samples of given size** n_D, equals the true value of the parameter of given population

$$\mathbb{E}\left[\hat{\theta} \mid n_D\right] = \theta. \tag{22}$$

Conservativeness of the variance estimator

Expected value of **a conservative variance estimator** $\hat{\mathbb{V}}(\hat{\theta})$ is higher than the true variance

$$\mathbb{E}\left[\hat{\mathbb{V}}(\hat{\theta})\right] > \mathbb{V}(\hat{Y}). \tag{23}$$

Analogously to unbiasedness, conservativeness can be defined and assessed conditionally or unconditionally on sample size in D.

Literature I

J. Anděl. *Matematická statistika*. SNTL - Státní nakladatelství technické literatury, Praha, 1978.

C. Cooper. Sampling and variance estimation on continuous domains. *Environmetrics*, 17:539–553, 2006.

B. Cordy, C. and M. Thompson, C. An application of the deterministic variogram to design-based variance estimation. *Mathematical Geology*, 27:173–205, 1995.

C. B. Cordy. An extension of the horwitz-thompson theorem to point sampling from a continuous universe. *Statistics and Probability Letters*, 18:353–362, 1993.

J. (eds. Kangas A. & Maltamo M.) Heikkinen. *Forest Inventory, Methodology and Applications*, chapter Assessment of Uncetainty in Spatially Systematic Sampling, pages 155–176. Springer Verlag, 2006.

Literature II

D. Mandallaz. *Sampling Techniques For Forest Inventories*. Chapman and Hall/CRC, 2007.

D. L. Jr. Stevens. Variable density grid-based sampling designs for continuous spatial populations. *Environmetrics*, 8:167–195, 1997.

C. E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer, 2003.

K. Wolter M. *Introduction to Variance Estimation*. Springer, 1985.