SAE course spring 2015 / Risto Lehtonen

VARIANCE ESTIMATION for HT and GREG

EXAMPLE: SRSWOR case

(1) Planned domains under STR-SRSWOR

SRSWOR sample from every domain U_d

Domain sample sizes n_d are fixed

Sample allocation schemes

- Optimal (Neyman) allocation
- Bankier allocation
- Equal allocation
- Proportional allocation

see e.g. Lehtonen and Pahkinen (2004) Practical methods for design and analysis of complex surveys. Wiley.

Sample s_d of size n_d elements is drawn from stratum U_d whose size is N_d elements, $d = 1,...,D$

Design weights are $a_k = N_d / n_d$ for all $k \in s_d$

NOTE: We assume that N_d are known

NOTE: SURVEYMEANS with BY statement

Domain totals (unknown parameters)

$$t_d = \sum_{k \in U_d} y_k \ , \ d = 1, \ldots, D$$

a) HT estimator (direct estimator)

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k = \frac{N_d}{n_d} \sum_{k \in s_d} y_k = N_d \bar{y}_d$$

Variance estimator for HT

$$\hat{v}_{str-srswor}(\hat{t}_{dHT}) = N_d^2 (1 - \frac{n_d}{N_d})(\frac{1}{n_d}) \sum_{k \in s_d} \frac{(y_k - \bar{y}_d)^2}{n_d - 1}$$

b) GREG estimator (direct estimator)

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$$

$$= \sum_{k \in U_d} \hat{y}_k + \frac{N_d}{n_d} \sum_{k \in s_d} (y_k - \hat{y}_k)$$

Variance estimator for GREG

$$\hat{v}_{str-srswor}(\hat{t}_{dGREG}) = N_d^2 (1 - \frac{n_d}{N_d})(\frac{1}{n_d}) \sum_{k \in s_d} \frac{(e_k - \bar{e}_d)^2}{n_d - 1}$$

where $e_k = y_k - \hat{y}_k$, $k \in s_d$ are residuals

$\bar{e}_d = \sum_{k \in s_d} e_k / n_d$ is mean of residuals in domain d

Assisting model: $y_k = \mathbf{x}_k' \boldsymbol{\beta}_d + \varepsilon_k$, $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}_d$, $k \in s_d$

(2) Unplanned domains

A single SRSWOR sample *s* of size *n* elements from population *U* whose size is *N* elements

Sample size n_d in domain U_d is a random variable with expectation $E(n_d) = nN_d / N$

Inclusion probability is $\pi_k = n / N$

Design weights are $a_k = N / n$ for all $k \in U$

Define

Domain y-variables $y_{dk} = I_{dk} y_k$

Domain residuals $e_{dk} = y_{dk} - \hat{y}_k$, $d = 1,\ldots,D$

where $I_{dk} = 1$ if $k \in U_d$, zero otherwise

$\hat{y}_k$ are fitted values from the specified model

NOTE: SURVEYMEANS with DOMAIN statement

a) HT estimator (direct estimator)

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k = \frac{N}{n} \sum_{k \in s} y_{dk} = \frac{N}{n} n_d \bar{y}_d$$

Variance estimator for HT:

$$\hat{v}_{srswor}(\hat{t}_{dHT}) = N^2 (1 - \frac{n}{N})(\frac{1}{n}) \sum_{k \in s} \frac{(y_{dk} - \bar{y}_d)^2}{n-1}$$

b) GREG estimator (indirect estimator)

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$$

$$= \sum_{k \in U_d} \hat{y}_k + \frac{N}{n} \sum_{k \in s_d} (y_k - \hat{y}_k)$$

Variance estimator for GREG

$$\hat{v}_{srswor}(\hat{t}_{dGREG}) = N^2 (1 - \frac{n}{N})(\frac{1}{n}) \sum_{k \in s} \frac{(e_{dk} - \bar{e}_d)^2}{n-1}$$

Assisting model: $y_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k$

Fitted values are $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}$, $k \in s$

Residuals are $e_{dk} = y_{dk} - \hat{y}_k$, $k \in s$

NOTE: Elements outside the domain *d* also contribute to the GREG variance estimator

This is because $e_{dk} = -\hat{y}_k$ for $k \notin s_d$ and $k \in s$