HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Small Area Estimation
Spring 2015

Topic 3: Direct estimators for domains

Risto Lehtonen, University of Helsinki

Topic 3

- **Direct estimators for domains**
  - Definitions and notation
  - Estimation of domain totals for planned and unplanned domains
  - Horvitz-Thompson estimator
  - Hájek estimator
  - Variance estimation
  - Example

Definitions and notation - 1

Fixed and finite population $U = \{1,2,...,k,...,N\}$, where *k* refers to the *label* of population element

The fixed population is said to be generated from a *superpopulation*.

Variable of interest *y*

For practical purposes, we are interested in one particular realized population U with $(y_1, y_2, ..., y_N)$, not in the more general properties of the process (or model) explaining how the population evolved.

NOTE: In the *design-based* approach, the values of the variable of interest are regarded as *fixed but unknown* quantities. The only source of randomness is the *sampling design*, and our conclusions should apply to hypothetical repeated sampling from the fixed population.

Definitions and notation - 2

Basic parameters for study variable *y* for the whole population:

$$\text{Total } t = \sum_{k \in U} y_k$$

$$\text{Mean } \bar{y} = \sum_{k \in U} y_k \, / \, N$$

In most cases we discuss the estimation of totals – Why?

In practice, the values y_k of *y* are observed in an *n* element sample $s \subset U$ which is drawn by a sampling design giving probability $p(s)$ to each sample s

NOTE: The sampling design can be *complex* involving stratification and clustering and several sampling stages – see e.g. the Survey sampling reference guidelines document by Lehtonen&Djerf (2008)

Definitions and notation - 3

The *design expectation* of an estimator $\hat{t}$ of population total t is determined by the probabilities $p(s)$:

Let $\hat{t}(s)$ denote the value of estimator that depends on y observed in sample s

Expectation is $E(\hat{t}) = \sum_s p(s)\hat{t}(s)$

Design unbiased estimator: $E(\hat{t}) = t$

Design variance: $Var(\hat{t}) = \sum_s p(s)\left(\hat{t}(s) - E(\hat{t})\right)^2$

NOTE: $Var(\hat{t})$ is an unknown parameter

An *estimator* of design variance is denoted by $\hat{V}(\hat{t})$

Definitions and notation - 4

Variance estimators are derived in two steps:

(1) The theoretical design-based variance $Var(\hat{t})$ (or its approximation if the theoretical design variance is intractable) is derived

(2) The derived quantity is estimated by a design unbiased or design-consistent estimator $\hat{V}(\hat{t})$

NOTE: An estimator is *design consistent* if its design bias and variance tend to zero as the sample size increases

Definitions and notation - 5

Inclusion probability: An observation k is included in the sample with probability $\pi_k = P\{k \in s\}$

The inverse probabilities $a_k = 1/\pi_k$ are called *design weights*

Sample membership indicator:
$I_k = I\{k \in s\}$ with value 1 if k is in the sample and 0 otherwise

Expectation of sample membership indicator $E(I_k) = \pi_k$

Probability of including both elements k and l $(k \neq l)$ is $\pi_{kl} = E(I_k I_l)$ with inverse $a_{kl} = 1/\pi_{kl}$ $(a_{kl} = a_k$ when $k = l)$

The covariance of I_k and I_l is $Cov(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$

Estimation for domains

Domain estimation of totals or averages of variable of interest y over D non-overlapping domains $U_d \subset U$, $d = 1,2,...,d,...,D$, with possibly known domain sizes N_d

Example: Population of a country is divided into D domains by regional classification, with N_d households in domain U_d

The aim is to estimate statistics on household income for the regional areas (domains)

The key parameter is **domain total**: $t_d = \sum_{k \in U_d} y_k$,

where y_k refers to measurement for household k

Why domain totals are important?

Totals are basic and the simplest descriptive statistics for continuous (or binary) study variables

Many other, more complex statistic are functions of totals

Domain ratio: $$R_d = \frac{t_{dy}}{t_{dz}} = \frac{\sum_{k \in U_d} y_k}{\sum_{k \in U_d} z_k}$$

Estimator: $$\hat{R}_d = \frac{\hat{t}_{dy}}{\hat{t}_{dz}} = \frac{\sum_{k \in s_d} a_k y_k}{\sum_{k \in s_d} a_k z_k}$$

Domain mean: $$\bar{y}_d = t_d / N_d$$

Estimator: $$\hat{\bar{y}}_d = \hat{t}_d / N_d \quad \text{or} \quad \hat{\bar{y}}_d = \hat{t}_d / \hat{N}_d$$

Estimation for planned domains - 1

Sample is divided into subsamples s_d, $d = 1,...,D$

Planned domains:

Stratified sampling with domains = strata

- The population domains U_d can be regarded as separate subpopulations
- Domain sizes N_d in domains U_d are assumed known
- Sample size n_d in domain sample $s_d \subset U_d$ is fixed in advance
- **Standard population estimators are applicable as such**

Estimation for planned domains - 2

NOTES

Stratified sampling with a suitable *allocation scheme* (e.g. optimal (Neyman) or power (Bankier) allocation) is advisable in practical applications, in order to obtain control over domain sample sizes

Singh, Gambino and Mantel (1994) describe allocation strategies to attain reasonable accuracy for small domains, still retaining good accuracy for large domains

Estimation for unplanned domains - 1

Unplanned domains: A single sample *s* of size *n* is drawn from population U.
Domain samples are $s_d \subset U_d$

Domain sample sizes n_d cannot be considered fixed but are *random*

Extended domain variable of interest y_d defined as:

$$y_{dk} = y_k \text{ for } k \in U_d \text{ and } y_{dk} = 0 \text{ for } k \notin U_d$$

In other words, $y_{dk} = I\{k \in U_d\} y_k$

Because $t_d = \sum_{k \in U_d} y_k = \sum_{k \in U} y_{dk}$, we can estimate domain total of y by estimating the population total of y_{dk}

Estimation for unplanned domains – 2

NOTES

Contribution of extra variance caused by random domain sample sizes can be incorporated in variance expressions and computation

SAS survey procedures:

SURVEYMEANS
SURVEYREG etc.

can handle the unplanned domains case by using the DOMAIN statement with extended domain y-variables and extended residuals

NOTE: This is not necessarily so in the R Survey package of

Thomas Lumley

Horvitz-Thompson estimator of domain totals

Horvitz-Thompson (HT) estimator (*expansion estimator*) is the basic *design-based direct* estimator of the domain total $t_d = \sum_{k \in U_d} y_k$, $d = 1,...,D$:

$$\hat{t}_{dHT} = \sum_{k \in U_d} I_k y_k / \pi_k = \sum_{k \in s_d} y_k / \pi_k = \sum_{k \in s_d} a_k y_k \qquad (1)$$

HT estimates of domain totals are additive: they sum up to the HT estimator $\hat{t}_{HT} = \sum_{k \in s} a_k y_k$ of the population total $t = \sum_{k \in U} y_k$

As $E(I_k) = \pi_k$, the HT estimator is design unbiased for t_d

Variance estimation for HT - 1

Standard *variance estimator* for $\hat{t}_{dHT}$ under **planned** domains:

$$\hat{V}\left(\hat{t}_{dHT}\right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) y_k y_l \qquad (2)$$

An alternative Sen-Yates-Grundy formula:

$$\hat{V}\left(\hat{t}_{dHT}\right) = \sum_{k \in s_d} \sum_{l<k; l \in s_d} \left(\frac{a_{kl}}{a_k a_l} - 1\right)(a_k y_k - a_l y_l)^2 \qquad (3)$$

NOTE: Both (2) and (3) are somewhat impractical... Why?

Variance estimation for HT - 2

Variance estimation for planned domains in practice

- SUDAAN: Standard formula (2)
- SAS macro CLAN: Sen-Yates-Grundy formula (3)

Variance estimators are impractical because of $a_{kl} = 1/\pi_{kl}$

Approximations to π_{kl} for fixed- size without-replacement (WOR) probability proportional-to-size (πPS) designs :

- Hájek (1964) and Berger (2004, 2005) approximation
- Särndal (1996) approximation
- Berger and Skinner (2005) jackknife variance estimator
- Kott (2006) delete-a-group jackknife variance estimator

Variance estimation for HT - 3

Variance estimation for planned domains in practice

$$\hat{V}_A\left(\hat{t}_{dHT}\right) = \frac{1}{n_d(n_d-1)}\sum_{k\in s_d}\left(n_d a_k y_k - \hat{t}_{dHT}\right)^2 \qquad (4)$$

For example, SAS Procedure SURVEYMEANS uses (4)

Variance estimation for HT - 4

Unplanned domains:

Variance estimator should account for random domain sizes

Approximate variance estimator by using *extended domain variables* y_{dk} :

$$\hat{V}_U\left(\hat{t}_{dHT}\right) = \frac{1}{n(n-1)}\sum_{k\in s}\left(na_k y_{dk} - \hat{t}_{dHT}\right)^2, \qquad (5)$$

where n is the total sample size

NOTE: e.g. SAS procedure SURVEYMEANS uses (5)

NOTE: Extended domain variables are $y_{dk} = I\{k \in U_d\}y_k$

Recall: $y_{dk} = y_k$ if $k \in U_d$, 0 otherwise

Hájek estimator of domain totals

Hájek type direct estimator:

$$\hat{t}_{dH(N)} = N_d \hat{\bar{y}}_d = \frac{N_d}{\hat{N}_d} \sum_{k \in s_d} a_k y_k \qquad (6)$$

where $\hat{\bar{y}}_d = \sum_{k \in s_d} a_k y_k \,/\, \hat{N}_d$ are estimated domain means

$\hat{N}_d = \sum_{k \in s_d} a_k$ are estimated sizes of population domains

Assuming domain sizes N_d are known we expect better results with the Hájek estimator (Särndal, Swensson and Wretman 1992)

The variance of $\hat{t}_{dH(N)}$ is estimated by

$$\hat{V}\left(\hat{t}_{dH(N)}\right) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) \left(y_k - \hat{\bar{y}}_d\right)\left(y_l - \hat{\bar{y}}_d\right) \qquad (7)$$

EXAMPLE: HT and Hájek estimators for domain totals

- Real population data from Western Finland (Statistics Finland)
- Domains: D = 12 regional areas = strata
- Planned domains for HT and Hájek
- Unplanned domains for HT
- Study variable y: Disposable income (registers)
- Auxiliary data: Sizes of population domains
- Sample size: n = 1,000 households (dwelling units)
- Sampling: stratified πPS (WOR type probability proportional to size sampling) with household size as the size variable
- Details: See separate pdf sheet (course website) and Table 2

Table 2. Mean absolute relative error MARE (%) and mean coefficient of variation MCV (%) of direct HT and Hájek estimators of totals for minor, medium-sized and major domains for **planned domains** (HT and Hájek) and **unplanned domains** (HT).

Domain sample size class	HT			Hájek	
	Auxiliary information				
	None			Domain sizes	
	MARE %	MCV1 %	MCV2 %	MARE %	MCV1 %
Minor $8 \le n_d \le 33$	11.5	**11.9**	**28.3**	5.3	**10.9**
Medium $34 \le n_d \le 45$	7.6	**9.0**	**20.3**	6.4	**9.0**
Major $46 \le n_d \le 277$	12.5	**5.2**	**9.6**	4.7	**5.6**
MCV1: Assuming planned domains for HT and Hájek MCV2: Assuming unplanned domains for HT					

References

- Berger, Y.G. (2004). A simple variance estimator for unequal probability sampling without replacement. Journal of Applied Statistics 31, 305-315.
- Berger, Y.G. (2005). Variance estimation with highly stratified sampling designs with unequal probabilities. Australian & New Zealand Journal of Statistics 47, 365-373.
- Berger, Y.G. and C.J. Skinner (2005). A jackknife variance estimator for unequal probability sampling. Journal of the Royal Statistical Society, Series B, 67, 79-89.
- Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. Annals of Mathematical Statistics 35, 1491-1523.

References(contd.)

- Kott, P.S. (2006). Delete-a-group variance estimation for the general regression estimator under Poisson sampling. Journal of Official Statistics 22, 759-767.
- Särndal, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. Journal of the American Statistical Association 91, 1289-1300.
- Singh, M.P., J. Gambino and H.J. Mantel (1994). Issues and strategies for small area data. Survey Methodology 20, 3-14.