Jean Monnet Chair

Analysis of European Data by Small Area Methods

Lecture 2: Estimation for planned and unplanned domains, Horvitz and Thompson estimates of domain total

http://sampleu.ec.unipi.it

Estimation for domains

Measures of *precision (MSE)* are usually computed to evaluate the quality of a population parameter estimate and to obtain valid inferences.

The estimation method and the sampling design determine the properties of the MSE and of the sampling error .

Estimation for domains: direct estimators

- The basic design-consistent *Horvitz-Thompson estimator* is the most natural estimator to use if there is no auxiliary information available at the estimation stage.

SAE generally include direct estimates where sampling weights, calibration, re-weighting had their effect

Estimation dor domains: direct estimators

- Proper estimation conforms to the sampling design.
- Sampling weights are incorporated in the estimation process
- Sampling weights derive from: stratification, clustering, and multi-phase or multi-stage information

Estimation for domains: direct estimators

Use auxiliary data whenever possible to improve the reliability of the estimates (decrease MSE).

Evaluate the use of the auxiliary data.

Estimation for domains: direct estimators

Stat Canada suggestion 1:

"Whenever auxiliary data are available for sample units, together with known population totals for such data, consider using calibration estimation ("evolution of HT estimator!") so that the weighted auxiliary data add up to these known totals. This may result in improved precision and lead to greater consistency between estimates from various sources."

Estimation for domains: direct estimators

Stat Canada suggestion 2:

"Incorporate the requirements of small domains of interest at the sampling design and sample allocation stages (Singh, Gambino and Mantel, 1994).

If this is not possible at the design stage, or if the domains are only specified at a later stage, consider special estimation methods (small area estimators) at the estimation stage. These methods "borrow strength" from related areas (or domains) to minimize the mean square error of the resulting estimator (Platek et al., 1987; Ghosh and Rao, 1994; Rao, 1999)."

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:

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

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

We discuss here the estimation of totals

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

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

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 - 3

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}

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 - 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 - 3

Unplanned domains:

Variance estimator should account for random domain sizes

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

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

where n is the total sample size

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

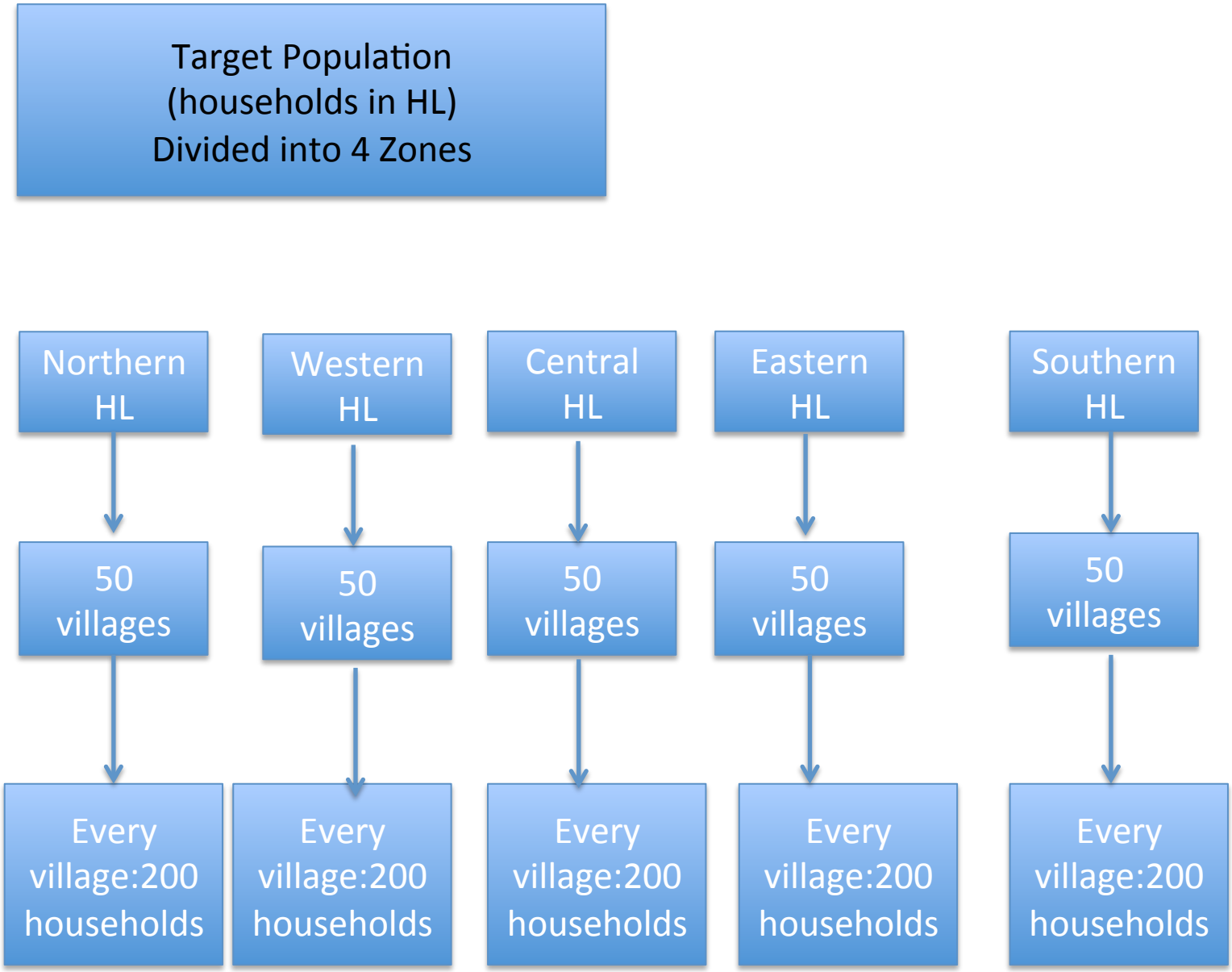
NOTE: Extended domain variables are $y_{dk} = I\{k \in U_d\} y_k$

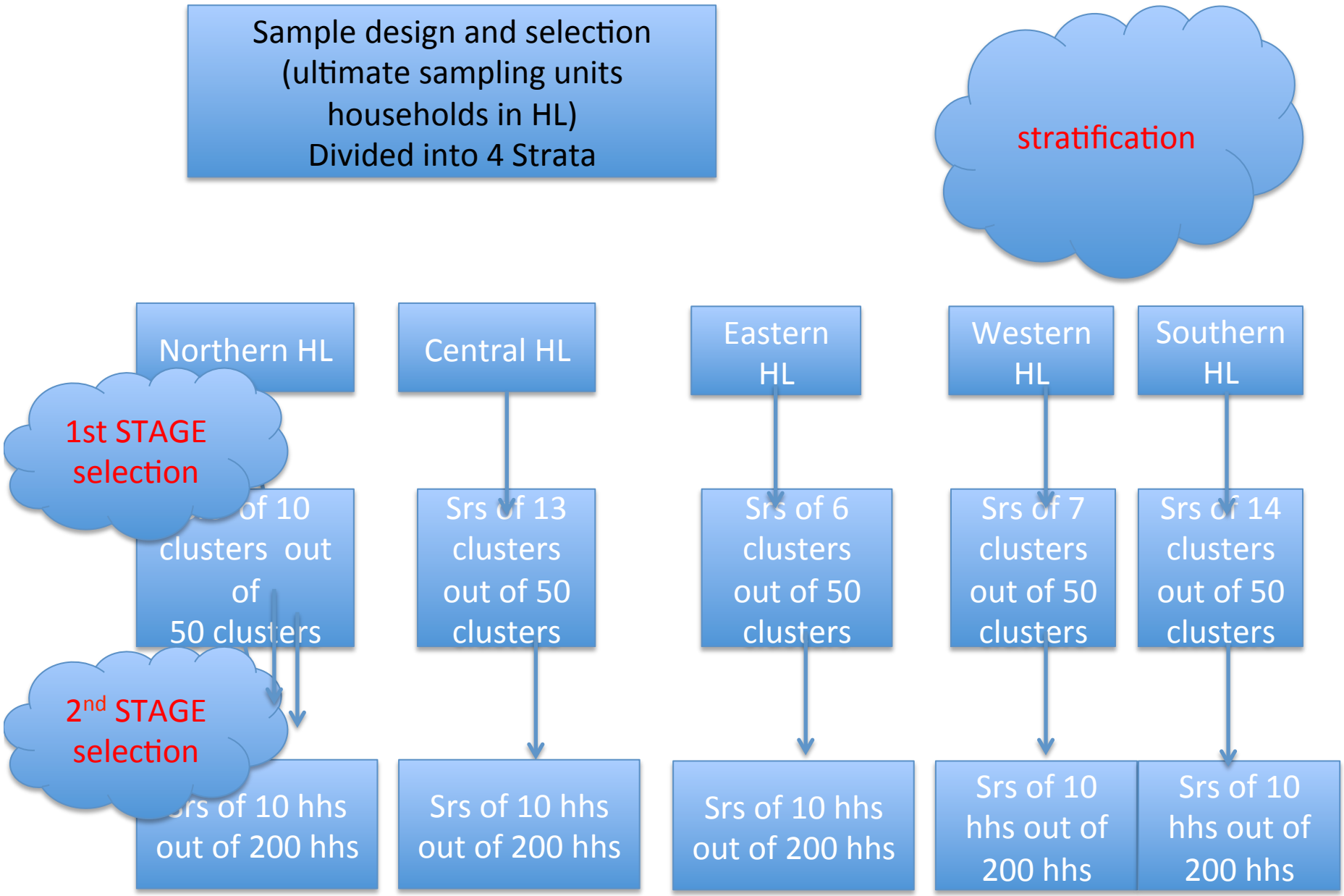
Recall: $y_{dk} = y_k$ if $k \in U_d$, 0 otherwise

Example: estimation for domains

Happy Land Food Survey
Stratified two stage sample survey
H=5 strata; N=20000 households
A= 200 villages in HL (clusters)
a=50 sampled villages

- n=500 households
- Complete coverage of the target population
- Full response of the interviewed households


Target Population
(households in HL)
Divided into 4 Zones
Northern HL
Western HL
Central HL
Eastern HL
Southern HL
50 villages
50 villages
50 villages
50 villages
50 villages
Every village:200 households
Every village:200 households
Every village:200 households
Every village:200 households
Every village:200 households


Sample design and selection
(ultimate sampling units
households in HL)
Divided into 4 Strata
stratification
Northern HL
Central HL
Eastern
HL
Western
HL
Southern
HL
1st STAGE
selection
of 10
clusters out
of
50 clusters
Srs of 13
clusters
out of 50
clusters
Srs of 6
clusters
out of 50
clusters
Srs of 7
clusters
out of 50
clusters
Srs of 14
clusters
out of 50
clusters
2nd STAGE
selection
rs of 10 hhs
out of 200 hhs
Srs of 10 hhs
out of 200 hhs
Srs of 10 hhs
out of 200 hhs
Srs of 10
hhs out of
200 hhs
Srs of 10
hhs out of
200 hhs

H code	Sa-weights	village	stratum	FoodExp	H-size	H-income	HH-educ	HH-age
1		1	north HL	12	1	2500	1	56
2		1	north HL	16,5	1	2800	1	70
3		1	north HL	18	1	2000	1	20
4		1	north HL	17	1	4500	1	60
5		1	north HL	46,5	1	8000	1	40
6		1	north HL	45	1	7000	1	51
7		1	north HL	15	1	3500	1	76
8		1	north HL	60	2	2800	1	20
9		1	north HL	15	2	2500	1	51
10		1	north HL	18	2	4000	1	32
11		2	north HL	22,5	2	5000	1	47
12		2	north HL	20	2	8000	1	35
13		2	north HL	97	2	5500	1	58
14		2	north HL	57	2	6000	1	27
15		2	north HL	39	2	3000	1	38
16		2	north HL	30	2	4000	1	40
17		2	north HL	42	2	3000	1	19

CENTRAL HL

village	hh x village	foodexp	h size	income	
5	10	171	16	26500	13 villages
6	10	181,5	20	19700	
7	10	212	20	31300	
8	10	429,5	26	50800	
9	10	460,5	42	54400	
10	10	537,5	52	70200	
11	10	439	38	60700	
12	10	266,5	15	70500	
13	10	317,5	20	81100	
14	10	394,5	24	82200	
15	10	380,5	31	57300	
16	10	438	48	81400	
17	10	370	39	77200	
total	130	4598	391	763300	

EASTERN HL

village	hh x village	foodexp	h size	income	
38	10	545	73	79000	6 villages
39	10	478	57	102500	
40	10	397	21	119500	
41	10	429	31	162000	
42	10	569,7	38	152000	
43	10	543	45	133000	
	60	2961,7	265	748000	

WESTERN HL

village	hh x village	foodexp	h size	income	
44	10	604	42	133000	7 villages
45	10	459,5	44	118000	
46	10	544	44	142500	
47	10	305,5	22	147500	
48	10	593	38	212000	
49	10	342	35	99500	
50	10	647	57	227000	
	70	3495	282	1079500	

NORTHERN HL

village	hh x village	foodexp	h size	income	
1	10	263	13	39600	10 villages
2	10	399,5	20	50000	
3	10	453,5	33	124400	
4	10	270,5	37	46000	
32	10	573,5	52	129500	
33	10	390,5	30	68000	
34	10	379,5	39	90000	
35	10	606,5	48	110000	
36	10	482	48	120000	
37	10	452,5	56	67500	
total	100	4271	376	845000	

SOUTHERN HL

village	hh x village	foodexp	h size	income	
18	10	404,5	43	61500	14 villages
19	10	540	47	96700	
20	10	472	65	100500	
21	10	541	55	74800	
22	10	382,5	10	118000	
23	10	496	17	125800	
24	10	392	20	154000	
25	10	282,5	20	119000	
26	10	395,5	25	112000	
27	10	449,5	36	115000	
28	10	406,5	30	99400	
29	10	525	40	114500	
30	10	446,5	40	103000	
31	10	561	50	119000	
total	140	6294,5	498	1513200	

Probability of inclusion of *k-th* hh

This is what I need*:* $1/\pi_k = a_k$

Sampling weight for the *k-th* household

$\pi_k = \pi_{hi} \times \pi_{hk|i}$

π_{hi}

probability of inclusion of the village (*h-th* stratum)

$\pi_{hk|i}$

probability of inclusion of the household, given that the *i-th* village is included (*h-th* stratum)