STA 304H1F-1003H Fall 2019

Week 11 - Two-Stage Cluster Sampling

Chapter 9

How to Draw a Cluster Sample

1. **Simple one-stage cluster sample:**

   List all the clusters in the population, and

   from the list, select the clusters – usually with simple random sampling (SRS) strategy.

   All units (elements) in the sampled clusters are selected for the survey.

2. **Simple two-stage cluster sample:**

   List all the clusters in the population, and

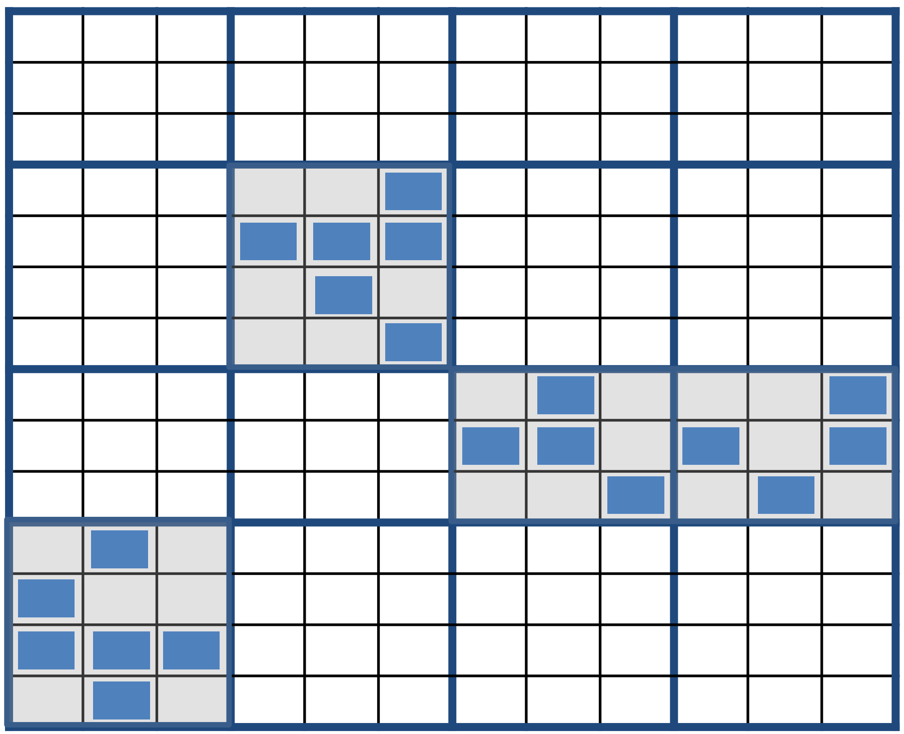
   First, select the clusters, usually by simple random sampling (SRS).

   The units (elements) in the selected clusters of the first-stage are then sampled in the second-stage, usually by simple random sampling (or often by systematic sampling).

Cluster Sampling, Two Stages (I)

Remainder on basics (I)

Basics: The population is divided into large number of (bigger) groups (clusters). **First stage**: sample of clusters selected. **Second stage**: sample of elements from each selected cluster. Sample: all selected elements.

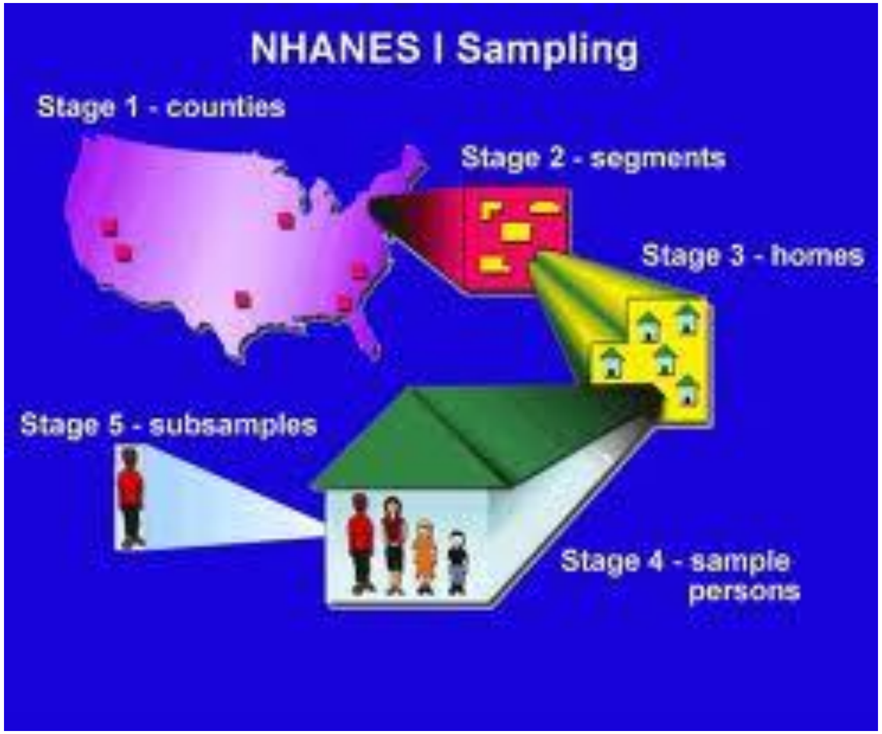


Population size: 168.
8 clusters of size 9,
8 clusters of size 12.
Sample size: 20.
4 clusters selected "at random" from 16 clusters.
6+6+4+4 elements selected.

Cluster Sampling, Two Stages (I)

Remainder on basics (II)

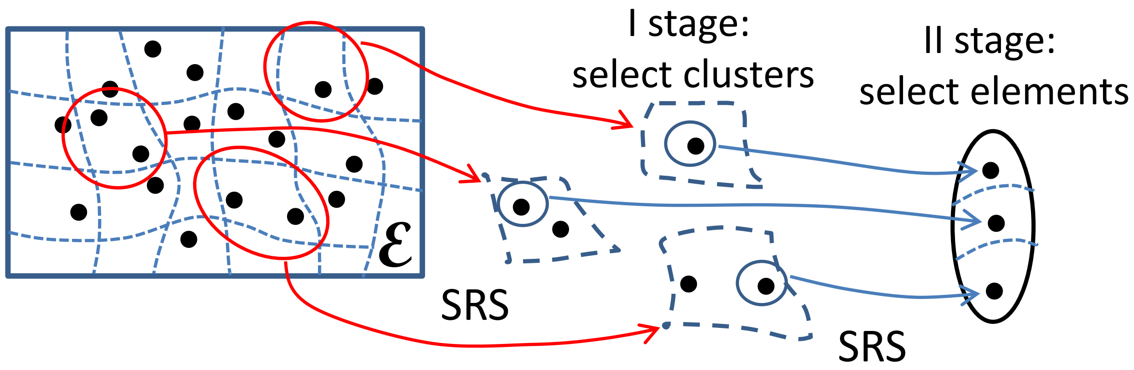
The population is divided into clusters on several levels/stages - primary sampling units (PSUs), secondary sampling units, ternary sampling units, … Sampling is performed at every stage. Sample: all sampling units selected at the last stage.



National Health and Nutrition Examination Survey (USA)
Five-stage cluster sampling
Stage 1: 1900 PSUs – cities, counties
Stage 2: Segments – each with 18 housing units/addresses
Stage 3: Households
Stage 4: Family members for basic examination (~ 28,000, 1-74 years old)
Stage 5: Subsample for detailed health examination (25-74 years old)

Cluster Sampling, Two Stages (I)

General considerations, definition (I)



Simplest case: SRS of sampling units at every stage (PSU, SSU, TSU, …)

Large number of (bigger or smaller) groups

– First stage: SRS of *n* groups – primary sampling units (PSU)

– Second stage: SRS of elements – secondary sampling units (SSU) from every PSU

– Sampling units: Clusters at every stage

– Design: Two-stage cluster sampling. All elements selected at the second stage are in the sample

Cluster Sampling, Two Stages (I)

General considerations, notation (II)

$N-$ number of clusters in the population (# of sampling units)

$n-$ number of clusters selected in the sample at stage I

$M_i - i$th clusters size

$m_i -$ sample size selected from ith clusters, $m_i \le M_i$

In one-stage $m_i = M_i$

$M -$ population size, $M = M_1 + M_2 + \cdots + M_N = \sum_{i=1}^{N} M_i$

$\overline{M} -$ average cluster size, $\overline{M} = \dfrac{M}{N} = \dfrac{1}{N}\sum_{i=1}^{N} M_i$

ith cluster elements: $y_{i1}, y_{i1}, \ldots, y_{iM_i}$

$\tau_i = \sum_{j=1}^{M_i} y_{ij}$ - cluster total, $\mu_i = \dfrac{\tau_i}{M_i}$ - cluster mean, $\tau_i = M_i \mu_i$

Cluster Sampling, Two Stages (I)

General considerations, notation (III)

Summary, population:

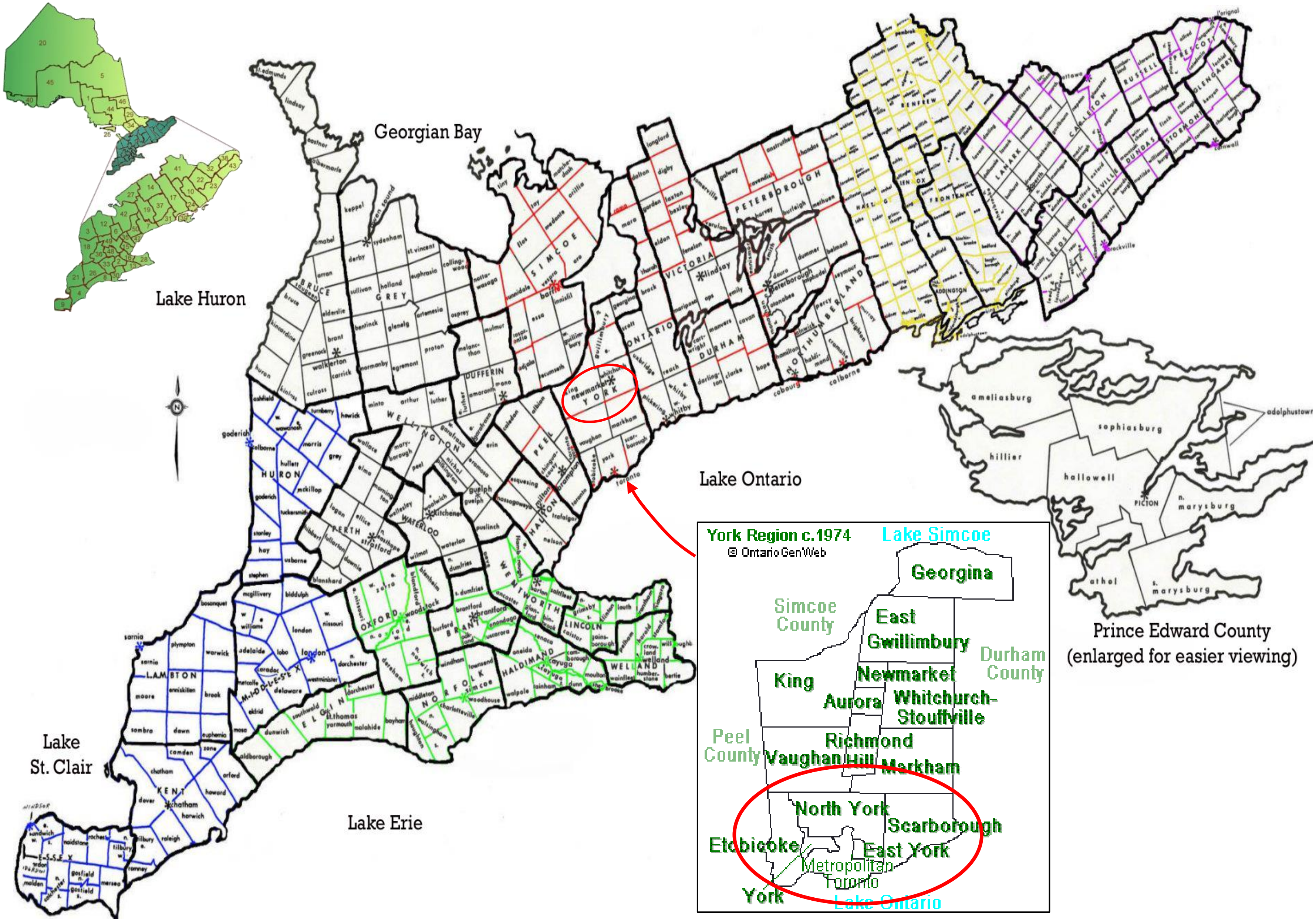
Cluster	1	2	$\cdots$	N	Pop
Size	M_1	M_2	$\cdots$	M_N	M
Total	τ_1	τ_2	$\cdots$	τ_N	τ
Mean	μ_1	μ_2	$\cdots$	μ_N	μ

$$\tau = \tau_y = M\mu_y, \mu = \mu_y = \frac{\tau_y}{M}$$

$$\tau_y = \sum_{i=1}^{N} \tau_i = \sum_{i=1}^{N} M_i \mu_i = N \frac{1}{N} \sum_{i=1}^{N} \tau_i = N\mu_t = N\bar{\tau}$$

$$\mu_t = \bar{\tau} = \frac{1}{N} \sum_{i=1}^{N} \tau_i$$ - average cluster total

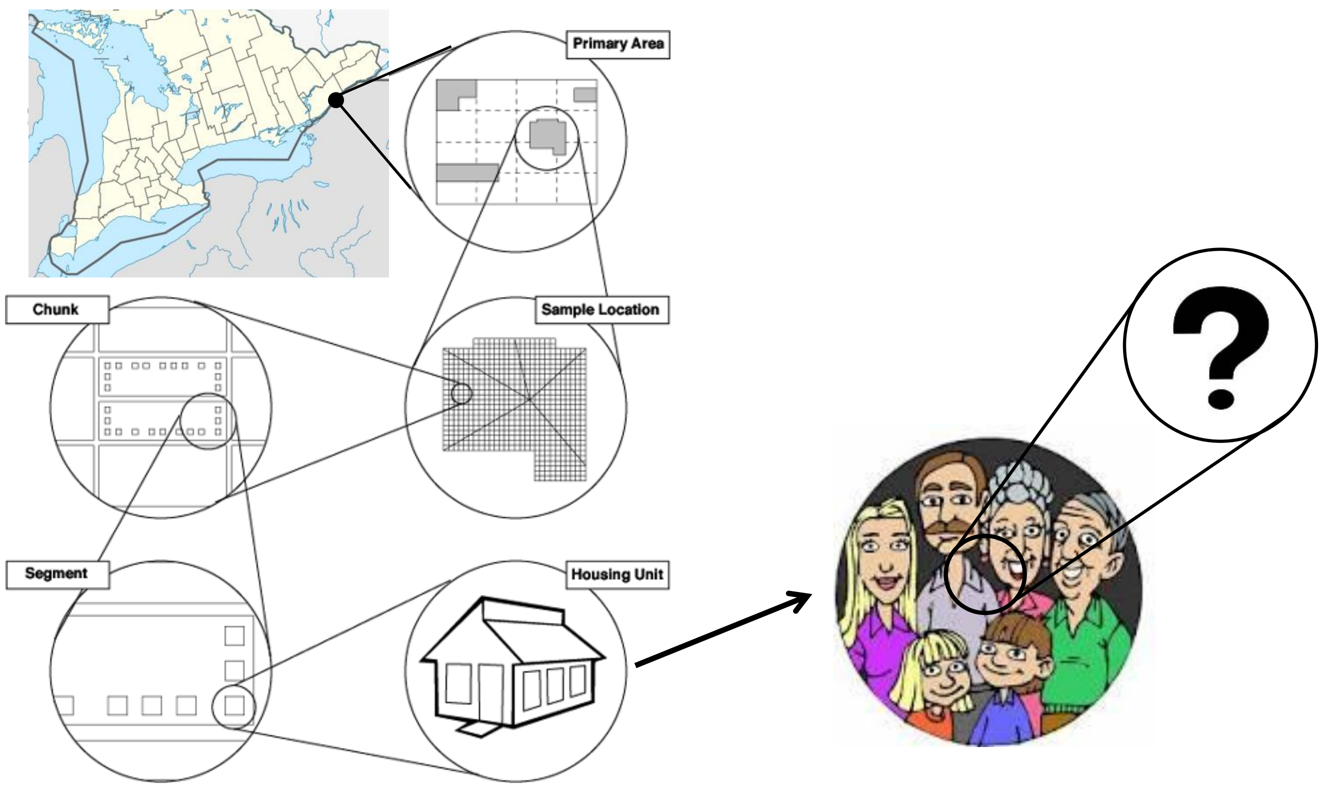
Example: Southern Ontario, map of counties and townships (I)



Example: Multi-stage cluster sample from Southern Ontario (VI)

First stage sampling units: Counties
Second stage sampling units: Townships
Third stage sampling units: Communities
Fourth stage sampling units: Neighbourhoods/wards
Fifth stage sampling units: ? Households/mailing addresses
Sixth stage sampling units: ? Persons

Two strata: Metropolitan Toronto, York region (why?)



Example: Multi-stage cluster sample from Southern Ontario (VII)

Selecting Respondents within Households using telephone surveys:

(a) The respondent (an adult) is asked to list all the eligible members of the household by gender and age; they are then ordered by age, first males and then females, and then one person is selected at random using preselected random number – *an unbiased but not very convenient procedure , creating higher refusal rate*

(b) *A less invasive procedure*: The respondent is asked two questions:

1. How many persons 18 years or older live in your household, counting yourself?
2. How many of them are men (women)?

One of four (or seven) selection matrices is randomly assigned in advance to each sampled telephone number and used depending on answers to Q. 1 and 2 – e.g., to speak with the “oldest” or “youngest” man or women

(c) *The simplest*: The respondent is asked to speak to the eligible member who had the “last” (or will have “the next”) birthday

An interviewer

Cluster Sampling, Two Stages (I)

General considerations, sample (IV)

I stage : SRS of n clusters

II stage : SRS of m_i elements selected from cluster i selected in stage I

Sample from cluster i : $y_{i1}, y_{i1}, \ldots, y_{im_i}$, $i = 1,2,\ldots,n$

$y_i = \sum_{j=1}^{m_i} y_{ij}$ - ith cluster sample total $(\neq \tau_i)$

$\hat{\mu}_i = \bar{y}_i = \frac{y_i}{m_i}$ - ith cluster sample mean

Unbiased estimator

$S_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2$ - ith cluster sample variance

Cluster Sampling, Two Stages (II)

Inference: Unbiased estimation of mean and total (I)

First estimate $\tau_i, i = 1,2,\ldots,n$ - totals of clusters in the sample

$$\hat{\mu}_i = \bar{y}_i \Rightarrow \hat{\tau}_i = M_i \hat{\mu}_i = M_i \bar{y}_i \Rightarrow \hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_n$$

$$\boxed{\hat{\mu}_t = \hat{\bar{\tau}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i = \frac{1}{n}\sum_{i=1}^{n} M_i \bar{y}_i \Rightarrow \hat{\tau} = N\hat{\bar{\tau}} = \frac{N}{n}\sum_{i=1}^{n} M_i \bar{y}_i,}$$

$$\boxed{\hat{\mu} = \frac{\hat{\tau}}{M} = \frac{N}{M}\frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i = \frac{1}{\overline{M}}\frac{1}{n}\sum_{i=1}^{n} M_i \bar{y}_i,}$$

All unbiased

$$\boxed{\hat{Var}(\hat{\mu}) = \frac{N-n}{N}\frac{S_b^2}{n\overline{M}^2} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2 \frac{M_i - m_i}{M_i}\frac{S_i^2}{m_i}}$$

First stage component Second stage component

Derivation of the formula for variance is not simple, see book.

$$\boxed{S_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(M_i \bar{y}_i - \overline{M}\hat{\mu})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\tau}_i - \hat{\bar{\tau}})^2}$$

Cluster Sampling, Two Stages (II)

Inference: Unbiased estimation of mean and total (II)

How to organize data: Calculation table

Cl	sample				$\bar{y}_i$	S_i^2	$\hat{\tau}_i = M_i \bar{y}_i$
1	y_{11}	y_{12}	$\cdots$	y_{1m_1}	$\bar{y}_1$	S_1^2	$\hat{\tau}_1 = M_1 \bar{y}_1$
2	y_{21}	y_{22}	$\cdots$	y_{2m_2}	$\bar{y}_2$	S_2^2	$\hat{\tau}_2 = M_2 \bar{y}_2$
$\vdots$	$\vdots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_{n1}	y_{n2}	$\cdots$	y_{nm_n}	$\bar{y}_n$	S_n^2	$\hat{\tau}_n = M_n \bar{y}_n$

$$\hat{\mu}_t = \hat{\bar{\tau}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i \Rightarrow \hat{\tau} = N\hat{\bar{\tau}}, \hat{\mu} = \frac{\hat{\tau}}{M}$$

$$S_b^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\hat{\tau}_i - \hat{\bar{\tau}})^2$$

Cluster Sampling, Two Stages (III)

Inference: Ratio estimation of mean and total (I)

We can use ratio estimation instead of unbiased.

We have to use it to estimate μ if M is not known.

$$\overline{M} = \frac{1}{N}\sum_{i=1}^{N} M_i \Rightarrow \boxed{\hat{\overline{M}} = \frac{1}{n}\sum_{i=1}^{n} M_i, \hat{M} = N\hat{\overline{M}} = \frac{N}{n}\sum_{i=1}^{n} M_i}$$

Unbiased

$$\boxed{\hat{\mu}_r = \frac{\hat{\tau}}{\hat{M}} = \frac{\frac{N}{n}\sum_{i=1}^{n} M_i \overline{y}_i}{\frac{N}{n}\sum_{i=1}^{n} M_i} = \frac{\sum_{i=1}^{n} M_i \overline{y}_i}{\sum_{i=1}^{n} M_i}, \hat{\tau}_r = M\hat{\mu}_r = M\frac{\sum_{i=1}^{n} M_i \overline{y}_i}{\sum_{i=1}^{n} M_i}}$$

M known

Biased

We can use $\overline{M}$ if we know it

$$\boxed{\hat{Var}(\hat{\mu}_r) = \frac{N-n}{N}\frac{S_r^2}{n\hat{\overline{M}}^2} + \frac{1}{nN\hat{\overline{M}}^2}\sum_{i=1}^{n} M_i^2 \frac{M_i - m_i}{M_i}\frac{S_i^2}{m_i}}$$

Two components of the variance

$$\boxed{S_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(M_i\overline{y}_i - M_i\hat{\mu}_r)^2 = \frac{1}{n-1}\sum_{i=1}^{n} M_i^2(\overline{y}_i - \hat{\mu}_r)^2}$$

Cluster Sampling, Two Stages (III)

Inference: Comparison of ratio and unbiased estimation

We want to compare variances of $\hat{\mu} = \hat{\mu}_{UNB}$ and $\hat{\mu} = \mu_r$.

Assuming $\hat{\overline{M}} \approx \overline{M}$ (or just using $\overline{M}$),

$$\boxed{\hat{V}ar(\hat{\mu}_r) < \hat{V}ar(\hat{\mu}_{UNB}) \Leftrightarrow S_r^2 < S_b^2}$$

$$\sum_{i=1}^{n}(M_i\bar{y}_i - M_i\hat{\mu}_r)^2 < \sum_{i=1}^{n}(M_i\bar{y}_i - \overline{M}\hat{\mu}_{UNB})^2$$

or $$\sum_{i=1}^{n}(\hat{\tau}_i - M_i\hat{\mu}_r)^2 < \sum_{i=1}^{n}(\hat{\tau}_i - \hat{\bar{\tau}}_{UNB})^2$$

Here we have usual discussion about deviations.
Ratio estimator usually performs better than unbiased.

Cluster Sampling, Two Stages (IV)

Science and Medicine Library: Collection of statistical books (I)

Sampling population: 161 shelves

Sampling design: Two-stage cluster sampling

PSU – clusters: N = 25 bookcases (vertical collections of shelves)

Sample size: n = 10 bookcases (clusters), and m = 2 shelves per bookcase, total 20 shelves

Bookcase

1	2	3	4	5
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6

	6	7	8	9	10	11
1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7

12	13	14	15	16
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6

	17	18	19	20
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6

	21	22	23	24	25
1	1	1	1	1	1
2	2	2	2	2	2
3	3	3	3	3	3
4	4	4	4	4	4
5	5	5	5	5	5
6	6	6	6	6	6
7	7	7	7	7	7

Cluster Sampling, Two Stages (IV)

Science and Medicine Library: Collection of statistical books (II)

Sample, first stage: SRS of 10 bookcases (clusters) from 25 bookcases

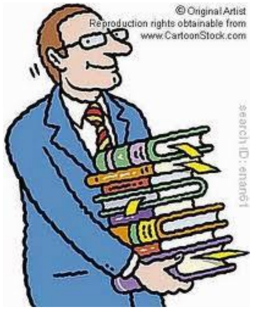
Table of random numbers (two digits, modulo 50):

46	32	74	60	48	22	35	47	47	75	63	13	11	96	20	98	16	56	87	48	38	30	35	78	21
		24	**10**		**22**				**25**	**13**	~~13~~	**11**	~~46~~	**20**	~~48~~	**16**	**6**	~~37~~					~~28~~	**21**

Bookcases selected in the first stage: 6, 10, 11, 13, 16, 20, 21, 22, 24, 25

Second stage: From each selected bookcase two shelves are selected (using one digit)

Bookcase	6	10	11	13	16	20	21	22	24	25
# of shelves	7	7	7	6	6	6	7	7	7	7
Shelves selected	1, 6	6, 4	2, 1	1, 6	4, 5	3, 4	7, 3	4, 3	7, 1	2, 1



Cluster Sampling, Two Stages (IV)

Science and Medicine Library: Collection of statistical books (III)

Sample values:
(first value y,
second value x)

1	2	3	4	5

	6	7	8	9	10	11
1	24,16					21,19
2						17,17
3						
4					30,17	
5						
6	21,17				20,19	
7						

12	13	14	15	16
	21,16			
				32,22
				28,17
	19,13			

	17	18	19	20
1				
2				
3				26,15
4				26,18
5				
6				

	21	22	23	24	25
1				21,8	17,10
2					17,7
3	18,5	15,7			
4		21,11			
5					
6					
7	18,16			14,0	

Notice: No bookcase was selected from first group of 5 bookcases, just by chance. And, what is that chance?

Cluster Sampling, Two Stages (IV)

Science and Medicine Library: Collection of statistical books (IV)

Number of clusters (bookcases) N = 25, number of elements (shelves) M = 161

Sample and calculation organized:

Cluster	1	2	3	4	5	6	7	8	9	10
M_i	7	7	7	6	6	6	7	7	7	7
m_i	2	2	2	2	2	2	2	2	2	2
y_{i1}	24	30	21	21	32	26	18	15	21	17
y_{i2}	21	20	17	19	28	26	18	21	14	17
$\bar{y}_i$	22.5	25	19	20	30	26	18	18	17.5	17
S_i^2	4.5	50	8	2	8	0	0	18	24.5	0
$\hat{\tau}_i$	157.5	175	133	120	180	156	126	126	122.5	119

Unbiased estimation:

$$\overline{M} = \frac{161}{25} = 6.44, \ \sum_{i=1}^{10} M_i = 7\times 7 + 3\times 6 = 67, \ \hat{\overline{M}} = \frac{67}{10} = 6.7$$

Average # of books per bookcase

$$\hat{\bar{\tau}} = \hat{\mu}_t = \frac{1}{n}\sum M_i \bar{y}_i = \frac{1}{10}(157.5 + 175 + 133 + 120 + \cdots + 122.5 + 119) = 141.5$$

$$\hat{\mu} = \frac{N}{M}\frac{1}{n}\sum M_i \bar{y}_i = \frac{N}{M}\hat{\mu}_t = \frac{25}{161}\times 141.5 = 21.97$$

Average # of books per shelve

Cluster Sampling, Two Stages (IV)

Science and Medicine Library: Collection of statistical books (V)

Accuracy, unbiased:

$$S_b^2 = \frac{1}{n-1}\sum(\hat{\tau}_i - \hat{\bar{\tau}})^2 = \frac{1}{10-1}\sum(\hat{\tau}_i - 141.5)^2 = 550.33$$

$$\sum M_i^2 \frac{M_i - m_i}{M_i}\frac{S_i^2}{m_i} = \frac{1}{2}\sum M_i(M_i - 2)S_i^2 = 1957.5$$

$$\hat{V}ar(\hat{\mu}) = \frac{N-n}{N}\frac{1}{n\overline{M}^2}S_b^2 + \frac{1}{nN\overline{M}^2}\sum M_i^2 \frac{M_i - m_i}{M_i}\frac{S_i^2}{m_i} \qquad \overline{M} = 6.44$$

$$= \frac{25-10}{25} \times \frac{1}{10\times(6.44)^2} \times 550.33 + \frac{1}{10\times 25\times(6.44)^2} \times 1957.5 = 0.985$$

$$\hat{S}d(\hat{\mu}) = 0.992$$

Comparison with SRS, sample size 20: $\hat{\mu} = 22.9, \hat{S}d(\hat{\mu}) = 1.507$

Cluster Sampling, Two Stages (IV)

Science and Medicine Library: Collection of statistical books (VI)

Ratio estimation

$$\hat{\mu}_r = \frac{\Sigma M_i \bar{y}_i}{\Sigma M_i} = \frac{\Sigma \hat{\tau}_i}{\Sigma M_i} = \frac{1415}{67} = 21.12 \qquad S_r^2 = \frac{1}{n-1}\sum(\hat{\tau}_i - M_i\hat{\mu})^2 = 802.243$$

$$\hat{V}ar(\hat{\mu}_r) = \frac{N-n}{N}\frac{1}{n\overline{M}^2}S_r^2 + \frac{1}{nN\overline{M}^2}\sum M_i^2 \frac{M_i - m_i}{M_i}\frac{S_i^2}{m_i} \qquad \overline{M} = 6.44$$

$$= \frac{25-10}{25}\times\frac{1}{10\times(6.44)^2}\times 802.243 + \frac{1}{10\times 25\times(6.44)^2}\times 1957.5 = 1.349$$

$$\hat{S}d(\hat{\mu}) = 1.162$$

Comparison

Unbiased cluster : $\hat{\mu} = 21.97, \hat{S}d(\hat{\mu}) = 0.992$

Ratio cluster : $\hat{\mu}_r = 21.12, \hat{S}d(\hat{\mu}_r) = 1.162$

Simple random : $\hat{\mu}_{SRS} = 22.90, \hat{S}d(\hat{\mu}_{SRS}) = 1.507$

Two-stage cluster sampling behaves better than SRS in this problem, but in general need not, if clusters are homogeneous.

Cluster Sampling, Two Stages (V)

Population proportion, summary (Ch. 9.5)

$$y_{ij} = \begin{cases} 0 \\ 1 \end{cases}, \quad \tau_i = \sum_j y_{ij}$$

Number of elements in cluster i with given property

$$y_i = a_i$$

Same, but in sample

$$\hat{\mu}_i = \hat{p}_i = \bar{y}_i = \frac{a_i}{m_i}$$ - sample proportion in ith cluster $\Rightarrow \hat{\tau}_i = M_i \hat{p}_i$

$$\Rightarrow \hat{\hat{\tau}}_{UNB} = \frac{1}{n}\sum_{i=1}^{n} M_i \hat{p}_i, \quad \hat{p}_{UNB} = \frac{\hat{\hat{\tau}}_{UNB}}{\overline{M}} = \frac{N}{M}\frac{1}{n}\sum_{i=1}^{n} M_i \hat{p}_i, \quad \hat{p}_r = \frac{\sum_{i=1}^{n} M_i \hat{p}_i}{\sum_{i=1}^{n} M_i}, \quad \hat{\hat{\tau}}_r = \overline{M}\hat{p}_r$$

$$\hat{Var}(\hat{p}) = \frac{N-n}{N}\frac{S_p^2}{n\overline{M}^2} + \frac{1}{nN\overline{M}^2}\sum_{i=1}^{n} M_i^2 \frac{M_i - m_i}{M_i}\frac{\hat{p}_i \hat{q}_i}{m_i - 1}$$

$$S_p^2 = \begin{cases} S_b^2, \text{unbiased} \\ S_r^2, \text{ratio} \end{cases}$$

$$S_r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(M_i\hat{p}_i - M_i\hat{p}_r)^2 = \frac{1}{n-1}\sum_{i=1}^{n} M_i^2(\hat{p}_i - \hat{p}_r)^2$$

$$S_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(M_i\hat{p}_i - \overline{M}\hat{p}_{UNB})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{\tau}_i - \hat{\hat{\tau}}_{UNB})^2$$

$\overline{M}$ with unbiased

$\hat{M}$ with ratio (?)

Examples: book

Cluster Sampling, Two Stages (V)

Equal cluster sizes, Ch. 9.6 (I)

$M_i = \overline{M}, i = 1,2,\ldots,N$ - equal cluster sizes, no ratio estimation

Use $m_i = m, i = 1,2,\ldots,N$ - equal sample sizes from each cluster

$$\hat{\mu} = \hat{\mu}_{UNB} = \frac{1}{\cancel{\overline{M}}}\frac{1}{n}\sum_{i=1}^{n}\cancel{M_i}\bar{y}_i = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i \Rightarrow$$

$$\boxed{\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i = \frac{1}{n\times m}\sum_{i,j}y_{ij} = \bar{y}}$$

Just sample mean

$$S_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(M_i\bar{y}_i - \overline{M}\hat{\mu})^2 = \frac{\overline{M}^2}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \hat{\mu})^2 = \overline{M}^2 S_{\bar{y}}^2$$

$$MSB = \frac{m}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \hat{\mu})^2 = mS_{\bar{y}}^2, \quad MSW = \frac{1}{n}\sum_{i=1}^{n}S_i^2$$

$$\boxed{\hat{Var}(\hat{\mu}) = \frac{N-n}{N}\frac{S_{\bar{y}}^2}{n} + \frac{\overline{M}-m}{\overline{M}}\frac{1}{mN}\frac{1}{n}\sum_{i=1}^{n}S_i^2 = (1-\frac{n}{N})\frac{MSB}{n\times m} + (1-\frac{m}{\overline{M}})\frac{1}{N}\frac{MSW}{m}}$$

Cluster Sampling, Two Stages (V)

Equal cluster sizes (II)

Summary:

From $\hat{Var}(\hat{\mu}) = (1-\frac{n}{N})\frac{MSB}{n\times m} + (1-\frac{m}{\overline{M}})\frac{1}{N}\frac{MSW}{m}$

1) N large : $\hat{Var}(\hat{\mu}) \approx (1-\frac{n}{N})\frac{MSB}{n\times m} \approx \frac{MSB}{n\times m}$

2) If $m = \overline{M}$ - one - stage : $\hat{Var}(\hat{\mu}) = (1-\frac{n}{N})\frac{MSB}{n\times m}$

3) If $n = N$ - stratified, $L = N$, equal allocation, $n_i = m$:

$$\hat{Var}(\hat{\mu}) = (1-\frac{m}{\overline{M}})\frac{MSW}{n\times m}$$

Cluster Sampling, Two Stages (V)

Equal cluster sizes: optimal sample size (I)

Two cases to select sample size n:

I Given cost of sampling, minimize variance (error bound)

II Given error bound (accuracy, standard error), minimize cost

Cost model – linear model: $\boxed{C = C(n) = c_0 + c_1 \times n + c_2 \times n \times m}$

c_0 – fixed cost,

c_1 – cost of sampling one cluster,

c_2 – cost of sampling one unit

$\boxed{\text{Find optimal } n \text{ and } m}$

Theoretical variance: $Var(\hat{\mu}) = \frac{N-n}{N}\frac{\tilde{\sigma}_\mu^2}{n} + \frac{1}{n}\frac{\bar{M}-m}{\bar{M}}\frac{\bar{\tilde{\sigma}}^2}{m}$

$$\tilde{\sigma}_\mu^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\mu_i - \mu)^2, \bar{\tilde{\sigma}}^2 = \frac{1}{N}\sum_{i=1}^{N}\tilde{\sigma}_i^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\bar{M}-1}\sum_{i=1}^{\bar{M}}(y_{ij} - \mu_i)^2$$

Cluster Sampling, Two Stages (VI)

PPS and two-stage cluster (I)

Sample design: First stage – PPS of clusters,
Second stage – SRS of elements

Selest cluster i with probability proportional to its size: $\pi_i = \dfrac{M_i}{M}$

$$\boxed{\hat{\mu}_{PPS} = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i \ , \quad \hat{\tau}_{PPS} = M\hat{\mu}_{PPS} \ , \ \bar{y}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} y_{ij}}$$

Unbiased

$$\boxed{\hat{V}ar(\hat{\mu}_{PPS}) = \frac{1}{n}S_{\bar{y}}^2 = \frac{1}{n}\times\frac{1}{n-1}\sum_{i=1}^{n}(\bar{y}_i - \hat{\mu}_{PPS})^2}$$

$$\boxed{\hat{V}ar(\hat{\tau}_{PPS}) = \hat{V}ar(M\hat{\mu}_{PPS}) = M^2\frac{1}{n}S_{\bar{y}}^2}$$

Cluster Sampling, Two Stages (VI)

Example: PPS and cattle farms (I)

Population: 2072 farms divided into 53 clusters of unequal size (by area), $\overline{M} =$ 2072/53 = 39.094, the average cluster size.

Goal: Estimate the average number of cattle per farm.

Variables: Number of cattle on farm (*y*).

Parameters to be estimated: Average number of cattle per farm (μ_y) and total number of cattle on farms.

Sampling design: First stage: 14 clusters with probabilities of selection proportional to cluster size (number of farms), with replacements.

Second stage: SRS of ¼ farms from the cluster.

Method of estimation: Unbiased PPS estimation.

Sample: PPS sample of 14 clusters selected from 53 clusters with probabilities proportional to number of farms (clusters in the sample ordered by cluster size). (see next slide)

Cluster Sampling, Two Stages (V)

Example: PPS and cattle farms (II)

Sample cluster, i	Number of farms, M_i	Number of farms in sample, m_i	Number of cattle in sample, y_i	Average cattle per sampled farm, $\bar{y}_i = y_i/m_i$	Selection Probability $\pi_i = M_i/M$
1	13	3	30	10.00	13/2072
2	15	3	58	19.33	15/2072
3	19	5	14	2.80	19/2072
4	28	7	73	10.43	.
5	39	10	162	16.20	.
6	41	11	88	8.00	.
7	46	12	102	8.50	.
8	46	12	102	8.50	.
9	48	12	203	16.92	.
10	51	13	134	10.31	.
11	59	14	195	13.93	.
12	74	19	272	14.32	.
13	83	20	242	12.10	83/2072
14	83	20	242	12.10	83/2072
Total	645	161	1917	163.43	

$$\bar{y}_1 = \frac{30}{3} = 10,$$

$$\bar{y}_2 = \frac{58}{3} = 19.33,$$

$$\ldots$$

$$m_i \approx M_i/4$$

Alternative first stage PPS sampling: Selection probabilities proportional to the cluster area.

Cluster Sampling, Two Stages (V)

Example: PPS and cattle farms (III)

Estimation:

$$\boxed{\hat{\mu}_{pps} = \frac{1}{n}\sum_{1}^{n}\frac{y_i}{m_i} = \frac{1}{n}\sum_{1}^{n}\bar{y}_i = \frac{163.43}{14} = 11.674}$$

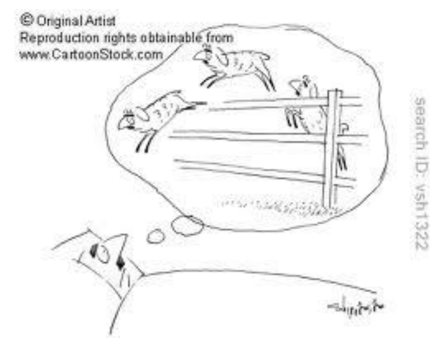
$$\boxed{\hat{\tau}_{pps} = M\hat{\mu} = 2072 \times 11.674 \;=\; 24{,}188.5}$$

$$\hat{V}ar(\hat{\mu}_{pps}) = \frac{1}{n}\left[\frac{1}{(n-1)}\sum_{1}^{n}(\bar{y}_i - \hat{\mu}_{pps})^2\right] = \frac{1}{n}S_{\bar{y}}^2 = \frac{1}{14}18.283 = 1.306$$

$$\hat{\sigma}(\hat{\mu}) = \sqrt{1.306} = 1.143$$

$$\hat{V}ar(\hat{\tau}_{pps}) = M^2\hat{V}ar(\hat{\mu}_{pps}) = 2072^2 \times 1.306$$

$$\hat{\sigma}(\hat{\tau}_{pps}) = 2072 \times 1.143 \;=\; 2{,}367.8$$



Compare with one stage PPS estimation: $\hat{\mu} = 11.713, \hat{\sigma}(\hat{\mu}) = 0.813$. Number of sampled farms: for one stage sample 662, for two stage sample 161. Two stage sample is four times smaller, but (slightly?) less efficient.