

DATA 606: Statistical Methods in Data Science

— Cluster sampling with equal probabilities

Wenjun Jiang

Department of Mathematics & Statistics
The University of Calgary

Lecture 5



An example

Goal: find out how many bicycles are owned by residents in a community of 10000 households.

- ▶ Take a SRS of 400 households.
- ▶ Or, divide the community into blocks of about 20 households each, then take a SRS of 20 blocks.
- ▶ The 2nd sampling method is called *cluster sampling*.
- ▶ The blocks are called *primary sampling units* (psus) or *clusters*, the households are called the *second sampling units* (ssus).

An example (cont.)

What's the problem with cluster sampling?

- ▶ Some blocks of the community are composed mainly of families with more bicycles, while the residents of other blocks are mainly retirees with fewer bicycles
- ▶ Twenty households in the same block are not as likely to mirror the diversity of the community as well as 20 households chosen at random.

Thus, cluster sampling in this situation will probably result in less information per observation than a SRS of the same size.

Why use cluster sampling?

Two main reasons:

- ▶ Constructing a sampling frame list of observation units may be difficult, expensive, or impossible.

Example 1

We cannot list all honeybees in a region or all customers of a store. We may be able to construct a list of all trees in a stand of northern hardwood forest or a list of individuals in a city for which we only have a list of housing units, but constructing the list will be time consuming and expensive.

Why use cluster sampling

- ▶ The population may be widely distributed geographically and it is less expensive to take a sample of clusters rather than a SRS of individuals.

Example 2

If the target population is residents of nursing homes in the United States, it is much cheaper to sample nursing homes and interview every resident in the selected homes than to interview a SRS of nursing home residents: With a SRS of residents, you might have to travel to a nursing home just to interview one resident.

The difference between cluster and stratified sampling

A cluster, like a stratum, is a grouping of the members of the population. However,

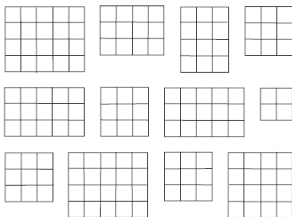
- ▶ **The sample selection process** is quite different.
- ▶ Generally, stratified sampling increases estimation precision while cluster sampling decreases precision.
- ▶ Units in the same stratum tends to have similar or common characteristics.

Example 3

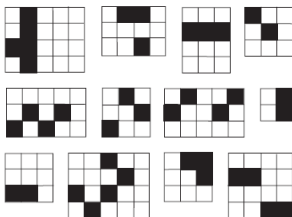
Survey the professors in a university. *Stratified sampling*: group professors according to the departments they belong to. *Cluster sampling*: group professors according to the buildings where their offices are in.

The difference between cluster and stratified sampling

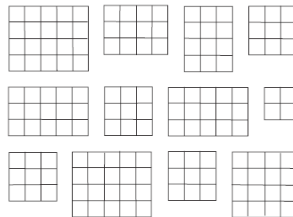
Population of H strata; stratum h has n_h elements:



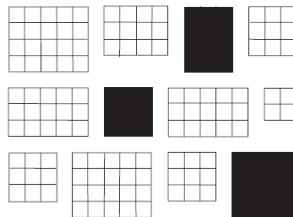
Take an SRS from *every* stratum:



One-stage cluster sampling; population of N clusters:



Take an SRS of clusters; observe all elements within the clusters in the sample:



Cluster sampling

Two main cluster sampling methods:

- ▶ **One-stage cluster sampling**, in which every element within a sampled cluster is included in the sample.
- ▶ **Two-stage cluster sampling**, in which we subsample only some of the elements of selected clusters.

In a cluster sampling

- ▶ U is the population of N psus (primary sampling units, or *clusters*).
- ▶ S is the sample of psus. S_i is the sample of ssus chosen from the i th psu.
- ▶ y_{ij} : measurement for j th element in i th psu.

Notations

Primary sampling unit (psu) level:

- ▶ N : number of psus in the population.
- ▶ M_i : number of ssus in psu i .
- ▶ $M_0 = \sum_{i=1}^N M_i$: total number of ssus in the population.
- ▶ $t_i = \sum_{j=1}^{M_i} y_{ij}$: total of psu i .
- ▶ $t = \sum_{i=1}^N t_i = \sum_{i=1}^N \sum_{j=1}^{M_i} y_{ij}$: population total.
- ▶ $V_t = \frac{1}{N-1} \sum_{i=1}^N \left(t_i - \frac{t}{N}\right)^2$: population variance of the psu totals.

Notations

Secondary sampling unit (ssu) level:

- ▶ $\bar{y}_U = \frac{t}{M}$: population mean.
- ▶ $\bar{y}_{iU} = \frac{t_i}{M_i}$: population mean in psu i .
- ▶ $V = \frac{\sum_{i=1}^N \sum_{j=1}^{M_i} (y_{ij} - \bar{y}_U)^2}{M-1}$: population variance.
- ▶ $V_i = \frac{\sum_{j=1}^{M_i} (y_{ij} - \bar{y}_{iU})^2}{M_i-1}$: population variance within psu i .

Notations

Sample quantities:

- ▶ n : number of psus in the sample.
- ▶ m_i : number of ssus in the sample from psu i .
- ▶ $\bar{y}_i = \frac{\sum_{j \in S_i} y_{ij}}{m_i}$: sample mean for psu i .
- ▶ $\hat{t}_i = M_i \cdot \bar{y}_i$: estimated total for psu i .
- ▶ $\hat{t} = N \cdot \frac{\sum_{i \in S} \hat{t}_i}{n}$: population total estimator.
- ▶ $v_t = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}}{N} \right)^2$: estimated variance of psu totals.
- ▶ $v_i = \frac{\sum_{j \in S_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1}$: sample variance within psu i .

One-stage: equal size

We take a SRS on psu level and take all the units within these psus as our sample.

What if $M_1 = M_2 = \dots = M_N$?

One-stage: equal size

We take a SRS on psu level and take all the units within these psus as our sample.

What if $M_1 = M_2 = \dots = M_N$?

$$\hat{t} = N \cdot \frac{\sum_{i \in S} \hat{t}_i}{n} = N \cdot M \cdot \frac{\sum_{i \in S} \hat{t}_i}{n \cdot M} = N \cdot \frac{\sum_{i \in S} t_i}{n}.$$

One-stage: equal size

We take a SRS on psu level and take all the units within these psus as our sample.

What if $M_1 = M_2 = \dots = M_N$?

$$\hat{t} = N \cdot \frac{\sum_{i \in S} \hat{t}_i}{n} = N \cdot M \cdot \frac{\sum_{i \in S} \hat{t}_i}{n \cdot M} = N \cdot \frac{\sum_{i \in S} t_i}{n}.$$

We hereby ignore the ssu level, and only keep $t_i (i \in S)$ in our record. Thus

$$\text{Var}(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{V_t}{n}.$$

One-stage: equal size

As we do not know V_t (population variance at psu level), we use v_t to replace:

$$v_t = \frac{1}{n-1} \sum_{i \in S} \left(t_i - \frac{\sum_{i \in S} t_i}{n} \right)^2.$$

One-stage: equal size

As we do not know V_t (population variance at psu level), we use v_t to replace:

$$v_t = \frac{1}{n-1} \sum_{i \in S} \left(t_i - \frac{\sum_{i \in S} t_i}{n} \right)^2.$$

To estimate \bar{y}_U , we have


$$\hat{\bar{y}} = \frac{\hat{t}}{N \cdot M}.$$

Its variance is given by

$$\text{Var}(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right) \frac{V_t}{n \cdot M^2}.$$

An example

A student wants to estimate the average grade point average (GPA) in his dormitory. Instead of obtaining a listing of all students in the dorm and conducting an SRS, he notices that the dorm consists of 100 suites, each with four students; he chooses 5 of those suites at random, and asks every person in the 5 suites what her or his GPA is. The results are as follows:



Person Number	Suite (psu)				
	1	2	3	4	5
1	3.08	2.36	2.00	3.00	2.68
2	2.60	3.04	2.56	2.88	1.92
3	3.44	3.28	2.52	3.44	3.28
4	3.04	2.68	1.88	3.64	3.20
Total	12.16	11.36	8.96	12.96	11.08

An example (cont.)

- ▶ $N = 100, n = 5, M = 4$.
- ▶ Population total estimate:

$$\hat{t} = \frac{100}{5}(12.16 + 11.36 + 8.96 + 12.96 + 11.08) = 1130.4.$$

$$\hat{\bar{t}} = 1130.4/100 = 11.304.$$

- ▶ Average estimate:

$$\hat{\bar{y}} = \frac{1130.4}{100 \cdot 4} = 2.826.$$

- ▶ Variance estimate of psu totals

$$v_t = \frac{1}{5-1}[(12.16 - 11.304)^2 + \cdot + (11.08 - 11.304)^2] = 2.256.$$

- ▶ The standard error of $\hat{\bar{y}}$:

$$SE(\hat{\bar{y}}) = \sqrt{\left(1 - \frac{5}{100}\right) \frac{2.256}{5 \cdot 4^2}} = 0.164.$$

Clusters of equal sizes: theory

In stratified sampling, we use ANOVA table. Similarly, here we have

Clusters of equal sizes: theory

In stratified sampling, we use ANOVA table. Similarly, here we have

Population ANOVA Table—Cluster Sampling

Source	df	Sum of Squares	Mean Square
Between psus	$N - 1$	$SSB = \sum_{i=1}^N \sum_{j=1}^M (\bar{y}_{iU} - \bar{y}_U)^2$	MSB
Within psus	$N(M - 1)$	$SSW = \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{y}_{iU})^2$	MSW
Total, about \bar{y}_U	$NM - 1$	$SSTO = \sum_{i=1}^N \sum_{j=1}^M (y_{ij} - \bar{y}_U)^2$	S^2

Clusters of equal sizes: theory

- ▶ To **improve precision**, we hope SSW (sum of square within) is small relative to SSTO (sum of square total)¹.

¹In other words, we hope these stratum are as heterogeneous as possible

Clusters of equal sizes: theory

- ▶ To **improve precision**, we hope SSW (sum of square within) is small relative to SSTO (sum of square total)¹.
- ▶ In cluster sampling, **the opposite situation occurs**. We hope these clusters are homogeneous (or as similar as possible).

¹In other words, we hope these stratum are as heterogeneous as possible

Clusters of equal sizes: theory

- ▶ To **improve precision**, we hope SSW (sum of square within) is small relative to SSTO (sum of square total)¹.
- ▶ In cluster sampling, **the opposite situation occurs**. We hope these clusters are homogeneous (or as similar as possible).
- ▶ With equal sizes, we have

$$V_t = \frac{\sum_{i=1}^N (t_i - t/N)^2}{N-1} = \frac{\sum_{i=1}^N M^2 (\bar{y}_{iU} - \bar{y}_U)^2}{N-1} = M \cdot \text{MSB}.$$

¹In other words, we hope these stratum are as heterogeneous as possible

Clusters of equal sizes: theory

Stratified sampling (within strata sum of squares)



- ▶ To **improve precision**, we hope SSW (sum of square within) is small relative to SSTO (sum of square total)¹.
- ▶ In cluster sampling, **the opposite situation occurs**. We hope these clusters are homogeneous (or as similar as possible).
- ▶ With equal sizes, we have

$$V_t = \frac{\sum_{i=1}^N (t_i - t/N)^2}{N-1} = \frac{\sum_{i=1}^N M^2 (\bar{y}_{iU} - \bar{y}_U)^2}{N-1} = M \cdot \text{MSB}.$$

- ▶ The variance of estimated total is

$$\text{Var}(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{M \cdot \text{MSB}}{n}.$$

¹In other words, we hope these stratum are as heterogeneous as possible

Clusters of equal sizes: theory

- ▶ If MSB/MSW is large in cluster sampling, then cluster sampling decreases precision.

Example 4

If we took a cluster sample of classes and sampled all students within the selected classes, we would likely find that average reading scores varied from class to class. An excellent reading teacher might raise the reading scores for the entire class; a class of students from an area with much poverty might tend to be undernourished and not score as highly at reading. Unmeasured factors, such as teaching skill or poverty, can affect the overall mean for a cluster, and thus cause MSB to be large.

Clusters of equal sizes: theory

1. Compare with SRS

Now, instead of taking n psus each with M units, we take $n \cdot M$ SRS.

$$\text{Var}(\hat{t}_{SRS}) = (NM)^2 \left(1 - \frac{nM}{NM}\right) \frac{S^2}{nM} = N^2 \left(1 - \frac{n}{N}\right) \frac{MS^2}{n}.$$

Thus, if $MSB > S^2$, then cluster sampling is more volatile than SRS.

2. Intraclass correlation coefficient (ICC)

Defined as the Pearson correlation coefficient for the $NM(M-1)$ pairs (y_{ij}, y_{ik}) for $i \in \{1, 2, \dots, N\}$ and $j \neq k$.

$$\text{ICC} = 1 - \frac{M}{M-1} \cdot \frac{\text{SSW}}{\text{SSTO}}.$$

Smaller ICC = better clustering!

Clusters of equal sizes: theory

ICC formula:

$$ICC = 1 - \frac{M}{M-1} \cdot \frac{SSW}{SSTO}$$

In light of ICC, we have

$$\frac{\text{Var}(\hat{t}_{\text{cluster}})}{\text{Var}(\hat{t}_{\text{SRS}})} = \frac{MSB}{S^2} = \frac{NM-1}{M(N-1)} [1 + (M-1) \cdot ICC].$$

- ICC ↗, SSW ↘, more stable intraclass, less efficient the cluster sampling method.

Clusters of equal sizes: theory

An alternative measure of intraclass correlation:

$$R_{\alpha}^2 = 1 - \frac{\text{MSB}}{S^2}$$

With R_{α}^2 , we have

$$\frac{\text{Var}(\hat{t}_{\text{cluster}})}{\text{Var}(\hat{t}_{\text{SRS}})} = \frac{\text{MSB}}{S^2} = 1 + \frac{N(M-1)}{N-1} R_{\alpha}^2.$$

An example

Table 1. psu level data.

	Population A			Population B		
psu 1	10	20	30	9	10	11
psu 2	11	20	32	17	20	20
psu 3	9	17	31	31	32	30

Tbale 2. Within-cluster mean and variance.

	Population A		Population B	
	\bar{y}_{iU}	S_i^2	\bar{y}_{iU}	S_i^2
psu 1	20	100	10	1
psu 2	21	111	19	3
psu 3	19	124	31	1

An example (cont.)

ANOVA Table for Population A:

Source	df	SS	MS
Between psus	2	6	3
Within psus	6	670	111.67
Total, about mean	8	676	84.5

For population A:

$$R_a^2 = 1 - \frac{111.67}{84.5} = -0.3215$$

$$\text{ICC} = 1 - \left(\frac{3}{2}\right) \frac{670}{676} = -0.4867$$

ANOVA Table for Population B:

Source	df	SS	MS
Between psus	2	666	333
Within psus	6	10	1.67
Total, about mean	8	676	84.5

For population B:

$$R_a^2 = 1 - \frac{1.67}{84.5} = 0.9803$$

$$\text{ICC} = 1 - \left(\frac{3}{2}\right) \frac{10}{676} = 0.9778$$

Clusters with unequal sizes

Unbiased estimate of population total:

$$\hat{t} = \frac{N}{n} \cdot \sum_{i \in S} t_i.$$

Clusters with unequal sizes

Unbiased estimate of population total:

$$\hat{t} = \frac{N}{n} \cdot \sum_{i \in S} t_i.$$

► If we know $M_1 \sim M_N$, then *an unbiased estimate of population mean is*

$$\hat{y}_U = \frac{\hat{t}}{M_0} = \frac{\hat{t}}{\sum_{i=1}^N M_i} \quad (\text{unbiased}).$$

Clusters with unequal sizes

Unbiased estimate of population total:

$$\hat{t} = \frac{N}{n} \cdot \sum_{i \in S} t_i.$$

► If we know $M_1 \sim M_N$, then an unbiased estimate of population mean is

$$\hat{y}_U = \frac{\hat{t}}{M_0} = \frac{\hat{t}}{\sum_{i=1}^N M_i} \quad (\text{unbiased}).$$

► If we do not know $M_1 \sim M_N$, then a ratio estimation has to be applied:

$$\hat{y}_U = \frac{\hat{t}}{\hat{M}_0} = \frac{\frac{N}{n} \cdot \sum_{i \in S} t_i}{\frac{N}{n} \cdot \sum_{i \in S} M_i} = \frac{\sum_{i \in S} t_i}{\sum_{i \in S} M_i}.$$

Two-stage cluster sampling

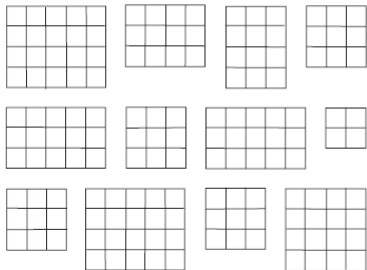
The steps:

1. Select SRS S of n psus from the population N psus.
2. Select SRS of ssus from each selected psu. The STS of m_i units from the i th psu is denoted S_i .

The steps

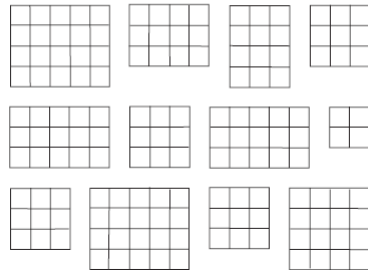
One-Stage

Population of N psu's:



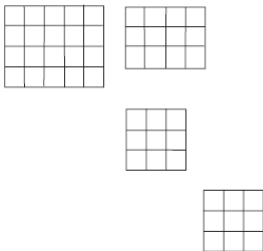
Two-Stage

Population of N psu's:

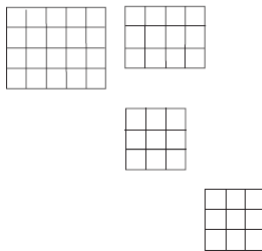


The steps

Take an SRS of n psu's:



Take an SRS of n psu's:

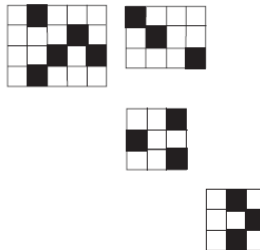


The steps

Sample all ssu's in sampled psu's:



Take an SRS of m_i ssu's in sampled psu i :



The estimations

- ▶ Since we do not observe all the units from the i th psu, we have to estimate the total of it

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} \cdot y_{ij} = M_i \cdot \bar{y}_i.$$

The estimations

- ▶ Since we do not observe all the units from the i th psu, we have to estimate the total of it

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} \cdot y_{ij} = M_i \cdot \bar{y}_i.$$

- ▶ An unbiased estimator of the population total is

$$\hat{t} = \frac{N}{n} \sum_{i \in S} \hat{t}_i.$$

The estimations

- ▶ Since we do not observe all the units from the i th psu, we have to estimate the total of it

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} \cdot y_{ij} = M_i \cdot \bar{y}_i.$$

- ▶ An unbiased estimator of the population total is

$$\hat{t} = \frac{N}{n} \sum_{i \in S} \hat{t}_i.$$

- ▶ Due to the uncertainty of estimating i th psu's total,

$$\text{Var}(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{V_t}{n} + \frac{N}{n} \sum_{i=1}^N \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{V_i}{m_i}.$$

The estimations

- To estimate $\text{Var}(\hat{t})$,

$$v_t = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\sum_{i \in S} t_i}{n} \right)^2,$$

$$v_i = \frac{1}{m_i - 1} \sum_{j \in S_i} (y_{ij} - \bar{y}_i)^2.$$

The estimations

- ▶ To estimate $\text{Var}(\hat{t})$,

$$v_t = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\sum_{i \in S} t_i}{n} \right)^2,$$

$$v_i = \frac{1}{m_i - 1} \sum_{j \in S_i} (y_{ij} - \bar{y}_i)^2.$$

- ▶ To estimate population mean, we use ratio estimation

$$\hat{y}_U = \frac{\hat{t}}{\hat{M}_0} = \frac{\sum_{i \in S} \hat{t}_i}{\sum_{i \in S} M_i}.$$