Second Regional Training Course on Sampling Methods for Producing Core Data Items for Agricultural and Rural Statistics

Module 2: Review of Basics of Sampling Methods

Session 2.4: Clustering and Single-Stage Cluster Sampling

9 – 20 November 2015,
Jakarta, Indonesia

BPS, Statistics Indonesia

Contents

- Clustering and stratification
- Single-Stage Cluster sampling
  - ➢ *epsem* selection and estimation
  - ➢ Selection method (single stage) – *epsem* and PPSWR and PPS systematic
  - ➢ Estimation under PPS cluster sampling

Clustering and Stratification

Strata and Clusters

- Both stratification and clustering involve subdividing the population into mutually exclusive groups.
- Sub-divisions of the population are called 'clusters' or 'strata' depending upon the sampling procedure adopted.
- The term 'cluster' is used in the context of cluster sampling and multi-stage (cluster) sampling.
- To understand the application of these in different situations, let us take a simple example.

Choice of Strata and Clusters - an Illustration

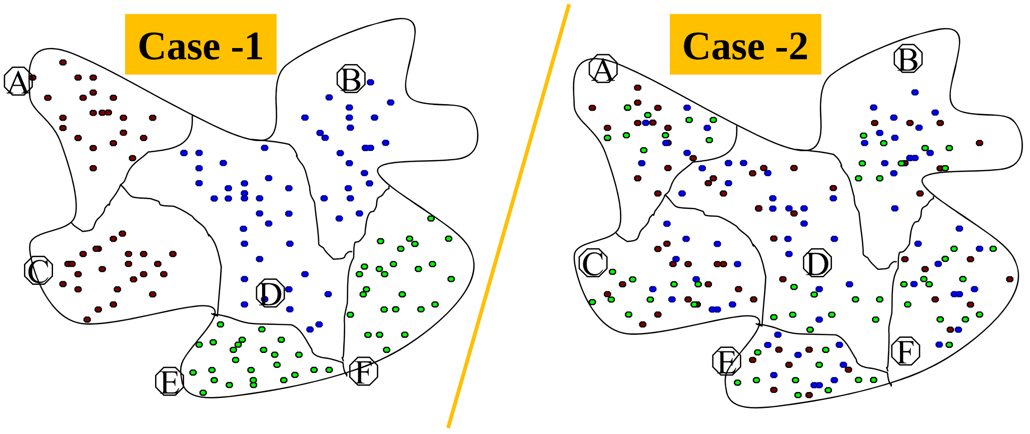
Using data of a Agricultural Census taken 5 years ago, a population of about 120 agricultural holdings is found to be distributed in six villages each having approximately equal number of holdings.

Task: to estimate the current proportion of marginal holdings in these six villages.

Resources permit only a sample size of 20.

Composition of villages in last Census

Case 1: each sub-division has only one type of units
Case 2: each sub-division has an uniform composition by type.



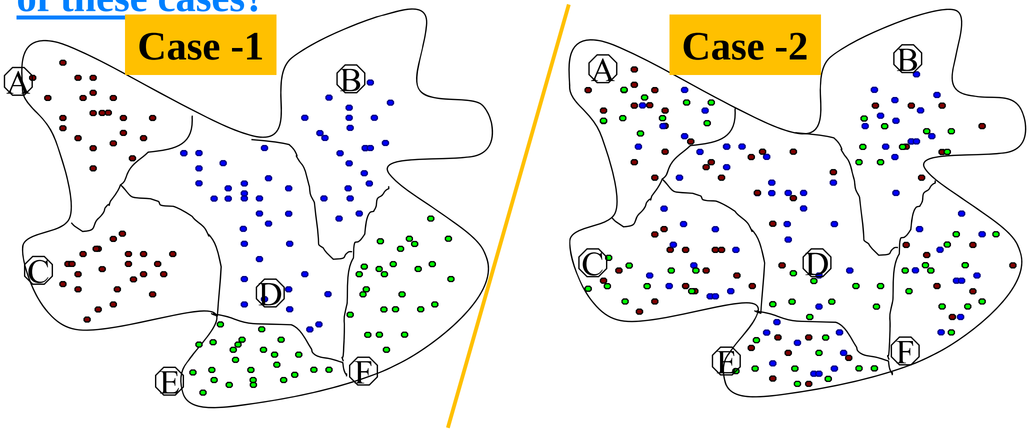
● marginal ● small & semi-medium ● medium & large

Suggest a strategy of selection

Task: to estimate the proportion of marginal holdings in these six villages.

Resources permit only a sample size of 20.

What will be your strategy of selecting a sample for each of these cases?

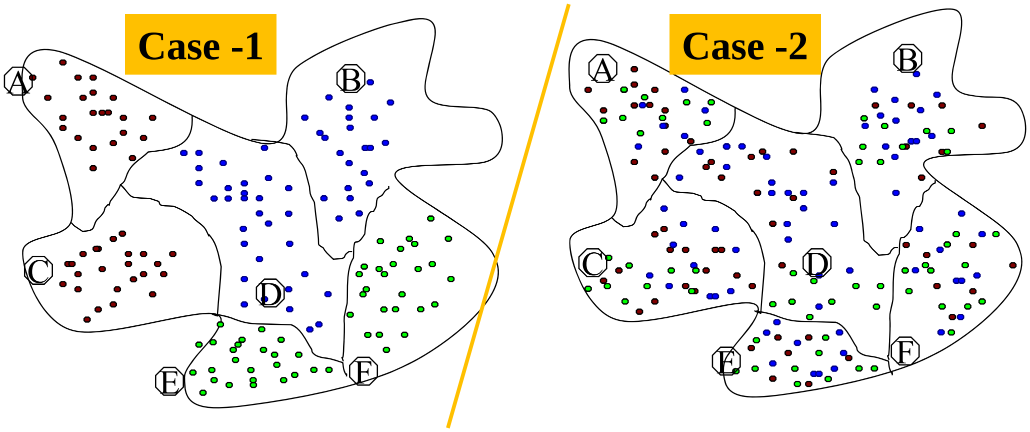


● marginal ● small & semi-medium ● medium & large

Homogeneity and clustering

Case 1: each village is homogenous - thus stratified sampling preferable. [$\rho = 1$]

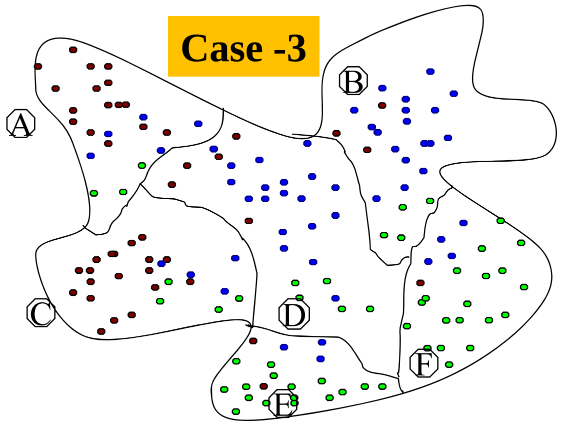
Case 2: each village a replica of the population – studying one cluster is sufficient. [ρ {= -1/(B-1)}< 0]



● marginal ● small & semi-medium ● medium & large

Composition commonly found

In practice, the villages are neither of uniform composition nor totally homogenous. It is usually like the case 3 below:



Value of ρ depends on the variable under consideration and nature of clusters.

Usually, (as in case 3)

$\rho > 0$

● marginal ● small & semi-medium ● medium & large

Clustering and Stratification

Clusters are …

- "Heterogeneous" groups of population elements

Examples:

- A village is a cluster of holdings
- A holding is a cluster of plots
- A dairy farm is a cluster of cattle (buffaloes?)
- An enterprise is a cluster of workers

Clustering and Stratification in Sample Design

- Typically, sample surveys conducted by NSOs involve subdividing the population into strata and clusters.
- Usually, the technique of stratifying the clusters and then further stratifying the units within clusters are applied to obtain the final sample.
- The sampler's objective is to get the right combination of stratification and clustering to get the required estimates at the desired level of accuracy with the given resources.

Clustering and Stratification in Sample Design (Contd.)

- The reliability or precision of the estimates depends on the degree to which the sample is *clustered.*
- Generally, *clustering* increases the *sampling variance* considerably.
- Usually, stratification is applied to decrease the *sampling variance,* but its effect is often not significant.
- Effects of *clustering* and *stratification* is measured by the *design effect* or *deff.*
- Primarily, *deff* indicates, how much *clustering* there is in the survey sample.

Single-Stage Cluster Sampling

Cluster sampling

- Cluster sampling - selection of a sample of clusters and survey all the units of each selected clusters.
- This is also called ‘Single-stage cluster sampling’.
- ‘Multi-stage cluster sampling’ or simply ‘multi-stage sampling’: Instead of doing survey of all the units of selected clusters, only a sample of units are taken from each selected clusters.

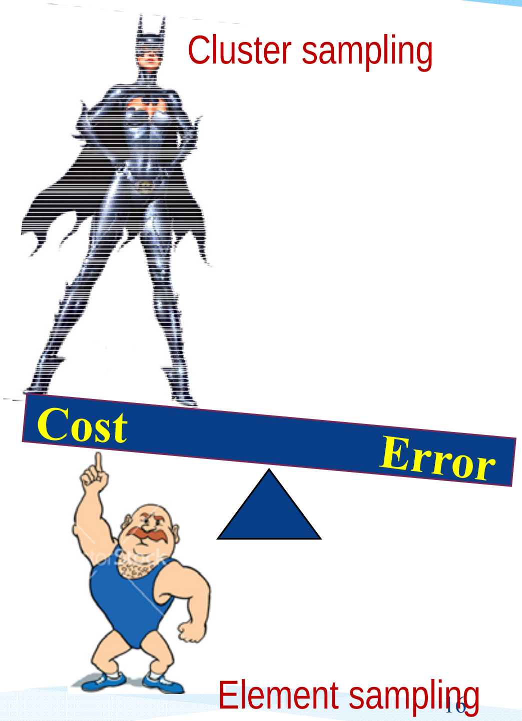
What is wrong with element sampling?

- Element sampling: Every "element" in the population is a "sampling unit"

Problem

- □ Need a complete list of elements
- □ Often an updated list of elements is not available
- □ Costly to make a list of population elements

Cluster Sampling



Solution

- □ Select a group of elements (Cluster)
- □ Then list elements only within the selected cluster

Cluster sampling

- Example: In a survey of dwellings, Select blocks first and then list dwellings only in the selected blocks

❑ Clusters: Blocks

❑ Elements: Dwellings

Cluster Sampling

Selecting a cluster sample involves

1) Create sampling frame: list of all clusters
2) From the list, select a sample of clusters – by using a selection method (e.g., SRS, Systematic,..)
3) List all population units within the selected clusters
4) Collect data from all units within the selected clusters

Cluster sampling - Advantages

Main advantage

- Exact knowledge of the size of the sub-divisions (clusters) not required, unlike that for stratified sampling.
- Often a complete list of clusters

  - defined by location or as social entities or by institutions –

  is available, but frame of population units is not available or is costly to obtain.

  In such cases, cluster sampling can be adopted.
- Reduced cost of personal interviews, particularly when the survey cost increases with the distance separating the sampled units.

Cluster sampling - Disadvantages

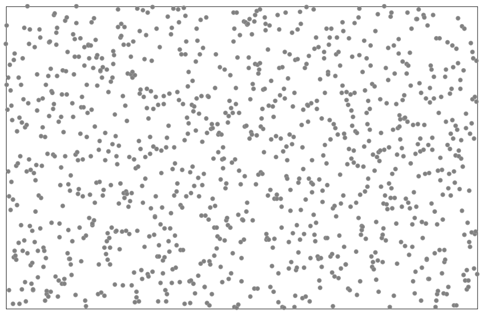
Main disadvantage

Increased sampling error due to a less representative sample, since:

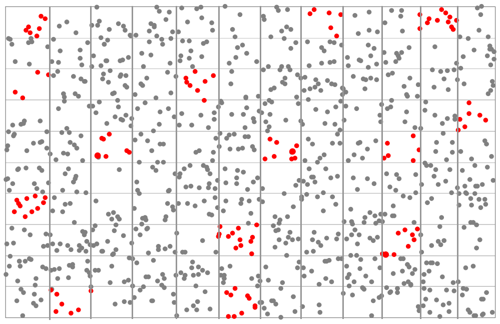
- in practice, units are typically homogeneous within normally defined clusters
- and the composition of clusters can not be altered, as they are pre-defined.

Cluster Sampling

Population



clustered sample



Estimation under Cluster sampling (1)

For a **quantitative variable**, the observed value (y_i) for a sampled (ith) cluster is

- ➢ sum of observed value of all units in the ith cluster

i.e. $y_i = \sum_{j=1}^{m_i} y_{ij}$ where

m_i : number of units in the ith cluster

y_{ij} : value of the jth unit of the ith cluster

Under *epsem*:

If n clusters are selected out of N clusters in the population, estimate of the variable Y is

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} y_i = \frac{N}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} y_{ij}$$

Estimation under Cluster sampling (2)

For a **categorical variable,** the observed value (y_i) for a sampled (ith) cluster is

- ➢ Number of units in a category in the ith cluster

We define

$y_{ij} = 1 \quad if\ the\ j^{th}$ of the i^{th} cluster is in the category

0 otherwise

Then the same formula applies for the estimate of units in the given category, i.e.

$$\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i = \frac{N}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij}$$

Selection of Clusters

- *epsem* and PPS

Cluster Sampling - *epsem*

Common alternatives

Commonly used *epsem* designs used for single stage cluster sampling:

- SRSWOR
- Circular systematic

Estimation under *epsem*:

As discussed earlier.

The efficiency of estimates from a *epsem* sample of clusters is often not very good.

Using auxiliary variables, if available, at the stage of selection or estimation often help improving the efficiency.

A Question

Suppose we have an auxiliary variable, Z, for a study variable Y.

Let, Z_i: the value of Z for the *i*th cluster

Y_i : *the value of* Y for the *i*th cluster

Z_i's are perfectly proportional to Y_i's.

You know the value of Z (Zi's) for all the clusters.

What will your strategy be for estimating Total of Y?

Cluster of Sub-units

Cluster	Sub-units	# times sub-units in the cluster
1	Y_1/Z_1	Z_1
2	Y_2/Z_2	Z_2
3	Y_3/Z_3	Z_3
…	…	…
i	Y_i/Z_i	Z_i
…	…	…
N	Y_N/Z_N	Z_N
Total	**Y**	**Z**

Sampling with Probability Proportional to Size (PPS)

- Probability of selection is related to an auxiliary variable, *Z*, that is a measure of "size"

  Example

  Number of households

  Area of farms

- "Larger" units are given higher chance of selection than "smaller" units
- Selection probability of *i*th unit is $p_i = Z_i \Big/ \sum_{i=1}^{N} Z_i$

  $i = 1,2, \ldots , N$

Estimation under PPSWR

If the sample size is *n*, estimate of total of *Y* is

$$\hat{Y} = {}^{1}\!/_{n} \sum_{1}^{n} \frac{\sum_{1}^{N} Z_i}{Z_i}\, y_i = {}^{1}\!/_{n} \sum_{1}^{n} \frac{Z}{Z_i}\, y_i$$

where

$$Z = \sum_{1}^{N} Z_i$$

PPS Selection Procedures

- Cumulative total method: with replacement
- Cumulative total method: without replacement
- PPS systematic sampling

PPS Selection Procedures

1) Cumulative total method: with replacement

2) Cumulative total method: without replacement

3) PPS systematic sampling

Cumulative Total Method

Select 5 villages using PPSWR sampling *(size is number of households)*

Village	No. of HHs (Measure of Size)	Selection Probability
1	47	0.067
2	45	0.064
3	28	0.040
4	29	0.041
5	45	0.064
6	36	0.051
7	58	0.083
8	29	0.041
9	31	0.044
10	21	0.030
11	47	0.067
12	17	0.024
13	28	0.040
14	41	0.059
15	22	0.031
16	32	0.046
17	25	0.036
18	41	0.059
19	33	0.047
20	45	0.064
Total	700	

Solution

- Sampling unit: **village**
- Measure of size: **number of households in village**
- Selection probability:

$$p_i = \frac{\text{number of HHs in village i}}{\text{total number of HHs}}$$

Cumulative Total Method (Contd.)

- Write down cumulative total for the sizes Z_i, i=1,2..N
- Choose a random number r such that $1 \le r \le Z$
- Select i^{th} population unit if
- $T_{i-1} \le r \le T_i$ where

$$T_{i-1} = Z_1 + Z_2 + .. + Z_{i-1}$$

and

$$T_i = Z_1 + Z_2 + .. + Z_i$$

Village	No. of HHs (Measure of Size) (Z_i)	Cumulative Size (T_i)	Assigned Random Numbers
1	47	47	1 - 47
2	45	92	48 - 92
3	28	120	93 -120
4	29	149	121 - 149
5	45	194	150 - 194
6	36	230	195- 230
7	58	288	231 - 288
8	29	317	289 - 317
9	31	348	318 - 348
10	21	369	349 - 369
11	47	416	370 - 416
12	17	433	417 - 433
13	28	461	434 - 461
14	41	502	462 - 502
15	22	524	503 - 524
16	32	556	525 - 556
17	25	581	557 - 581
18	41	622	582 - 622
19	33	655	623 - 655
20	45	700	656 - 700
Total	700		

Cumulative Total Method (Contd.)

1) A random number r, $1 \le r \le 700$, is selected
2) Suppose $r = 259$
3) $231 \le 259 \le 288$, the 7th village is selected.
4) Suppose the next 4 random numbers are 548, 170, 231, 505.
5) Samples selected using PPSWR are 16th, 5th, 7th, 15th .

Note: The 7th village is selected twice.

Village	No. of HHs (Measure of Size) (Z_i)	Cumulative Size (T_i)	Assigned Random Numbers
1	47	47	1 - 47
2	45	92	48 - 92
3	28	120	93 -120
4	29	149	121 - 149
5	45	194	150 - 194
6	36	230	195- 230
7	58	288	231 - 288
8	29	317	289 - 317
9	31	348	318 - 348
10	21	369	349 - 369
11	47	416	370 - 416
12	17	433	417 - 433
13	28	461	434 - 461
14	41	502	462 - 502
15	22	524	503 - 524
16	32	556	525 - 556
17	25	581	557 - 581
18	41	622	582 - 622
19	33	655	623 - 655
20	45	700	656 - 700
Total	700		

Cumulative Total Method (Contd.)

1) For a PPSWR the sample would be: 16th, 5th, 7th, 15th , with 7th village repeated.
2) For a PPSWOR selection, we have to continue further to get 5 distinct units in the sample.
3) Suppose the next random selected is r = 375,

The required PPSWOR sample would be 16th, 5th, 7th, 15th & 11th .

Village	No. of HHs (Measure of Size) (Z_i)	Cumulative Size (T_i)	Assigned Random Numbers
1	47	47	1 - 47
2	45	92	48 - 92
3	28	120	93 -120
4	29	149	121 - 149
5	45	194	150 - 194
6	36	230	195- 230
7	58	288	231 - 288
8	29	317	289 - 317
9	31	348	318 - 348
10	21	369	349 - 369
11	47	416	370 - 416
12	17	433	417 - 433
13	28	461	434 - 461
14	41	502	462 - 502
15	22	524	503 - 524
16	32	556	525 - 556
17	25	581	557 - 581
18	41	622	582 - 622
19	33	655	623 - 655
20	45	700	656 - 700
Total	700		

PPS systematic

1) Derive cumulative totals for the sizes Z_i, i=1,2..N, and allot random numbers to different units.
2) Calculate interval $k = Z_N / n$ (in this case 700/5 = 140)
3) Select a random number r (say 101) from 1 to k; and obtain $r+k$, $r+2k$, $r+3k$, …, $r+(n-1)k$
4) In this case, the selected cumulative sizes are 101, 241, 382, 523 & 664.

Village	No. of HHs (Measure of Size) (Z_i)	Cumulative Size (T_i)	Assigned Random Numbers
1	47	47	1 - 47
2	45	92	48 - 92
3	28	120	93 -120
4	29	149	121 - 149
5	45	194	150 - 194
6	36	230	195- 230
7	58	288	231 - 288
8	29	317	289 - 317
9	31	348	318 - 348
10	21	369	349 - 369
11	47	416	370 - 416
12	17	433	417 - 433
13	28	461	434 - 461
14	41	502	462 - 502
15	22	524	503 - 524
16	32	556	525 - 556
17	25	581	557 - 581
18	41	622	582 - 622
19	33	655	623 - 655
20	45	700	656 - 700
Total	700		

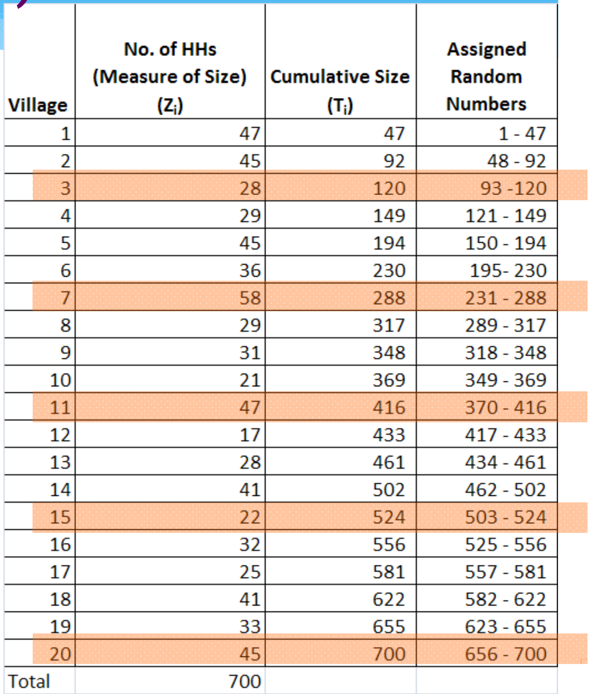
Cluster Sampling - Selection

PPS Systematic (Contd.)

- The selected units are:

  3rd (for 101),
  7th (for 241),
  11th (for 382),
  15th (for 523) &
  20th (for 664)

- *Note:* If any unit has size greater than *k*, it may be selected more than once.



Village	No. of HHs (Measure of Size) (Z_i)	Cumulative Size (T_i)	Assigned Random Numbers
1	47	47	1 - 47
2	45	92	48 - 92
3	28	120	93 -120
4	29	149	121 - 149
5	45	194	150 - 194
6	36	230	195- 230
7	58	288	231 - 288
8	29	317	289 - 317
9	31	348	318 - 348
10	21	369	349 - 369
11	47	416	370 - 416
12	17	433	417 - 433
13	28	461	434 - 461
14	41	502	462 - 502
15	22	524	503 - 524
16	32	556	525 - 556
17	25	581	557 - 581
18	41	622	582 - 622
19	33	655	623 - 655
20	45	700	656 - 700
Total	700		

Thanks