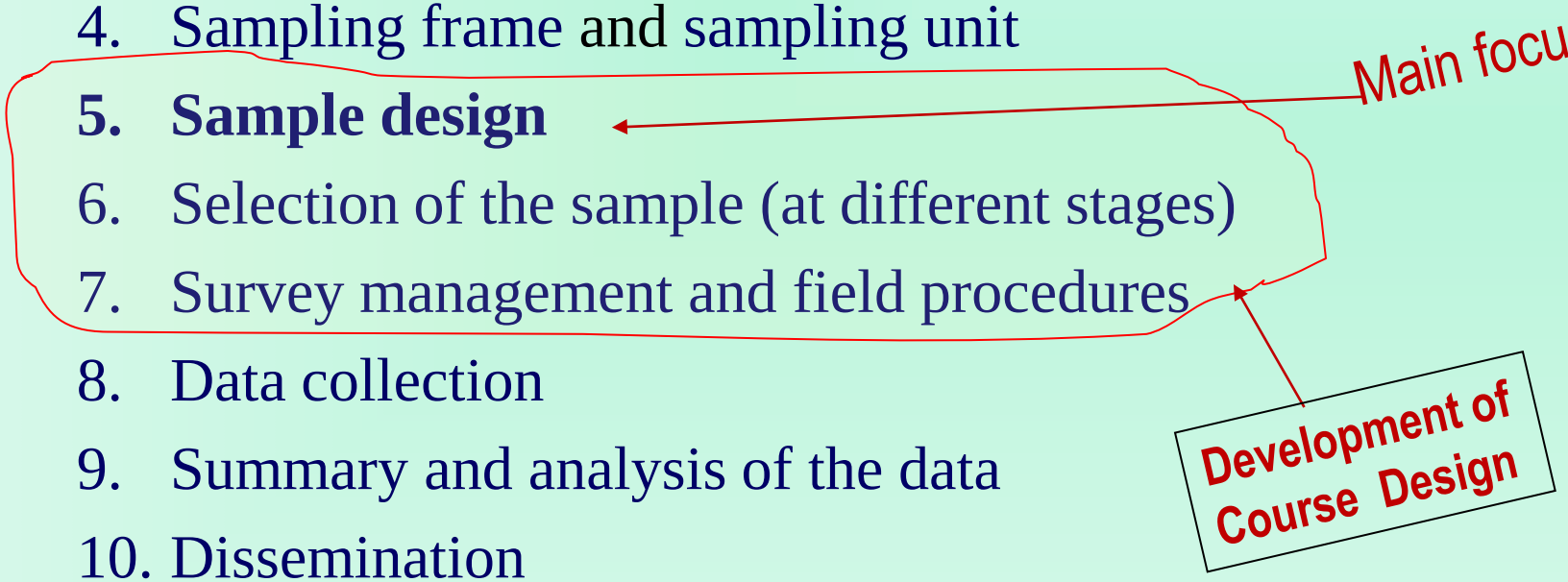
Basic Concepts of Sampling-Brief Review: Sampling Designs

Dr. A.C. Kulshreshtha
U.N. Statistical Institute for Asia and the Pacific (SIAP)

Second RAP Regional Workshop on
Building Training Resources for Improving Agricultural & Rural Statistics
Sampling Methods for Agricultural Statistics-Review of Current Practices
SCI, Tehran, Islamic Republic of Iran
10-17 September 2013

Survey Design – Issues involved

1. Determining survey objectives and data requirements
2. The population of interest or the target population
3. Reference period; Geographic and demographic boundaries
4. Sampling frame and sampling unit
5. **Sample design**
6. Selection of the sample (at different stages)
7. Survey management and field procedures
8. Data collection
9. Summary and analysis of the data
10. Dissemination



Introduction to Survey Design

Sample is a subset of the population on which observations are taken for obtaining information about the population.

Since studying a sample we wish to draw valid conclusions about the population, sample should desirably be 'representative' of the target population.

Sampling design specifies how to select the part of the population to be surveyed

Probability Sampling Designs

The most common sampling techniques used for official surveys are :

- **Simple Random Sampling**
- **Systematic Sampling**
- **Stratified Sampling**
- **Probability Proportional to Size (PPS) sampling**
- **Cluster Sampling**
- **Multi-Stage Sampling**

All are examples of probability sampling

Basic Sampling Schemes

Simple random sampling (SRS): is a probability selection scheme where each unit in the population is given an equal probability of selection

Systematic sampling: A method in which the sample is obtained by selecting every k^{th} element of the population, where k is an integer > 1. Often the units are ordered with respect to that auxiliary data

Stratified sampling: Uses auxiliary information (stratification variables) to divide the sampling units of the population into groups called 'strata' and increase the efficiency of a sample design

Basic Sampling Schemes (Contd.)

Probability Proportional to Size (PPS): The procedure of sampling in which the units are selected with probability proportional to a given measure of size

The size measure is the value of an *auxiliary variable* ***X*** related to the characteristic ***Y*** under study

Simple Random Sampling (SRS)

What is SRS ?

- SRS is simplest method of probability sampling
- SRS is special type of equal probability selection method (*epsem*)
- Rarely used in practice for large scale surveys but is theoretical basis for other sample designs
- SRS selection can be made
  - With Replacement (*SRSWR*) or
  - Without Replacement (*SRSWOR*)
- Selection probability: the probability that a population unit is selected at any given draw is the same, namely $\frac{1}{N}$

  for both SRSWR and SRSWOR

  N: number of units in the population (*Population size*)

Selection Procedure- Steps involved:

- Get a list (sampling frame) which uniquely identifies each unit in the population
- Allocate a serial number to each unit of the frame
- Generate random numbers [in the range of 1 to *N*] using Random Number Table/ Random Number Generator on computer:
  - For SRSWR: select the units with the serial numbers same as the first *n* random numbers generated, even if there be repetitions
  - For SRSWOR: select the units with the serial numbers same as the first *n* distinct random numbers generated

Estimators of Mean, Variance, Standard Error, CV in SRS

Estimator of variance of sample mean

In SRSWR $s^2_{\bar{y}} = \frac{s^2_y}{n}$ and in SRSWOR $v(\bar{y}) = (1-f)\frac{s^2}{n}$

Where, sample mean $\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$, and estimated sample variance is $s^2_y = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$

Estimated standard error of $\bar{y}$ will be $\sqrt{v(\bar{y})}$

Estimated CV of sample mean $\hat{CV}(\bar{y}) = \sqrt{\frac{(1-f)}{n}}\left(\frac{s}{\bar{y}}\right)\times 100$

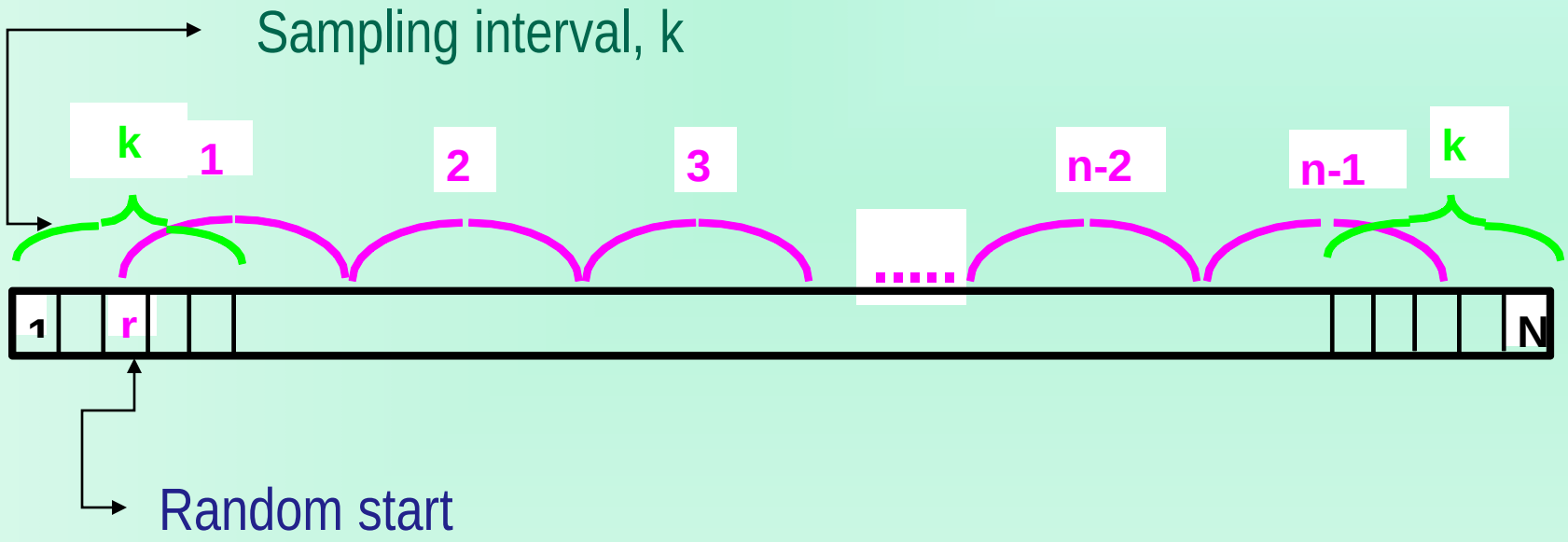
Estimator of variance of proportion *p*: $v(p) = (1-f)\frac{p(1-p)}{n-1}$

Systematic Sampling

- Systematic Sampling (SYS), like SRS, involves selecting n sample units from a population of N units
- Instead of randomly choosing the n units in the sample, a skip pattern is run through a list (frame) of the N units to select the sample
- The *skip* or *sampling interval*, $k = N/n$

Systematic Sampling (Contd.)

Linear Systematic Sampling



Systematic Sampling (Contd.)

Selection Procedure - *Linear Systematic Sampling*

Steps involved:

- Form a **sequential list** of population units
- Decide on a sample size *n* and compute the skip (*sampling interval*), $k = N/n$
- Choose a random number, *r* (*random start*) between *1* and *k* (inclusive)
- Add "*k*" to selected random number to select the second unit and continue to add "*k*" repeatedly to previously selected unit number to select the remainder of the sample

Systematic Sampling (Contd.)

Problem - *Linear Systematic Sampling*

If *N* is a multiple of *n*, then the number of units in each of the *k* possible systematic samples is *n*

In this case systematic sampling amounts to grouping the *N* units into *k* samples of exactly *n* units each in a systematic manner and selecting one of them with probability 1/*k*

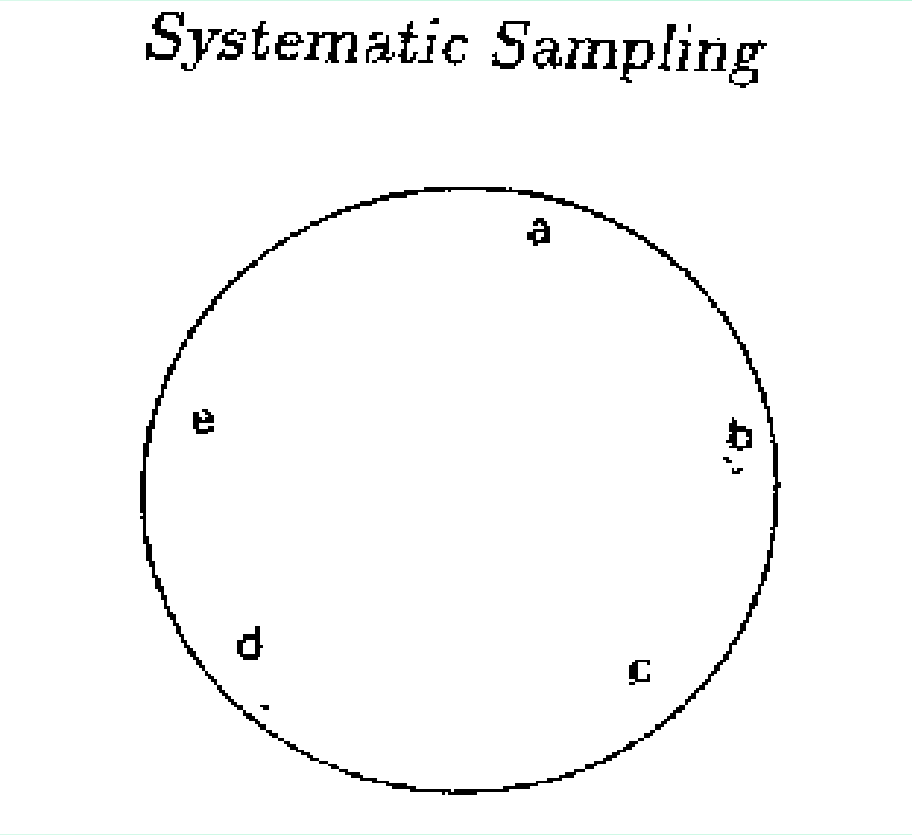
In this case, the sampling scheme is *epsem*

But, if *N/n* is not an integer, then the number of units selected systematically with the sampling interval *k* [= nearest integer to *N/n*] – no longer *epsem*

This problem may be overcome by adopting a device, known as *circular systematic sampling*

Systematic Sampling (Contd.)

Circular Systematic Sampling



K=5/2=2.5

a) If k=2 possible samples are:

ac; bd; ce; da and eb

b) If k=3 possible samples are:

ad; be; ca; db and ec

Systematic Sampling (Contd.)

Circular Systematic selection

- Useful when ***N/n*** is not integer
- Determine the interval ***k*** – rounding down to the integer nearest to ***N/n***

  [If ***N*** = 15 and ***n*** = 4, then ***k*** is taken as 3 and not 4]
- Take a random start between 1 and ***N***
- Skip through the circle by ***k*** units each time to select the next unit until ***n*** units are selected
- Thus there could be ***N*** possible distinct samples instead of ***k***
- This method is termed ***Circular Systematic Sampling*** (CSS)

Systematic Sampling – Important Features

- Often used as an alternative to SRS
- Requires ordering of the population units
  - Ordering enables SYS sample to be more representative
  - Ordering done by geographical location (say of dwellings) ensures fair spread of sample
  - Ordering done by industry type ensures fair representation of industries
- Ensures each population unit equal chance of being selected into sample

Systematic Sampling (Contd.)

Advantages and Disadvantages

Advantages:

- Operationally convenient - easier to draw a sample.
- SYS distributes the sample more evenly over the population – thus likely to be more efficient than SRSWOR,
  particularly when the ordering of the units in the list is related to characteristics of the variable of interest

Disadvantages :

- Requires complete list of the population
- A bad arrangement of the units may produce a very inefficient sample
- Variance estimates cannot be obtained from a single systematic sample

Systematic Sampling (Contd.)

Conceptual Framework (General)

Row Number	Random start (cluster number)					
	1	2	….	r	….	k
1	1	2	….	r	….	k
2	1+k	2+k	….	r+k	….	2k
.	.	.	….	.	….	.
.	.	.	….	.	….	.
.	.	.	….	.	….	.
j	1+(j-l)k	2+(j-l)k	….	r+(j-l)k	….	jk
.	.	.	….	.	….	.
.	.	.	….	.	….	.
n	1+(n-l)k	2+(n-1)k	….	r+(n-l)k	….	nk

The above table shows serial numbering of the *nk* population units. There are k possible samples (clusters).

We select **one cluster** by a random start between 1 and k.

Estimation of Variance of Systematic Sample

- SYS is a random sample of one cluster only.
- No estimate of variance can be formed from the sample
- Approximately treating the SYS as a random sample of n units, estimate of variance would be

$$v(\bar{y}_{sy}) = \left(\frac{1}{n} - \frac{1}{nk}\right) S_{wc}^{2}$$

- where S_{wc}^{2} is the mean square within the selected cluster (systematic sample) among units
- This is, however, not an unbiased estimator

Design Variance of Systematic Sample

- Design Variance of Sample Mean in Systematic Sampling *is*

$$V(\bar{y}_{sys}) = \sigma_b^2 = \frac{1}{k}\sum_{r=1}^{k}(\bar{y}_r - \bar{Y})^2 = \frac{k-1}{k}S_c^2$$

where S_c^2 denotes the mean square between the clusters

- Variance of systematic sample mean is the between sample (cluster) means variance $\sigma_b^2 = \sigma^2 - \sigma_w^2$

$$\sigma^2 = \frac{1}{nk}\sum_{r=1}^{k}\sum_{j=1}^{n}(y_{rj} - \bar{Y})^2 = \frac{1}{k}\sum_{r=1}^{k}(\bar{y}_r - \bar{Y})^2 + \frac{1}{nk}\sum_{r=1}^{k}\sum_{j=1}^{n}(y_{rj} - \bar{y}_r)^2$$

- Correlation coefficient between pairs of sampling units in a 'cluster' of study variable Y is,

$$\rho_c = \frac{\sum_{r=1}^{k}\sum_{j\neq j'=1}^{n}(y_{rj} - \bar{Y})(y_{rj'} - \bar{Y})}{kn(n-1)\sigma^2}$$

- Sampling variance σ_b^2 of the systematic sample mean can also be expressed in terms of ρ_c

$$\sigma_b^2 = \frac{\sigma^2}{n}\left[1 + (n-1)\rho_c\right]$$

Design Efficiency *(DEFF)*

$$DEFF_{sys}(\bar{y}) = \frac{V_{sys}(\bar{y})}{V_{srswr}(\bar{y})} = 1 + (n-1)\rho$$

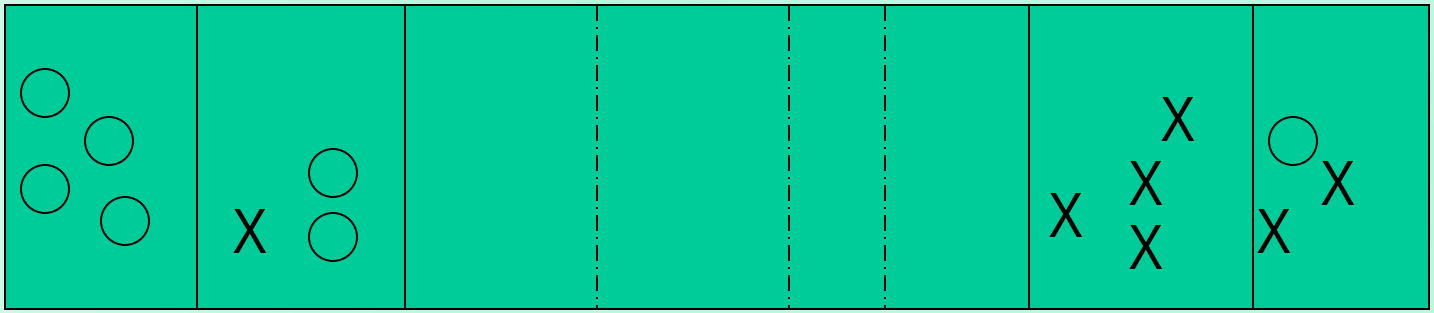
- Systematic sampling compared to simple random sampling with replacement is
  - More efficient, if $-\frac{1}{n-1} < \rho < 0$
  - Equally efficient, if $\rho = 0$
  - Less efficient, if $0 < \rho < 1$

Stratified Sampling- *Stratification*

- Divides the population into a number of distinct groups (strata) based on auxiliary information - referred to as *stratification variables* - relating to study variable(s)
- Each stratum is composed of units that satisfy the condition set by the values of the stratifying variable
- Main objectives:
  - Improve the sample estimations, i.e. to reduce the standard error of the estimates or higher efficiency for given per unit of cost
  - Provide separate estimates required for each sub-division of the population – "domain" estimates
  - Using different sampling procedures for different sub-population, to (i) increase efficiency of the estimates (ii) organize the field work

Stratified Sampling (Contd.)

Stratification
– Mutually Exclusive subsets



Stratum no.	1	2	h	L
Stratum size	N_1	N_2	N_h	N_H

Stratified Sampling (Contd.)

Stratified sampling involves:

- division or stratification of the population into homogeneous (similar) groups called strata
- selecting the sample using a selection procedure
  - like SRS or systematic sampling or PPS within each stratum
  - independent of the other strata
- Sampling in each stratum is carried out independently
  - Sampling fractions may differ
  - Selection procedures may also be different
- The total sample size is distributed over all the strata – **allocation**
- At the end of the survey, the stratum results are combined to provide an estimate for entire population

Stratified Sampling – in practice

- In most surveys - household or establishment surveys - stratification is used
- Stratification can be used with any type of sampling design
- Stratified sampling can be used in
  - Single - stage designs
  - Multi - stage designs

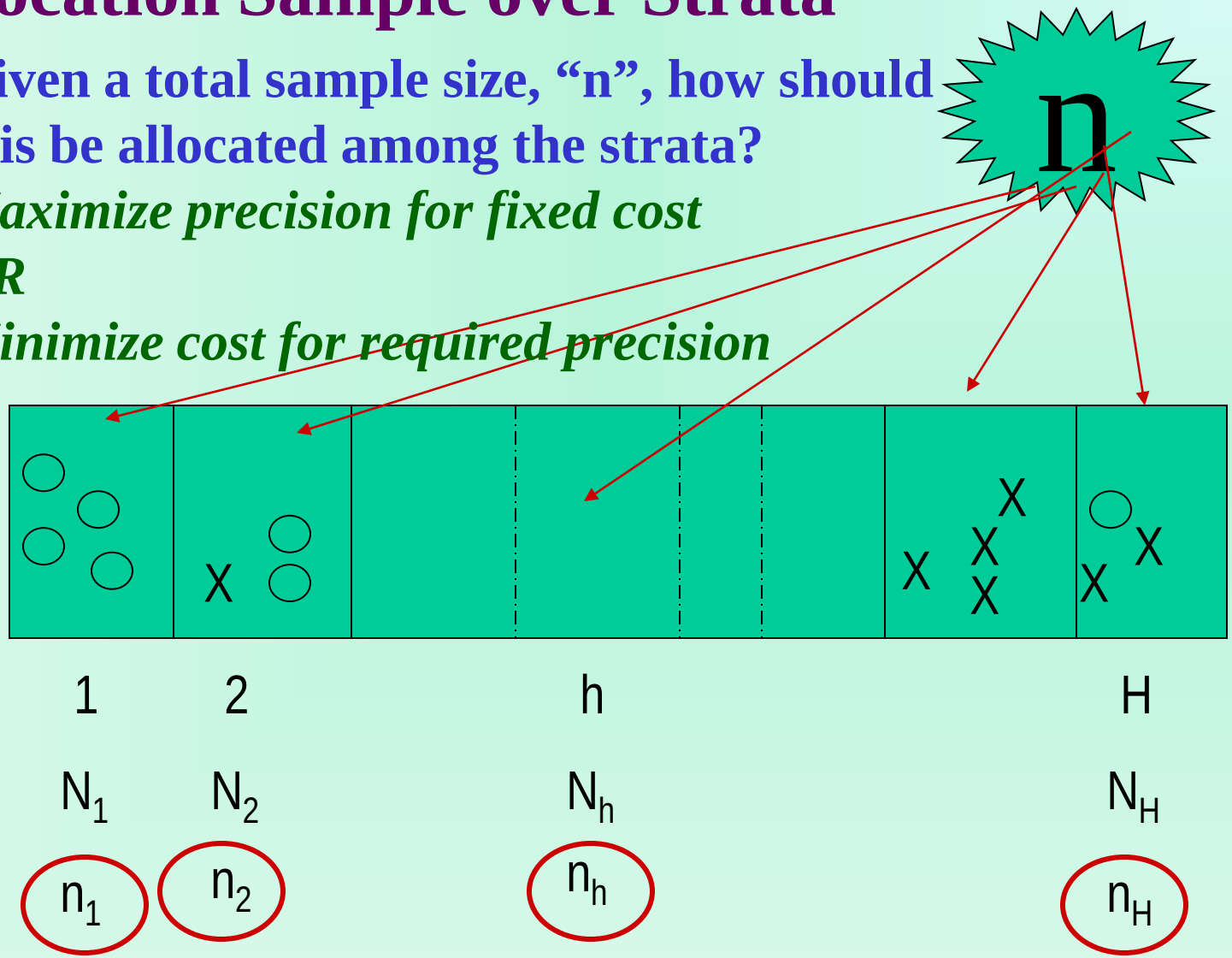
Clustering and Stratification

Defining Strata

1. Choice of stratification variables (location, output, etc):
   - Homogeneous within strata; Heterogeneous across strata
   - Highly correlated with study variables (output with profit or number of employees, etc.)
2. Number of strata
   - Depends on availability of stratifying information in sampling frame: less information, fewer strata
   - At least two sampling units per stratum to be able to compute sampling error

Allocation Sample over Strata

- **Given a total sample size, "n", how should this be allocated among the strata?**
  Maximize precision for fixed cost
  OR
  Minimize cost for required precision



Sample Allocation to Strata

Alternatives Methods:

- Proportionate allocation
  - Uniform or equal allocation
- Disproportionate allocation
  - Optimum allocation (minimum variance), fixed sample size
  - Cost optimum allocation

Sample Allocation to Strata

- In *proportionate stratification*, an uniform sampling fraction is applied to each strata; that is, the sample size selected from each stratum is made proportionate to the population size of the stratum

- In *disproportionate stratification*, different sampling rates are used deliberately in different strata

Proportionate Allocation

- In *proportionate stratification*, $\frac{n_h}{N_h}$

  is specified to be the same for each stratum

  This implies that the overall sampling fraction is $\frac{n}{N}$

$$\frac{n_h}{N_h} = \frac{n}{N}$$

The number of elements taken from the h^{th} stratum is

$$n_h = (N_h)\frac{n}{N}$$

Proportionate Allocation

$$V_{SRS} \geq V_{prop}$$

Thus, for proportionate stratified *deff < 1*

For a given total variability in the population, the gain is greater if:

- the ***strata mean are more heterogeneous*** (more unequal strata mean)

OR

- ***the element values within the strata are more homogeneous***

Optimum Allocation

- Uses widely different sampling rates for the various strata
- Objective: to achieve the ***least variance for the overall mean*** for the given sample size (Neyman's allocation); as well as given per unit of *cost in different strata*
- Without cost consideration, the allocation is

$$n_h = n\frac{N_h\sigma_h}{\sum N_h\sigma_h}$$

- This gives better efficiency as compared to proportionate allocation:

$$V_{SRS} \geq V_{prop} \geq V_{opt}$$

Stratified Sampling (Contd.)

Implicit Stratification

- This refers to a systematic sampling with the units arranged in a certain order
- Prior to sample selection, all the units are sorted with respect to one or more variables that are deemed to have a high correlation with the variable of interest
- Implicit stratification guarantees that the sample of units will be spread across the categories of the stratification variables

Variance of Stratified Sample Mean

Stratified sample mean, $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$

Variance of the estimate $\bar{y}_{st}$ is,

$$V(\bar{y}_{st}) = \sum (1 - f_h) \frac{W_h^2 S_h^2}{n_h}$$

If sampling fractions f_h are negligible,

$$V(\bar{y}_{st}) = \sum \frac{W_h^2 S_h^2}{n_h}$$

Stratified Sampling (Contd.)

Estimated Variance of Stratified Sample Mean

If a SRS is taken in each stratum, an unbiased estimate of S_h^2 is s_h^2 ,

$$s_h^2 = \frac{1}{(n_h - 1)} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$$

Estimated variance of the estimate $\bar{y}_{st}$ is,

$$v(\bar{y}_{st}) = \sum (1 - f_h) \frac{W_h^2 s_h^2}{n_h}$$

If sampling fractions f_h are negligible,

$$v(\bar{y}_{st}) = \sum_{h=1}^{H} \frac{W_h^2 s_h^2}{n_h}$$

Stratified Sampling (Contd.)

Precision of proportionate stratified sample over SRS

For a given total variability in the population, the gain is greater if the *strata mean are more heterogeneous* (more unequal strata mean) or equivalently the *more homogeneous are the element values within the strata.* The variance of a stratified sample is given by

$$V(\bar{y}_{st}) = \sum_{h=1}^{H} W_h^2 (1 - f_h) \frac{S_h^2}{n_h} = \sum_{h=1}^{n_h} (1 - f_h) \frac{W_h^2 S_h^2}{n_h}$$

The variance for stratified sample under optimum allocation

$$V_{opt} = \frac{1}{n} (\sum W_h S_h)^2 - \frac{1}{N} \sum W_h S_h^2$$

$$= \frac{1}{nN^2} (\Sigma N_h S_h)^2 - \frac{1}{N^2} \sum N_h S_h^2$$

$$V_{ran} \geq \quad V_{prop} \geq \; V_{opt}$$

Probability Proportional to Size (PPS) Sampling

- Probability of selection is related to an auxiliary variable, Z, that is a measure of "size"
- **Example** Information on Number of households, Area of farms
- "Larger" units are given higher chance of selection than "smaller" units
- Selection probability of i^{th} unit is $p_i = Z_i \Big/ \sum_{i=1}^{N} Z_i \;,\; i = 1,2,\ldots,N$

Selection Procedures:

- Cumulative total method: with replacement
- Cumulative total method: without replacement
- PPS systematic sampling
- Lahiri's method

Cumulative Total Method

Select a sample of 5 villages using varying probability WR sampling, the size being the number of households

Solution

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

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

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	

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.)

- To select a village, a random number r, $1 \le r \le 700$, is selected
- Suppose $r = 259$,
  Since $231 \le 259 \le 288$, the 7^{th} village is therefore selected. The next 4 random numbers to be considered are 548, 170, 231, 505 Hence the required sample selected using PPS with replacement are 16^{th}, 5^{th}, 7^{th}, 15^{th}

Note: The 7^{th} 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.)

- For a PPSWR selection therefore the sample would be: 16^{th}, 5^{th}, 7^{th}, 15^{th} , with 7^{th} village repeated.
- For a PPSWOR selection, we have to continue further to get 5 distinct units in the sample.
- Suppose the next random selected is $r = 375$,

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

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		

PPP Systematic

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

PPS Systematic (Contd.)

- Thus 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		

Lahiri's Method

- A procedure which avoids the need of calculating cumulative totals for each unit has been given by Lahiri (1951)

Steps involved:

1. Select a random number i from 1 to N
2. Select another random number j, such that $1 \leq j \leq M$, where M is either equal to the maximum of sizes Z_i, $i = 1,2,.. N$, or is more than the maximum size in the population.
3. If $j \leq Z_i$, the i^{th} unit is selected, otherwise, the pair (i, j) of random numbers is rejected and another pair is chosen by repeating the steps (1) and (2)

Lahiri's Method

Select a sample of 2 villages using varying probability WR sampling, the size being the number of households

Solution

- $N = 20$, $n = 2$, $M = 58$
- Select a random number i, $1 \le i \le 14$,
- Then a second random number j, $1 \le j \le 58$,
- Suppose the 1st pair of random number is (2, 30). Since $30 \le 45$ thus 2nd village is selected .

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	

Lahiri's Method

Solution (continued)

- Similarly we find the next pair of random *number (12, 47) since 47 >30, the 12th village* is not selected The 3rd pair ot random numbers (7, 40) results in the selection of 7th village since $40 \leq 58$
- Hence, the selected sample are 2nd and 7th villages

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	

PPS Sampling (Contd.)……Probability of unit in the sample

$$p_i = \frac{X_i}{\sum_{i=1}^{N} X_i}, \quad i = 1, \ldots, N$$

i.e. Sampling is done with a varying probability

In epsem the probability of ith unit being in the sample is:

SRSWR, $p_i = 1 - \left[1 - \frac{1}{N}\right]^n$

SRSWOR, $p_i = \frac{n}{N}$

Simple cluster, $p_i = \frac{m}{M}$

} fixed probabilities

But in PPS, p_i depends on X_i, size of the unit which is variable

The larger the X_i, the more probable ith unit to be chosen

PPS Sampling (Contd.)……Probability of selection unit in sample

- If values y_i were known before sampling and sampling is carried out with probability proportional to y_i ,

$$p_i = \frac{y_i}{\sum_{i=1}^{N} y_i}$$

- If instead of drawing a unit with probability proportional to its actual value, we draw with probability proportional to an auxiliary variable whose size (x_i) is related to y_i by relation, $x_i = ky_i$ where k is positive constant, the probability of selection

$$ppx_i = \frac{x_i}{\sum_{i=1}^{N} x_i} = \frac{ky_i}{k\sum_{i=1}^{N} y_i} = \frac{y_i}{\sum_{i=1}^{N} y_i} = ppy_i$$

remains the same, and would give the same results as pps of y

PPS Sampling (Contd.)……Estimator of Y from i^{th} unit

In SRS from a population of size N , population total Y

- Probability of selection of a unit at any draw is *1/N*
- Unbiased estimator of Y from i^{th} unit is $N.y_i$

  Or, $\hat{Y} = N.y_i = \frac{y_i}{\frac{1}{N}}$

- Similarly, the unbiased estimator of the population total Y in varying probability sampling from the i^{th} draw is $\hat{Y}_{pps} = \frac{y_i}{p_i}$, where p_i (which varies from unit to unit in the universe) is probability of selection of y_i

PPS Sampling (Contd.)……Estimated variance of estimator of Y

- If we draw a sample of *n* units with replacement out of *N* units, with the initial probability of selection of the i^{th} unit as p_i ,the combined unbiased estimator of Y is

$$\hat{Y}_{pps} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}$$

- The sampling variance of this estimator is

$$\sigma^2_{y\,pps} = \frac{1}{n}\sum_{i=1}^{N}\left(\frac{Y_i}{p_i} - Y\right)^2 p_i = \frac{1}{n}\left(\sum_{i=1}^{N}\frac{Y_i^2}{p_i} - Y^2\right)$$

- Estimated variance of $\hat{Y}_{pps}$ is

$$v(\hat{Y}_{pps}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{Y}_{pps}\right)^2$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i}\right)^2 - n\left(\hat{Y}\right)^2_{PPS}$$

PPS Sampling (Contd.)……Estimation in PPSWR Sampling

$$\hat{T}(Y)_{PPS} = \hat{Y}_{PPS} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i}; \text{Unbiased estimator}$$

$$\text{var}\left[\hat{T}(Y)_{PPS}\right] = \frac{1}{n}\sum_{j=1}^{N}\left[\frac{Y_j}{p_j} - T(Y)\right]^2 p_j$$

$$= \text{an unknown value}$$

Unbiased Estimator of $Var\left(T(Y)_{PPS}\right)$:

$$\text{var}\left[\hat{\text{T}}(\text{Y})_{\text{pps}}\right] = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{T}(Y)_{PPS}\right)^2$$

$$= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left[\left(\frac{y_i}{p_i}\right)^2 - n\hat{T}(Y)^2_{PPS}\right] = \hat{\text{var}}\left(\hat{T}(Y)_{PPS}\right)$$

Therefore, unbiased estimator of population mean $\bar{Y}$ is

$$\bar{y}_{PPS} = \frac{1}{N}\hat{T}(Y)_{PPS}$$

with variance

$$Var(\bar{y}_{PPS}) = \frac{1}{N^2}\text{var}\left[\hat{T}(Y)_{PPS}\right] = \frac{1}{nN^2}\sum_{j=1}^{N}\left[\frac{Y_i}{p_j} - T(Y)\right]^2$$

whose unbiased estimator is

$$\text{var}(\bar{y}_{PPS}) = \frac{1}{N^2}\text{var}(T(Y)_{PPS}) = \frac{1}{n(n-1)N^2}\sum_{i=1}^{n}\left(\frac{y_i}{p_i} - \hat{T}(Y)_{PPS}\right)^2$$

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

Clustering and Stratification

Naturally occurring clusters

Clusters are usually defined as groups of units that are found naturally ‘clustered’ together - by location or socially defined entities like households or by institutions like schools and enterprises

Cluster	Population Unit
Census Enumeration Area	Dwelling
household	Person
Day	Hour
School	Student
Employer	Employee

Clustering and Stratification in Sample Design

- Typically, sample surveys by NSOs involve subdividing population into strata and clusters
- Technique of stratifying clusters and then further stratifying the units within clusters are applied to obtain the final sample
- Sampler's objective is to get right combination of stratification and clustering to get required estimates at desired level of accuracy with given resources
- Reliability or precision of estimates depend on degree to which sample is *clustered*
- Generally, *clustering* increases *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

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

Cluster Sampling (Contd.)

Selecting a (single-stage) cluster sample

- Require sampling frame: list of all the clusters
- From the list, a sample of clusters is selected - using a selection scheme (e.g., SRS, Systematic)
- All population units within the selected clusters are listed
- The information is then collected from all the units of the selected clusters

Cluster Sampling- Advantages/ disadvantage

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
- Reduced cost if personal interviews, particularly when the survey cost increases with the distance separating the sampled units

Main disadvantage

- Increased sampling error due to a less representative sample
- in practice, units are homogeneous within normally defined clusters
- composition of clusters can not be altered, as they are pre-defined

Clusters of equal size- Estimates

- Assume a SRS of n clusters from a population of N clusters

Total $\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i$

Mean $\hat{\bar{\bar{Y}}} = \frac{1}{nM}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$

N = # clusters in pop

n = # clusters in sample

M **= # units in cluster**

y_i **= ith cluster total**

$\bar{y}_i$ **= ith cluster mean**

Variance of estimators

$$V(\hat{\bar{\bar{Y}}}) = \frac{1}{n}\left(1-\frac{n}{N}\right)S_b^2$$

$$S_b^2 = \sum_{i=1}^{N}\frac{(\bar{y}_i - \bar{\bar{y}})^2}{N-1} = \frac{S^2}{M}\left(1+(M-1)\rho\right)$$

Where ρ is intra class correlation and S^2 is population mean square

Clusters of equal size- Estimates (contd.)

$$(NM-1)S^2 \equiv N(M-1)S_w^2 + M(N-1)S_b^2 \qquad S^2 \cong S_b^2 + S_w^2$$

$$\rho = \frac{M(N-1)S_b^2 - NS_w^2}{N(M-1)S_w^2 + M(N-1)S_b^2} \cong \frac{MS_b^2 - S_w^2}{(M-1)S_w^2 + MS_b^2}$$

Estimating the variance $\hat{V}(\hat{Y}) = \frac{\mathbf{N^2M^2}}{n}(1-\frac{n}{\mathbf{N}})\mathbf{s_b^2}$

$$\hat{V}(\hat{\bar{\bar{Y}}}) = \frac{1}{n}(1-\frac{n}{\mathbf{N}})\mathbf{s_b^2} \quad \text{where} \quad \mathbf{s_b^2} = \sum_{i=1}^{n} \frac{(\bar{y}_i - \hat{\bar{\bar{y}}})^2}{n-1}$$

Clusters of unequal size- Estimating

Assume that a SRS of n clusters from a population of N clusters is selected

Estimating Total $\hat{Y} = \frac{N}{n}\sum_{i=1}^{n} y_i$

N = # clusters in pop

n = # clusters in sample

M_i = # units in i^{th} cluster

y_i = i^t cluster total

Estimating the variance

Total

$$\hat{V}(\hat{Y}) = \frac{N^2}{n}(1-\frac{n}{N})\mathbf{s_b'^2}*$$

Where,

$$\mathbf{s_b'^2}* = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\right)^2$$

Clusters of unequal size (contd.)

Estimating Mean

$$\text{Total} = \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} M_i \bar{y}_i \quad ; \quad \text{Mean} = \bar{Y} = \frac{\sum_{i=1}^{N} M_i \bar{y}_i}{\sum_{i=1}^{N} M_i}$$

$$\hat{\bar{Y}}_1 = \frac{1}{n} \sum_{i=1}^{n} \bar{y}_i$$

$$\hat{\bar{Y}}_2 = \frac{1}{n\bar{M}} \sum_{i=1}^{n} M_i \bar{y}_i = \frac{1}{n} \sum_{i=1}^{n} u_i \bar{y}_i \quad ; \quad \bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i \, , u_i = \frac{M_i}{\bar{M}}$$

$$\hat{\bar{Y}}_3 = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$$

Clusters of unequal size (contd.)

Estimating variances

$$\hat{V}(\hat{\bar{Y}}_1) = \frac{1}{n}\left(1-\frac{n}{N}\right)\mathbf{s_b^2} \;;\; \mathbf{s_b^2} = \frac{1}{n-1}\sum\left(\bar{y}_i - \hat{\bar{Y}}_1\right)^2$$

$$\hat{V}(\hat{\bar{Y}}_2) = \frac{1}{n}\left(1-\frac{n}{N}\right)\mathbf{s_b'^2} \;;\; \mathbf{s_b'^2} = \frac{1}{n-1}\sum\left(\bar{u}_i\bar{y}_i - \hat{\bar{Y}}_2\right)^2$$

$$\hat{V}(\hat{\bar{Y}}_3) = \frac{1}{n}\left(1-\frac{n}{N}\right)\mathbf{s_b''^2} \;;\; \mathbf{s_b''^2} = \frac{1}{n-1}\sum\frac{M_i^2}{\bar{M}_n^2}\left(\bar{y}_i - \hat{\bar{Y}}_3\right)^2$$

Where $\bar{M}_n = \frac{1}{n}\sum M_i$

Multi-stage (Cluster) Sampling

- In a *single-stage cluster sampling*, a sample of cluster is selected and all the population units of each selected cluster are surveyed
- When clusters are to large to cover all their population units in the survey, a sample of population units from each selected cluster is surveyed
- Such a design is 'Multi-stage cluster' / 'Multi-stage' sampling
- Multi-stage sampling involves multiple stages of sampling
- The number of stages can be numerous, although it is rare to have more than 3 stages
- We will concentrate only on two-stage sampling
- Process of selecting a sample of population units from selected clusters is known as *Sub-sampling*

Multi-stage (Cluster) Sampling (Contd.)

Stage-wise Selection

- Stage sampling is an extension of cluster sampling. For a two-stage sampling
- we select the clusters at the first stage

  – selected clusters are called *first stage units* (FSUs) or *primary stage units* (PSUs)
- then select a sample of units from within each selected cluster – selected units are called second stage units (SSUs)

Multi-stage (Cluster) Sampling (Contd.)

Sampling Units at different Stages

Examples of two-stage sampling

Stage 1	*Stage 2*
villages	households
Dwellings	People
Hospitals	Patients
Businesses	Employees
Coconut trees	Coconuts

Multi-stage (Cluster) Sampling (Contd.)

Advantages of multistage sampling

- Sampling frames normally available at higher stages, may be prepared at lower stages
- Cost considerations
- Flexibility in choice of sampling units and methods of selection at different stages
- Contributions of different stages towards sampling variance may be estimated separately

Multi-stage (Cluster) Sampling (Contd.)

Sampling at two stages

- In practice, many multi-stage designs involve complex sub-sampling of and within PSUs
- The selection at the two stages are done independently and may employ different sampling schemes like:
  - SRSWOR
  - Systematic
  - Probability Proportional to Size (PPS)

Different methods of selection

Method 1

Stage	Selection
I	N srs i^{th} n
II	M_i srs m_i

Multi-stage (Cluster) Sampling (Contd.)

Method 1 (contd.)

Estimating Total

$$\hat{Y} = \frac{\mathrm{N}}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \hat{\mathrm{Y}}_{\mathrm{i}} = \frac{\mathrm{N}}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{M}_{\mathrm{i}} \bar{y}_i$$

$$V\left(\hat{Y}\right) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) S_b'^2 + \frac{N}{n} \sum_{i=1}^{N} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2$$

$$S_b'^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i - \frac{1}{N} \sum Y_i\right)^2 \text{ and } S_i^2 = \frac{1}{M_i - 1} \sum_{j=1}^{M_i} \left(Y_{ij} - \frac{1}{M_i} \sum Y_{ij}\right)^2$$

Multi-stage (Cluster) Sampling (Contd.)

Method 1 (contd.)

Estimating Variance

$$\hat{V}\left(\hat{\bar{Y}}\right) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_b'^2 + \frac{N}{n}\sum_{i=1}^{n} M_i^2\left(\frac{1}{m_i}-\frac{1}{M_i}\right)s_i^2$$

$$s_b'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{Y}_i - \frac{1}{n}\sum \hat{Y}_i\right)^2$$

and

$$s_i^2 = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(y_{ij} - \frac{1}{M_i}\sum y_{ij}\right)^2$$

Multi-stage (Cluster) Sampling (Contd.)

Method 1 (contd.)

Estimating Mean (equal PSU's)

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i \quad ; \quad V\left(\hat{\bar{Y}}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2$$

Where,

$$S_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{Y}_i - \frac{1}{N}\sum \bar{Y}_i\right)^2$$

and

$$\bar{S}_w^2 = \frac{1}{N}\sum_{i=1}^{N} S_i^2 \quad ; \quad S_i^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(y_{ij} - \bar{Y}_i\right)^2$$

Multi-stage (Cluster) Sampling (Contd.)

Method 1 (contd.)

Estimating variance (equal PSU's)

$$\hat{V}\left(\hat{\bar{Y}}\right)=\left(\frac{1}{n}-\frac{1}{N}\right)s_b^2+\frac{1}{N}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{s}_w^2$$

$$s_b^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{y}_i-\hat{\bar{Y}}\right)^2$$

and

$$\bar{s}_w^2=\frac{1}{n}\sum_{i=1}^{n}s_i^2\ ;\ s_i^2=\frac{1}{m-1}\sum_{j=1}^{m}\left(y_{ij}-\bar{y}_i\right)^2$$

Multi-stage (Cluster) Sampling (Contd.)

Estimation of population mean Unequal PSU's

$$\hat{\bar{Y}}_1 = \frac{N}{nM_0}\sum_{i=1}^{n} M_i \bar{y}_i$$

$$V(\hat{\bar{Y}}_1) = \left(\frac{1}{n} - \frac{1}{N}\right)S_{1b}^2 + \frac{1}{nN}\sum_{i=1}^{N}\frac{M_i^2}{\bar{M}^2}\left(\frac{1}{m_i} - \frac{1}{Mi}\right)S_i^2$$

$$\hat{V}(\hat{\bar{Y}}_1) = \left(\frac{1}{n} - \frac{1}{N}\right)s_{1b}^2 + \frac{1}{nN}\sum_{i=1}^{n}\frac{M_i^2}{\bar{M}^2}\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_i^2$$

Multi-stage (Cluster) Sampling (Contd.)

Estimation of population mean Unequal PSU's (Contd.)

$$\hat{\bar{Y}}_2 = \frac{1}{n}\sum_{i=1}^{n}\bar{y}_i$$

$$B(\hat{\bar{Y}}_2) = -\frac{1}{N\bar{M}}\sum_{i=1}^{N}(M_i - \bar{M})\bar{Y}_i$$

$$V(\bar{Y}_2) = (\frac{1}{n} - \frac{1}{N})S_{2b}^2 + \frac{1}{nN}\sum_{i=1}^{N}(\frac{1}{m_i} - \frac{1}{M_i})S_i^2$$

$$\hat{V}(\bar{Y}_2) = (\frac{1}{n} - \frac{1}{N})s_{2b}^2 + \frac{1}{nN}\sum_{i=1}^{N}(\frac{1}{m_i} - \frac{1}{M_i})s_i^2$$

Estimating Population mean unequal clusters (Contd.)

$$\hat{\bar{Y}}_3 = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$$

$$V(\hat{\bar{Y}}_3) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{3b}^2 + \frac{1}{\bar{M}^2 nN} \sum_{i=1}^{N} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2$$

$$\hat{V}(\hat{\bar{Y}}_3) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{3b}^2 + \frac{1}{\bar{M}^2 nN} \sum_{i=1}^{n} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2$$

Multi-stage (Cluster) Sampling (Contd.)

Different methods of selection

Method 2

Stage	Selection
I	N srs i^{th}
II	n M systematic m

Multi-stage (Cluster) Sampling (Contd.)

Method 2 (cont)

Estimating Mean

$$\hat{\bar{Y}} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$$

$$V\left(\hat{\bar{Y}}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 + \frac{1}{nm}\left(1 - \frac{1}{M}\right)\frac{1}{N}\sum_{i=1}^{N} S_i^2\left(1 + \rho_i\left(m-1\right)\right)$$

$$\hat{V}\left(\hat{\bar{Y}}\right) \cong \frac{s_b^2}{n}$$

Multi-stage (Cluster) Sampling (Contd.)

Different methods of selection

Method 3

Stage	Selection
I	N ppswr n
II	i^{th} M_i srs m_i

Multi-stage (Cluster) Sampling (Contd.)

Method 3 (Contd)

Estimating Total

$$\hat{Y} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{Y}_i}{p_i} \quad ; \quad \hat{Y}_i = M_i \bar{y}_i$$

$$\hat{V}\left(\hat{Y}\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{\hat{Y}_i}{p_i} - \hat{Y}\right)^2$$

Optimum sample sizes

- Equal PSU's
- SRSWOR at both the stages
- Variance function

$$V \cong \frac{S_b^2}{n} + \frac{S_w^2}{nm}$$

- Cost function

$$C = C_0 + nc_1 + nmc_2$$

Optimum sample sizes (Contd.)

- Minimize V subject to given cost C*
- Optimum m and n are obtained as

$$m^* = \left(c_1 S_w^2 / c_2 S_b^2\right)^{1/2}$$

$$n^* = \frac{(C^* - C_0)\sqrt{S_b^2 / c_1}}{\left(\sqrt{c_1 S_b^2} + \sqrt{c_2 S_w^2}\right)}$$

Two stage: Optimum Sample Sizes (1)

- Goal: get the most information (and hence, more statistically efficient) for the least cost
- Illustrative example: PSUs with equal sizes, SRSWOR at both stages

Two stage: Optimum Sample Sizes (2)

- Variance function $V \cong \frac{S_b^2}{a} + \frac{S_w^2}{ab}$
- Cost function $C = C_0 + ac_1 + abc_2$
- ***Problem***: Minimize V subject to given cost C*

Two stage: Optimum Sample Sizes (3)

- Minimize V subject to given cost C*
- Optimum a=a* and b=b*

$$b^* = \sqrt{c_1 S_w^2 / c_2 S_b^2}$$

$$a^* = \frac{(C^* - C_0)\sqrt{\frac{S_b^2}{c_1}}}{\sqrt{c_1 S_b^2} + \sqrt{c_2 S_w^2}}$$

Two stage: Optimum Sample Sizes (4)

- Optimum b=b*

$$b^* = \sqrt{\frac{c_1}{c_2}\frac{S_w^2}{S_b^2}} = \sqrt{\frac{c_1}{c_2}\frac{1-roh}{roh}}$$

Thanks