Household Survey Design Techniques

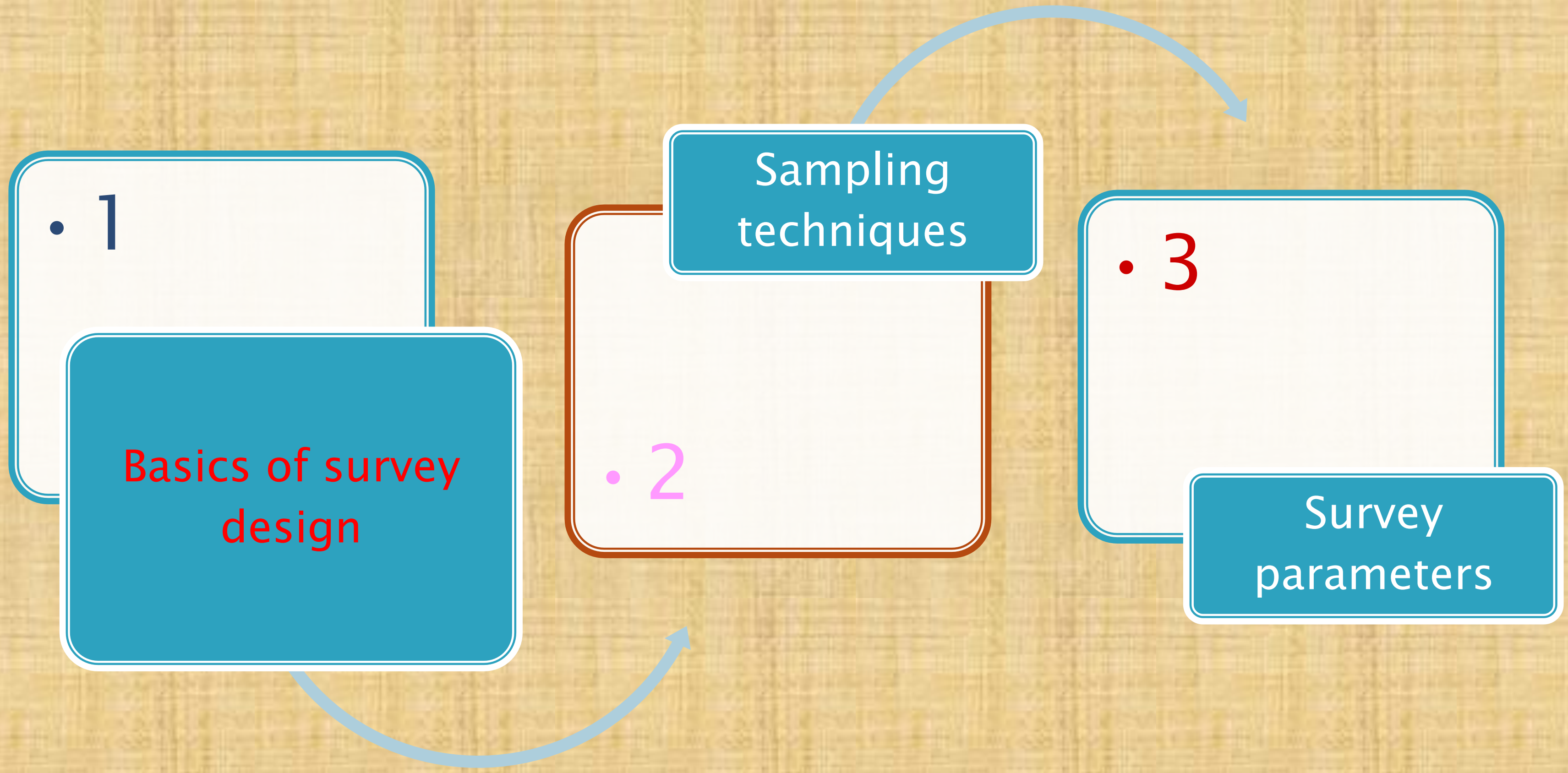
Workshop on Measuring Sustainable Agriculture, Food Security and Poverty Alleviation for enhancing Accountability in the Post 2015 Development Agenda.

24-28 November 2014, Bogor Indonesia

Alick Nyasulu
United Nations Statistical Institute for Asia and the Pacific (SIAP)

- Sampling Methods
- Survey Design
- Estimation of parameters

The need for sampling



- 1

Basics of survey design

- 2

Sampling techniques

- 3

Survey parameters

Key Definitions

- Why sample?

- To make an inference about a population

- Studying entire pop is impractical or impossible
- Select a few households and make an inference about all households

Sampled population is a collection of elements that were actually available for selection into the sample

Key Definitions

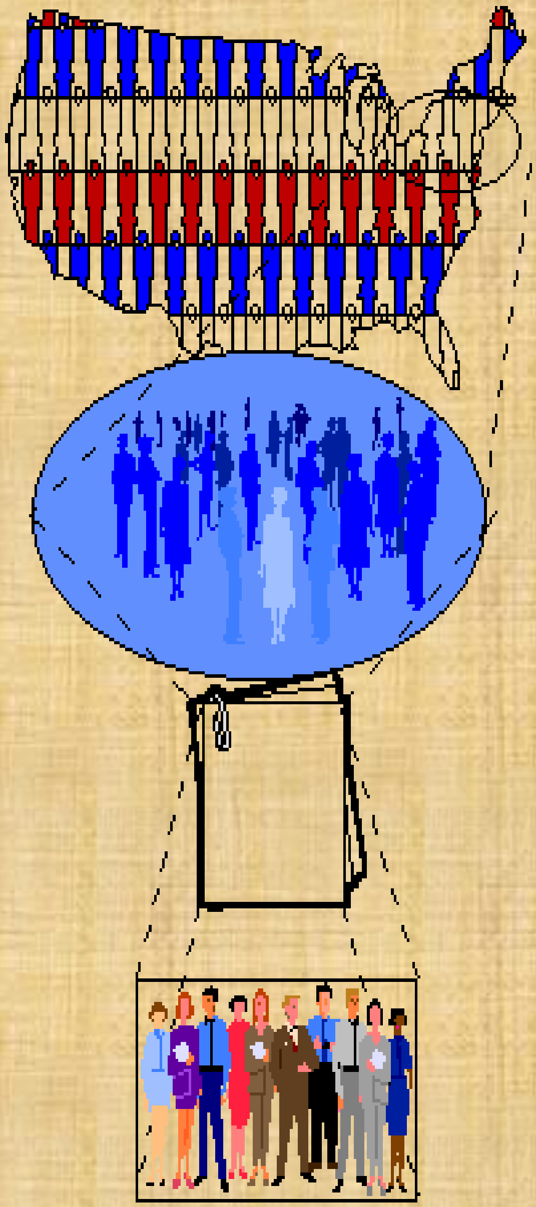
- **Sampling Unit**
  - *A sampling unit is an entity that is selected in the sampling process.*
- Observational unit

- **Characteristic**

- A quantitative variable like age of a person, income, land size!

- Qualitative variable like employment status of a person

Definitions



Who do you want to generalize to?

The Theoretical Population

What population can you get access to?

The Study Population

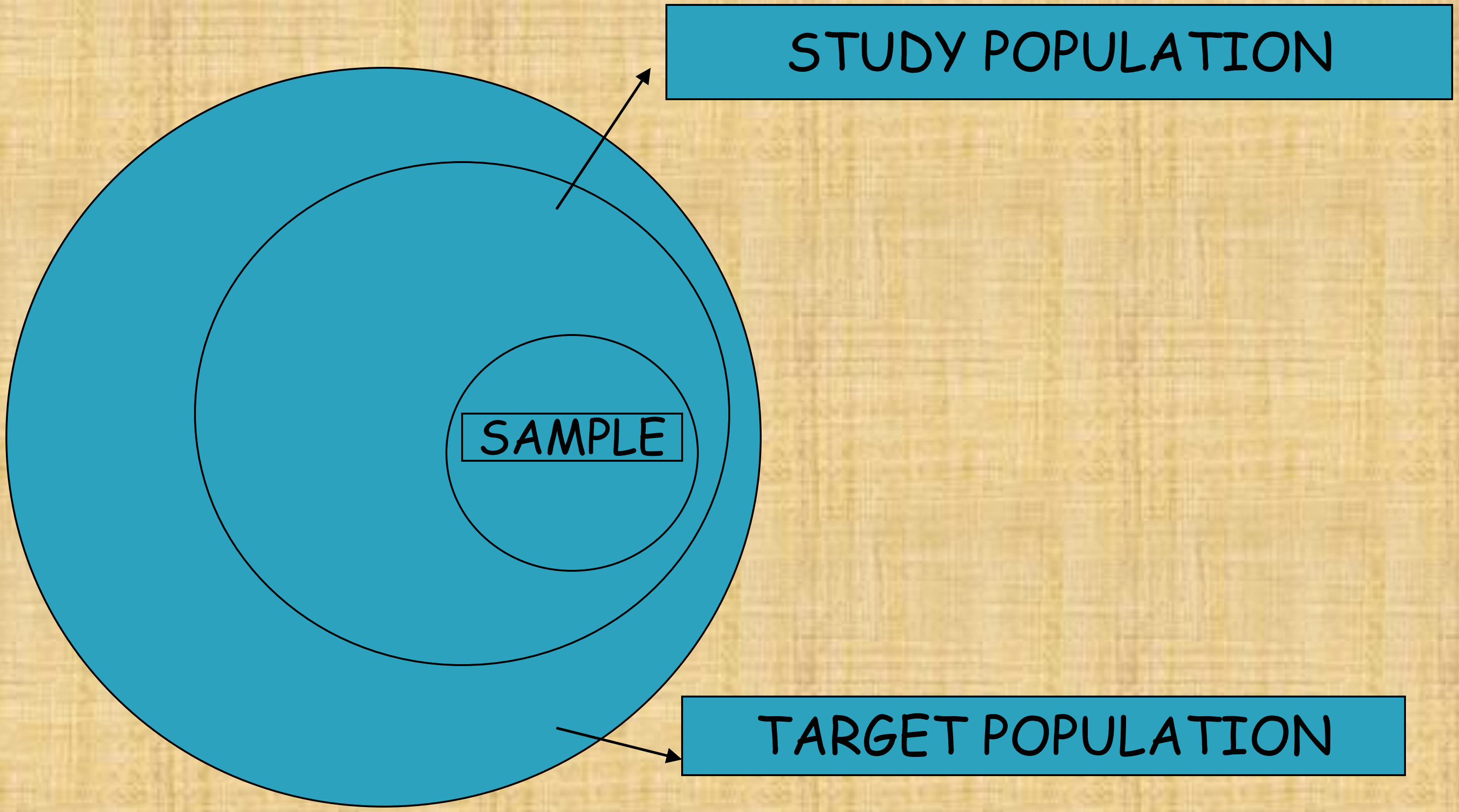
How can you get access to them?

The Sampling Frame

Who is in your study?

The Sample

SAMPLING…….



STUDY POPULATION

SAMPLE

TARGET POPULATION

Population

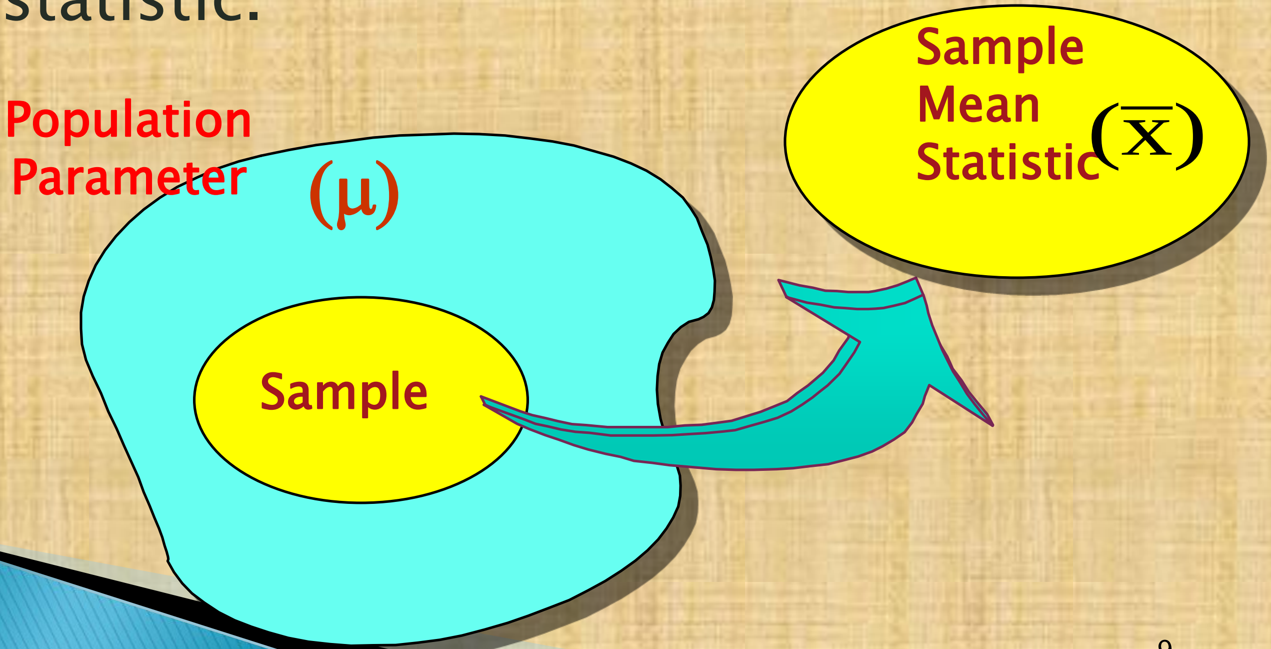
- The population or universe is the entire group of all the units of analysis whose characteristics are to be estimated

- A **target population** is a collection of elements of interest as defined by survey objectives

Definitions

Population Parameter and Sample Statistic

- A population parameter is a numerical summary of a population
- Any numerical measure computed from a subset of the population (typically a sample) is a statistic.



Sample Design – two broad kinds

- Probability sampling

each element of the population is assigned a non-zero chance of being included in the sample

[our focus]

- Non-Probability sampling

consists of a variety of procedures, including judgment-based and ‘purposive’ choice of elements – considered “representative” of the population

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.

Basic Sampling Schemes

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

- **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)

- 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
- Theoretical basis for other sample designs

SRS with replacement (SRSW)

SRS without replacements (*SRSWOR*)

SRS selection procedures

1. Get a list (sampling frame) which uniquely identifies each unit in the population
2. Allocate a serial number to each unit of the frame
3. 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

Systematic Sampling

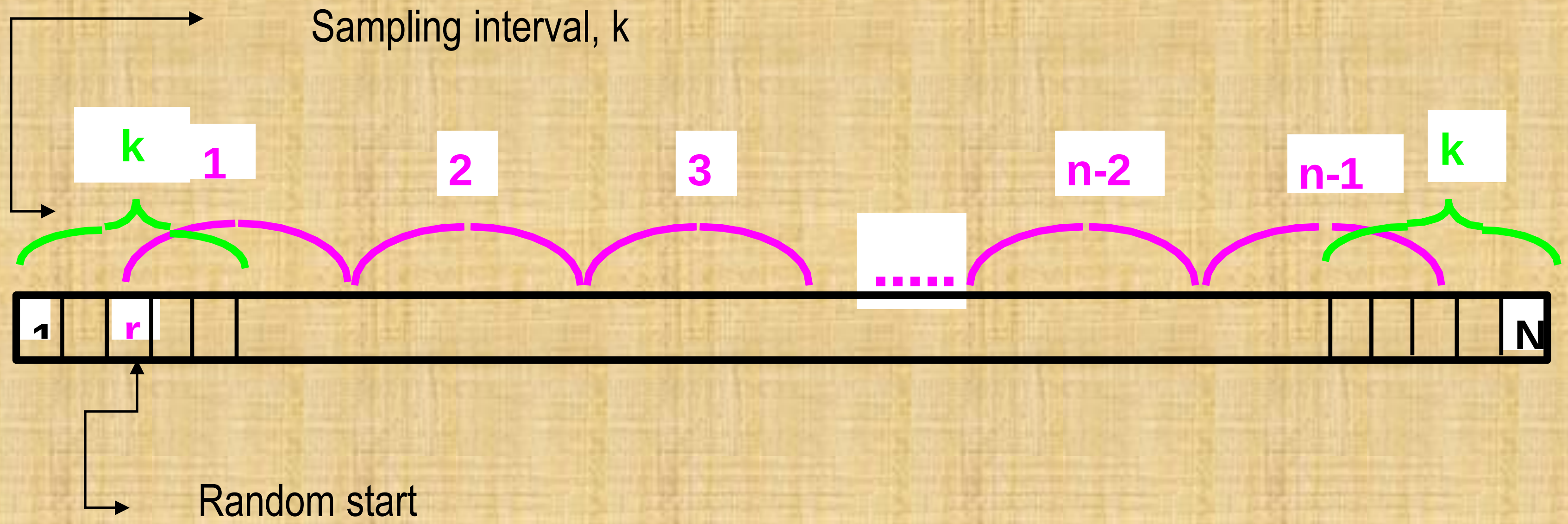
- Systematic Sampling (SYS), like SRS, involves selecting *n* sample units from a population of *N* units

- 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

Linear Systematic Sampling



Systematic Sampling

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

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

  i = 1,2, … , *N*

$$p_i = Z_i \Big/ \sum_{i=1}^{N} Z_i$$

PPS Selection

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	

PPS Selection

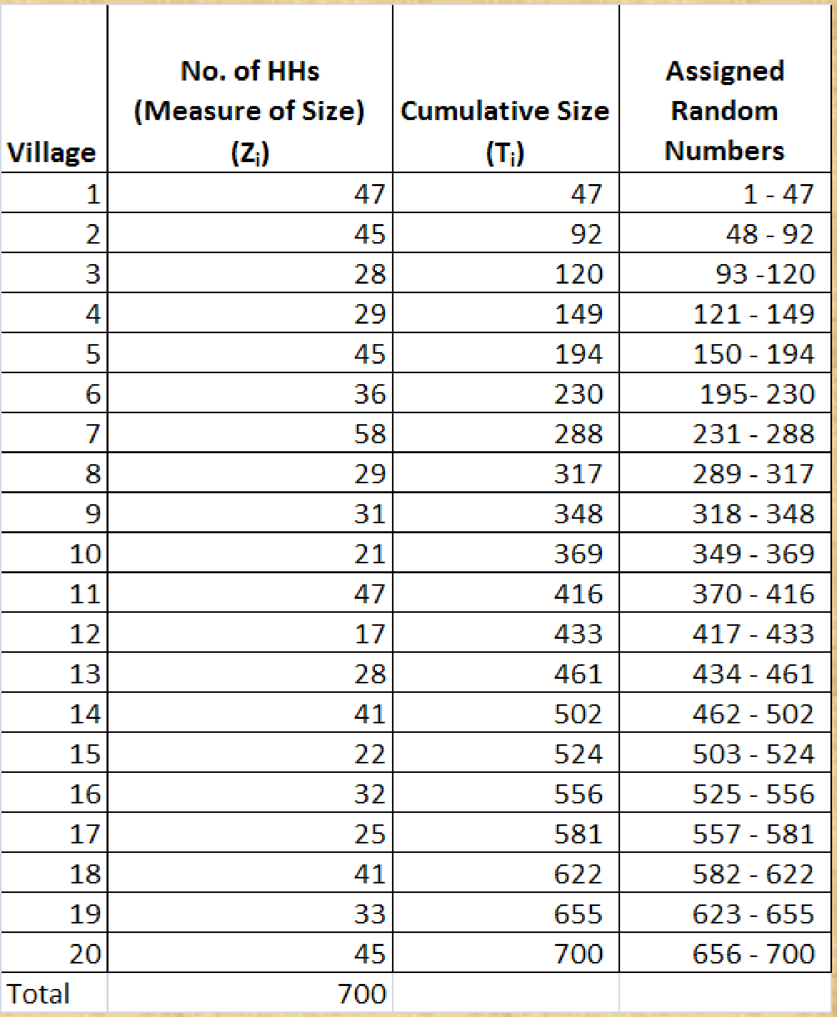
Cumulative Total Method (Contd.)

- 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 7th 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 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		

Survey Design

- Structure of population is very diverse
  - High income or income groups?
  - Location issues?

- Possible solutions??
  - Apply PPS techniques but consider segmenting the population
  - Stratification and clustering

Impact on estimates

Systematic sampling is just a selection procedure

Stratification

- Divide the population into a number of distinct groups (strata) based on auxiliary information – referred to as *stratification variables* – relating to study variable(s)

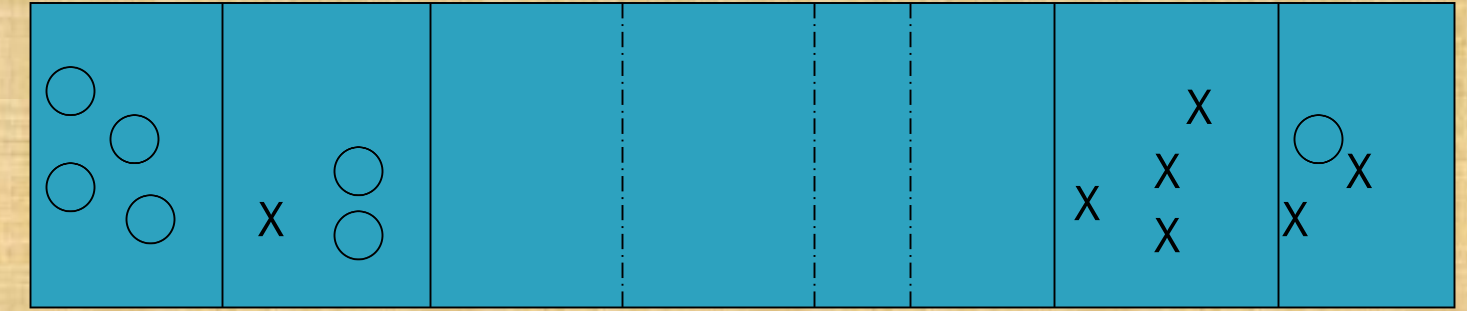
- The division of the population into strata is termed stratification

- Each stratum is composed of units that satisfy the condition set by the values of the stratifying variable.

- Main purpose: to improve the sample estimations, i.e. to reduce the standard error of the estimates.

Stratified Sampling

Stratification
– Mutually Exclusive subsets



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

Stratification

Stratified sampling involves:

- division or stratification of the population into homogeneous (similar) groups called strata; and
- selecting the sample using a selection procedure
  - like SRS or systematic sampling or PPS within each stratum and
  - independent of the other strata

Clustering

- Subsets of the listing units in the population
- Set of clusters must be mutually exclusive and collectively exhaustive
    - councils
    - townships
    - regions
    - Institutions
    - villages

Clustering

- **Single Stage**

- There are 400 dairy farmers located in 20 districts in Province A

- We wish to interview a sample of these farmers
  - select a simple random sample of 5 districts
  - interview all farmers in the 5 selected districts

- **Two Stage**

- Select a sample of clusters, as in the single-stage method

- From each selected cluster, select a subsample of listing units

Clustering

- **Single Stage**
- Districts are the PSUs
- Frames are the listing units
- Sampling probability for each farmer is 5/20
- Thus, this is an EPSEM sample
- Sampling frame is the list of 20 districts

- **Two Stage**
- We want to interview a sample of 50 farmers
- We can afford to visit 10 different districts
- Thus, we need to interview 50/10 = 5 farmers at each district

Two Stage Clustering

- PSUs are the districts
- Listing units are the farmers
- Sampling frames
  - Stage 1: List of 20 districts
  - Stage 2: Lists of farmers in each selected district

- **Stage 1**: select a sample of 10 districts
  - Selection prob. proportional to "size"
- **Stage 2:** select a sample of 5 farmers from each selected districts
- At each stage, use one of the simple sampling methods

Clustering reduces logistical costs, lists of all 400 farmers may not be available

The estimates are less precise due to possible homogeneity

ESTIMATION

- An estimator is a sample statistic
- A sample statistic is a summary value of a variable calculated from the sample.
-
- An estimator is any quantity calculated from the sample data
  - a function of sample observations – which is used to give information about an unknown quantity of the population.
- *Example:* sample mean is an estimator of the population mean.

- **Desirable characteristics**
- Unbiasedness=sample estimate equal to true population value
- Consistency=as sample increases estimate gets closer to true population value
- Efficiency=estimate with least variance

Base (sampling) weight: basic concept

Base weight…

- Is the inverse of the probability of selection
  - ○ Thus, depends on the sample selection plan

- Number of units in the population being represented by the sample unit
  - ○ In ideal conditions, the design weights take care of "*representativeness*"
  - ○ But, this is not true in *less than ideal conditions*

Sampling weight: basic concept

In a SRS design: N=10 and n=5

Population:

Inclusion probability/probability of selection (chance to be selected in the sample)=

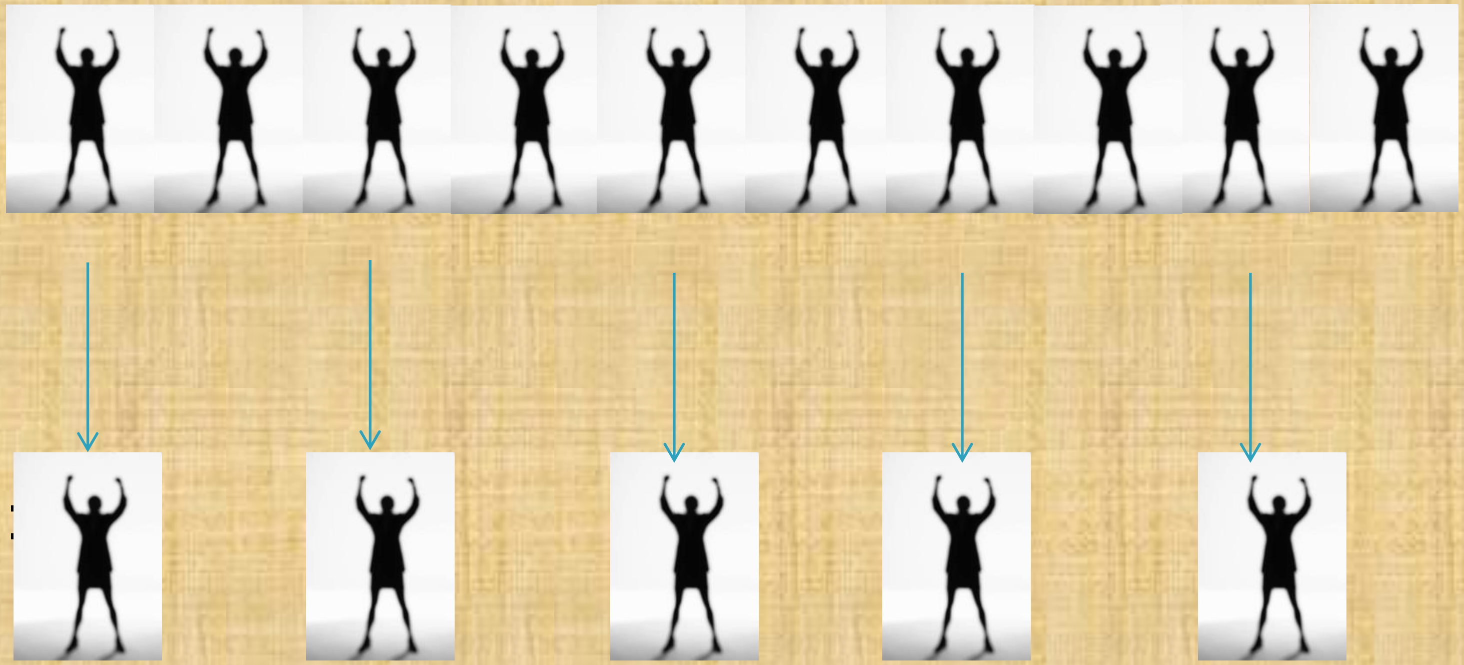
$$\pi = \frac{n}{N} = \frac{5}{10} = \frac{1}{2}$$

Each individual has 50% chance to be selected in the sample

Sampling weight: basic concept

Population:

SRS sample:



Sampling weight= inverse of inclusion probability:

$$w=\frac{1}{\pi}=\frac{1}{1/2}=2 \quad OR \quad w=\frac{1}{\left({}^{n}\!/\!{}_{N}\right)}=\frac{N}{n}=\frac{10}{5}=2$$

Final weight: Illustration

Weighted values

ID	Stratum	probability of selection	Base weight (w_b)	y	Adjustment for non-response (w_r)	Calibration weight (w_c)	Final weight (w_F)	w_F*y
01	1	0.0025	400.0000	8				
02	2	0.0035	285.7143	-				
03	1	0.0018	571.4286	6				
04	2	0.0031	322.5806	10				
05	1	0.0016	625.0000	5				
06	2	0.0035	285.7143	18				
07	2	0.0038	266.6667	18				
08	1	0.0015	666.6667	-				
09	1	0.0028	363.6364	9				

$$w_F = w_b \cdot w_r \cdot w_c$$

- Thank You