SIAP

Statistical Institute for Asia and the Pacific

UNITED NATIONS

siap

Statistical Institute for Asia and the Pacific

SAMPLING BASIC CONCEPTS

BY

ALICK MJUMA NYASULU-SIAP

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

Jakarta, Indonesia ,29Sep-10 October 2014.

UNITED NATIONS
siap
Statistical Institute for Asia and the Pacific

LEARNING OBJECTIVES

At the end of the of this session participants are expected to:

1. Demonstrate knowledge of basic sampling theory
2. Apply use of simple random sampling techniques in sample selection
3. Explain and use systematic sampling as a tool for sample selection

OVERVIEW OF THE PRESENTATION

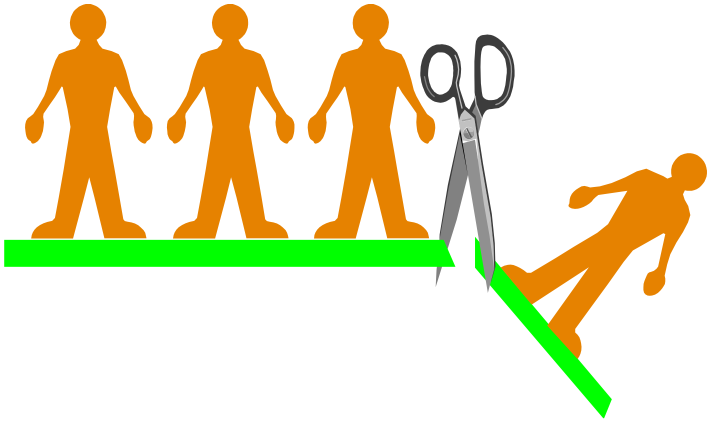
1. What is sampling?
2. Basic Concepts
3. Simple Random Sampling
4. Systematic Sampling

WHAT IS SAMPLING?

If all members of a population were identical, the population is considered to be *homogenous*.

That is, the characteristics of any one individual in the population would be the same as the characteristics of any other individual (little or no variation among individuals).

So, if the human population on Earth was homogenous in characteristics, how many people would an alien need to abduct in order to understand what humans were like?



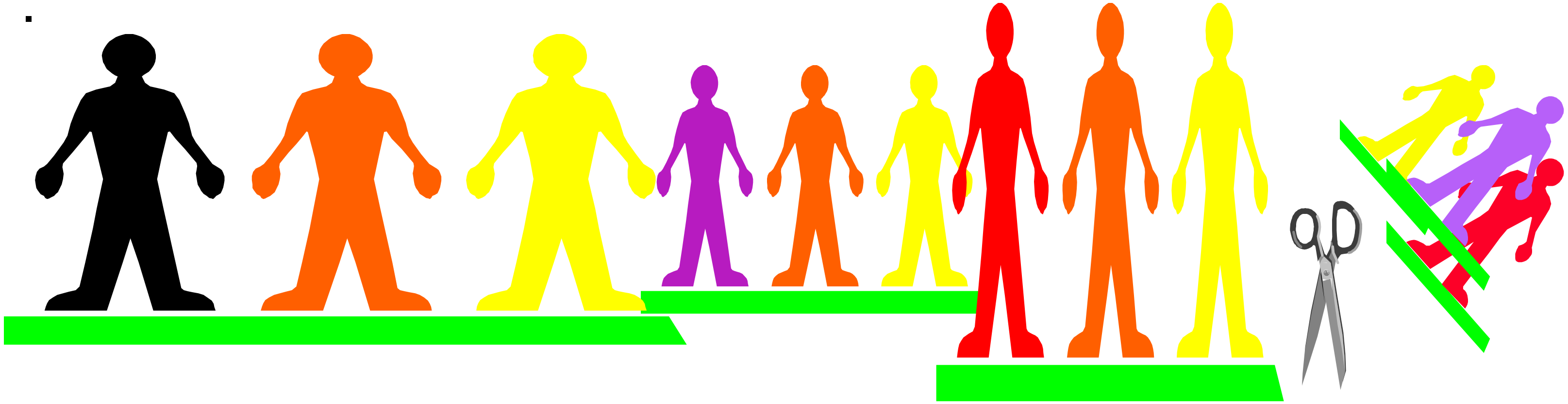
WHAT IS SAMPLING?

When individual members of a population are different from each other, the population is considered to be *heterogeneous* (having significant variation among individuals).

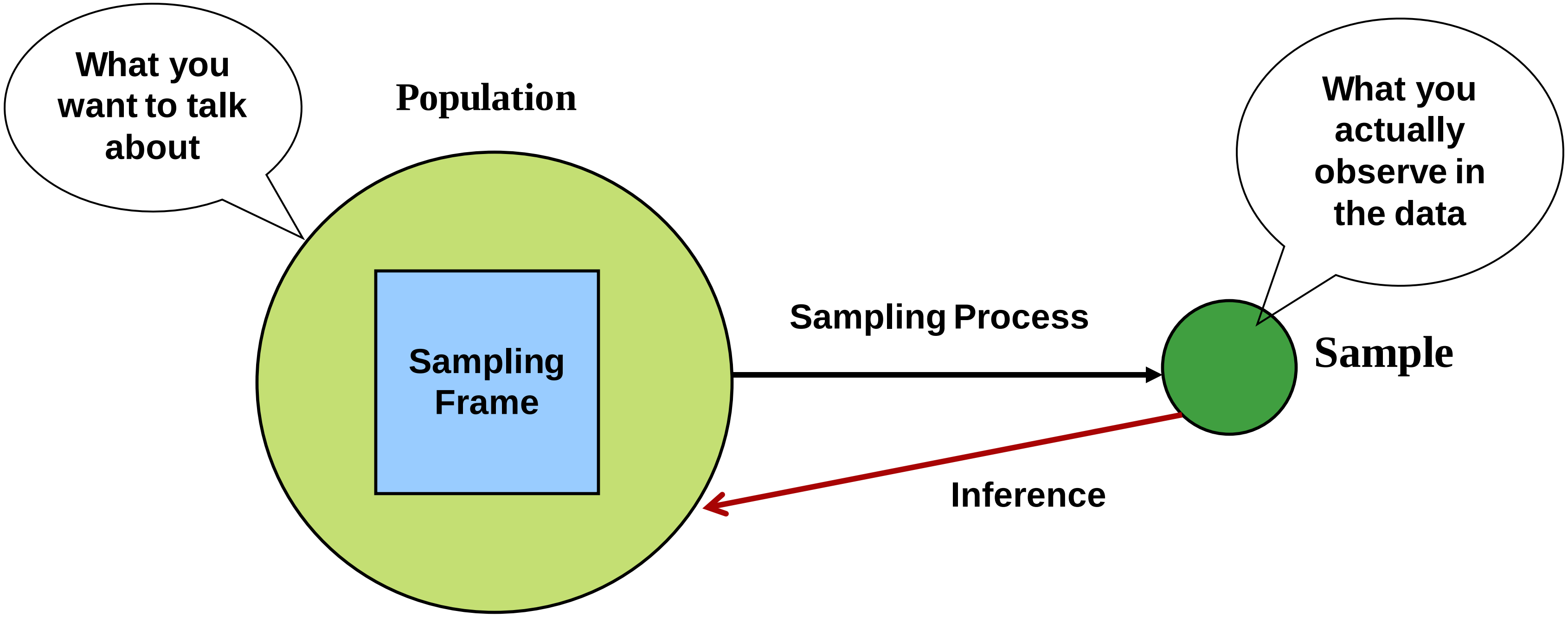
How does this change an alien's abduction scheme to find out more about humans?

In order to describe a heterogeneous population, **observations of multiple individuals** are needed to account for all possible characteristics that may exist

.



WHAT IS SAMPLING?


What you want to talk about
Population
Sampling Frame
Sampling Process
Inference
What you actually observe in the data
Sample

Using data to say something (*make an **inference***) with confidence, about a whole (population) based on the study of a only a few (sample).

WHAT IS SAMPLING?

If a sample of a population is to provide useful information about that population, then the sample must contain essentially the same variation as the population.

The more heterogeneous a population is…

- The greater the chance is that a sample may not adequately describe a population→ we could be wrong in the inferences we make about the population.

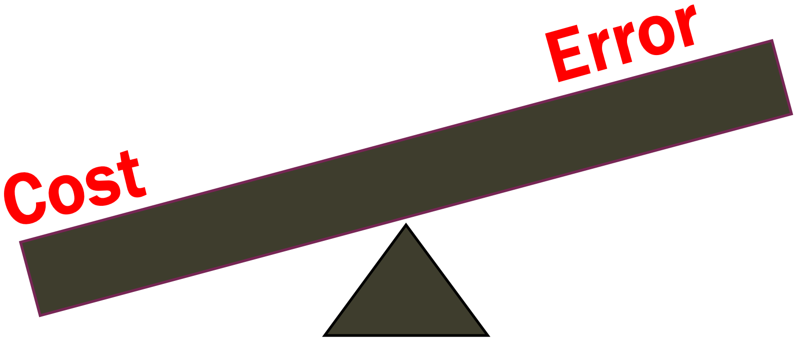
And…

- The larger the sample needs to be to adequately describe the population→ we need more observations to be able to make accurate inferences.

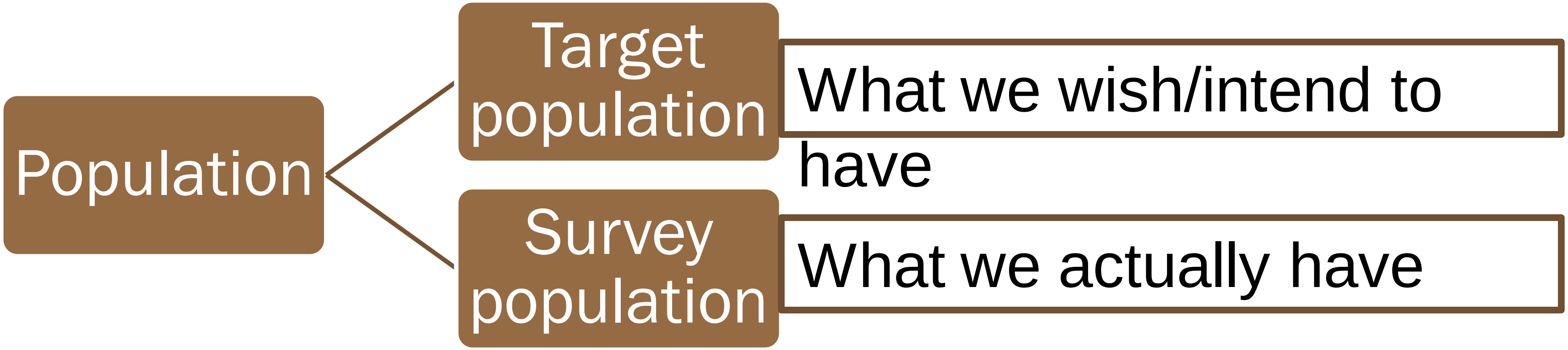
What is sampling?

Ultimate goal in sampling

Select a "representative" sample, to estimate population parameters with "lowest possible cost" and "error"



Basic Concepts



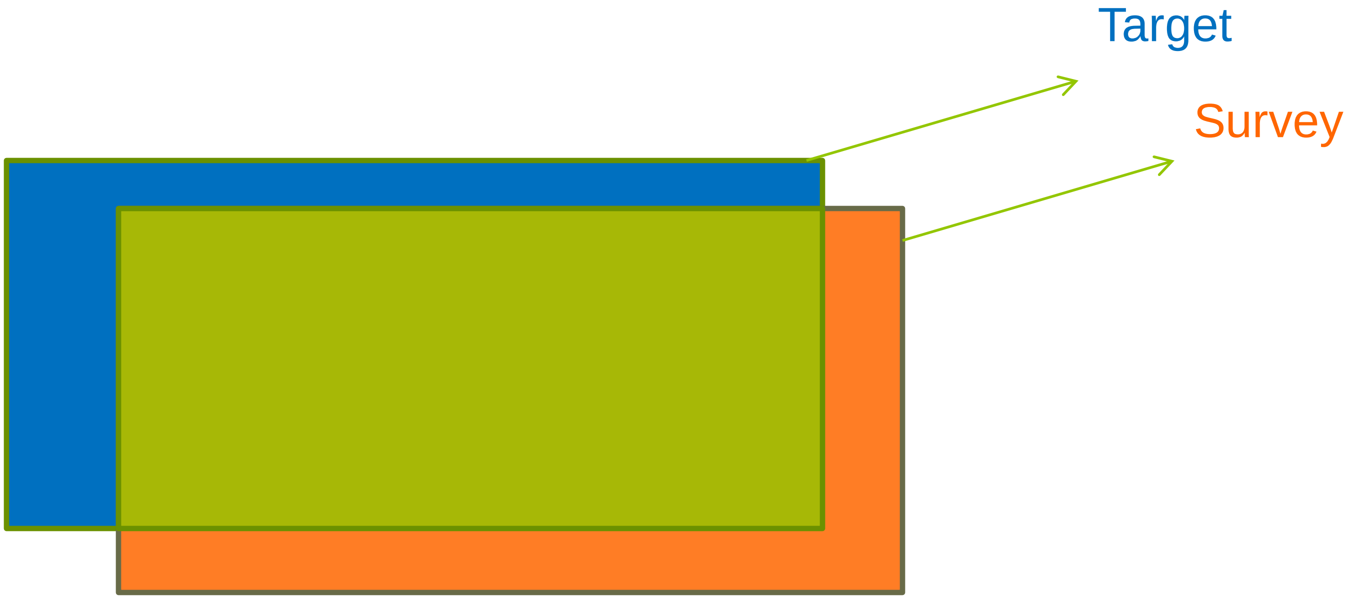
Population

Target population

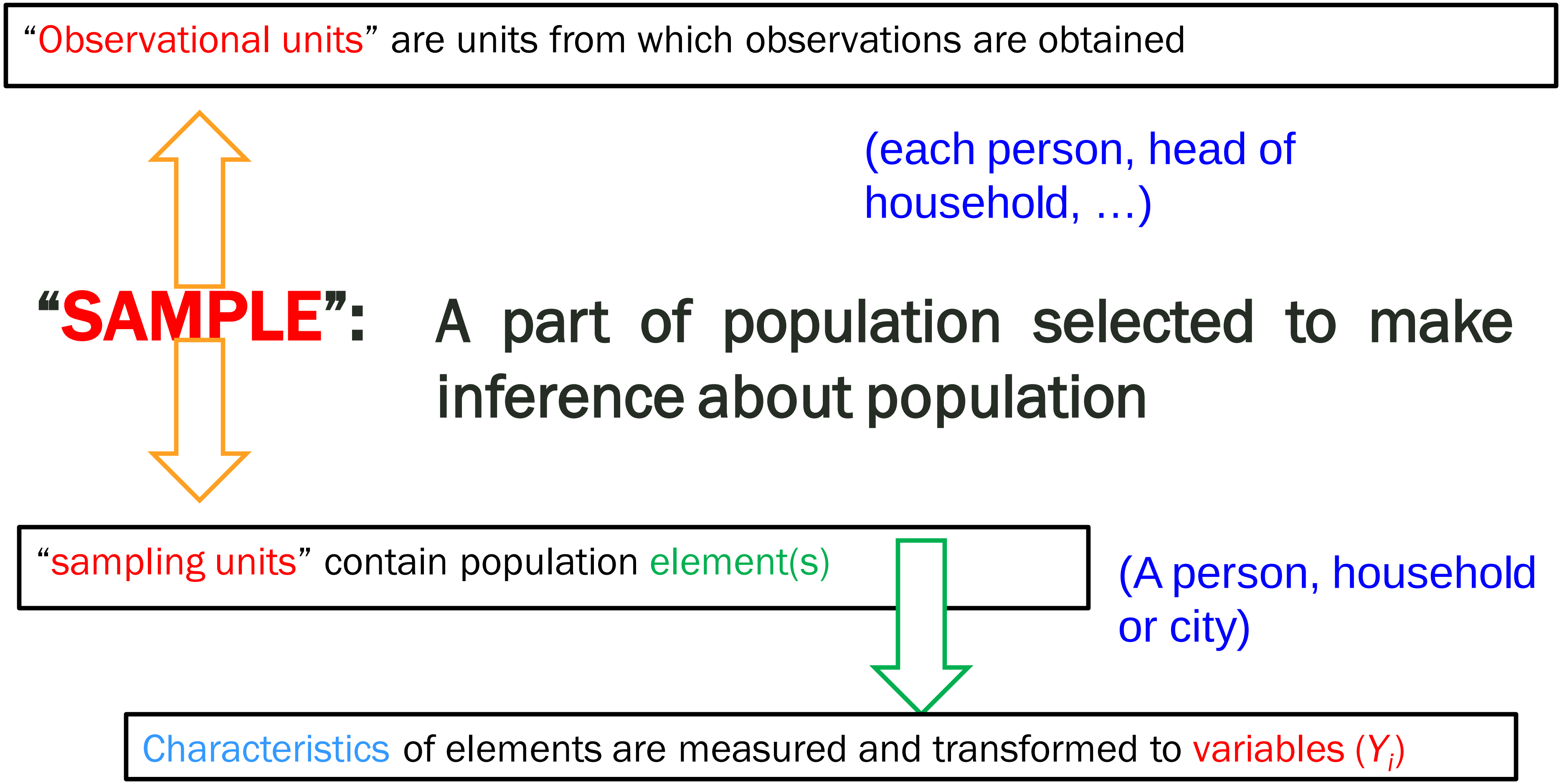
What we wish/intend to have

Survey population

What we actually have



Basic Concepts



"Observational units" are units from which observations are obtained

(each person, head of household, …)

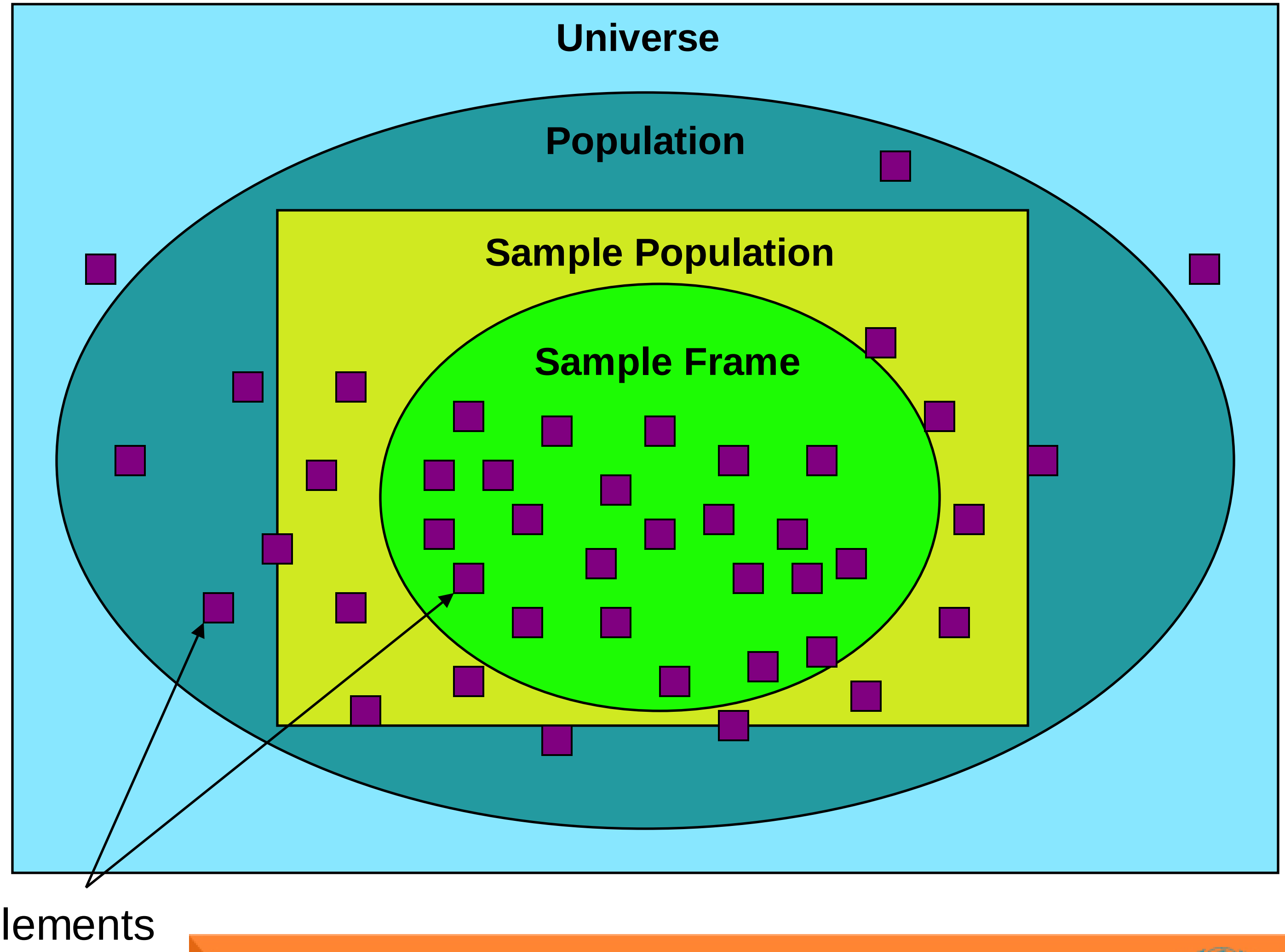
"SAMPLE": A part of population selected to make inference about population

"sampling units" contain population element(s)

(A person, household or city)

Characteristics of elements are measured and transformed to variables (Y_i)

BASIC CONCEPTS



BASIC CONCEPTS

- *Unit* is an element on which observations can be made. These are the units of analysis.[Examples: households, farms/ plots of agriculture crop]
- *Reporting unit* is one that actually supplies the required statistical information
- *Observation unit* is one about which data are reported

BASIC CONCEPTS- SAMPLING UNITS

Sampling Unit is an element of the population selected in the sampling process on which we collect data

Example:

When we select a sample of households , target units of observations may be persons living in the households

Female members of household at age 15-49 in reproductive health survey

In multi-stage sampling plan, one has *first stage sampling unit (fsu), second stage sampling unit (ssu),* etc.

Basic Concepts- : Characteristic

Characteristic: Different kinds of information on elements of the population are collected in a survey. Each of these items of information is called a characteristic

Each characteristics has different values for different individual units

Observations on several characteristics of the units are collected in a survey

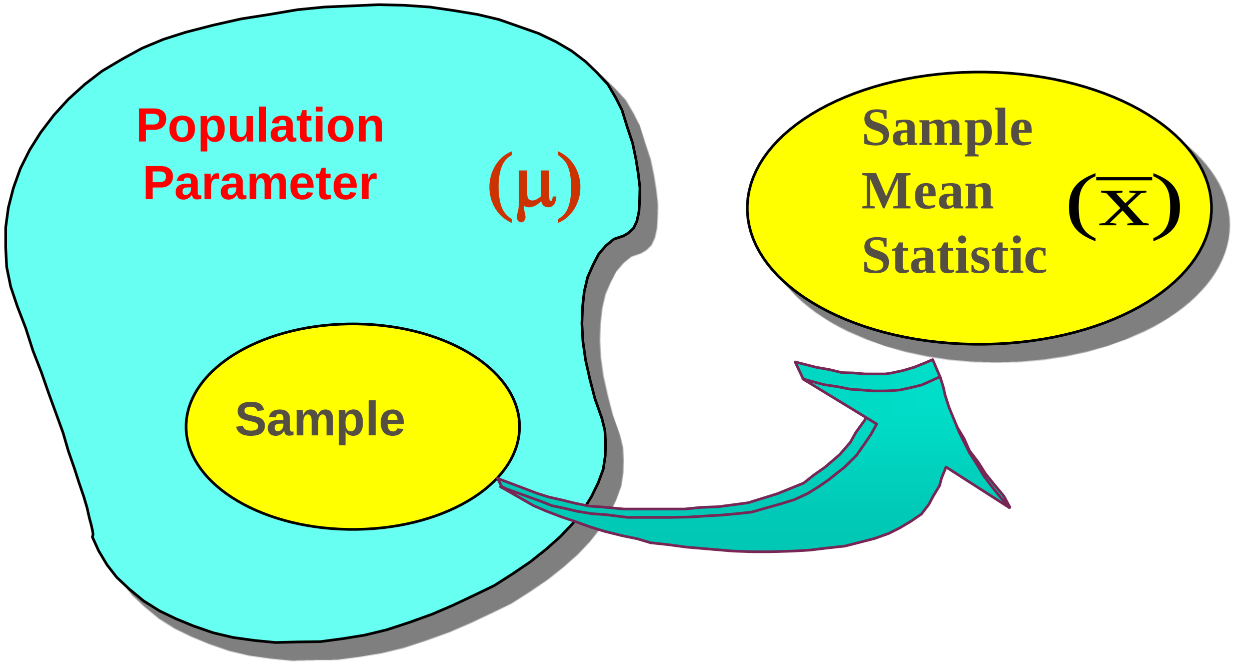
Characteristic can be a ***quantitative variable*** like income of a household, number of cattle on a farm, area of land under rice crop in an agricultural holding

or an ***attribute or categorical variable*** like gender, employment status of a person, economic activity code of a production unit

Basic Concepts- Parameter, Statistic, Estimator, Estimate

➔ A population **parameter** is a numerical summary of a population, a of elements in the population function

➔ A **Statistic** is a function of elements in the sample (a subset of the population) It is called **estimator** if indicating parameter



- **Estimate:** numerical value of an **estimator** that is obtained from a particular sample of data and used to indicate the value of a parameter

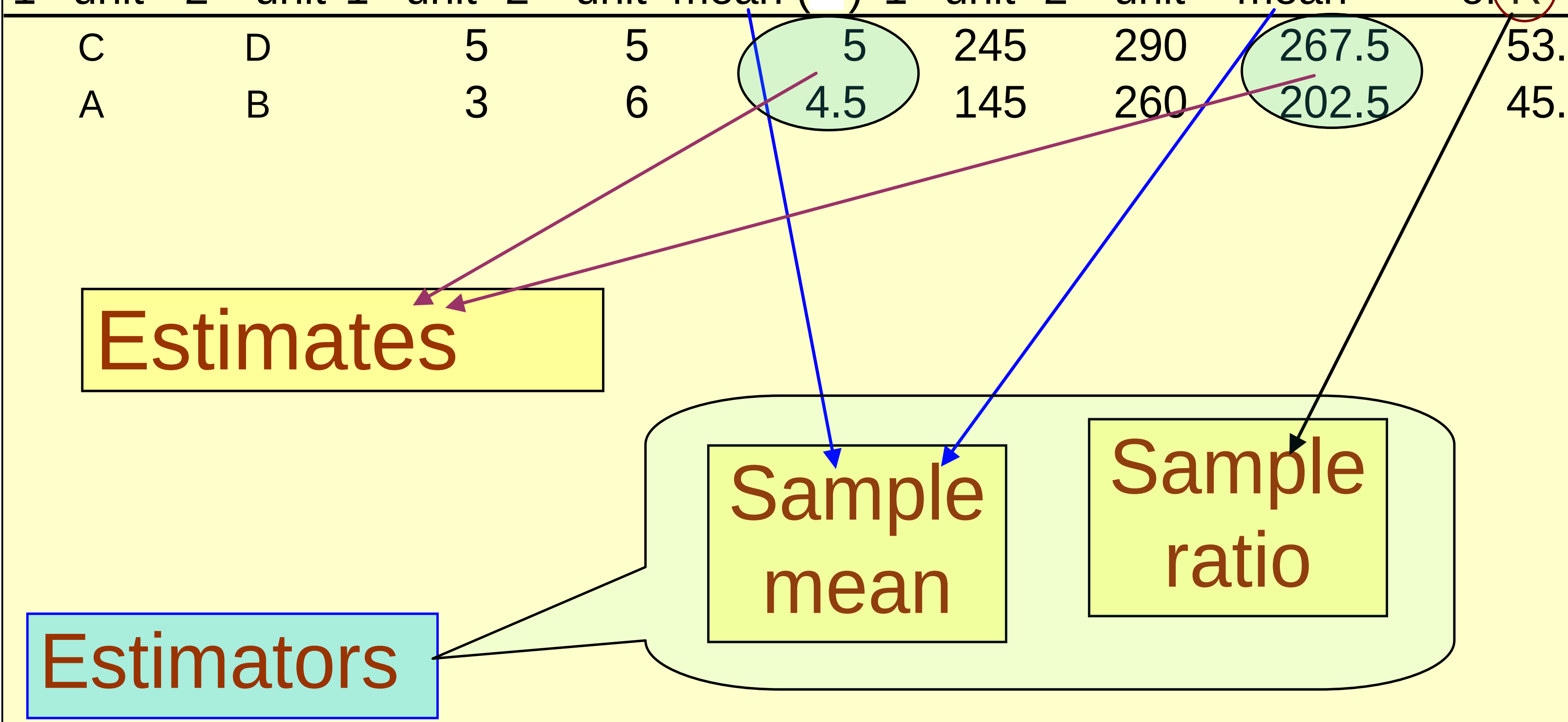
EXAMPLE

THE VALUES OF X AND Y SHOWN IN THE TABLE BELOW ARE THE ACTUAL VALUES (NOT KNOWN TO THE SAMPLER)

Milk Producers	# milch animals (X)	Milk output (Y)	average yield (R)
A	3	145	48.3
B	6	260	43.3
C	5	245	49.0
D	5	290	72.5
E	2	140	70.0
F	4	180	45.0

IN THE EXAMPLE

Samples		sample values of X			sample values of Y			sample ratio - estimate
1st unit	2nd unit	1st unit	2nd unit	mean ($\bar{x}$)	1st unit	2nd unit	mean	of R
C	D	5	5	5	245	290	267.5	53.5
A	B	3	6	4.5	145	260	202.5	45.0



Estimates

Estimators

Sample mean

Sample ratio

Qualities of Estimators: Unbiased, Consistent, Efficient

- ***Unbiased estimator*** of a population parameter is an estimator whose *expected value* is equal to that parameter

  In the example, Sample means of ‘average number of milch animals’ and ‘average output’ are unbiased estimators of the respective population parameters (Population means) But, sample yield rate (which is a ratio) is not an unbiased estimator of the corresponding population parameter

- ***Consistent estimator*** is one where the difference between the estimator and the parameter grows smaller as the sample size grows larger

  Sample ratio (in the example) is not unbiased but is a consistent estimator

- ***Efficiency*** is defined as the reciprocal of sampling variance

  If there are two unbiased estimators of a parameter, the one whose variance is smaller is said to be *relatively efficient*

UNITED NATIONS SIAP
Statistical Institute for Asia and the Pacific

Selection process

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

Simple Random Sampling (SRS)

What is a simple random sampling (SRS)?

Simplest method of probability sampling

Special type of equal probability selection method (*epsem*)

Rarely used in practice for large scale surveys

Theoretical basis for other sample designs

What is a simple random sampling (SRS)?

TYPES OF SRS

SRS selection can be made

- With Replacement (*SRSWR*) or
- Without Replacement (*SRSWOR*)

What is a simple random sampling (SRS)?

SELECTION PROCEDURE

1) Get a list (sampling frame) which uniquely identifies each *"sampling unit"* in the population

2) Allocate a serial number to each *"sampling unit"* of the frame

3) Generate random numbers [in the range of *1 to N*] using Random Number Table/ Random Number Generator

   - 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

What is a simple random sampling (SRS)?

EXAMPLE (N=9 , N=3)

Sampling units	Serial No	Selection order (SRSWOR)	Selection order (SRSWR)
1000050001	1		
3000050002	2		
1004050003	3		
1023050004	4	2nd	2nd & 3rd
1000054002	5		
1011050005	6	3rd	
1110050001	7	1st	1st
1030051020	8		
1025051201	9		

Random No (between 1-9)
7
4
4
6

What is a simple random sampling (SRS)?

SELECTION PROBABILITY

Probability that a population unit is selected at any given draw

Selection probability is the same for both SRSWR and SRSWOR:

$$\frac{1}{N}$$

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

Systematic Sampling

- Linear systematic sampling
- Circular systematic sampling

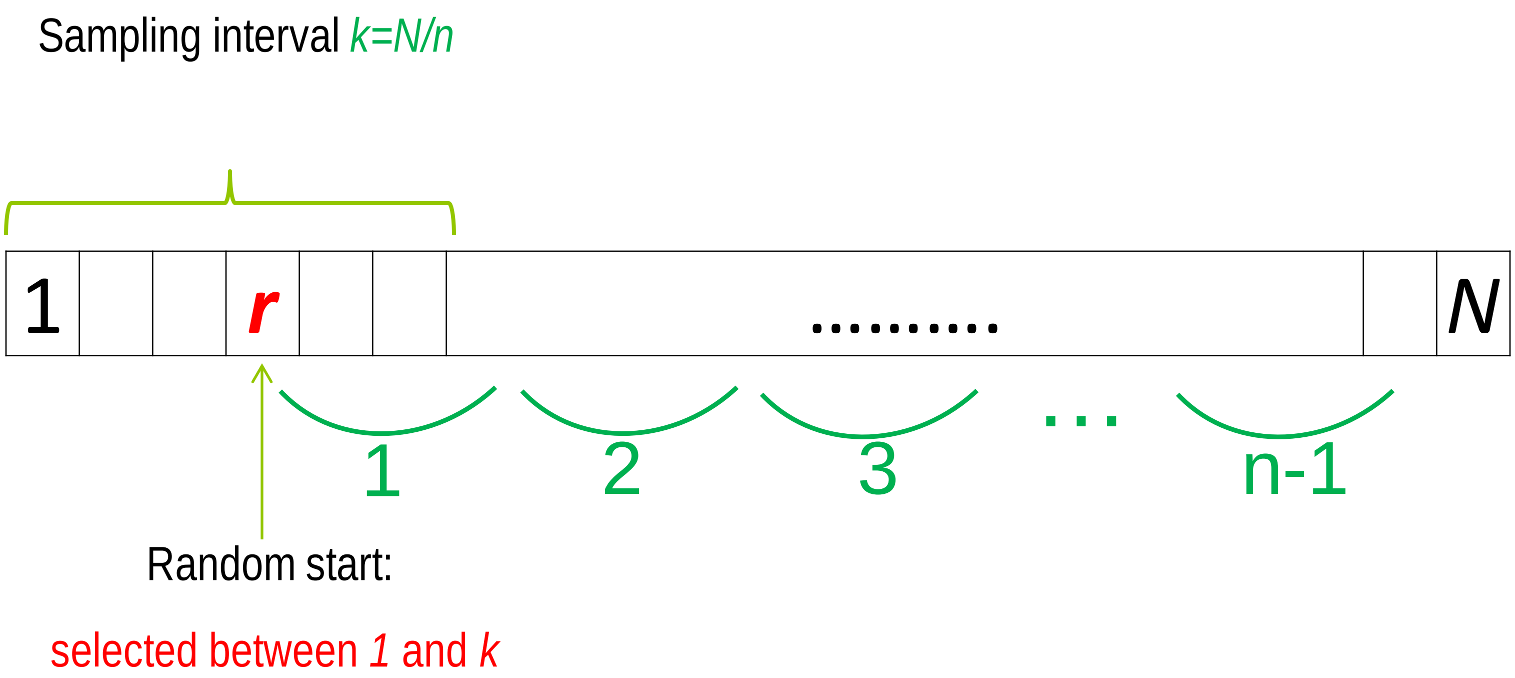
Systematic Sampling

Systematic Sampling (SYS), like SRS, involves selecting *n* sampling 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*, $k = N/n$

Linear Systematic Sampling



Linear Systematic Sampling

Selection Procedure

1) Form a sequential list of population units

2) Decide on a sample size n and compute the skip (*sampling interval*), $k = N/n$

3) Choose a random number, r (*random start*) between 1 and k (inclusive)

4) 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

Linear Systematic Sampling

Problem

***k* =*N*/*n* is integer**

- ❑ *N* is a multiple of *n*
- ❑ *N* units can be grouped into *k* samples of exactly *n* units each
- ❑ Sampling design is *epsem*.

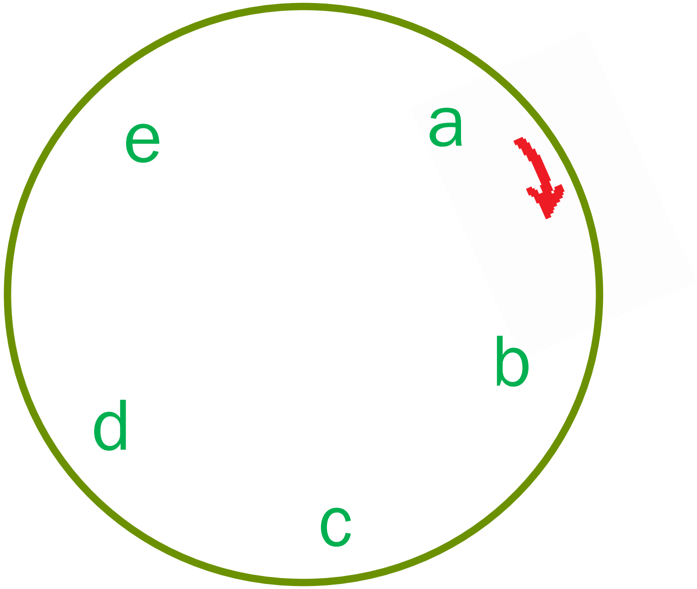
***k* =*N*/*n* is NOT integer**

- ❑ Number of units selected with the sampling interval *k*

  [= nearest integer to N/n] – no longer *epsem*.

Circular Systematic Sampling

SOLUTION



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.

Circular Systematic Sampling

SELECTION PROCEDURE

1) 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)

2) Take a random start between 1 and *N*

3) Skip through the circle by *k* units each time to select the next unit until *n* units are selected

4) Thus there could be *N* possible distinct samples instead of *k*

Systematic Sampling

TO REMEMBER THAT SYS ....

Often used as an alternative to SRS.

Requires ordering of the population units

- For SYS sample to be more representative
- Geographical ordering ensures fair spread of sample
- Ordering by age ensures representativeness of all ages

Ensures each population unit equal chance of being selected into sample

Systematic Sampling

Advantages

- ❑ Easier to draw a sample
- ❑ Distributes sample more evenly
- ❑ Likely to be more efficient than SRSWOR, particularly when ordered by characteristics related to 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**

THANK YOU