Basic Concepts of Sampling-Brief Review; Sampling Frame

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

Sampling – Basic concepts and definitions

- Surveys and Sampling
- Population and Sampling unit
- Population Parameters and Statistic/ Estimators
- Random Samples and Probability Sampling
- Estimator and estimate
- Unbiasedness, Consistency, Efficiency
- Sampling Frame

Basic Concepts and Definitions- Surveys

Survey: A set of statistical activities designed to obtain data

Types of survey: Census surveys, Sample surveys

Census survey: A complete enumeration of the population of interest. Data are collected from all elements of the population

Examples: Population Census, Agriculture Census, Economic Census, Census of housing, Census of establishments, Livestock Census

Sample survey: A survey of a subset of population employing selection procedure

Examples: Labour Force survey, Household Income & Expenditure survey, Survey of Establishments and Enterprises, Demographic health survey, Living conditions survey

Basic Concepts - Sampling

Sampling is selection process of a part (sample) of an aggregate material (technically termed as Population) to represent the whole aggregate

The *process of sampling* provides information about the population characteristics on the basis of the representative sample observations

Advantages of sampling:

- Reduced cost (economy)
- Greater Speed and Timeliness
- Feasibility (if observations are destructive)
- Greater Quality and accuracy
- Detailed/specialized information, etc.

Basic Concepts- Units and Population

- *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
- ***Population (or universe)*** is defined as totality or collection of all units (elements) under study
- ***Finite population*** has limited number of elements and an ***infinite population*** has unlimited elements

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- Population(Contd.)

- *Variable:* a characteristic of an element under study whose measure called *value*, differs from unit to unit
- *Uni-variate population:* one where only one characteristic is observed on an element
- *Multi-variate population:* one where more than one characteristic is observed on an element (for two characteristics, it is *bi-variate population*)
- *Target Population:* the set of elements about which information is sought and estimates of the population parameters are obtained on the basis of the sample. Also known as *coverage universe*

  [some units may be excluded (e.g. the homeless)]

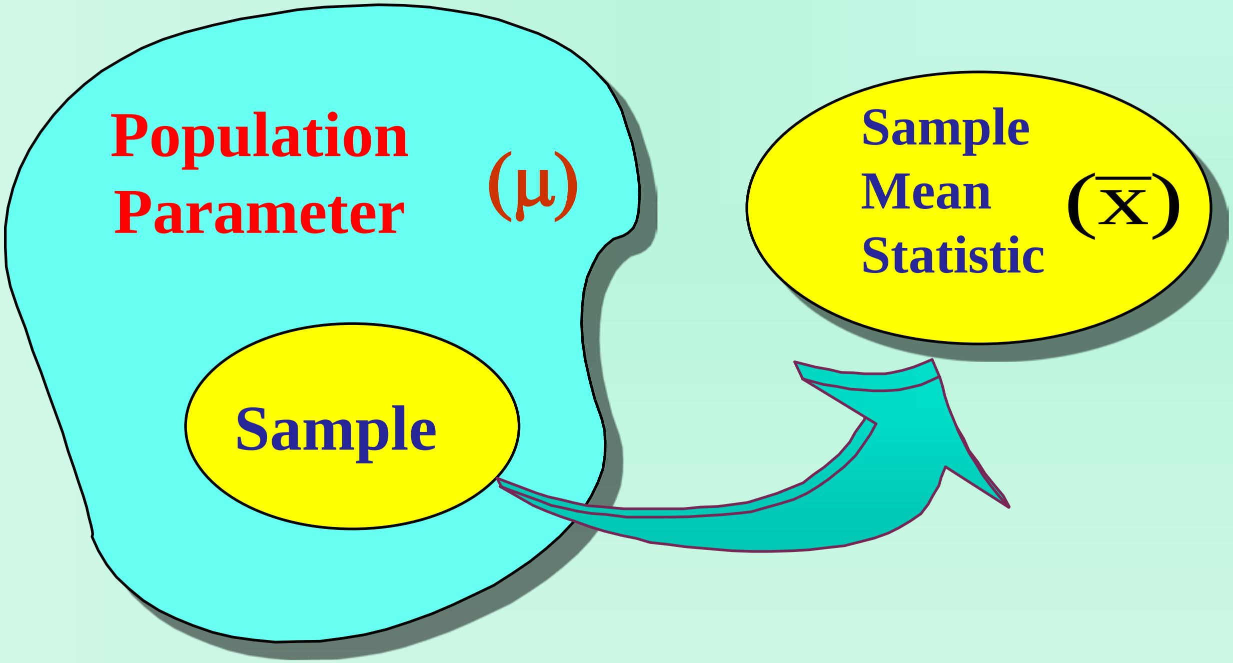
Basic Concepts- Random Sample, Probability Sampling

- *Random sample* is a sample based on a probability scheme
- *Probability sampling* theory aims to make statistical inference about the population parameters on the basis of random samples. Every population element has a known, non-zero, probability of selection for the sample
- *Probability Sampling Ingredients*
  - Defined set of samples that are possible to obtain with the sampling procedure
  - A known probability of selection is associated with each possible sample (sampling design)
  - A known nonzero probability of selection (inclusion probability) of each element in the population
  - One of the possible samples is selected by a random mechanism according to the **sampling design**
- *Non-probability Sampling* makes use of purposive sample (Quota, Convenience, Snowball, etc.), results are not statistically valid

[We will focus only on probability *sampling]*

Basic Concepts- **Parameter, Statistic, Estimator, Estimate**

- A population **parameter** is a numerical summary of a population, a function of elements in the population (Pop mean μ, Pop variance σ^2)
  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

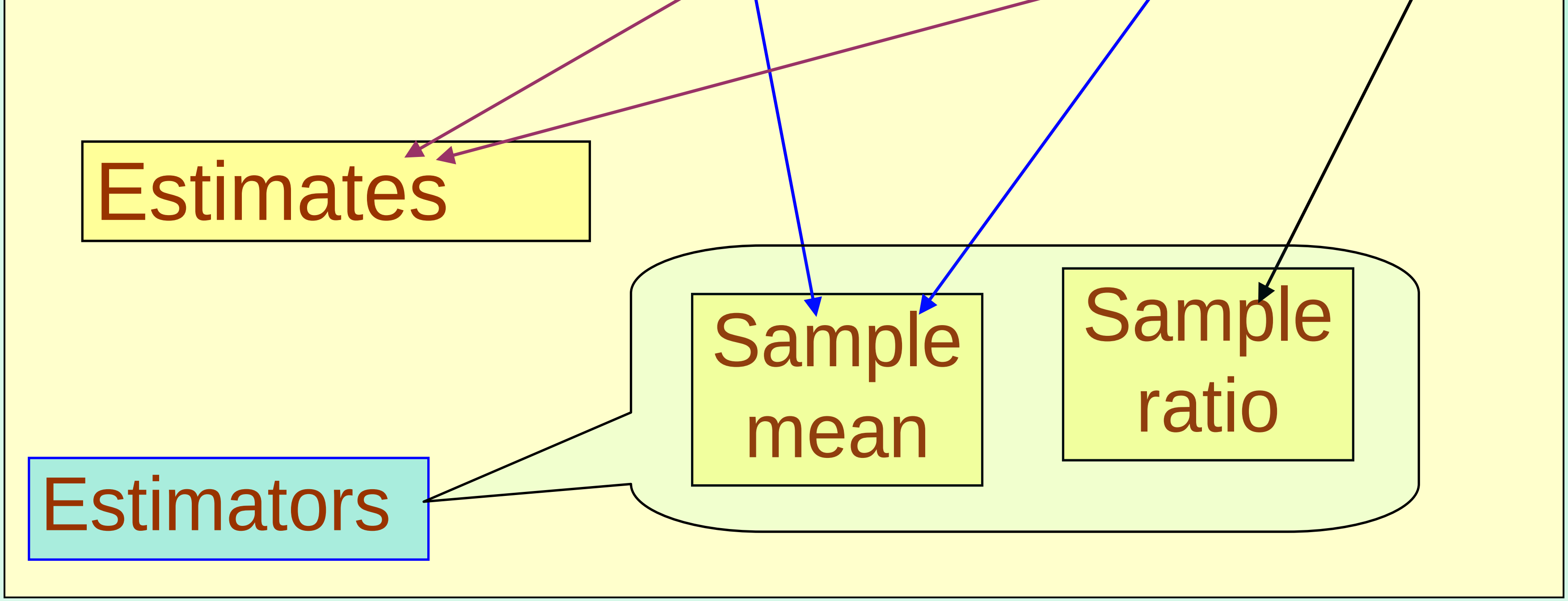
Example

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



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*

Properties of Estimators

- Sampling error
- Sampling Distribution and Sampling Variance
- Standard error and Design effect (*Deff*)

Properties of Estimators- Sampling Error

Sampling Error: represents the difference between the estimate $\hat{\theta}$ and the value of the population parameter *θ*

($\hat{\theta}$ – *θ*)*,* the error in a sample estimate that owes to the selection of only a subset (sample) of the total population rather than the entire population. All sample estimates are subject to sampling error

The most commonly used measure of sampling error is

Sampling Variance: $V(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$

Where *E* denotes expected value, Sampling variance depicts Precision of the estimator

Bias of estimator : $B(\hat{\theta}) = E(\hat{\theta}) - \theta$

Efficiency of the estimator: $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = [Bias(\hat{\theta})]^2 + Var(\hat{\theta})$

Properties of Estimators- Sampling Variance

- ***Sampling Variance*** is the average of squares of (value of the survey estimator obtained from a sample *minus* the value of the population parameter) over all possible samples that can be drawn from the population

$$V(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$$

- The variance of an estimator contains information regarding how close the estimator is to the population parameter
- Sampling variance is a measure of sampling error
- Sampling error reflects the difference between an estimate derived from a sample and the "true value"

Properties of Estimators (Cond.)

Sampling Distribution: A frequency distribution of the values of an estimator for each sample that can possibly be drawn from the population

- The sample design and sample size remaining unchanged, the **higher** the **population variance** (measure of variability of a population) the **higher** is the **sample variance**
- The sample design and population (variance) remaining unchanged, the **higher** the **sample size** the **lower** is the **sample variance**
- For a given population and sample size, the **sample variance** depends on the sample design adopted

Properties of Estimators- Standard Error, RSE

- *Standard error* of an estimator $\hat{Y}$ is the squared root of the sampling variance of the estimator

$$s.e.(\hat{Y}) = \sqrt{Var(\hat{Y})}$$

- *Relative Standard Error / Coefficient of Variation* of an estimator is Standard Error / value of Parameter *Y*

$$RSE = {SE\left(\hat{Y}\right)}\Big/{Y}$$

Sept.2013

Measure of Efficiency - Design Effect (*Deff*)

- The relative efficiency of a sample design w.r.t. the simplest sample design – SRSWOR – is measured by ***Design effect*** (*Deff*)
- ***Design effect*** of a sample design, say *D*, is defined as the ratio the standard errors of *D* and SRSWOR

$$Deff\ = \frac{s.e.(design\quad D)}{s.e.(SRSWOR)}$$

- Estimates of ***Deff*** are often used for determining the required sample size for a given design

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

Main focus

Development of Course Design

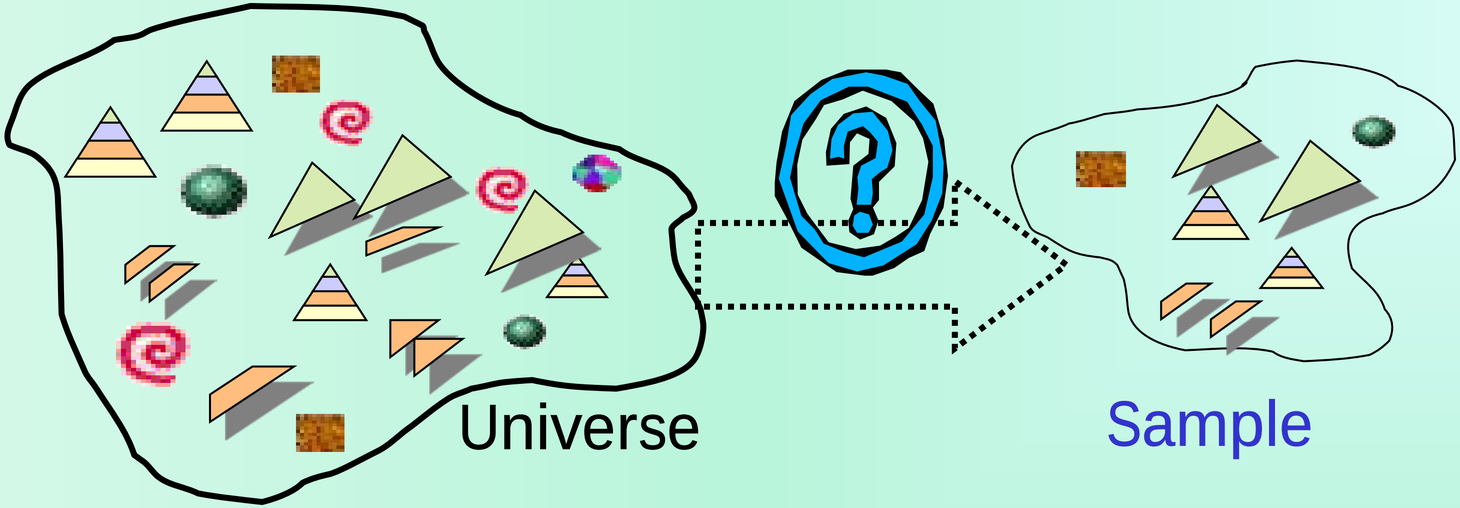
Sampling Frame

- ***Sampling frame*** is a list or device that delimits, identifies, and allows access to the elements of the target population. Sample elements are selected from ***frame*** using appropriate probability scheme

- Frame Units: Examples

- Area units
  - Administrative Subdivisions
  - Census Enumeration Areas
  - Areas Delineated on Maps
  - Agricultural Holding

- Non-area units
  - Housing Units
  - Households
  - Persons
  - Enterprises
  - Establishments

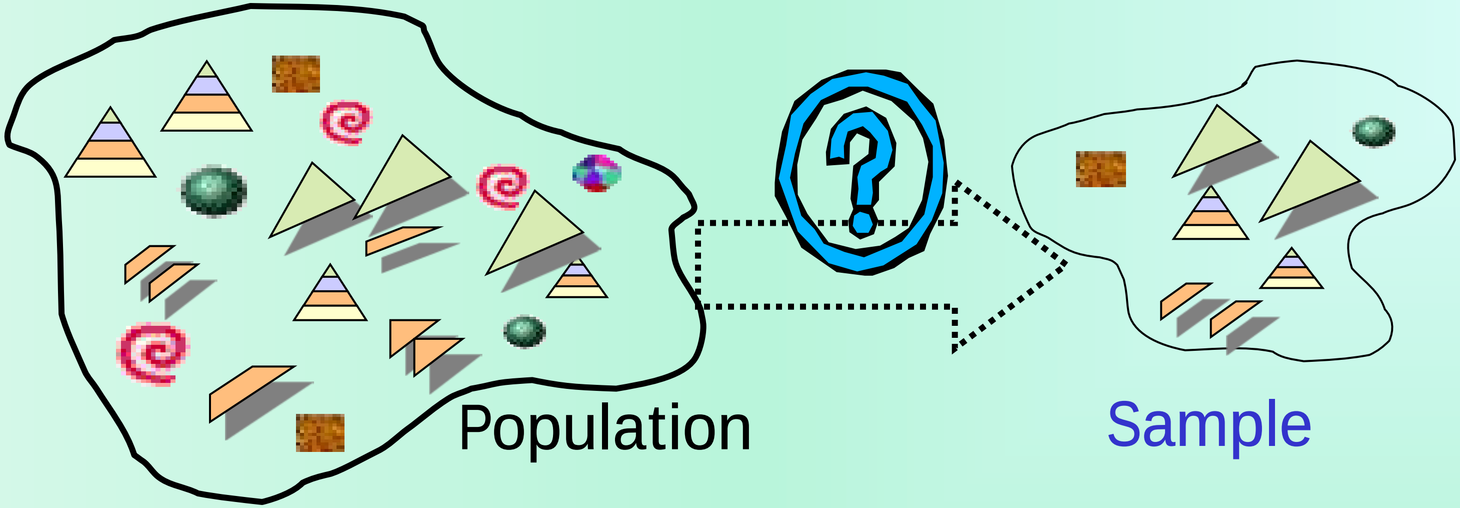


Sampling FRAME: Concept

- Survey Population has to be represented in a physical form from which sample can be selected
- Sampling Frame is such a representation
  - Explicit list of the elementary units to be surveyed
  - Account for all units in some implicit form

Access to Population

- Set of source materials from which the *sample* is selected
- Provides a means for selecting the *elements* of the *target population* that are to form the sample



Sampling FRAME: Importance

- Related to cost of survey: availability of sampling frame suited to best option among alternative sample designs
- Affects data quality: ‘faulty’ sampling frames lead to *non-sampling errors*

Types of Frames

- List Frames
- Area/ Area-based Frames
- Multi-Stage Frames
- Frames for Series of Surveys
- Multiple Frames (combination)
- Master Sampling Frame:
  - Multi-purpose
  - Multi characteristics

Considerations for choosing a Frame

- Long term view of all possible surveys
- Based upon available infrastructure
- Possibility of updating
- Statistical Registers: a type of Frame
  - difficult to keep them updated, if not linked to an administrative process

Sampling Frames: List Frame

- *List frame* enlists all the elements of the population Each frame unit is identified and listed with its identification particulars, such as--
  - Listing of household addresses from Census
  - Listing of names, addresses and telephone numbers in telephone directory
  - Listing of establishment names, addresses and type of business produced from a business register

List Frame Sources: Population Registers, Census Listings, Administrative Lists (e.g., Business Registers)

Listing by operations (e.g., Listing of households within selected enumeration areas in 2-stage sampling)

List Frames (Contd.)

- Completeness is most critical requirement
- Accurate information of the size and characteristics of individual units
- Different situations require separate treatment
  - Large and few units for which good lists may be available
  - Medium sized units covered by a list frame that is more difficult to construct and maintain
  - Medium to small units which may require a combination of list and area based frames
  - Small units which can only be covered with area based frames
  - For unevenly distributed units, specially constructed area frame, taking account of the patterns of concentration (e.g., fishing households, mining worker households); Rare and dispersed population also require special methods (tribal households)

Problems with List Frames

- Blanks (L=0): Listing represents no real unit
- Duplications (L, L) =U: Same unit is represented by more than one unit
- Clustering of elements L= (U, U): More than one unit represented by same listing
- Under coverage U=0: Units not represented in the frame
- Failure to locate units L=?: Failure to identify which unit a selected listing represents
- Change in units and unit characteristics L = U*: Unit itself (or characteristics of the unit associated with the unit) has changed

List Frames

- Strengths:
  - can use inexpensive data collection methods (mail, telephone)
  - can target specific or rare commodities
  - can reduce variability due to sampling
  - cost-efficient: could be built on available resources
- Weaknesses:
  - does not cover entire population (threshold criteria)
  - goes out-of-date quickly
  - increased non-sampling errors due to data collection methods
  - requires on-going maintenance
    - ➢ build
    - ➢ update
    - ➢ remove duplication
    - ➢ remove out-of-scope records

Sampling Frames: Area-based

- **Area frame** is a geographic frame consisting of area units. Every population element belongs to an area unit
- Frame units are the geographical units in a hierarchical arrangement
  - Cover the entire country
  - Boundaries are well delineated
  - Population figures are available
  - Units are mapped
- For household surveys, in developing countries, frame generally consists of
  - one or more stages of area units
  - followed by the list of households or enterprises within the selected ultimate area units
- Durability of frame declines as we move down the hierarchy of the units

Sampling Frames: Area-based (Contd.)

- Choice of the type of units to be used as the primary stage units (PSU) is important
- Coverage errors in area based frames arise from
  - failures to define boundaries of area units correctly
  - poor quality of the lists of ultimate units
- Sources for area-based frames are generally the
  - Census blocks or enumeration areas
  - Maps of administrative sub-divisions
  - Aerial photographs of housing units
  - Satellite images of land cover

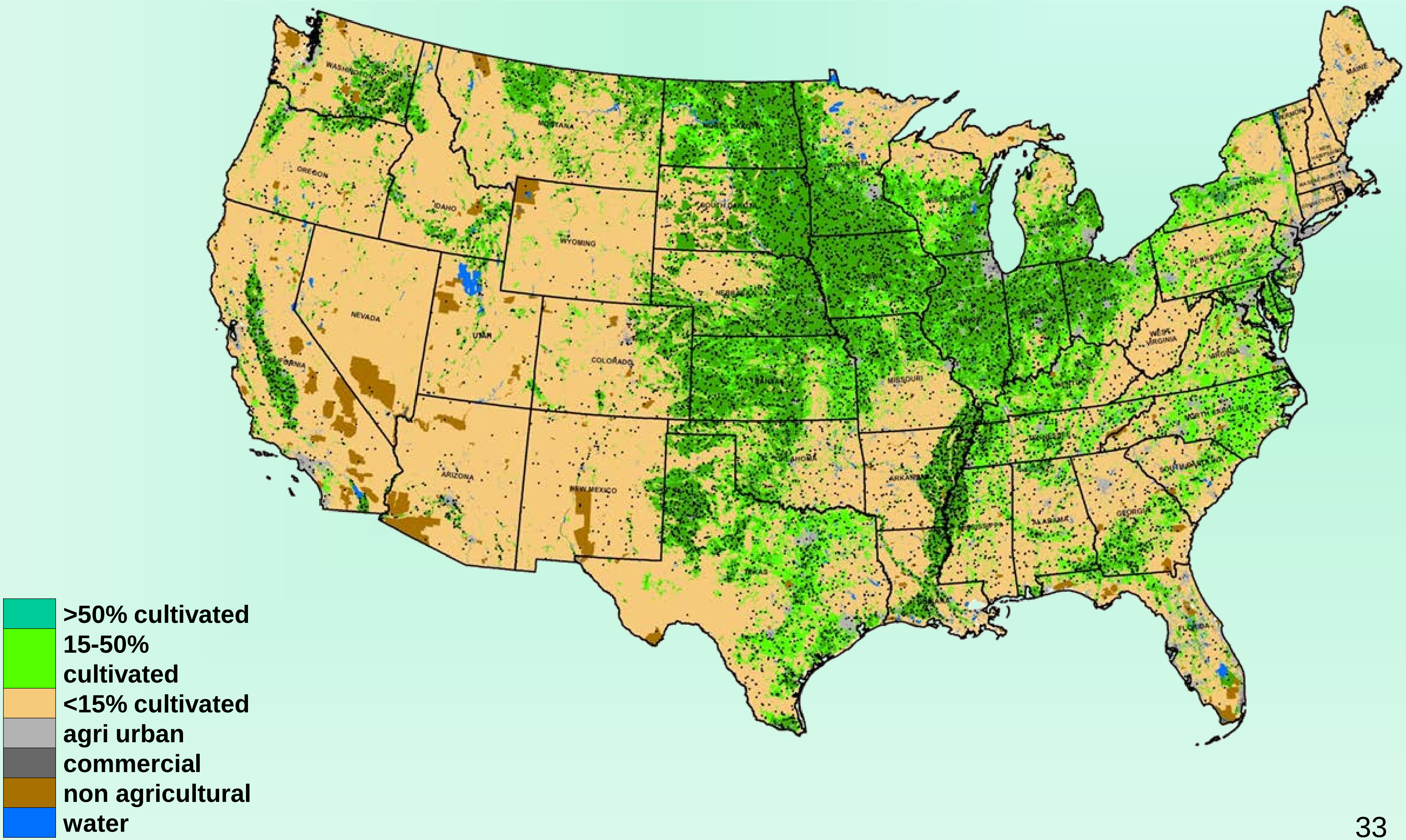
Area Frame...*More..*

- Land is divided in segments on the basis of land use
- A unique ID is assigned to each segment
- List of segments
- Segments (or blocks) are Stratified

  (Stratum: a group of homogeneous units)
- Area Frame: linking of segment with the holding is possible

[National Agric. Statist Service] NASS Area Frame

constructed using satellite imagery, digital maps, GIS software, aerial photography: (i) divide land area into strata based on land use, (ii) subdivide land use strata into strata blocks, (iii) select a sample of strata blocks (iv) subdivide selected strata blocks into segments



Common Problems with Area Frames

- Though areas as units are larger, more easily identified and more stable than dwellings, households which appear as units in lists, there are problems, like
  - Failure to cover the population exhaustively
  - Errors in area boundaries
  - Inappropriate type and size of units
  - Lack of auxiliary information
  - High cost (useful only for repetitive surveys)

Area Frame (Contd.)

Strengths:

- complete coverage
- reduced non-sampling errors
- estimates well for commonly produced commodities
- versatility
- longevity (desertification/ greening)

Weaknesses:

- expensive (frame construction & data collection)
- difficult to target specific or rare commodities
- sensitive to outliers
- can be inefficient
- requires definable physical boundaries

Multi-stage Frame

- Usually area based
  - Primary sampling frame (frame for the first stage of sampling) has to cover the entire population
  - Following first stage selection, the list of units at any lower stage is required only with in the larger units selected at the preceding stage
- Frame from a single source / number of sources
- Different types of frames may be used for different parts of the population

Multi-stage Frame (Contd.)

- Used in Multi-Stage Sampling
  - Primary sampling frame for first stage selection
  - Intermediate sampling frames for intermediate stages of selection
  - PSU, SSU, TSU,…….,FSU

Example: 2-stage Sampling

- First stage frame
  - Sampling frame units are areas (e.g., blocks, EAs)

    [Sampling frame is an area frame]
- Second stage frame
  - Sampling frame units are elements (e.g., households; informal sector enterprises)

    [Sampling frame is a list frame of households for each of the area units sampled in the first stage]

Multiple Frame

- Technique for combining several incomplete frames in order to capture completeness of the frames
- Union of all the frames constitutes the entire population
- Samples are selected independently from each frame
- Optimum combination of estimates coming from non-overlap
- Overlap domains provides an overall estimate for the multiple frame situation

Multiple Frame (Contd.)

- Overlap between frames provides independent estimates from samples coming from different frames
- When overlapping frames used, ensure that unit's probabilities of selection remain definite & known
- Various approaches:
  - Make frames non-overlapping
  - If a list and an area frame are used in combination, any unit selected from area frame must be excluded from list (whether or not selected)

Frame for Series of Surveys- Master Sample

- In multi-stage sampling design, each stage involves task of frame preparation and sample selection, till finally a sample of ultimate units is obtained
- For economy/convenience one or more stages of these tasks may be combined or shared among a number of surveys
  - Sample resulting from shared stages is called a *Master Sample*

Master Sampling Frame

- Master Sampling Frame is basically a list of area units that covers the whole country
- For each frame unit there may be information on
  - urban/rural classification
  - identification of higher level units (for example, the district and province to which the unit belongs)
  - population counts
  - stratification variables

Master Sampling Frame- Concept

- From a master sampling frame, it is possible to select the samples for different surveys entirely independently
- In many instances, there are substantial benefits resulting from selecting one large sample, a master sample of area units for the first stage of sampling, and then selecting sub-samples of this master sample to service different (but related) surveys
- It is a sample from which sub-samples can be selected to serve the needs of more than one survey or survey round

Master Sample- Objectives

- A common sample of units down to a certain stage from which further sampling done for individual surveys
- Economize, by sharing among surveys, on costs of developing sampling frame, design and selection
- Facilitate linkages between different surveys or between successive rounds of a continuing survey
- Control drawing of multiple samples for various surveys from the same frame

Master Sample (Contd.)

- Master sample for household surveys (multi-stage)
  - Same sample of primary sampling units (usually, area units) are used for different surveys
  - Next stage units may be same or different
- Periodic (monthly, quarterly, annual) surveys of establishments make use of a sample of establishments maintained over some interval of time
- Sample of units designed for multiple use (different or same surveys repeated over time)

Master Sampling Frame- Advantages

- Cost of developing a good sampling frame is usually high and the establishment of a continuous survey programme makes it possible for the NSO to spread the costs of construction of a sampling frame over several surveys
- Facilitates quick and easy selection of samples for surveys of different kinds and it could meet different requirements for the sample from the surveys

Integrated Household Survey Programme

- Concept
  - long range plan rather than *ad hoc*
  - coordinated planning
  - integration of survey design and operations
- Integration can be achieved by
  - Use of same concepts and definitions for variables occurring in several surveys
  - Sharing of survey personnel and facilities among the surveys to secure effective use of staff and facilities
  - Use of common sampling frames and samples for all the surveys in the survey programme

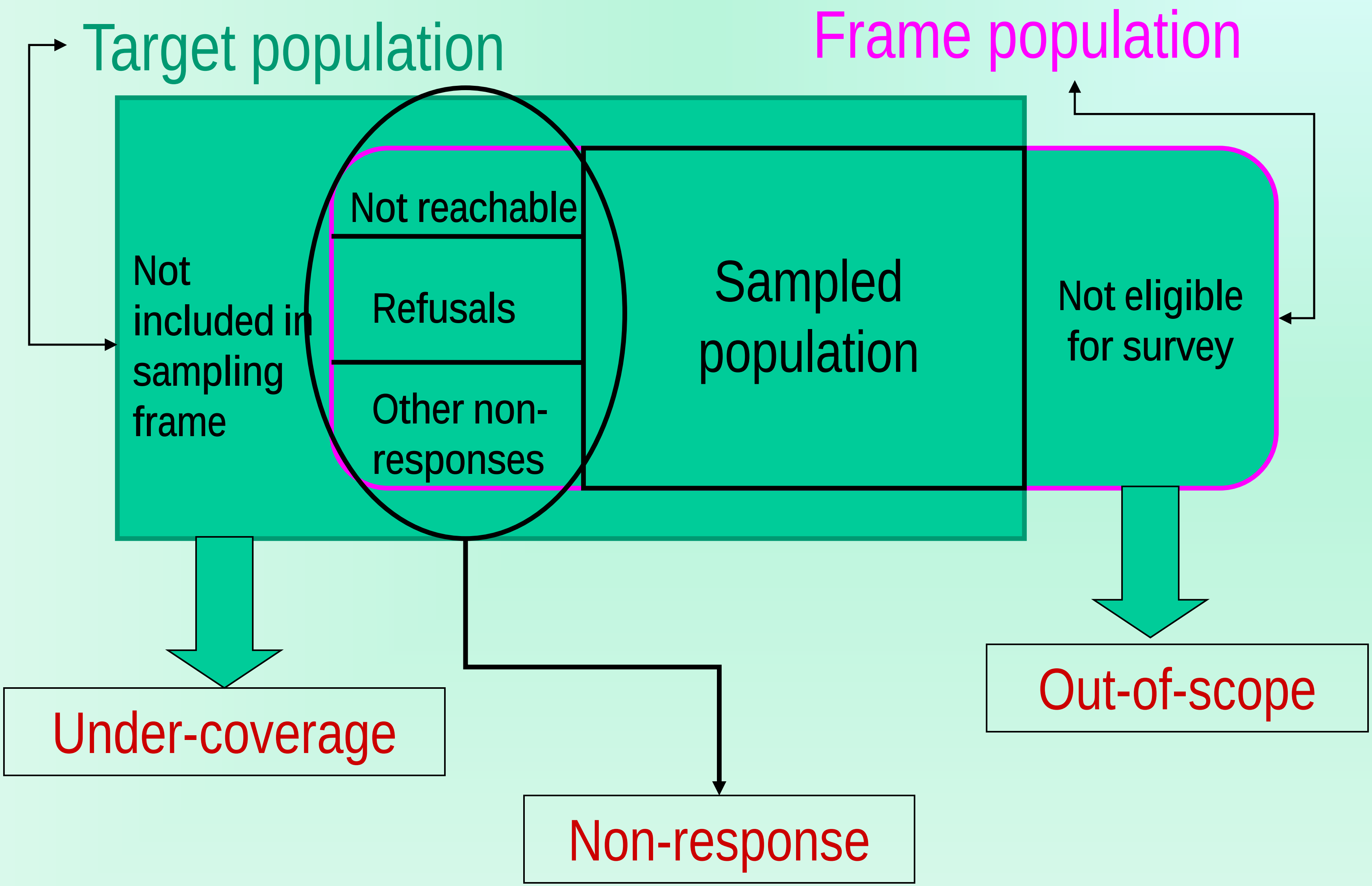
Desirable Frame Properties

- Quality-related
- Efficiency-related
- Cost-related

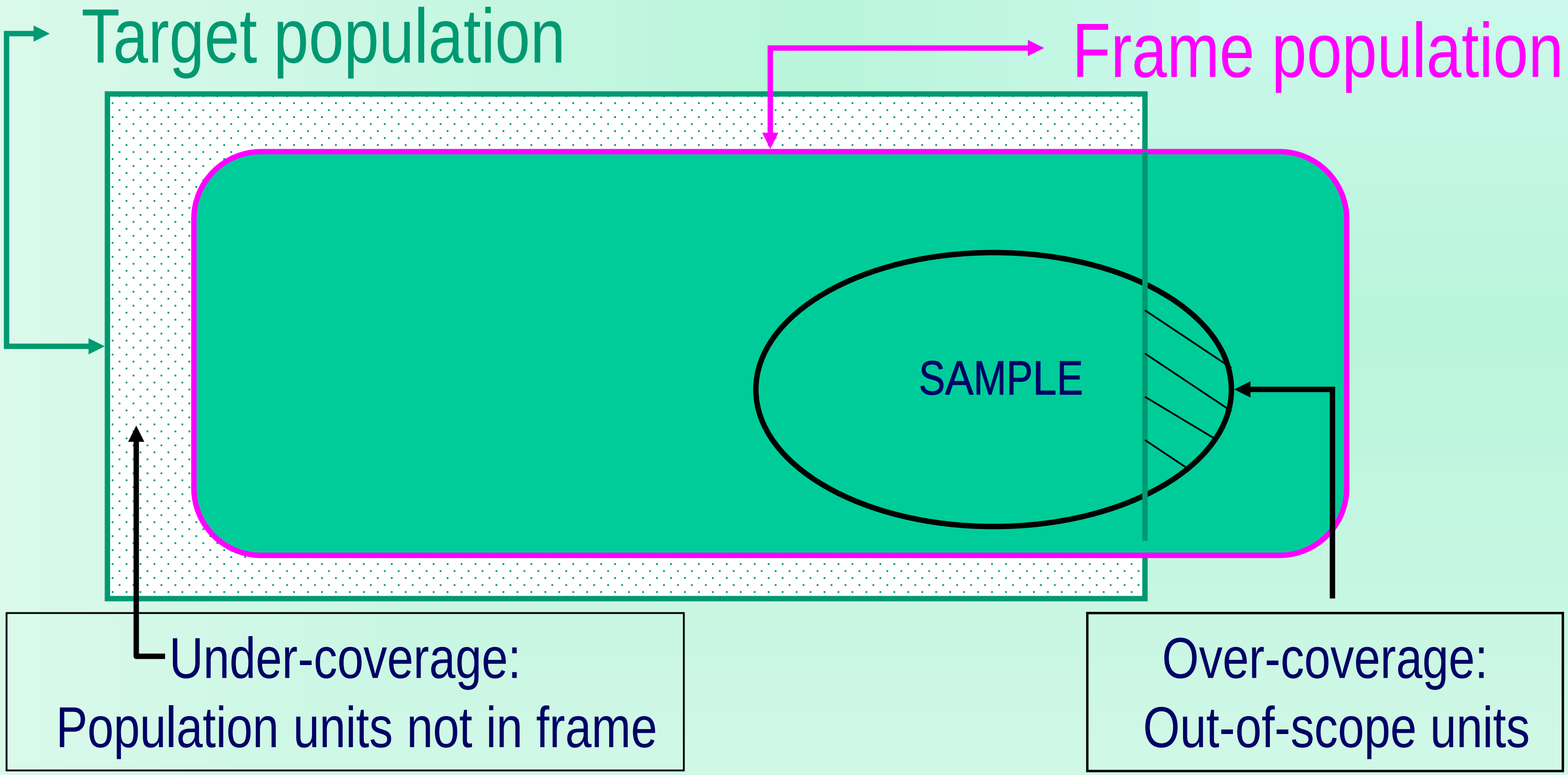
Frame Quality

- Criterion: as a result of the frame, every member of the target population has a known, non-zero chance of being selected
- Completeness: all elements are covered
- Accuracy: each element is included once and only once
- Up-to-date: updating of frame must be regular plan activity

In Practice ...


Target population
Frame population
Not included in sampling frame
Not reachable
Refusals
Other non-responses
Sampled population
Not eligible for survey
Under-coverage
Non-response
Out-of-scope

Frame Imperfections



Frame Imperfections (Contd.)

- Over-coverage
- Duplications
- Under-coverage, non-coverage, incomplete coverage
- Not enough information to provide access to some survey population units

Efficiency

- Inclusion of accurate and up-to-date auxiliary information
- Good quality maps of units available
- Easy access; Easy to process
- Choice of sampling units available (if to be used for more than one survey or survey round)
- Enabling production of summary statistics

Cost

Low cost of acquisition / preparation
Low cost of use
Low cost of maintenance

Sampling Frames: Contents

- Key contents
- Operational Considerations
- Construction
- Administration and Maintenance

Sampling Frames: Contents

- One record per frame unit
  - Primary identifier
  - Secondary identifier(s)

    Identifiers
  - Stratification variable(s)
  - Measure(s) of size

    Unit Characteristics
  - Sample selection indicator
  - Change indicator(s)

    Operational data

The Choice of Frame

- Depends upon related on going activities in other sectors
  - Resource availability
- Nature of agriculture
  - Extensive, mono-crop, or
  - Intensive multi-crop
- Scope of surveys
  - just the crop area, crop monitoring, land degradation or
  - many economic characteristics e.g. fertilizers, cost of production

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.

Sample 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

[These will be discussed in next session]

Thanks