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

Module 2: Review of Basics of Sampling Methods

Session 2.1: Terminology, Concepts and Definitions

9 – 20 November 2015,
Jakarta, Indonesia

UNITED NATIONS SIAP Statistical Institute for Asia and the Pacific

Improving Agricultural Statistics Global Strategy

BPS, Statistics Indonesia

Contents

- Probability Sampling in Statistical Surveys
- Sampling unit and Sampling Frame
- Characteristics – quantitative variables and categorical variable
- Important Definitions: Population parameter, statistic, estimator & estimate, sampling with and without replacement, Simple Random Sampling (SRS)
- Sampling error, Sampling variance and Sampling Distribution
- Unbiasedness, Efficiency and Consistency

Probability Sampling
in
Statistical Surveys

Statistical 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

Content of the present course

Sample

A subset of the population on which observations are taken for the purpose of obtaining information about the population.

- By studying a sample we hope to draw valid conclusions about the population.
- Thus, a sample should desirably be 'representative' of the target population.

Sample Design - two broad kinds

Sample design specifies how to select the part of the population, i.e. sample, to be surveyed.

Two broad kinds:

- **Probability sampling** or **Random sampling**:
  In which each element of the population is assigned a non-zero chance of being included in the sample.
  [focus of this course]
- **Non-probability sampling**:
  Consists of a variety of procedures, including judgment-based and 'purposive' choice of elements – considered "representative" of the population.

Use of Random Numbers for Sampling (1)

- Consider a population whose *population size* is 1000.

  [Number of units in a population is called *Population size*]

- In the simplest kind of probability sampling, each unit is assigned equal chance of selection. In this case, the chance, i.e. the *selection probability*, of all the units is 1/1000.
- *Selection probability* of a unit is the probability of selecting it at the first draw, when the sample is required to contain more than one unit.
- A sample usually consists of more than 1 unit. The number of unit a sample should consist of is called its *sample size*.

Use of Random Numbers for Sampling (2)

- Before selection, all the units in the population are first given a running serial number from 1 to 1000. This serial number is called *Sampling Serial Number*.
- For selecting a unit, a random number between 1 and 1000 is drawn. The unit having the sampling serial number same as the random number is selected for the sample.
- This ensures that every unit in the population has a non-zero chance of selection and that all their *selection probability* is the same.
- We will later see how units from a population can be drawn with *unequal probability*.

A Few Questions

Which of the following are *random* or *probability* samples?

A. From a class of 50 students, a sample consisting of those with roll numbers that are multiple of 7?

B. From the same class, a sample of those with roll numbers same as first seven random numbers drawn by you?

C. In a ranch with 35 grazing cows, a sample of the first five cows that you come across as you walk in?

D. A ladle full of rice being cooked (in the traditional way) in a pot , after having stirred its contents well?

E. A random sample of holdings from a randomly selected sample of villages?

With and Without Replacement Sampling (1)

With Replacement Sampling: in which, at each draw, the member of the population selected for the sample is returned to the population (i.e. replaced) before the next draw is made.

In these, a unit may appear more than once (repeated) in a sample. Commonly not used in practice.

Without Replacement: in which, once a unit is selected in one draw is excluded from the population for subsequent draws.

In these, a unit can at most appear once in a sample.

With and Without Replacement Sampling (2)

Conceptually, 'with replacement' sampling schemes are same as 'without replacement' sampling from a notional population that has infinite number of each unit of the original population.

Thus, 'with replacement' schemes are notionally sampling from *infinite population*, while 'without replacement' schemes are sampling from *finite population*.

Simple Random Sampling (SRS)

- SRS is the simplest method of probability sampling
- Special type of equal probability selection method (*epsem*)
- SRSWR and SRSWOR
  - SRSWR rarely used in practice for large scale surveys
  - Theoretical basis for other sample designs

Probability Sampling in Statistical Surveys

Statistical surveys like those conducted government agencies are always conducted to obtain information about a population, such as

- Resident (human) population / households
- All villages
- Livestock population
- All agricultural holdings
- All fruit bearing mango trees etc.

Note that the *population* in a sample surveys is always finite.

Sampling Unit

and

Sampling Frame

Target population and Sampling frame

Target population: The population intended to be studied or covered in the survey;

also known as *coverage universe.*

Sampling frame: A list of units (possibly with other information on each unit) from which selection of sample is made.

Note that it is desirable that the sampling frame be exactly the same as the target population, but it is not always so.

Sampling Unit

- The units constituting the sampling frame are called *sampling units*.
- These are entities that are selected in the sampling process.
- Each sampling unit in the frame should be distinguishable and have a unique identification number.

Examples of sampling units:

- Individual in the sampling frame for a Labor force survey (LFS)
- Cattles in a livestock survey
- Holdings in agricultural surveys
- We will later see, that a village can be sampling unit in an Agricultural Survey with a multi-stage sampling design.

Sampling Unit and Unit of Observation

A **sampling unit** may be different from *unit of observation*.

When we

- select a sample of households for LFS, target units of observation are persons living in the household.
- select a sample of households (target unit of observation) by selecting a sample of villages first, then selecting a sample of households within the sampled villages.

Characteristics –
quantitative variables and categorical variable

Characteristic (1)

Different kinds of information on the elements of the population (or populations) are collected in a survey.

Each of these items of information is called a **characteristic**.

* Each of the characteristics have different possible values for different individual units.
* Observations on several characteristics of the units are collected in a survey.

Characteristic (2)

A characteristic can be a **quantitative variable** like

* age of a person
* income of a household
* number of cattle on a farm
* area of land under rice crop in an agricultural holding
* value of output of a manufacturing unit

or an **attribute** or **categorical variable** like

* sex of a person aged 15+
* employment status of a person
* economic activity code of a production unit
* whether agricultural produced sold

Characteristic (3)

For a **quantitative variable**, we can talk of total and mean.

For a **categorical variable**, the total and averages are NOT meaningful.

For a categorical variable, what we are interested in is

* proportion of a category in the population or
* Percentage distribution of different categories in the population.

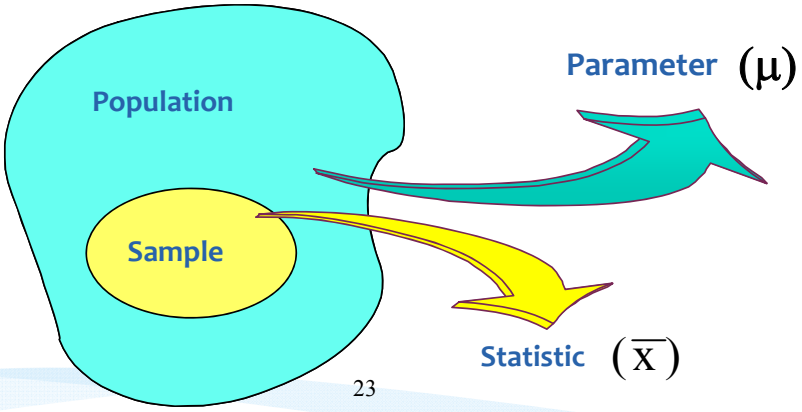
Important Definitions:

Population parameter, statistic, estimator & estimate

With & without replacement and SRS

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 sample ***statistic***.



Population Parameter

A population *parameter* is a summary measure of a population, the value of which helps to describe a population.

For example:

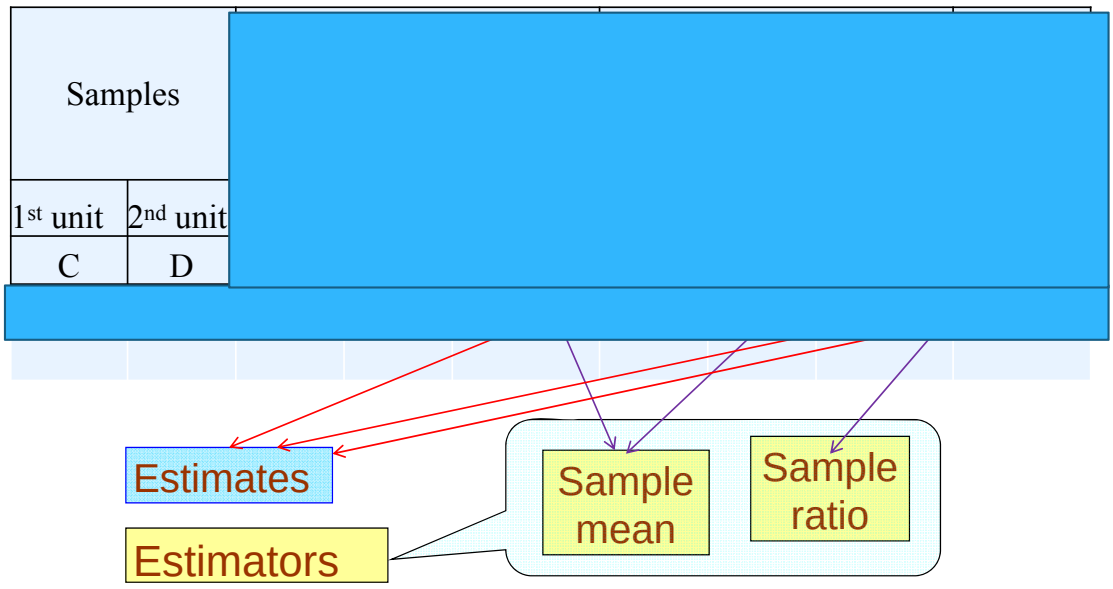
- Total population
- Average size of agricultural holdings
- Literacy rate

An 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	2	65	32.5
B	7	71	10.1
C	21	265	12.6
D	12	113	9.4
E	18	166	9.2
F	3	35	11.7
Average =	10.5	119	11.3

In this example



Definitions

Estimator

An *estimator* is a sample statistic - a function of sample observations -

that gives information about an unknown *population parameter*.

Examples:

* *sample mean* is an estimator of the *population mean*.
* *Sample proportion* is an estimator of *population proportion*
* *Sample ratio* is an estimator of *population ratio*.

Definitions

Estimate

An **estimate** is an indication of the value of an unknown quantity based on observed data.

More formally, an estimate is the particular value of an *estimator* that is obtained from a particular sample of data and used to indicate the value of a *parameter*.

Sampling error, Sampling variance and Sampling Distribution

Sampling Error

- 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.
- Sampling error represents the difference between the estimate and the value of the population parameter.
- All sample estimates are subject to sampling error.
- The most commonly used measure of sampling error is ***sampling variance***.

Sampling Variance (1)

- Sampling variance is a measure of sampling error.
- Sampling error reflects the difference between an estimate derived from a sample and the "true value".
- It can be measured from the population values, if they are known.

  (But these are unknown - otherwise there would be no need for a survey).

Sampling Variance (2)

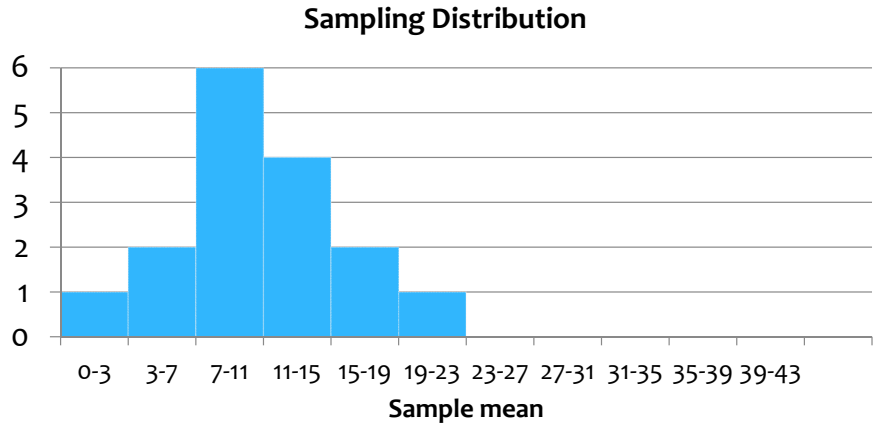
Definition: The average of squares of (value of the survey estimator obtained from a sample *minus* the value of the population parameter) over all possible samples that can be drawn from the population.

- The variance of an estimator contains information regarding how close the estimator is to the population parameter.
- Estimates from a survey may have different sampling variance.

32

Sampling Distribution

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



33

Sampling Variance – determining factors

- Sampling error (variance) is affected by a number of factors:
  - variability within the population.
  - sample size - sampling fraction
  - sample design, and
- If sampling principles are applied carefully
  - within the constraints of available resources,

sampling error can be accurately measured and kept to a minimum.

Unbiasedness, Efficiency and Consistency

Desirable Qualities of an Estimator

* *Unbiasedness*
* *Consistency*
* *Efficiency*

Unbiased Estimator

An ***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.
- But, sample yield rate (which is a ratio) is not an unbiased estimator of the corresponding population parameter.

Consistency

An estimator is said to be ***consistent*** if 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

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***.

Thanks