Sampling and Sampling Distribution

-Dr. Umesh R A

Aegis
SCHOOL OF BUSINESS
SCHOOL OF DATA SCIENCE
SCHOOL OF TELECOMMUNICATION

Sampling Methods

- Sample is nothing but part of population.
- **Sampling method** refers to the way that observations are selected (extracted) from a population.
- The reason for conducting a sample survey is to estimate the value of some attribute of a population.
- **Parameter**: It is a statistical constant of the population.

  Ex. Population mean, Population standard deviation
- **Statistic**: It is a function of sample value.

Sampling Unit:

- The constituents of a population which are individuals to be sampled from the population and cannot be further subdivided for the purpose of the sampling at a time are called sampling units.
- For example
  - –to know the average income per family, the head of the family is a sampling unit.
  - –to know the average yield of rice, each farm owner’ s yield of rice is a sampling unit.

Sampling Frame:

- For adopting any sampling procedure it is essential to have a list identifying each sampling unit by a number. Such a list or map is called sampling frame.
- Ex.
  - A list of voters
  - A list of house holders
  - A list of villages in a district,
  - A list of farmers

Advantages of sampling:

- Sampling saves time and labour.
- It results in reduction of cost in terms of money and man hour.
- Sampling ends up with greater accuracy of results.
- If the population is too large, or hypothetical or destroyable sampling is the only method to be used.

Disadvantages:

- Sampling is to be done by qualified and experienced persons. Otherwise, the information will be unbelievable.
- Sample method may give the extreme values sometimes instead of the mixed values.
- There is the possibility of sampling errors. Census survey is free from sampling error.

Probability vs. Non-Probability Samples

- **Probability sampling:** With probability sampling methods, each population element has a known (non-zero) chance of being chosen for the sample.
- **Non-probability sampling:** With non-probability sampling methods, we do not know the probability that each population element will be chosen, and/or we cannot be sure that each population element has a non-zero chance of being chosen. Non-probability sampling methods offer two potential advantages - convenience and cost.

Non-Probability Sampling Methods

1. Convenience sampling:

It is used in exploratory research where the researcher is interested in getting an inexpensive approximation of the truth. As the name implies, the sample is selected because they are convenient. This non-probability method is often used during preliminary research efforts to get a gross estimate of the results, without incurring the cost or time required to select a random sample.

Non-Probability Sampling Methods

2. Judgment sampling:

It is a common non-probability method. The researcher selects the sample based on judgment. This is usually and extension of convenience sampling. For example, a researcher may decide to draw the entire sample from one "representative" city, even though the population includes all cities. When using this method, the researcher must be confident that the chosen sample is truly representative of the entire population.

Non-Probability Sampling Methods

3. Quota sampling:

It is the non-probability equivalent of stratified sampling. Like stratified sampling, the researcher first identifies the stratums and their proportions as they are represented in the population. Then convenience or judgment sampling is used to select the required number of subjects from each stratum. This differs from stratified sampling, where the stratums are filled by random sampling.

Non-Probability Sampling Methods

4. Snowball sampling

It is a special non-probability method used when the desired sample characteristic is rare. It may be extremely difficult or cost prohibitive to locate respondents in these situations. Snowball sampling relies on referrals from initial subjects to generate additional subjects. While this technique can dramatically lower search costs, it comes at the expense of introducing bias because the technique itself reduces the likelihood that the sample will represent a good cross section from the population.

Probability Sampling Methods

- The main types of probability sampling methods are simple random sampling, stratified sampling, cluster sampling, multistage sampling, systematic random sampling, etc. The key benefit of probability sampling methods is that they guarantee that the sample chosen is representative of the population. This ensures that the statistical conclusions will be valid.

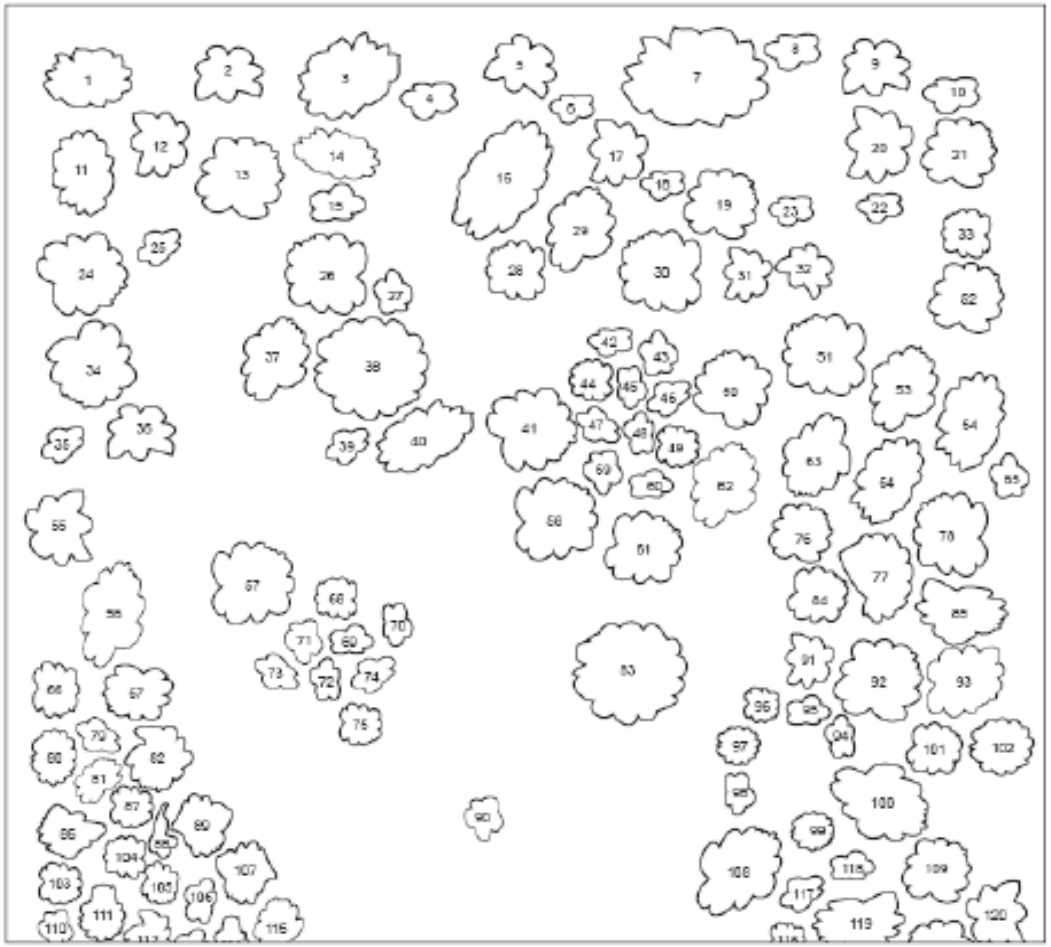
1. Simple Random Sampling (SRS).

- Simple random sampling refers to any sampling method that has the following properties.
  - –The population consists of N objects.
  - –The sample consists of n objects.
  - –If all possible samples of n objects are equally likely to occur, the sampling method is called simple random sampling.

There are many ways to obtain a simple random sample. One way would be the lottery method. Each of the N population members is assigned a unique number. The numbers are placed in a bowl and thoroughly mixed. Then, a blind-folded researcher selects n numbers. Population members having the selected numbers are included in the sample.

Example: simple random sampling

- Survey of insect population living in woodland
- Trees numbered 1 to 200
- 10 trees chosen at random



Merits

- Personal bias is eliminated as a selection depends solely on chance .
- A random sample is in general a representative sample for a homogenous population.
- There is no need for the thorough knowledge of the units of the population.
- The accuracy of a sample can be tested by examining another sample from the same universe when the universe is unknown.
- This method is also used in other methods of sampling

Demerits

- Preparing lots or using random number tables is tedious when the population is large.
- The size of the sample required under this method is more than that required by stratified random sampling.
- It is generally seen that the units of a simple random sample lie apart geographically. The cost and time of collection of data are more.

2. Stratified Sampling

- With stratified sampling, the population is divided into groups, based on same characteristic. Then, within each group, a probability sample (often a simple random sample) is selected. In stratified sampling, the groups are called **strata**.

  As a example, suppose we conduct a national survey. We might divide the population into groups or strata, based on geography - north, east, south, and west. Then, within each stratum, we might randomly select survey respondents.

Example: Stratified sampling

- Foot measurement study of the population of Taiwan
- Total sample size of 1,000
- Sample for each category selected randomly from the population

Age Group	Population (000s)			Sample		
	Male	Female	Total	Male	Female	Total
0-4	830	772	1602	41	38	79
5-9	1005	945	1950	50	47	97
10-14	1016	958	1974	51	48	99
15-19	929	885	1814	46	44	90
20-29	1993	1895	3888	99	94	193
30-49	2744	2635	5379	137	131	268
50+	1882	1618	3500	94	80	174
Total	**10399**	**9708**	**20107**	**518**	**482**	**1000**

Merits:

- It is more representative.
- It ensures greater accuracy.
- For heterogeneous population it gives good results.
- When the population is skewed this method is appropriate.

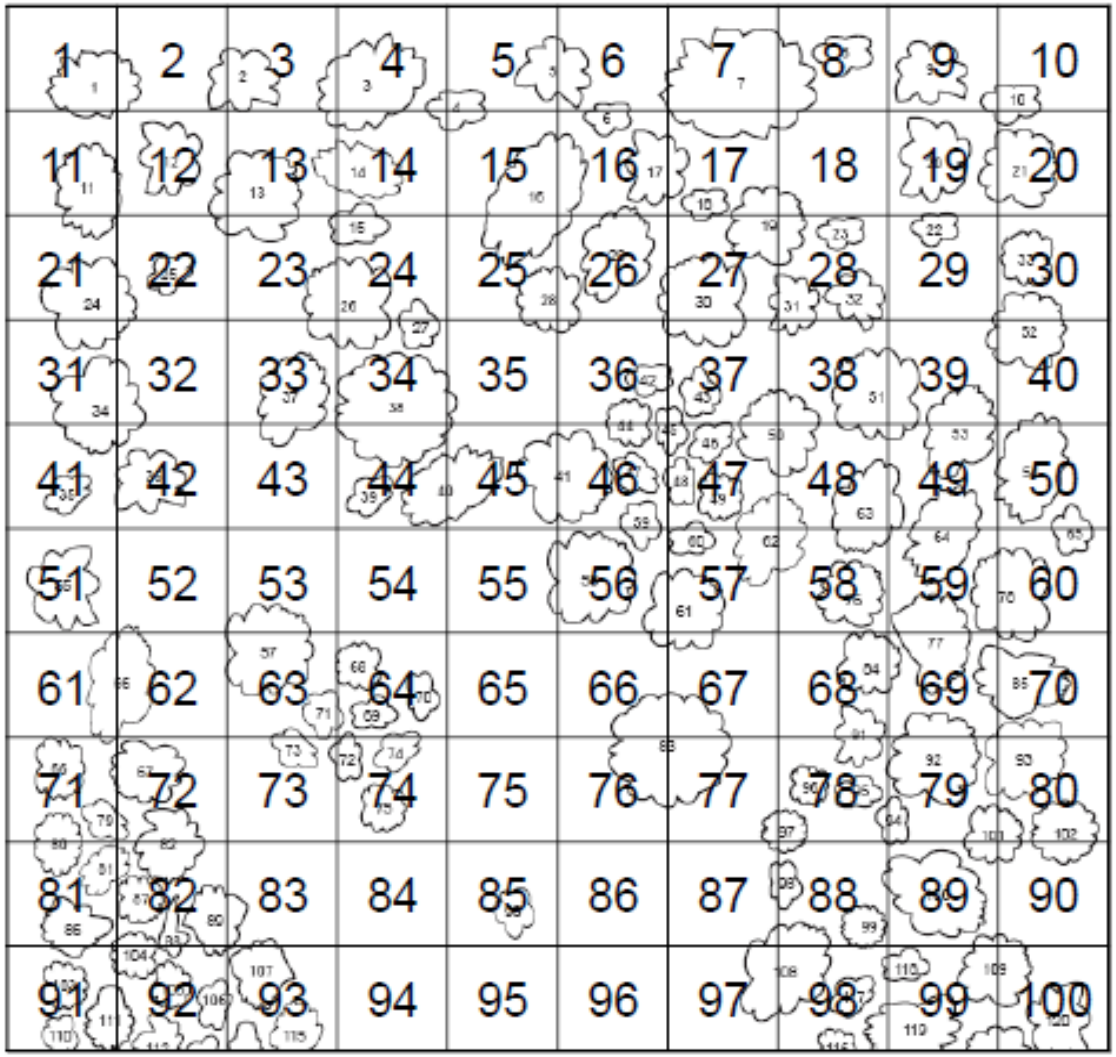
3. Cluster Sampling

- With cluster sampling, every member of the population is assigned to one, and only one, group. Each group is called a cluster. A sample of clusters is chosen, using a probability method (often simple random sampling). Only individuals within sampled clusters are surveyed.

  Note the difference between cluster sampling and stratified sampling. With stratified sampling, the sample includes elements from each stratum. With cluster sampling, in contrast, the sample includes elements only from sampled clusters.

Example: cluster sampling

- Survey of insect population living in woodland
- Squares chosen a random on the grid
- Trees lying within the squares chosen until 10 chosen



4. Two-stage/Multistage sampling

- With multistage sampling, we select a sample by using combinations of different sampling methods.

  For example, in Stage 1, we might use cluster sampling to choose clusters from a population. Then, in Stage 2, we might use simple random sampling to select a subset of elements from each chosen cluster for the final sample.

5. Systematic Sampling

- With systematic random sampling, we create a list of every member of the population. From the list, we randomly select the first sample element from the first *k* elements on the population list. Thereafter, we select every k^{th} element on the list.

  This method is different from simple random sampling since every possible sample of *n* elements is not equally likely.

- A frequently used method of sampling when a complete list of the population is available is **systematic sampling**.
- It is also called **Quasi-random sampling.**

Selection procedure:

- The first unit is selected with the help of random numbers and the rest get selected automatically according to some pre designed pattern is known as **systematic sampling**.
- With systematic random sampling every Kth element in the frame is selected for the sample, with the starting point among the first K elements determined at random.

- How to choose K element. K is called sampling interval

$$K = \frac{Population\ Size}{Sample\ Size} = \frac{N}{n}$$

- Ex. If we want to select 50 persons form 500 persons under this method K^{th} sample is picked such as

$$K = \frac{500}{50} = 10$$

- We have to choose every 10 sample from the 500 persons.

- In systematic sampling first sample (r) is selected using simple random sampling method. Next samples are chosen using some systematic rule such as every “K”th sample have to choose such as

$r, r+k, r+2k, r+3k$

- Systematic sampling is preferably used when the information is to be collected from trees in a forest, house in blocks, entries in a register which are in a serial order etc

Sampling Error and Non-sampling Error

Sampling Error

- Sampling Error denotes a statistical error arising out of a certain sample selected being unrepresentative of the population of interest. In simple terms, it is an error which occurs when the sample selected does not contain the true characteristics, qualities, or figures of the whole population.
- Moreover, they can also arise out of defective sample design, faulty demarcation of units, wrong choice of statistic, substitution of sampling unit done by the enumerator for their convenience.

Sampling Error and Non-sampling Error

Non-sampling Error

- Non-Sampling Error is an umbrella term which comprises of all the errors, other than the sampling error. They arise due to a number of reasons, i.e. error in problem definition, questionnaire design, approach, coverage, information provided by respondents, data preparation, collection, tabulation, and analysis.

There are two types of non-sampling error:

- **Response Error**
- **Non-Response Error**

- **Response Error**: Error arising due to inaccurate answers was given by respondents, or their answer is misinterpreted or recorded wrongly. It consists of researcher error, respondent error and interviewer error which are further classified as under.
  - –Researcher Error
    - •Measurement Error
    - •Data Analysis Error
    - •Population Definition Error
  - –Respondent Error
    - •Inability Error
    - •Unwillingness Error
  - –Interviewer Error
    - •Questioning Error
    - •Recording Error
    - •Respondent Selection Error
    - •Cheating Error
- **Non-Response Error**: Error arising due to some respondents who are a part of the sample does not respond.

Key Differences between Sampling and Non-Sampling Error

1. Sampling error is a statistical error happens due to the sample selected does not perfectly represent the population of interest. Non-sampling error occurs due to sources other than sampling while conducting survey activities is known as non-sampling error.
2. Non-sampling error can be random or non-random whereas sampling error occurs in the random sample only.
3. Sample error arises only when the sample is taken as a representative of a population. As opposed to non-sampling error which arises both in sampling and complete enumeration.
4. Sampling error increases as the sample size decreases. On the contrary, the non-sampling error increase as sample size increases, it won't be reduced.

Sampling Distribution

- **Parameter**: It is a statistical constant of the population.
- **Statistic**: It is a function of sample value.
- **Sampling Distribution:** The distribution of values of a statistics for different samples of the same size is called Sampling Distribution of the statistics.
- **Standard Error:** S.E. of a statistic is the standard deviation of the sampling distribution of the statistics.

What is a 'Sampling Distribution‘?

- A sampling distribution is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population. The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.

Sampling Distribution of the Mean

Suppose we draw all possible samples of size *n* from a population of size *N*. Suppose further that we compute a mean score for each sample. In this way, we create a sampling distribution of the mean.

Then The mean of the sampling distribution (μ_x) is equal to the mean of the population (μ).

$$\mu_x = \mu$$

And the standard error of the sampling distribution (σ_x) is determined by the standard deviation of the population (σ), the population size (N), and the sample size (n). These relationships are shown in the equations below:

finite population correction or fpc

$$\sigma_x = [\ \sigma / sqrt(n)\ ] * sqrt[\ (N - n\ ) / (N - 1)\ ]$$

When the population size is very large relative to the sample size, the fpc is approximately equal to one; and the standard error formula can be approximated by:

$$\sigma_x = \sigma / sqrt(n)$$

Central Limit Theorem - CLT

- Sampling distribution of the sample means approaches a normal distribution as the sample size gets larger - no matter what the shape of the population distribution.

- It's a pretty useful phenomenon that can help accurately predict characteristics of a population.

THANK YOU