[Academic Script]

Sample and Surveys

Subject:	Business Economics
Course:	B. A. (Hons.), 1st Semester, Undergraduate
Paper No. & Title:	Paper – 102 Statistics For Business Economics
Unit No. & Title:	Unit – 5 Probability and Distribution
Lecture No. & Title:	Lecture – 4 Sample and Surveys

Academic Script

1. Introduction

Information can be collected by complete enumeration or by sample survey. In case of destructive type survey complete enumeration is not advisable because of heavy economic loss. Sample data can be obtained like simple random sampling, stratified random sampling, multistage sampling etc.

2. Complete Enumeration survey

The method of obtaining the required information at regional or national levels by collecting the data for each and every unit like: household, factory, agricultural farm etc. belonging to the population, which is the aggregate of all units of a given type under consideration is termed complete enumeration survey or census.

To obtain different types of data mentioned above by complete enumeration survey the time money and effort generally be extremely large. However, if the information is required for each and every unit of the population under study A complete enumeration survey is must. For example, preparation of voters, list for election purposes, the census of industrial enterprises in the urban areas of the country. Due to difficulties arises in organizing a survey on such a large scale, the errors in a complete enumeration survey may arise mainly due to incomplete coverage, observational and tabulation errors etc., but some margin of error is permissible in the data needed for practical purposes. The drawback of the complete enumeration survey is that it is not practicable in in the case of destructive

type survey. For example, to obtain the actual life of all the electronic components of a produced lot.

3. Sample survey

In case of destructive type surveys the observations on the life of components are collected for a part of the population or a produced lot to infer about the whole population. That is when inference about the whole population is drawn on the basis of the survey conducted for some selected units of the population, the units are selected by some suitable manner, inplace of complete enumeration survey, such alternative is known as sample survey. When trained investigators and large amount of resources in terms of fund needed for the conducting a survey is not available a complete enumeration survey is not at all possible. The inferences made from the sample survey are likely to be different from the population values. The difference would be due to the sample. Thus information obtained from the sample survey is subject to a kind of error, which we call sampling error. But due to part of the population is surveyed observational errors may be eliminated by employing proper trained investigators to conduct sample survey. For example to estimate total rice of production in the Gujarat state instead of conducting complete enumeration a sample survey would be more effective with respect to time and cost. Sample survey remains the only way when population contains infinitely many number of units.

Thus Complete enumeration or a sample surveys are considered on the basis of certain types of facilities available such as funds, trained investigators, supervisors, sufficient number of

investigators, a complete list of units of the population, transportation, etc.

4. Some basic concepts in sampling

Frame:

A list of all units of a population is known as frame.

Parameter:

Different types of summary measures that describe any given characteristics of the population are called parameter of the population which are usually denoted by Greek or capital letters. For example, population total is denoted by Y, population mean is denoted by μ, standard deviation by σ and variance by σ^2, population proportion by P.

Statistic:

A summary measure that describes the characteristic of the sample is known as statistic. The statistic is usually denoted by lower case roman letters. For example, sample mean is denoted by $\bar{x}$, sample standard deviation by s, sample proportion by p etc. The statistic is a random variable because it varies from sample to sample.

5. Sampling Theory

There are two basic types of sampling. Various methods of sampling can be grouped under these two types namely,

1) Probability sampling or random sampling
2) Non-probability sampling or non-random sampling

1) Probability sampling or random sampling

Under this method, each unit of the population has some nonzero probability of being included in the sample. It helps to estimate the mean, variance etc. of the population.

Under probability sampling there are two procedures

a) Sampling with replacement (SWR)

b) Sampling without replacement (SWOR)

When the successive draws are made with placing back the units selected in the Preceding draws, it is known as sampling with replacement. When such replacement is not made it is known as sampling without replacement.

Usually, when the population is finite sampling with replacement is adopted otherwise SWOR is adopted.

Many kinds of random sampling are used in the sample survey. Some of them are.

1. Simple random sampling
2. Stratified random sampling
3. Multistage sampling
4. Cluster sampling

1. Simple Random sampling (SRS)

This is the basic probability sampling method. This method is the simplest of all the probability sampling methods. It is used when the population is homogeneous, that is units of the population possess almost same types of characteristics. For example, a population of farms having same farm size, a population of students of a tenth standard class of a school, a population of blood of a human body, etc.

When the units of the sample are drawn independently with equal probabilities, the sampling method is known as Simple

Random Sampling (SRS). Thus if the population consists of N units, the probability of selecting any unit is 1/N.

A theoretical definition of SRS is as follows

Suppose we draw a sample of size n from a population of size N without replacement, There are $\binom{N}{n}$ = ${}_N C_n$ possible number of samples of size n. If all possible samples have an equal probability $1/{}_N C_n$ of being drawn, the sampling is said be simple random sampling.

In case of simple random sampling with replacement there are N^n possible samples of size n and probability of selecting any of such samples is $1/ N^n$.

Simple random sample can be drawn by lottery method or random number table method.

i) Lottery method

This is the most popular method and simplest method. In this method all the items of the universe are numbered on separate chits of paper of same size, shape and colour. They are folded and mixed up in a bowl or a container. A selection of one chit is made randomly from the bowl to represent one member of the sample. For a sample of size n we have to repeat this process n time.

For example, if we want to select 10 farms out of 100 farms, we number the 100 farms first. We write the numbers from 1-100 on chits of the same size, fold them and mixed up in a bowl. Then we make a blindfold selection of 10chits. As the population size increases, it becomes more and more difficult to work with chits, or if population size is infinite the this method is inapplicable. There is a lot of possibility of personal bias if the size , shape and colour of the chits are not identical.

ii) Random number table method

As the lottery method cannot be used when the population is infinite, the alternative method is the use of random number tables.

There are several tables of random numbers prepared by Tippet (1927), Fishers and Yates (1938), Kendall and Smith (1939) , Rand corporation (1955) etc., In the table of random numbers each digit (from 0 to 9) has equal probability to appear in any particular position. Suppose we have a population of size 1000 with each number being a serial number starting from 000 to 999 and we draw the first three digits from the top left hand number and move to right . For example the first three- digited number is say 230, second number is 094 etc. thus, write down 100 numbers from the table of random numbers, then the units in the population having these selected numbers would be the sample units. If the same three- digited number repeats in the random number table then we ignore it for a simple random sample under without replacement and if we allow the repeated random number , it gives a simple random sample under with replacement.

Let consider an example for simple random sampling with replacement and without replacement.

Consider a population having five units A, B, C, D, and E. The possible simple random samples without replacement of size two will be $\binom{5}{2} = 10$, viz:

AB, AC, AD, AE, BC, BD, BE, CD, CE, and DE.

The possible simple random samples with replacement of size two will be $5^2 = 25$, viz:

AA, AB, AC, AD, AE,

BA, BB, BC, BD, BE,

CA, CB, CC, CD, CE,

DA, DB, DC, DD, DE,

EA, EB, EC, ED, EE

Note that a complete list of all the members of the population i.e. frame is required before a simple random sampling is adopted. In some situation neither the frame nor is it practical to prepare the frame in time and cost effective manner, under such situations simple random sampling method is not a suitable method.

2. Stratified random Sampling:

When the population is not homogenous that is, it possess the different units with different characteristics, such population we will call heterogeneous population with respect to the characteristic in which we are interested. In such situation the population is divided into more homogeneous groups, such that each group becomes more homogenous within with respect to the study characteristic of our interest and heterogeneous between the groups. Such group is called stratum and all such groups are called strata. For example, total consumption of a family usually depends on the income of the family and the families can be divided into three categories like: high income, middle income and low-income families. The groups of families with such types of categories are known as strata.

For such heterogeneous population random samples are drawn from each strata of the population. The sample size for each strata are not necessarily same, but total units selected from all the strata must be equal to our required sample size say n. This sampling method is known as stratified random sampling.

There are different methods for deciding the sample size for each strata. The number of units to be selected may be uniform in all strata or may vary from stratum to stratum.

Usually the following methods of allocation of strata are used.

- Equal allocation
- Proportional allocation
- Neyman's fixed sample size allocation
- Fixed cost optimum allocation

If the number of units to be selected is uniform in all strata it is known as equal allocation of samples.

If the number of units to be selected from a stratum is proportional to the size of stratum, it is known as proportional allocation of samples. Let us consider the following example to understand the above methods.

Stratum number (h)	Stratum size (N_h)	Sample size under equal allocation (n_h)	Sample size under proportional allocation (n_h)
1	200	40	20
2	400	40	40
3	600	40	60
Total	1200	120	120

When the sample size of the survey is fixed but allocation is made in such a way that the variance of the estimator of the population characteristic becomes minimum, it is known as Neyman's allocation or optimum allocation.

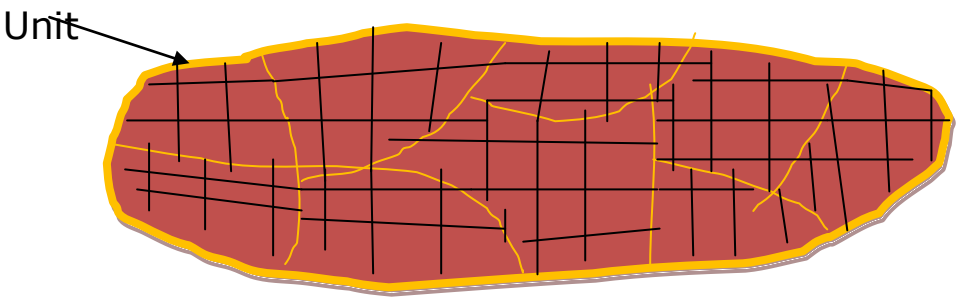
When the total cost of conducting survey is fixed and allocation is made in such a way that the variance of the estimator of the population characteristic becomes minimum, it is known as fixed cost optimum allocation.

This method provides more representative and statistical efficiency compared to simple random sampling. But, total cost of the study goes up due to the additional costs of stratification. If the proper stratification is not done, the sample will have an effect of bias and decreases the statistical efficiency.

3. Multi-Stage Sampling:

Sometimes the units of the population under study may not be a single element, but may be a group of small elements. For example, In a population of families, family is a unit having some family members within each unit, thus a unit of family is a group of family members. Sometimes the primary units are divisible in to secondary units and divisions of secondary units are possible into elements and so on. For example, Districts are divided into talukas; talukas are divided into small villages etc. For such population multi-stage sampling method is quit suitable.

Elements



To estimate the total production of wheat in Gujarat State, the state is divided into different districts. First of all by random sampling method a sample of district is selected. After selecting a sample of districts, a random sample of talukas is selected from each selected districts. Again from each selected talukas, a random sample of villages is drawn, and then production of wheat for selected villages would be recoded. Such method of selecting sample is known as multi stage sampling. At each stage not only simple random sampling method but also any other probability sampling methods may be applied.

4. Cluster sampling:

In multi stage sampling, after selecting primary units, a sample is drawn from all the selected primary units. But sometimes instead of selecting the sub sample from the selected primary units, all the elements of the primary units are considered in the sample, that is the complete selected primary units are used. For example, to know the education level in a given locality, First of all a random sample of primary units (Families) is drawn and all the members of the selected families are considered as sample elements for the study and no further selection is made. Here the family is called a cluster. Thus, a cluster means a group of elements. Suppose a population consist such N clusters then a random sample of n clusters is taken from the given population

and all the elements of the selected clusters will be the elements of the sample, such sample is called a cluster sample. And the method of selecting such a cluster sample is called cluster sampling.

Note that stratification is done to make the strata homogenous within and different from other strata where as in cluster sampling clusters should be heterogeneous within and the different clusters should be similar to each other. We consider cluster as mini population which possess almost all the characteristics of the population.

In stratified sampling there are fewer strata and we select a random sample from each stratum while in cluster sampling, there are many clusters and we select only a few clusters by random sampling, the selected clusters are completely enumerated.

Cluster sampling is very much cost effective with respect to simple random sampling although, it is statistically less efficient than simple random sampling in most of the cases. But this deficiency can be compensating in terms of total cost of the survey by adopting cluster sampling method.

6. Statistical measure and degrees of freedom

In statistics, the number of degrees of freedom (df) is the number of independent pieces of the data being used to make estimation of the population parameter. It is usually denoted with the greek letter ϑ (nu) or v.

The number of degrees of freedom is a measure of how certain we are that our sample data is representative of the entire population - the more degrees of freedom, usually the more certain we can be that we have accurately sampled the entire

population. For statistical analysis, this is usually the number of independent observations or measurements made in the experiment.

The degrees of freedom can be viewed as the number of independent observations available to fit a model to data. Generally, the more degrees of freedom you have, the more accurate your fit will be. However, for each estimate made in the fitting of the model, we remove one degree of freedom. This is because each assumption or approximation you make puts one more restriction on how many parameters are used to generate the model.

In calculating the variance of a sample data with n observations, the following formula is used:

$$S^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n - 1}$$

In this case, the degrees of freedom becomes v = n-1, because an estimate was made that the sample mean is a good estimate of the population mean, so we have one less degrees of freedom than the number of independent observations.

In many statistical calculations such as linear regression, chi – square test, t-tests, F- tests, etc, we calculate different degrees of freedom.

7. Summary

In this talk we have discussed when and how sample survey and complete enumeration is done to get the data. We have also covered how to select a sample from the population. Different

types of sampling methods are used to draw a random sample from the population. On basis of the nature of the population and the units in the population, various types of methods can be apply like: Simple random sampling, Stratified random sampling, multistage sampling and cluster sampling. Sample measures and degrees of freedom were also described.