

(A) Personal Details

Role	Name	Affiliation
Principal Investigator	Prof. Masood Ahsan Siddiqui	Jamia Millia Islamia, New Delhi
Paper Coordinator	Prof. Aslam Mahmood (Retd.)	Jawaharlal Nehru University, New Delhi
Content writer/Author (CW)	Prof. Aslam Mahmood (Retd)	Jawaharlal Nehru University, New Delhi
Content Reviewer (CR)	Prof. Aslam Mahmood (Retd)	Jawaharlal Nehru University, New Delhi
Language Editor (LE)		

(B) Description of Module

Items	Description of Module
Subject Name	Geography
Paper Name	Quantitative Techniques in Geography
Module Name	Sampling Methods
Module ID	QT 12
Pre- Requisites	Introduction to probability theory, permutation and combination are needed
Objectives	To understand the basic principle of sampling
Key words	Census enumeration, sampling distribution, standard error, random sample, stratification, multi stage sampling, probability and non-probability sampling

Sampling Methods

Aslam Mahmood

(1) E-Contents

Statistics being the empirical science is basically concern with the analysis of real world data, which is collected either from the Primary source or from the Secondary source. When the data is collected from the primary source a researcher has to make a choice of data collection, either through the method of Census enumeration or through the method of Sample surveys.

In Census enumeration all the units of observations which are parts of the study are enumerated. Group of observations in a Census is generally large and is known as the Universe of the study. Some time when the universe is large, Census enumeration may not be possible due to lack of resources, time or trained personal etc. In such cases the researcher can use method of Sampling in which only a small representative part of the Universe is studied known as Sample and with the help of the theory of sampling inferences are made about the Universe.

Effective generalizations about the Universe are possible only when the Sample is a good representative of the Universe. It is therefore most essential to ensure that when a sample is drawn from the Universe it should be drawn in such a manner that it represents the Universe in all the aspects. This can be achieved by selecting the suitable method for drawing the sample. There is no single method of drawing a representative sample from the Universe. There are several ways in which a representative sample can be drawn from a Universe depending on its structure. Methods of sampling which ensure best representation of the universe fall under two broad categories:

1. Probability Sampling Methods ; and
2. Non-probability Sampling Methods

In probability sampling methods samples are drawn using the principles of probability. In probability sampling, whether any observation will be selected in a sample will depend on chance. The method of drawing a sample is designed in such a manner that all the members of the universe are given a proper chance. A sample selected through probability sampling therefore is supposed to be an unbiased sample. Neither any observation is discriminated nor any observation is given any favour.

Probability Sampling Methods suitable for different situations are given as below:

1. Simple Random Sampling,
2. Systematic Sampling,
3. Stratified Random Sampling,
4. Cluster Sampling and
5. Multi-stage Sampling

Simple Random Sampling

This is the simplest method of drawing a sample in which each and every member of the Universe has equal chance of being selected in the Sample. It is like drawing a lottery ticket. It is the most suitable method of drawing a representative Sample when the observations of the universe are fairly Uniform or homogeneous.

Method of Simple random Sampling is very much facilitated by the use of Random Number Table prepared by Tippet, Kendall, Smith, Fisher and Yeats or by Rand Corporation etc. These numbers are given in tabular form in books on statistics. Any number on any page is either selected from any row or column or to constitute a random sample of numbers.

How to select a Random Sample

In practical exercise of sampling, first the members of the universe are arranged in a table and assigned a serial number. Second stage of drawing a simple random sample is to select a set of random numbers. Number of selected random numbers should be equal to the size of the sample to be selected. Next step is to consider the set of chosen random numbers as the serial number of the members of the Universe to be selected in the sample. If the total number of observations in the Universe are between 1 and 99, two adjoining columns have to be selected in the Random Number Tables. In case number of observations are more than 99 but less than 1000, three adjoining columns of the Random Number tables have to be selected, and so on. For example ,if we take a universe and number the units from 1 to 700, we have to select three consecutive columns of a random number table. In case the number of items to be selected are 35, we select a combination of first three value, then next three values and so on until we get 35 combinations of serial numbers between 001 and 700.

Now a days random number tables have been replaced by the computers and software are used to generate random numbers directly.

Systematic Sampling

A simple random sample is an ideal method of ensuring the unbiased selection of the sample observations. However, it requires sufficient preparations which takes considerable time specially when the universe is large and a larger random sample is to be selected. In most of the national surveys the researcher face this problem. Census of India also after collecting the information of the total population of India on Census slips, find it difficult to do tabulation on the basis of all the slips. Often, it is decided to do tabulation on the basis of 10 % or 20 % slips only and then after tabulation the figures are inflated to give data on 100 %.

This is a special kind of sampling in which the selection of the first unit of the sample is selected randomly. The remaining units are selected from the population at a fixed interval of n where n gives the percentage of the universe to be selected in the sample. For example if it is decided to select 20 %

sample after initial selection every fifth observation in the list or in the slips is selected so that in the end 20 % of total slips will get selected in the sample.

Even if the total number of sample units are decided we can use systematic sampling after dividing the range of serial number into appropriate size and selecting one unit after every fixed interval. For example, if the size of a universe is 800 and we want to draw a sample of size 40. In such a case the unit of the universe are divided into 40 intervals of size $800/40 = 20$. The units in the universe have serial number from 1 to 800 which is divided into 20 intervals of range from 1 – 20, 21 – 40, 41 – 60, 61 – 80, and so on. There will be 20 numbers of such intervals which is equal to the size of the sample to be drawn. First one number is selected randomly from serial number 1 – 20. Suppose it is 7, so the seventh item is selected from the first interval. Adding 20 to it will generate the second random number, thus second item to be selected will be 27. Third number will be 47 and so on.

The sample serial numbers will be:

007, 047, 067, 087, 107, 127, 147, 167, 187, 297

227, 247, 267, 287, 307, 327, 327, 347, 367, 387,

407, 427, 247, 467, 487, 507, 527, 547, 567, 587,

607, 627, 647, 667, 687, 707, 727, 747, 767, 787.

One of the advantage of this method is that : once the first unit of the sample is selected randomly all other sample units will be selected automatically. However, if there is some bias in selecting a serial number it will persist in all other selections also.

Stratified Random Sampling

A random sampling is good only when there are not marked differences in the population. Problem arises when the population is composed of highly varied observations. Suppose in a family planning survey of a district a sample of 1000 household are to be selected from a total universe of 25000 households living both in rural and urban areas. 20000 households live in rural areas (80 %) and 5000 live in urban areas (20 %). In a simple random sample there is no surety that the selected sample units will also represent this diversity of the universe.

However, if we divide the universe into two strata of rural households and urban households and then selecting 80 % households from rural areas and 20 % of them from urban areas the selected sample will be not only random but also represent the rural urban diversity in the same proportion.

A stratified random sampling is therefore a method of sampling used when the universe of the study consists of diverse units of observations. In order to get the diversity reflected in the sample also, first the universe is divided into homogeneous group of observations known as strata and then from each stratum draw a random sample of appropriate size in proportion to the share of the stratum in the

universe. Such a stratified random sample will be random as well as be representing the diversity in the Universe also.

Proportionate and Disproportionate Stratified random Sampling.

In a stratified random sampling the samples can be drawn either proportional to the size of each strata or it could be disproportional to its share in the universe. When the sampling units are all homogeneous a sample of size proportional to the size of each strata will be drawn. However, when strata are quite heterogeneous, the size of the sample has to be disproportional. The stratum having higher variability will require greater size to represent the strata. On the other hand the strata having uniform values with low variability can be represented by even a smaller size of the sample also.

Cluster Sampling

One of the problem with Simple random sampling and with systematic and stratified random sampling is that all the items of the Universe have to be listed with a serial number and the selected sample numbers are to be located in the universe for data collection. This involves quite a good running around specially when the universe covers sufficient geographic space. In some situation when observations of a study are composed of similar groups scattered over space, it is sufficient to study one or two instead of all of them. These similar groups of observations are known as clusters and it is assumed that observations between each cluster are similar and within each cluster these are heterogeneous. Each cluster within itself is like a universe. In a cluster sampling therefore few clusters are selected and studied in totality. All the shades of the universe are represented in each cluster so a cluster is the best representative of the universe.

Cities are suppose to have clusters of slums. Each slum has considerable variations of the quality of life within itself. These variations will get repeated when we move from one slum to another slum. It is therefore, suggested not to go from one slum to another slum and increase the physical burden of sample survey. On the contrary, it is advised to select one cluster (Slum) on sample basis and study it in full. Cluster sampling saves time and resources, at the same time is as effective as any other random or stratified random sample could be.

Multi – stage Sampling

In a large scale survey covering a country, the sampling frame will be large and selecting a small sample of few hundred or thousand villages out of six or seven lakh total villages of India will be difficult as large sampling frame will require more time and cost. Multi-stage sampling technique can help in designing a procedure of sampling which will start with a big sampling frame and will keep on becoming smaller and smaller as we move from one stage of sample selection to another stage. Multi stage sampling will make the study more practicable in terms of cost and time as well as ensure the unbiased selection of the sample units through the properties of the theory of probability.

It involves more than one stage of sampling the population which will depending of the problem. Consider a problem of selecting a sample of 1000 villages from 600000 villages of India. A stratified

random sample will not be sufficient to represent all variations of the village life of India. We have to have many layers of stratification.

Multi stage sampling suggest to divide the country into similar regions and select a random sample of one or two states from each region. After that each selected state is further sub-divided into districts and a set of representative district is selected randomly at the second stage. In the third stage from the selected districts villages are select randomly. There also we can go to the fourth stage of selecting the sample households from each selected village and so on.

Non-probability Sampling

Sometime due to study of a specific nature the choice of selecting a sample a random sample may not give the required results. Researcher may not have any idea about the location of the sample units which are not easily traceable also. The researcher has to depend on their availability, there is no sampling framework possible in such cases. In such cases the researcher has to consider non probability methods of sampling. Their first concern is the convenience of the availability of the sample units which is known as "**Convenience sampling**".

Sometime only few specific group of the respondent have to be studied as only they serve the purpose of the study. For example among the household of an area only large size households are required to study the factors behind their choice of the size of the family etc. Such a situation to serve a particular purpose may force the researcher to go for "**Purposive sampling**" only. In some of the cases of a stratified random sampling kind of situation, researcher may be interested to give numerical representation to each group by fixing some quota to each group but may not be particular about their unbiased representation of their group. Only their inclusion may be sufficient in form of some quota of each group may be enough. This kind of sampling is known as "**Quota sampling**". There is yet another kind of non-probability sampling known as "snowball" sampling. Here one has very limited information about the availability of the respondents for the sample. When somehow one or two respondents are located, with the help of them researcher may come across few more respondents. Now he has a bigger group which can also provide some more references to increase the list of respondents further. Such method of sampling is known as "**Snowball sampling**".

All such non-probability sample methods are listed below:

1. Convenience sampling.
2. Judgement Sampling or purposive.
3. Quota sampling and
4. Snowball Sampling.

Sampling Error and determining the size of the sample

Before a sampling exercise is undertaken the size of the sample is a big question before a researcher. If the sample size is too small it will not achieve the purpose and if it is too large it will require huge cost

and time that will amount to waste of resources. Therefore as a general rule it should neither be very large nor very small. Following factors will be considered while deciding about the size of a sample:

1. **NATURE OF THE UNIVERSE** A homogeneous universe will require smaller size of the sample than a heterogeneous Universe.
2. **Number of classes Proposed** In a situation where large number of classes are subgroups are formed a larger sample will be required, because a small sample will not give good number of cases in each class.
3. **In depth/ continuous study groups.** Such samples will require detailed technical observations over time which is possible only in cases of small samples.
4. **Quality of Sampling.** A small sample properly selected is better than a large sample poorly collected.
5. **Requirement of accuracy and accepted level of confidence.** A more accurate and precise estimate will require a higher size of the sample.
6. **Availability of finances.** Cost of the sample survey will increase with the size of the sample.
7. **Other considerations.** Nature of units, size of the universe, size of questionnaire, availability of trained investigators, conditions under which a sample is conducted, time availability etc are few other considerations which will determine the size of the sample.

Sampling Error and Size of the Sample (Confidence Level Approach)

Sample size can also be determined on the basis of the theory of sampling as given below:

If a large number of samples of size n are drawn from a universe with mean M and Standard Deviation as S , different values of the sample means will form another distribution known as sampling distribution of means. This distribution will be Normal with mean equal to the universe mean M and Standard Deviation as S/\sqrt{n} . Thus if we get a sample mean as \bar{X} from the above universe the interval within which the population mean will lie can be estimated. Such an interval is known as confidence interval with a probability.

The 99 % confidence interval in the above case will be :

$$\bar{X} - 2.56 S/\sqrt{n} \text{ to } \bar{X} + 2.56 S/\sqrt{n}. \text{ (the range in this case will be } = 5.12 S/\sqrt{n})$$

Similarly 95 % confidence limits will be:

$$\bar{X} - 1.96 S/\sqrt{n} \text{ to } \bar{X} + 1.96 S/\sqrt{n}. \text{ (the range in this case will be } = 3.92 S/\sqrt{n})$$

Sample size can be worked out using the above confidence intervals also.

If the range is fixed for 99% CL as R , the size of the sample will be :

$$\sqrt{n} = 5.12 S / R \text{ or}$$
$$n = (5.12 S / R)^2$$

Similarly for 95% CL we can get = $n = (3.92 S / R)^2$

We can use the above equation to determine the size of the sample provided the value of R and level of confidence is determined by the researcher. The researcher has the liberty to decide about them, but still he has to know the value of S, the standard deviation of the size of the landholdings. This is a very tricky question for the researcher. If he or she can know standard deviation they can know the value of the mean also. The answer is that for the value of the required standard deviation any rough estimate of it can also be used.

Example

For a sample survey of estimating the average size of the land holdings of the farmers of a region, what should be the size of the sample so that the lower limit and the upper limit of the estimate does not differ by 3.0 acres. (given the rough estimate of the standard deviation of the size of the land holdings = 10 acres) .

For 99 % confidence level the estimated size of the land holdings the size of the sample will be :

$$n = (5.12 S / R)^2$$
$$= (5.12 \times 10 / 3.0)^2 = (51.2/3.0)^2 = 17.1 \times 17.1 = 292 \text{ after rounding.}$$

For 95 % confidence level the estimated size of the land holdings the size of the sample will be :

$$n = (3.92 S / R)^2$$
$$= (3.92 \times 10 / 3.0)^2 = (39.2/3.0)^2 = 13.07 \times 13.07 = 170 \text{ after rounding.}$$