**Chapter 3: Elements of Sampling**

It is usually not possible to undertake a census or a complete enumeration of all items in the population particularly when the population is large. So one resort to a sample survey is to generate the data required for the investigation. A sample survey saves money and time and effort. Sampling aims at obtaining consistent and unbiased estimates of the population parameters. The interest in a sample survey is not just the characteristics of the sample, but also the characteristics of the population from which the sample has been drawn. So sampling has the advantage of simplicity, cost reduction and timeliness.

**Sampling Concepts and Terminologies**

Researchers and statisticians have developed specialized terminologies to deal with the various sampling issues. Some of these terminologies are discussed next.

Population to be sampled: the first step in the sampling process is to define the population of interest accurately. This is the theoretically specified aggregation of the survey elements. It may refer to a group of people or to non-human entitles such as objects, institutions, time, geographical area, price, salaries, etc.

Survey population: a survey population is that aggregation of elements from which the survey sample is actually selected.

Sampling elements: this is the unit of analysis or case in the population. It is that unit about which information is collected and which provides the basis for the analysis. It can be a person, a group or an organization that is being measured. This is the subject on which measurement is being taken. Example: people, families, corporations, etc.

Sampling Unit: for the purpose of sampling a population can be thought of as consisting of sampling units. A sampling unit is that element or set of elements considered for selection in some stage of sampling. In simple single stage sampling, the sampling units are the same as the elements. If the sampling process involves many stages then we may have primary sampling units, secondary sampling units and final sampling units to designate the successive stages.

Example: if all jobless in Addis Ababa city represent the population then the 28 woredas and the 256 kebeles may be the sampling units.

Sampling Frame: A sampling frame is the actual list of sampling units from which the sample or some stage of the sample is selected. In a single stage sampling design the sampling frame is a list of the elements comprising the survey population.

Example: if a simple sample of students were selected from a student roster, then the roster would be sampling frame. Telephone directories, tax records, and PA household list are ether examples.

Sample Ratio: is the size of the sample divided by the size of the target population. Example: if the population consists of 50,000 elements, and a sample of 150 is draw then the sampling ratio would be 150/50,000=0.003 or 0.3%.

**Reasons for Sampling**

The basic idea of sampling is that some of the elements in a population provide useful information on the entire population. The economic advantages of taking a sample rather than a census are massive.

1. Resource limitations compared to a census a sample survey requires less time and financial resources. In other words it is less expensive. Sampling can save time and money. A sample study is usually less expensive than a census and produces results at a restively faster period.

Why should we interview 5000 individuals if one is able to get the same information by interviewing some of them only? Thus because of resource limitations it is difficult to conduct a census and the fore resort to sampling.

2. Superior quality of Results: the quality of a study is often better with sampling than with a census. Sampling may enable more accurate measurement for a sample study is generally conducted by trained and experienced investigations. Several reasons could be mentioned as to why this is so.

a) A census would require a very large staff for the interview and hence one can not limit himself only to the best available enumerators. In a census everyone in sight is employed with the result that the overall quality of the interview is reduced.

b) Interviewing every member of the population also require lengthy interviewing period. As a result it would be different if not impossible to specify the time to which the data refer so sampling produces quicker results than does a census.

 c) The managerial requiring of a very large population would be far greater than normally faced by surveys. Supervision record, keeping, training and so forth would be more difficult in a census. So the quality of the data obtained from a large survey would be lower than that obtained from a smaller more manageable survey. Better interviewing, more through investigation of missing wrong and suspicious information, better supervisions, better data processing, is possible from a sample survey.

3) Infinite population: sampling is also the only process possible if the population is infinite.

4) Destructive nature of some tests: to test the quality of some products or items it may be necessary to consume it or destroy it. Under such circumstances a census would mean consuming everything or destroying it. Sampling remains the only choice when a test involves the destruction of the items under study.

Example: testing the quality of commodity (beer, cigarette, coffee, etc.)

5. Information on sampling errors: sampling usually enables to estimate the sampling errors and thus assists in obtaining information concerning some characteristics of the population.

**Steps in Sampling Design (Critical Questions in Sampling Design)**

Five decisions must be made in designing a sample. While they are presented here sequentially their order is not fixed. The critical sampling questions are:

* Identifying the relevant population: when one wants to undertake a sample survey the relevant population from which the sample is going to be drawn need to be identified. If population is not properly defined a researcher does not know what issues to consider when selecting the sample. So good operational definition is critical for any study.

 Example: if the study concerns income, then the definition of the population elements as individuals or household can make a difference.

* Determining the method of sampling: what methods should be used to select a sample is also very important. Whether a probability sampling procedure or a non-probability sampling procedure has to be used is also very important.
* Secure a sampling frame: the concept of a sampling frame is closely related to the population. It is defined as the list of elements from which the sample is actually drawn. A list is important and necessary.
* Identifying parameters of interest: in designing a study it is important to determine what specific population that may be of interest. These characteristics may be divided into variables and attributes.
* Determining the sample size: once the parameters of interest are identified and the population is clearly defined then the next task of the researcher would be to determine the sample size for the study.

**How large a sample should be?**

One way to determine the size of a probability sample is to follow what others have done. Generally how large a sample should be a function of the variation in the population parameters under study, and the estimating precision needs by the researcher? A sample of 400 may be appropriate for one case while a sample of 2000 may be required for others. The question is how to determine the sample size.

Degree of homogeneity: The simple most important factor in determining the size of the sample need for estimating a parameter is the size of the population variance. The greater the dispersion or variance in the population the larger the sample must be to provide a given estimation precession.

Degree of confidence/confidence required: since a sample can never reflect in population for certain, the researcher must determine how much precision they need precision is measured in terms of:

1. An interval range in which they would expect to find the parameter estimate.
2. The degree of confidence they wish to have in the estimate.

Number of sub groups to be studied: The researcher may be interested in making estimates concerning various sub groups of the population then the sample must be large enough for each of these subgroups to meet the desired quality level. One achieves this by making the total sample large enough to assure that each critical subgroup meets the minimum size criterion.

Cost: cost considerations have a major impact on decision about the size and type of sample, as well as the data collection methods. All studies have some budgetary constraint and hence cost dictates the size of the sample.

**Characteristics of a good sample**

The logic of the theory of sampling is the logic of induction that is one proceeds from particular (i.e. sample) to the general (i.e. population) and all results are expressed in terms of probability. If all members of the population were identical to one another in all respects, there would be no need for careful sampling procedures. Under such extreme cases any sample would indeed be sufficient. In this extreme case of homogeneity one case would have sufficient as a sample to study. But when faced with variation or heterogeneity in the population under study, more controlled sampling procedures are required. For example, social groups differed in many respects, people differed in many ways. So a good sample must be, as representative of the entire population as possible and ideally it must provide the whole of the information about the population from which the sample has been drawn. So the ultimate test of a sample design is how well it represents the characteristics of the population it purports to represent. In measurement terms the sample must be valid. Validity of the sample depends on two factors.

*1. Accuracy*

Accuracy is defined as the degree to which bias is absent from the sample. An accurate (unbiased) sample is one which the underestimations and overestimations are balanced among the members of the sample. There is no variance with an accurate sample.

2*. Precision*

A second criterion of a good sample design is precision of estimates. No sample will fully represent its population in all respects. A sample statistics may be expected to differ from its parameters as a result of random fluctuations inherent in the sampling process. This is referred to as the error variance or sampling error and reflects the influence of chance in drawing the sample members. Precision is measured by the standard errors of estimates a type of standard deviation measurement. The smaller the standard error of estimates the higher is the precision of the sample. So it is desirable that a sample design produces minimum standard error of estimates.

**Probability and non- probability Sampling**

There are a variety of sampling methods that could be used to draw a sample. Which one is selected depends on the requirements of the project, its objectives, and the funds available. The different methods may be classified by their representation basis and the elements selection techniques. On the basis of representation the members of the sample can be selected either on a probability basis or by non-probability means.

Probability sampling is based on the concept of random selection- a controlled selection procedure that assures that each population element is given a known non-zero chance of selection. In contrast non-probability selection is non random i.e., each member does not have a known non-zero chance of being included.

* **Probability Sampling**

In probability sampling researchers use a randomization process in order to reduce or eliminate sampling bias. Under such circumstances one would have substantial confidence that the sample is representative of the population from which it is drawn. So, all probability samples must provide a known chance of selection for each population element. Probability sampling provides an efficient method for selecting a sample that should adequately reflect the variation that exists in the population. A basic principle of sampling is that a sample will be representative of the population from which it is drawn if all members of the population have an equal chance of being included in the sample.

Probability samples, while never perfectly representative are typically more representative than any other type of sample because the biases are avoided. Secondly, probability theory permits the researchers to estimate the accuracy or representativeness of his sample.

Probability sampling has considerable advantages over all other forms of sampling. First, sampling errors can be calculated for probability samples.

Example: for a sample of 100 children the average height was 1.2 meters. If the average height for all the children in the population from which the sample was drawn was 1.1 meters, then the error attributable to drawing the sample rather than measuring the whole population could be 0.1 meters. However, the population parameter is usually unknown in practice.

Second, probability samples rely on random process, i.e. the selection process operates in a truly random method (no pattern) and the researcher can calculate the probability of outcome. One can calculate the confidence interval within which the true population parameter lies.

Finally, in a true random process, each element has an equal chance or probability of being selected. It is possible to get consistent and unbiased estimate of the population parameter.

***Types of Probability Sampling***

There are variety of sampling techniques available which one is selected depends on the requirements of the project, its objectives and funds available. Generally speaking we could distinguish between five types of sampling designs.

*1****. Sample Random Sampling (SRS****)*

 The SRS is the simples and easiest method of probability sampling. It is sampling procedure in which each element has an equal chance of being selected into the sample. It assumes that an accurate sampling frame is developed in which all the population elements are listed.

Usually two methods are adopted to pick a sample (a) write the name of each person on a piece of paper, put them into a hat and mix them up, and draw a sample one by one (b) number all individuals in the population- assign one and only one number to each and not skipping any number in the process. Then use a list of random numbers or a table of random numbers to decide which individuals to select the sample.

Most probability samples do not however, use SRS in practice for some reasons.

It is too expensive to interview a national face to face sample based on SRS. The cost of interviewing randomly selected individuals dawn from a list of the entire population is extremely high.

It requires listing of the entire population of interest. This is impossible for a national survey in Ethiopia. So SRS can only be applied in situations where the population size is small.

***2. Systematic Sampling Technique (SST)***

This is also one of the most widely used probability sampling. If every element in the population is sampled beginning with a random start with an element from 1 to k is called systematic sampling. Under systematic sampling procedures, instead of a list of random number the researcher calculates a sampling interval. The sampling interval tells the researcher how to select elements from the sampling frame. When a frame is available or when a population can be properly listed and finite a systematic selection would be helpful. The major advantages of SST are its simplicity and flexibility.

In order to select a systematic sample, there is no need to number the cards. Merely determine the total number of cases in the population, the sampling interval to use, and the random start, and then begin drawing the sample by choosing every kth card or element. The sampling interval is the standard distance between elements selected in the sample. The sampling ratio is the proportion of elements in the population, which are selected.

Example: to select a sample of 300 from a population of 900 every third element have to be selected. So the sampling interval is 3.

It the population consists of 100 items and if we want to draw a random sample of 20 item then even fifth item would give use the required sample.

To insure against any possible human bias in using this method, the researcher selects the first element at random. Thus in the above example he would begin by selecting a random number between 1 and 3 and between 1 and 5 respectively.

Systematic sampling is slightly more accurate than SRS though the two seems to be identical.

Some of the problems associated with the SST are (a) possible periodically and (b) monotonic trends

***3. Stratified Sampling***

Most populations can be segregated into a number of mutually exclusive sub populations or strata. The stratified sampling technique is particularly useful when heterogeneous populations have. After a population is divided into the appropriate strata, a simple random sample can be taken from each stratum.

The researcher divided the population into sub population (strata) on the basis of supplementary information. After dividing the population into strata, the researcher draws a random sample from each sub population either using the SRS or the SST. The researcher controls the relative size of each stratum, letting the random process control it.

*The reasons for stratifying:*

There are three major reasons why a researcher chooses a stratified random sampling

1. To increase a sample’s statistical efficiency.

Stratification is almost always more efficient statistically than simple random sampling and at worst equal to it. With stratification each stratum is homogenous internally and heterogeneous with other strata.

1. To provide adequate data for analyzing the various sub population. Stratification is also useful when the researcher wants to study the characteristics of certain population sub groups.
2. To enable different research methods and procedures to be used in different strata. Stratification is also useful when different methods of sampling or data collection are called for in different parts of the population.

*How to Stratify*

Three major decisions must be made in order to stratify the given population into some mutually exclusive groups.

**What stratification base to use**: stratification would be based on the principal variable under study such as income, age, education, sex, location, religion, etc. if there are several variables about which we want to draw conclusions one should seek some basis for stratification that correlates well with the major variables.

**How many strata to use**: there is no precise answer as to how many strata to use. The more strata the closer one would be to come to maximizing inter-strata differences and minimizing intra- strata variables.

**What strata sample size to draw**: with regard to the strata sample size one condition how large the total sample should be and how the total sample shall be allocated among the strata one could use different approaches. One could adopt a proportional to the total number of units of the strata then we have proportionate sampling. Under proportionate sampling procedure each stratum is properly represented if the sample drawn from it is in proportion to the stratum’s share of the total population.

The procedure is to use disproportionate sampling which allocates elements on the basis of some bias.

***4. Cluster Sampling***

If the total area of interest happens to be a big one, a convenient way in which a sample can be taken is to divide the area into a number of smaller non-overlapping areas and then to randomly select al number of these areas. This is method of sampling, which entails the division of elements of a population into groups so that the units sampled contain more than one individual of the group is called cluster sampling. So, if the total area can be divided into groups (clusters) of elements and some of the groups (clusters) are selected randomly we have cluster sampling. Total population is divided into a number of relatively small subdivisions which are they clusters of still smaller units. In the case of stratified sampling the population is divided into subgroups and elements are selected from each group randomly. In the case of cluster sampling a number of the subgroups are selected for the study. Cluster sampling reduces cost by concentrating surveys in selected clusters. In that respect it may less precise than random sampling. Cluster sampling is a variation of the SRS techniques and is mostly used when the population is infinite.

*Reasons for fostering the use of cluster sampling*

Cluster sampling is used when it is either impossible or impractical to compile an exhaustive list of the elements comprising the target population. Cluster sampling is used because of the economic advantages it posses. Two main reasons are mentioned as justifications for using cluster sampling.

1. The need for more economic efficiency than can be provided by simple random sampling is one of the reasons.
2. The frequent un-availability of a practical sampling frame also necessitates the use of cluster sampling.
3. When a population is infinite we use cluster sampling.
4. When the geographical distribution of the units is highly scattered.
5. When sampling of the individual units is not convenient for several administrative reasons.

While statistical efficiency for cluster sampling is usually lower than for simple random samplings, chiefly because clusters tend to be homogenous, economic efficiency is often great enough to overcome this weakness. The criterion is the net relative efficiency resulting from the tradeoffs between economic and statistical factors.

*Multistage area sampling*

In cluster sampling researchers often prefer multistage cluster (area) sampling to single stage cluster (area) sampling. Multistage Area Sampling (MAS) is a cluster sampling with several stages;

First take a sample of a set of geographical regions or clusters--randomly select X number of clusters/regions.

Next a subset of geographical is sampled within each of those regions and so on.

 Finally a sample of elements is drawn from smaller areas.

Example: suppose a researcher wants a sample of households from a given city.

First districts are randomly selected

Those follows by a random selection of blocks or kebeles on the second stage and finally households within each block are randomly selected for the interview.

Note although there may not be an accurate list of residents in the city, there is an accurate list of districts and blocks or kebeles in the city .After selecting the districts the researchers lists all households in the selected districts to create a sample frame for each of the blocks or kebels. He or she uses the list to draw a random sample of households.

Accuracy declines in cluster sampling. First people who live in the same area tend to have similar characteristics and thus taking interview in the same area yields less information than would be gained by spreading the same number of interviews across a wider area. Second, each stage in MAS introduces sampling errors. A multistage sampling has more sampling errors than a one-stage sample. While statistical efficiency for cluster sampling is usually lower than for simple random samplings, chiefly because clusters tend to be homogenous, economic efficiency is often great enough to overcome this weakness. The procedure reduces costs. The criterion is the net relative efficiency resulting from the tradeoffs between economic and statistical factors.

***5. Hybrid sampling***

Where there is no single way to sample a particular population some researchers use multiage frame designs or a combination of the four different methods discussed above.

* **Non-Probability Sampling**

Despite accepted superiority of probability sampling methods in survey researcher non-probability methods are sometimes used instead for situations in which probability sampling would be prohibitively expensive and when precise representation is not necessary. Generally three conditions need to be met in order to use non-probability sampling.

First, one may use probability sampling because, such a procedure satisfactorily meets the sampling objectives. If there is no desire to generalize to a population parameter, then there is much less concern about whether is not the sample fully reflects the population.

Secondly, it is used because of cost and time requirements. Probability sampling clearly calls for a more planning and repeated callbacks to assure that each selected sample number is contacted.

Thirdly, though probability sampling may be superior in theory there are breakdowns in its applications. The ideal probability sampling may be only partially achieved because of the human element. The total population may not be available for the study in certain.

Four, different non probability methods could be identified

**Haphazard or Convenience Sampling**

The method selects anyone who is convenient. It can produce ineffective, highly unrepresentative samples and is not recommended. Such samples are cheap. However, the bias and systematic errors that easily occur make them worse than no sample at all.

Example: the person on the street interview conducted by television program is an example of a haphazard sample.

**Quota Sampling**

Quotas are assigned to different strata group. Interviewers are given quotas to be filled from different strata. The actual selection of the items is left to the interviewer’s discretion. Quota sampling is an improvement over the haphazard sampling, but it too is a weak type of sampling. In quota sampling a research first identifies categories of people (male, female) then decides how many to get in each category. Thus the number of people in various categories of the sample is fixed. The researcher interviews the first 5 male that he or she encounters. Similarly, the first 5 females encountered are interviewed.

The idea that quotas on some variables assure representativeness is an argument by analogy. It gives no real assurance that the sample is a representative on the variable being studied. Not only is misrepresentation is possible because haphazard sampling is used within the categories, but nothing prevents the researcher from selecting people who act friendly or who want to be interviewed. As there is no element of randomization, the extent of sampling error cannot be estimated.

In spite of its limitations, quota sampling is used in opinion pollsters, marketing sampling and other similar researcher areas.

**Purposive or Judgment sampling**

When one draws a non-probability sample that conforms to a certain criteria, it is called purposive sampling. It occurs when one picks sample members to conform to some criteria. It uses the judgment of the expert in selecting cases or it selects cases with specific purpose in mind. The researcher does not know whether the case selected represents the population. It is mainly used in exploratory research.

Example: In the study of standard of living; the cost of electricity, refrigerator, video reorder, stateliest dish, etc, can not be included for all people in Ethiopia.

**Snowball Sampling**

Also called network, chain referral, or computational sampling is a method for identifying and sampling (or selecting) the cases in a network. Snowball sampling is based on an analogy to a snowball, which begins small but becomes larger as it is rolled on wet snow and pick up additional snow. Snowball sampling begins with one or a few people or cases and spread out on the basis of links to the initial case.

The network could be scientists around the world investigating the same problem, the cities of a medium size city , the members of an organized family, persons who sit on the boards of directors of major corporations, etc each person is linked on connected with the other directly or indirectly.

*Problems in Sampling*

Errors are a word with a special meaning in sampling theory. It is not synonymous with mistake. However, a mistake by an interviewer or a wrong answer to a question would each contribute to error in a survey whether sample or a census. Field work problems, interviewer induced bias, clerical problems in managing amounts of data, etc would also contribute to error I a surrey, irrespective of whether a sample is drawn or a census is taken. Biases or errors due to such reasons or sources are known as non sampling errors. On the other hand, error which is attributable to sampling, and which therefore, is not present in census gathered information is called sampling error.

Since a sample has both kinds of errors where a census only has the former, you might conclude that the advantage really does rest with the census. However, the scale of census taking makes it difficult to reduce the risk of non-sampling error. Many sources of errors such as management problems, faulty measurement, lost or corrupted data, will be easier to control in a tightly constructed and managed survey than in a full census.

Moreover sampling error can be controlled (at least the extent of it can be estimated) with sample survey. Thus there are occasions when a sample survey could produce less error overall than a full census.

A). *Non-Sampling Error*

 Non sampling error refers to:-

* Non-coverage error
* Wrong population is being sampled
* Non response error
* Instrument error
* Interview error

*Non-Coverage Sampling Error*

This refers to sample frame defect. Omission of part of the target population for instance, soldiers, students living on campus, people in hospitals, prisoners, etc, are typically excluded from the national surveys. These omissions are unlikely to affect the national survey by more than 1 percent. But non-coverage errors could be serious say in telephone surveys since households without telephone are not included in the sampling. Non-coverage error also occurs when the list used for the sampling are incomplete or are outdated.

*The wrong population is sampled*

Researchers must always be sure that the group being sampled is drawn from the population they want to generalize about or the intended population. e.g. drawing a sample of college students to generalize about all college age persons.

*The response rate is low or non response error*

Some people refuse to be interviewed because they are ill, are too busy, or simply do not trust the interviewer. One should use many kinds of arguments to reduce the incidence of non-response errors. Non-response error cannot be a problem if researchers can show that people who refuse to respond do not differ much from those who do respond. The higher the refusal rate, the more important it is to ascertain whether the refusals are concentrated among a certain group. The demographic data can be compared with the census data to determine how representative the sample is and the data can be reweighed if need be. Usually non-response errors are ignored although they need more attention.

*Instrument error*

The word instrument in sampling survey means the device in which we collect data usually a questionnaire filled out b the respondents. Different wordings of a question is badly asked of worded, the resulting error is called instrument error. E.g. leading questions or carelessly worded questions may be misinterpreted by some researchers.

*Interviewer error*

This occurs when some characteristics of the interviewer such as age, sex, affects the way in which the respond answer questions. E.g. questions about racial discriminations might be differently answered depending on the racial group of the interviewer.

To sum up researcher must have to ensure that non sampling error are avoided as far as possible, or is evenly balanced (non systematic) and thus cancels out in the calculation of the population estimates, or else brought under statistical control. For example, in large survey the error related to small differences in technique on the randomly spread out data and will cancel out. Many of the non-sampling errors can be controlled by proper randomization in sampling; any residual effect should be small and would be lost in the estimates of the standard errors, possibly balancing with other small residual effects.

b) Sampling Errors

Sampling errors are random variations in the sample estimates around then true population parameters. Sampling errors can be calculated only for probability samples. Random sampling allows unbiased estimates of sampling error. The difference between sampling and non-sampling error is the extent of the former can be estimated from the sample variations, whereas the latter cannot. The measurement of sampling error is usually called the precision of the sampling plan.

Sampling error is related to confidence intervals. The confidence interval for the true population mean is given by:

Mean =

Means the sample mean, z is the value of the standard deviation at a given confidence level (to be reads from the table giving the area under the normal curve) n is the sample size, and o is the standard deviation of the sample mean. The sampling error is give by

Z=

If two samples are identical except that one is larger, the one with more cases will have a smaller error, i.e. a narrower confidence interval. Likewise if two samples are identical expect that the cases in one or more are similar to each other, the one with greater homogeneity will have a smaller sampling error i.e., narrower confidence interval. A narrower confidence interval means more precise estimates of the population for a given level of confidence. In general increasing the sample size is one of the major instruments to reduce the extent of the sampling error, a part from the sampling technique employed (Cluster vs.SRS).