LECTURE 11 SAMPLING METHODS

DR. GAURAV SANKALP

PROBABILITY AND NON-PROBABILITY SAMPLING

A probability sampling is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Example: We want to estimate the total income of adults living in a given street. We visit each household in that street, identify all adults living there, and randomly select one adult from each household. (For example, we can allocate each person a random number, generated from a uniform distribution between 0 and 1, and select the person with the highest number in each household). We then interview the selected person and find their income. People living on their own are certain to be selected, so we simply add their income to our estimate of the total. But a person living in a household of two adults has only a one-in-two chance of selection. To reflect this, when we come to such a household, we would count the selected person's income twice towards the total. (The person who is selected from that household can be loosely viewed as also representing the person who isn't selected.) In the above example, not everybody has the same probability of selection; what makes it a probability sample is the fact that each person's probability is known. When every element in the population *does* have the same probability of selection, this is known as an 'equal probability of selection' (EPS) design. Such designs are also referred to as 'self-weighting' because all sampled units are given the same weight.

Probability sampling includes: Simple Random Sampling, Systematic Sampling, and Stratified Sampling, Probability Proportional to Size Sampling, and Cluster or Multistage Sampling. These various ways of probability sampling have two things in common:

1. Every element has a known nonzero probability of being sampled and

2. Involves random selection at some point.

Non-probability sampling is any sampling method where some elements of the population have *no* chance of selection (these are sometimes referred to as 'out of coverage'/'under covered'), or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is nonrandom, non-probability sampling does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population.

Example: We visit every household in a given street, and interview the first person to answer the door. In any household with more than one occupant, this is a non-probability sample, because some people are more likely to answer the door (e.g. an unemployed person who spends most of their time at home is more likely to answer than an employed housemate who might be at work when the interviewer calls) and it's not practical to calculate these probabilities.

Non-probability sampling methods include accidental sampling, quota sampling and purposive sampling. In addition, non-response effects may turn *any* probability design into a non-probability design if the characteristics of non-response are not well understood, since non response effectively modifies each element's probability of being sampled.

SAMPLING TECHNIQUES

Within any of the types of frame identified above, a variety of sampling methods can be employed, individually or in combination. Factors commonly influencing the choice between these designs include:

- Nature and quality of the frame
- Availability of auxiliary information about units on the frame
- Accuracy requirements, and the need to measure accuracy
- Whether detailed analysis of the sample is expected
- Cost/operational concerns

PROBABILITY OR RANDOM SAMPLING

Probability sampling is based on the theory of probability. It is also known as random sampling. It provides a known nonzero chance of selection for each population element. It is used when generalization is the objective of study, and a greater degree of accuracy of estimation of population parameters is required. The cost and time required is high hence the benefit derived from it should justify the costs.

SIMPLE RANDOM SAMPLING

This sampling technique gives each element an equal and independent chance of being selected. An equal chance means equal probability of selection. An independent chance means that the draw of one element will not affect the chances of other elements being selected. The procedure of drawing a simple random sample consists of enumeration of all elements in the population.

1. Preparation of a List of all elements, giving them numbers in serial order 1, 2, B, and so on, and
2. Drawing sample numbers by using (a) lottery method, (b) a table of random numbers or (c) a computer.

Suitability: This type of sampling is suited for a small homogeneous population.

STRATIFIED SAMPLING

Where the population embraces a number of distinct categories, the frame can be organized by these categories into separate "strata." Each stratum is then sampled as an independent sub population, out of which individual elements can be randomly selected. There are several potential benefits to stratified sampling.

First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample.

Second, utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). Even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than would simple random sampling, provided that each stratum is proportional to the group's size in the population.

Third, it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified sampling approach may be more convenient than aggregating data across groups (though this may potentially be at odds with the previously noted importance of utilizing criterion-relevant strata).

Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can increase the cost and complexity of sample selection, as well as leading to increased complexity of population estimates. Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods (although in most cases, the required sample size would be no larger than would be required for simple random sampling.

Advantages: Stratified random sampling enhances the representativeness to each sample, gives higher statistical efficiency, easy to carry out, and gives a self-weighing sample.

Disadvantages: A prior knowledge of the composition of the population and the distribution of the population, it is very expensive in time and money and identification of the strata may lead to classification of errors.

SYSTEMATIC SAMPLING

Systematic sampling relies on arranging the study population according to some ordering scheme and then selecting elements at regular intervals through that ordered list. Systematic sampling involves a random start and then proceeds with the selection of every k^{th} element from then onwards. In this case, k = (population size/sample size). It is important that the starting point is not automatically the first in the list, but is instead randomly chosen from within the first to the k^{th} element in the list. A simple example would be to select every 10th name from the telephone directory (an 'every 10th' sample, also referred to as 'sampling with a skip of 10').

Suitability: Systematic selection can be applied to various populations such as students in a class, houses in a street, telephone directory etc.

Advantages: The advantages are it is simpler than random sampling, easy to use, easy to instruct, requires less time, it's cheaper, easier to check, sample is spread evenly over the population, and it is statistically more efficient.

Disadvantages: The disadvantages are it ignores all elements between two kth elements selected, each element does not have equal chance of being selected, and this method sometimes gives a biased sample.

CLUSTER SAMPLING

Sometimes it is more cost-effective to select respondents in groups ('clusters'). Sampling is often clustered by geography, or by time periods. (Nearly all samples are in some sense 'clustered' in time - although this is rarely taken into account in the analysis.) For instance, if surveying households within a city, we might choose to select 100 city blocks and then interview every household within the selected blocks.

Clustering can reduce travel and administrative costs. In the example above, an interviewer can make a single trip to visit several households in one block, rather than having to drive to a different block for each household.

Suitability: The application of cluster sampling is extensive in farm management surveys, socio-economic surveys, rural credit surveys, demographic studies, ecological studies, public opinion polls, and large scale surveys of political and social behaviour, attitude surveys and so on.

Advantages: The advantages of this method is it is easier and more convenient, cost of this is much less, promotes the convenience of field work as it could be done in compact places, it does not require more time, units of study can be readily substituted for other units and it is more flexible.

Disadvantages: The cluster sizes may vary and this variation could increase the bias of the resulting sample. The sampling error in this method of sampling is greater and the adjacent units of study tend to have more similar characteristics than do units distantly apart.

AREA SAMPLING

This is an important form of cluster sampling. In larger field surveys cluster consisting of specific geographical areas like districts, talluks, villages or blocks in a city are randomly drawn. As the geographical areas are selected as sampling units in such cases, their sampling is called area sampling. It is not a separate method of sampling, but forms part of cluster sampling.

PROBABILITY-PROPORTIONAL-TO-SIZE SAMPLING

In some cases the sample designer has access to an "auxiliary variable" or "size measure", believed to be correlated to the variable of interest, for each element in the population. These data can be used to improve accuracy in sample design. One option is to use the auxiliary variable as a basis for stratification, as discussed above.

Another option is probability-proportional-to-size ('PPS') sampling, in which the selection probability for each element is set to be proportional to its size measure, up to a maximum of 1.

In a simple PPS design, these selection probabilities can then be used as the basis for Poisson sampling. However, this has the drawback of variable sample size, and different portions of the population may still be over- or under-represented due to chance variation in selections. To address this problem, PPS may be combined with a systematic approach.

Example: Suppose we have six schools with populations of 150, 180, 200, 220, 260, and 490students respectively (total 1500 students), and we want to use student population as the basis fora PPS sample of size three. To do this, we could allocate the first school numbers 1 to 150, the second school 151 to 330 (= 150 + 180), the third school 331 to 530, and so on to the last school(1011 to 1500). We then generate a random start between 1 and 500 (equal to 1500/3) and count through the school populations by multiples of 500. If our random start was 137, we would select the schools which have been allocated numbers 137, 637, and 1137, i.e. the first, fourth, and sixth schools.

Advantages: The advantages are clusters of various sizes get proportionate representation, PPS leads to greater precision than would a simple random sample of clusters and a constant sampling fraction at the second stage, equal-sized samples from each selected primary cluster are convenient for field work.

Disadvantages: PPS cannot be used if the sizes of the primary sampling clusters are not known.

DOUBLE SAMPLING AND MULTIPHASE SAMPLING

Double sampling refers to the subsection of the final sample form a preselected larger sample that provided information for improving the final selection. When the procedure is extended to more than two phases of selection, it is then, called multi-phase sampling. This is also known as sequential sampling, as sub-sampling is done from a main sample in phases. Double sampling or multiphase sampling is a compromise solution for a dilemma posed by undesirable extremes. “The statistics based on the sample of ‘n’ can be improved by using ancillary information from a wide base: but this is too costly to obtain from the entire population of N elements. Instead, information is obtained from a larger preliminary sample n_L which includes the final sample n.

NON-PROBABILITY OR NON RANDOM SAMPLING

Non-probability sampling or non-random sampling is not based on the theory of probability. This sampling does not provide a chance of selection to each population element.

Advantages: The only merits of this type of sampling are simplicity, convenience and low cost.

Disadvantages: The demerits are it does not ensure a selection chance to each population unit. The selection probability sample may not be a representative one. The selection probability is unknown. It suffers from sampling bias which will distort results.

QUOTA SAMPLING

In **quota sampling**, the population is first segmented into mutually exclusive sub-groups, just a sin stratified sampling. Then judgement is used to select the subjects or units from each segment based on a specified proportion. For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.

It is this second step which makes the technique one of non-probability sampling. In quota sampling the selection of the sample is non-random. For example interviewers might be tempted to interview those who look most helpful. The problem is that these samples may be biased because not everyone gets a chance of selection. This random element is its greatest weakness and quota versus probability has been a matter of controversy for several years.

Suitability: It is used in studies like marketing surveys, opinion polls, and readership surveys which do not aim at precision, but to get quickly some crude results.

Advantage: It is less costly, takes less time, non-need for a list of population, and field work can easily be organized.

Disadvantage: It is impossible to estimate sampling error, strict control if field work is difficult, and subject to a higher degree of classification.

CONVENIENCE OR ACCIDENTAL SAMPLING

Accidental sampling (sometimes known as **grab**, **convenience** or **opportunity sampling**) is a type of non-probability sampling which involves the sample being drawn from that part of the population which is close to hand. That is, a population is selected because it is readily available and convenient. It may be through meeting the person or including a person in the sample when one meets them or chosen by finding them through technological means such as the internet or through phone. The researcher using such a sample cannot scientifically make generalizations about the total population from this sample because it would not be representative enough. For example, if the interviewer were to conduct such a survey at a shopping center early in the morning on a given day, the people that he/she could interview would be limited to those given there at that given time, which would not represent the views of other members of society in such an area, if the survey were to be conducted at different times of day and several times per week.

Suitability: Though this type of sampling has no status, it may be used for simple purposes such as testing ideas or gaining ideas or rough impression about a subject of interest.

Advantage: It is the cheapest and simplest, it does not require a list of population and it does not require any statistical expertise.

Disadvantage: The disadvantage is that it is highly biased because of researcher's subjectivity, it is the least reliable sampling method and the findings cannot be generalized.

PURPOSIVE (OR JUDGMENT) SAMPLING

This method means deliberate selection of sample units that conform to some pre-determined criteria. This is also known as judgment sampling. This involves selection of cases which we judge as the most appropriate ones for the given study. It is based on the judgment of the researcher or some expert. It does not aim at securing a cross section of a population. The chance that a particular case be selected for the sample depends on the subjective judgment of the researcher.

Suitability: This is used when what is important is the typicality and specific relevance of the sampling units to the study and not their overall representativeness to the population.

Advantage: It is less costly and more convenient and guarantees inclusion of relevant elements in the sample.

Disadvantage: It is less efficient for generalizing, does not ensure the representativeness, requires more prior extensive

5.7.13 SNOW-BALL SAMPLING

This is the colourful name for a technique of Building up a list or a sample of a special population by using an initial set of its members as informants. This sampling technique may also be used in socio-metric studies.

Suitability: It is very useful in studying social groups, informal groups in a formal organization, and diffusion of information among professional of various kinds.

Advantage: It is useful for smaller populations for which no frames are readily available.

Disadvantage: The disadvantage is that it does not allow the use of probability statistical methods. It is difficult to apply when the population is large. It does not ensure the inclusion of all the elements in the list.