**KAMPALA INTERNATIONAL UNIVERSITY**

**COLLEGE OF ECONOMICS AND MANAGEMENT**

**DEPARTMENT OF ECONOMICS AND STATISTICS**

**YEAR TWO SEMESTER TWO**

**COURSE OUTLINE**

**COURSE TITLE: SAMPLING METHODS**

**COURSE CODE: STA 2203**

**COURSE CONTENT**

* **INTRODUCTION**
* Definition of terms and concepts
* Objectives of the survey
* Data collected
* Methods of measurement
* Sampling frame, types of sampling frames and sampling defects.
* Sampling theory
* Sampling methods
* Types of samples (non-random and random samples)
* survey design
* types of surveys
* census versus sample surveys
* advantages and disadvantages of censuses and surveys.
* Organization of surveys
* Desirable properties of estimators
* Unbiasedness
* Efficiency, mean square error
* Sufficiency
* Consistency
* Errors in surveys
* Sampling errors, their causes and remedies
* Non-sampling errors, causes and remedies
* Sampling procedures/ techniques
1. Simple random sampling
* With replacement and without replacement
* Computation of means and variances
* Regression and ratio estimators
* Sample size determination
1. Stratified sampling
* Definition
* Basis of stratification
* Sample selection (proportionate allocation, optimum allocation)
* Means and variances
1. Systematic sampling
* Definition
* Advantages and disadvantages
* Mean and variance computation
1. Cluster sampling
* Definition
* Advantages and disadvantages
* Cluster selection
1. Multi-stage cluster sampling
* Intra-cluster correlation
* Design effect
* Sample selection procedures
* Application

**REFERENCES**

* Survey sampling by Leslie Kish
* Sampling techniques by Cochran
* The statistical analysis of time series by Anderson
* Survey methods in social investigation by C.A Moser and G. Kalton
* Sample surveys with special reference to Africa by Prof. Ben Kiregyera
* Survey methods and practices; statistics Canada (2003).
* Statistics in practice, sample surveys; Duke Energy.

CHAPTER ONE

1. INTRODUCTION/ BASIC CONCEPTS AND DEFINITIONS

Sampling is the selection of part of an aggregate to represent the whole and is frequently used in surveys.

1. Population: is the aggregate from which the sample is chosen OR a population is a well defined collection of all units of interest.
2. Sample: is a portion of the population used for analysis purposes to draw conclusions about the whole population. There are two types of samples i.e
3. subjective (non-random) samples whose composition is determined by the sampling practice without using any random or chance mechanism. Such samples are easy and cheap to select and not generally subject to non-response as substitution can be done at will e.g expert chosen samples (samples chosen by subject matter experts according to how representative they deem them to be representative of the population of interest) and quota samples (samples of designated composition).
4. Random (probability) samples: selected using a probability or chance mechanism and is not influenced by the sampling practitioner. Each unit in the population is assigned a non-zero but not necessarily the same probability of selection.
5. Sampling unit: a finite number of distinct and identifiable units subdivided from the population to be sampled.
6. Sampling frame/frame: is a list of distinct and distinguishable units of a given population which is used for selecting units from the population into a sample. there are two types of sampling frames;
7. Area (map) frames: these are land segments that are used as first stage sampling units with descriptions showing area boundaries and area size. These frames have a problem of incomplete maps not showing all sampling units due to wars, mapping costs are high and updating maps is expensive, sometimes maps do not exist, use of physical features as boundaries in survey work which may lead to double counting and maps change overtime.
8. List frames: consist of sampling units which can be identified and selected without reference to a map. e.g lists of households, employees.

A sampling frame should be adequate, that is

* Accurate. It is accurate if the units in the frame are defined precisely and information about them is correct.
* Complete. It lists every unit of the population once and only once without omission or duplication.
* Up-to-date. It is accurate and complete at the time when a sample is to be selected.
1. enumeration unit: unit on which data are collected according to well defined procedure e.g in a household survey, the household is the unit of enumeration.
2. Reporting unit; the unit that actually supplies the required data or on which information can be easily ascertained.
3. Unit of Analysis: unit used at the stage of tabulation.
4. Parameter: numerical value describing the characteristic of a population.
5. Statistic: numerical value describing the characteristic of a sample.
6. Survey design: relates to the whole process of conducting a survey including sample design, field strategy, field logistics, data collection, processing, analysis, report writing and dissemination of survey results.
7. Sample design: relates to the techniques for selecting a random sample and the methods for obtaining estimates of survey variables from sample data. Sample design involves identification and selection of sampling units, determining rules for assigning selection probabilities to sampling units, determining stratification and clustering procedures and deciding on different types of estimation methods.

1.2 C**ENSUS AND SURVEYS**

**1.21 CENSUS**

**Is** a statistical inquiry that involves complete enumeration of the whole population under consideration. The advantages of a census include;

* Provide benchmark data for planning, social and economic development.
* Provide small area statistics which is basic data disaggregated to the lowest administrative units e.g villages.
* No sampling errors in the data collected
* Provision of supplementary information that is required for efficient planning of sample surveys.
* If the population size is small, a census is ideal.

The disadvantages include;

* Expensive because of large coverage.
* Data from census are untimely due to delays as a result of large coverage
* Limited accuracy because census involves many people who may not be skilled and adequately trained in data collection techniques.
* There are short and brief questions thus do not generate indepth data and information for detailed analysis of certain aspects.

1.22 **SAMPLE SURVEYS**

A sample survey is a statistical inquiry based on a small part of a population from which inferences are made about the population characteristics. The advantages of sample surveys include;

* Less costly because data is collected from a small portion but cost per unit surveyed may be higher than in censuses.
* Greater speed. Data can be collected and summarized more quickly.
* Increased scope. Possible to use specialized equipment and ask detailed questions.
* Increased accuracy. Personnel of high quality can be employed and given intensive training and there can be more careful supervision of field work.
* Information collected is up to date because surveys can be carried out annually or at shorter periods.
* Errors other than sampling errors e.g those arising through non-response, incomplete, inaccuracy of returns are termed as non-sampling errors and are likely to be more wide spread in a census than in a sample survey.
* They are manageable and can be carried out at any stage of development of many countries.

The disadvantages include;

* Do not generate adequate small area statistics
* Overhead costs are usually higher in sample surveys than in census.
* They are subject to sampling errors and bias.
* When measures of small change from one period to the next are required, surveys are not the best.

1.23 **RELATIONSHIP BETWEEN SAMPLE SURVEYS AND CENSUSES**

These include;

* Sample surveys are a substitute for censuses’
* Surveys used to test census methods, procedures and materials.
* Sample surveys supplement census data
* Sample surveys used in census tabulation work
* Census used as a basis for subsequent surveys.

1.3 **TYPES OF SURVEYS**

 They can be classified into 3 depending on the purpose and objectives for which the data may be used. That is;

1. Descriptive surveys: their objective is to obtain descriptive measures with respect to the characteristics of the entire population under study. For example required for national planning and social economic development to collect data on agriculture production, utilization of land and water resources, industrial production, unemployment and size of labour force.
2. Analytical surveys: the objective is to obtain descriptive information for different sub groups of the population in order to test hypotheses concerning possible relationships between sub-groups.
3. Mixed surveys: involves both analytical and descriptive methods.

**CHAPTER TWO**

1. **SAMPLING DESIGN**

The planning phase the statistical agency must decide whether to conduct a census or sample survey. If the decision is a sample survey, then the agency needs to plan how to select the sample. ***Sampling is a means of selecting a subset of units from a population for the purpose of collecting information for those units to draw inferences about the population as a whole****.*There are two types of sampling: non-probability and probability sampling. The one chosen depends primarily on whether reliable inferences are to be made about the population. Non-probability sampling, uses a subjective method of selecting units from a population. It provides a fast, easy and inexpensive way of selecting a sample. However, in order to make inferences about the population from the sample, the data analyst must assume that the sample is representative of the population. This is often a risky assumption to make in the case of non-probability sampling.

Probability sampling, involves the selection of units from a population based on the principle of randomisation or chance. Probability sampling is more complex, time consuming and usually more costly than non-probability sampling. However, because units from the population are randomly selected and each unit’s inclusion probability can be calculated, reliable estimates can be produced along with estimates of the sampling error, and inferences can be made about the population.

There are several different ways in which a probability sample can be selected. The design chosen depends on a number of factors such as:

* the available survey frame
* how different the population units are from each other (i.e., their variability)
* how costly it is to survey members of the population.

 For a given population, a balance of sampling error with cost and timeliness is achieved through the choice of design and sample size.

**2.1 Non-Probability Sampling**

Non-probability sampling is a method of selecting units from a population using a subjective (i.e., nonrandom) method. Since non-probability sampling does not require a complete survey frame, it is a fast, easy and inexpensive way of obtaining data. The problem with non-probability sampling is that it is unclear whether or not it is possible to generalise the results from the sample to the population. The reason for this is that the selection of units from the population for a non-probability sample can result in large biases.

For example, a common design is for the interviewer to subjectively decide who should be sampled. Since the interviewer is most likely to select the most accessible or friendly members of the population, a large portion of the population has no chance of ever being selected, and this portion of the population is likely to differ in a systematic manner from those selected members. Not only can this bias the results of the survey, it can falsely reduce the apparent variability of the population due to a tendency to select ‘typical’ units and eliminate extreme values. By contrast, probability sampling avoids such bias by randomly selecting units.

Due to selection bias and (usually) the absence of a frame, an individual’s inclusion probability cannot be calculated for non-probability samples, so there is no way of producing reliable estimates or estimates of their sampling error. In order to make inferences about the population, it is necessary to assume that the sample is representative of the population. This usually requires assuming that the characteristics of the population follow some model or are evenly or randomly distributed over the population. This is often dangerous due to the difficulty of assessing whether or not these assumptions hold.

Non-probability sampling is often used by market researchers as an inexpensive and quick alternative to probability sampling, but it is not a valid substitute for probability sampling for the reasons delineated above. So, why bother with non-probability sampling? Non-probability sampling can be applied to studies that are used as:

- an idea generating tool

- a preliminary step towards the development of a probability sample survey

- a follow-up step to help understand the results of a probability sample survey.

For example, non-probability sampling can provide valuable information in the early stages of an investigation. It can be used for exploratory or diagnostic studies to gain insights into people’s attitudes, beliefs, motivations and behaviours. Sometimes non-probability sampling is the only viable option.

* 1. The **advantages and disadvantages** of non-probability sampling are that:
* It is quick and convenient. As a general rule, non-probability samples can be quickly drawn and surveyed.
* It is relatively inexpensive. It usually only takes a few hours of an interviewer’s time to conduct such a survey.
* Non probability samples are generally not spread out geographically, therefore travelling expenses for interviewers are low.
* It does not require a survey frame.
* It can be useful for exploratory studies and survey development.

The **disadvantages** of non-probability sampling are that:

* In order to make inferences about the population it requires strong assumptions about the representativeness of the sample. Due to the selection bias present in all non-probability samples, these are often dangerous assumptions to make. When inferences are to be made, probability sampling should be performed instead.
* It is impossible to determine the probability that a unit in the population is selected for the sample, so reliable estimates and estimates of sampling error cannot be computed.

**2.2 Types of non-probability sampling schemes**

These include; haphazard sampling, volunteer sampling, judgement sampling, quota sampling, modified probability sampling and Network or snowball sampling, which is less commonly used.

1. **Haphazard Sampling;** Units are selected in an aimless, arbitrary manner with little or no planning involved. Haphazard sampling assumes that the population is homogeneous: if the population units are all alike, then any unit may be chosen for the sample. An example of haphazard sampling is the ‘man in the street’ interview where the interviewer selects any person who happens to walk by. Unfortunately, unless the population is truly homogeneous, selection is subject to the biases of the interviewer and whoever happened to walk by at the time of sampling.
2. **Volunteer Sampling;** With this method, the respondents are volunteers. Generally, volunteers must be screened so as to get a set of characteristics suitable for the purposes of the survey (e.g., individuals with a particular disease). This method can be subject to large selection biases, but is sometimes necessary. For example, for ethical reasons, volunteers with particular medical conditions may have to be solicited for some medical experiments. Another example of volunteer sampling is callers to a radio or television show, when an issue is discussed and listeners are invited to call in to express their opinions. Only the people who care strongly enough about the subject one way or another tend to respond. The silent majority does not typically respond, resulting in a large selection bias. Volunteer sampling is often used to select individuals for focus groups or in-depth interviews (i.e., for qualitative testing, where no attempt is made to generalise to the whole population).
3. **Judgement Sampling;** With this method, sampling is done based on previous ideas of population composition and behaviour. An expert with knowledge of the population decides which units in the population should be sampled. In other words, the expert purposely selects what is considered to be a representative sample. Judgement sampling is subject to the researcher's biases and is perhaps even more biased than haphazard sampling. Since any preconceptions the researcher has are reflected in the sample, large biases can be introduced if these preconceptions are inaccurate. However, it can be useful in exploratory studies, for example in selecting members for focus groups or in-depth nterviews to test specific aspects of a questionnaire.
4. **Quota Sampling;** This is one of the most common forms of non-probability sampling. Sampling is done until a specific number of units (quotas) for various subpopulations has been selected. Quota sampling is a means for satisfying sample size objectives for the subpopulations. The quotas may be based on population proportions. For example, if there are 100 men and 100 women in the population and a sample of 20 are to be drawn, 10 men and 10 women may be interviewed. Quota sampling can be considered preferable to other forms of non-probability sampling (e.g., judgement sampling) because it forces the inclusion of members of different sub populations .Quota sampling is somewhat similar to stratified sampling in that similar units are grouped together. However, it differs in how the units are selected. In probability sampling, the units are selected randomly while in quota sampling a non-random method is used – it is usually left up to the interviewer to decide who is sampled. Contacted units that are unwilling to participate are simply replaced by units that are, in effect ignoring non response bias. Market researchers often use quota sampling (particularly for telephone surveys) instead of stratified sampling to survey individuals with particular socio-economic profiles. This is because compared with stratified sampling, quota sampling is relatively inexpensive and easy to administer and has the desirable property of satisfying population proportions. However, it disguises potentially significant selection bias. As with all other non-probability sample designs, in order to make inferences about the population, it is necessary to assume that persons selected are similar to those not selected. Such strong assumptions are rarely valid.
5. **Modified Probability Sampling;** Modified probability sampling is a combination of probability and non-probability sampling. The first stages are usually based on probability sampling. The last stage is a nonprobability sample, usually a quota sample. For example, geographical areas may be selected using a probability design, and then within each region, a quota sample of individuals may be drawn

.

* 1. **Probability Sampling**

Probability sampling is a method of sampling that allows inferences to be made about the population based on observations from a sample. In order to be able to make inferences, the sample should not be subject to selection bias. Probability sampling avoids this bias by randomly selecting units from the population (using a computer or table of random numbers). It is important to note that random does not mean arbitrary. In particular, the interviewers do not arbitrarily choose respondents since then sampling would be subject to their personal biases. Random means that selection is unbiased – it is based on chance. With probability sampling, it is never left up to the discretion of the interviewer to subjectively decide who should be sampled. There are two main criteria for probability sampling: one is that the units be randomly selected, the second is that all units in the survey population have a non-zero inclusion probability in the sample and that these probabilities can be calculated. It is not necessary for all units to have the same inclusion probability, indeed, in most complex surveys, the inclusion probability varies from unit to unit.

**2.31 Types of probability sample designs**.

 The most basic is simple random sampling and the designs increase in complexity to encompass systematic sampling, probability-proportional-tosize sampling, cluster sampling, stratified sampling, multi-stage sampling, multi-phase sampling and replicated sampling. Each of these sampling techniques is useful in different situations. If the objective of the survey is simply to provide overall population estimates and stratification would be inappropriate or impossible, simple random sampling may be the best. If the cost of survey collection is high and the resources are available, cluster sampling is often used. If subpopulation estimates are also desired (such as estimates by province, age group, or size of business), stratified sampling is usually performed. Most of the more complex designs use auxiliary information on the survey frame to improve sampling. If the frame has been created from a previous census or from administrative data, there may be a wealth of supplementary information that can be used for sampling. For example, for a farm survey, the statistical agency may have the size of every farm in hectares from the last agricultural census. For a survey of people, information (e.g., age, sex, ethnic origin, etc.) may be available for everyone from the last population census. For a business survey, the statistical agency may have administrative information such as the industry (e.g., retail, wholesale, manufacturing), the type of business (e.g., food store), the number of employees, etc. In order for the auxiliary information to improve sampling, there must be a correlation between the auxiliary data and the survey variables.

The main **advantage** of probability sampling is that since each unit is randomly selected and each unit’s inclusion probability can be calculated reliable estimates and an estimate of the sampling error of each estimate can be produced. Therefore, inferences can be made about the population. In fact, with a probability design, a relatively small sample can often be used to draw inferences about a large population.

The main **disadvantages** of probability sampling are that it is more difficult, takes longer and is usually more expensive than non-probability sampling. In general, the expense of creating and maintaining a good quality frame is substantial. And because probability samples tend to be more spread out geographically across the population than non-probability samples, sample sizes are generally much larger and data collection is often more costly and difficult to manage. However, for a statistical agency, the ability to make inferences from a probability sample usually far outweighs these disadvantages.

**2.3 PLANNING AND ORGANISATION OF A SURVEY**

**Steps of a Survey**

It may appear that conducting a survey is a simple procedure of asking questions and then compiling the answers to produce statistics. However, a survey must be carried out step by step, following precise procedures and formulas, if the results are to yield accurate and meaningful information. In order to understand the entire process it is necessary to understand the individual tasks and how they are interconnected and related.

The steps of a survey are:

1. formulation of the Statement of Objectives; This establishes not only the survey’s broad information needs, but the operational definitions to be used, the specific topics to be addressed and the analysis plan.
2. selection of a survey frame;The survey frame provides the means of identifying and contacting the units of the survey population. The frame is in the form of a list.
3. determination of the sample design; There are two kinds of surveys: sample surveys and census surveys. In a sample survey, data are collected for only a fraction (typically a very small fraction) of units of the population while in a census survey, data are collected for all units in the population.
4. Methods of collecting data. Methods should be selected considering aspects like cost, literacy, timing e.g questionnaire design;A questionnaire (or form) is a group or sequence of questions designed to obtain information on a subject from a respondent. Questionnaires play a central role in the data collection process since they have a major impact on data quality and influence the image that the statistical agency projects to the public. Questionnaires can either be in paper or computerised format.
5. The pretest. The questionnaire and field methods should be tried out on a small scale. This improves on the results and may reveal other troubles that may occur on a large scale.
6. data collection;Data collection is the process of gathering the required information for each selected unit in the survey. The basic methods of data collection are self-enumeration, where the respondent completes the questionnaire without the assistance of an interviewer, and interviewer-assisted (either through personal or telephone interviews). Other methods of data collection include direct observation, electronic data reporting and the use of administrative data.
7. Precision. Specify the degree of precision desired with respect to various population values to be estimated. Large samples may be considered and superior instruments used to avoid or reduce errors in measurement.
8. data capture and coding; After the data are collected, they are coded and, if a computer-assisted collection method was not used,captured. Coding is the process of assigning a numerical value to responses to facilitate data capture and processing in general.
9. editing and imputation; Editing is the application of checks to identify missing, invalid or inconsistent entries that point to data records that are potentially in error. The purpose of editing is to better understand the survey processes and the survey data in order to ensure that the final survey data are complete, consistent and valid. Imputation is a process used to determine and assign replacement values to resolve problems of missing, invalid or inconsistent data.
10. estimation; Estimation is the means by which the statistical agency obtains values for the population of interest so that it can draw conclusions about that population based on information gathered from only a sample of the population. An estimate may be a total, mean, ratio, percentage, etc.
11. data analysis; Data analysis involves summarising the data and interpreting their meaning in a way that provides clear answers to questions that initiated the survey. Data analysis should relate the survey results to the questions and issues identified by the Statement of Objectives.
12. data dissemination; Data dissemination is the release of the survey data to users through various media, for example, through a press release, a television or radio interview, a telephone or facsimile response to a special request, a paper publication, electronic media including the Internet or a microdata file on a CD, etc.. Results from the survey should be summarized and the strengths and weaknesses of the data indicated, with important details highlighted through a written report that includes tables and charts.
13. documentation. Documentation provides a record of the survey and should encompass every survey step and every survey phase. It may record different aspects of the survey and be aimed at different groups, such as management, technical staff, designers of other surveys and users. During implementation, documentation of procedures for staff helps to ensure effective implementation.

2.2 **BASIC PRINCIPLES OF SAMPLE SURVEY DESIGNS**

These include;

1. Principle of validity: a sample should be selected in such a way that the objective interpretation of the sample results is possible.
2. Optimization principle: this includes;
* Efficiency. The ability to collect more reliable information per unit cost .
* Cost: in terms of finances and time spent on data collection from a sampling unit. That is; minimize cost for a given level of efficiency.

2.3 **ESIMATORS AND THEIR DESIRABLE PROPERTIES**

 An estimator is a statistical measure used to obtain a sample characteristic. It is a function of the sample observation.

An estimator is a random variable while an estimate is a point.

**2.31 DESIRABLE** **PROPERTIES OF ESTIMATORS.**

These are properties of unbiasedness, consistency and efficiency.

* Unbiasedness: an estimator $\hat{θ}$ is an unbiased estimator for the population parameter $θ$ if its expected value is equal to the unknown population parameter. That is; $E\left(\hat{θ}\right)=θ.$
* Consistency: an estimator is consistency if it approaches the population parameter as sample size increases. That is, $\hat{θ}\rightarrow θ as n\rightarrow \infty $.
* Efficiency: an estimator is efficient if it has minimum variance among all the unbiased estimators. Efficiency $\left(\hat{θ}\right)=\frac{1}{MSE(\hat{θ})}$ , Where $MSE\left(\hat{θ}\right)=E(\hat{θ}-θ)^{2}.$ let $\hat{θ}\_{1} and \hat{θ}\_{2}$ be two estimators of the parameter θ, $\hat{θ}\_{1}$ is said to be uniformly better (more efficient) than $\hat{θ}\_{2}$ if and only if $MSE(\hat{θ}\_{1})\leq MSE(\hat{θ}\_{2})$ and $\hat{θ}\_{1}$ is more precise than $\hat{θ}\_{2}$ if $var \left(\hat{θ}\_{1}\right)\leq var\left(\hat{θ}\_{2}\right).$

Efficiency of $\hat{θ}\_{1}$ compared to $\hat{θ}\_{2}$ is referred to as the relative efficiency of $\hat{θ}\_{1}$ with respect to $\hat{θ}\_{2}$ and is given as $R.E \left(\hat{θ}\_{1},\hat{θ}\_{2}\right)=\frac{var (\hat{θ}\_{2})}{var (\hat{θ}\_{1})}$.

2.4 **SAMPLING AND NON-SAMPLING ERRORS IN SURVEYS**

2.41 SAMPLING ERRORS

These arise because observations are taken on a sample rather than the whole population. A sampling error is the difference between the said sample estimate and the population values that will be obtained by enumerating all the units in the population.

The magnitude of the sampling errors depends on;

* **Sample size**. The bigger the sample used, the lower the sampling error.
* **Sample design**. Some sample designs will lead to lower magnitude of sampling errors than others.
* **Estimator used**. Some estimators are more efficient than others e.g ratio and regression estimators are more efficient than the simple mean.
* **Nature of the population**. Sparsely distributed populations will have larger sampling errors than less sparsely distributed populations.

The most commonly used measure to quantify sampling error is sampling variance. Sampling variance measures the extent to which the estimate of a characteristic from different possible samples of the same size and the same design differ from one another.

Sampling errors are estimated by the standard error of the estimator, that is $S.E\left(\hat{θ}\right)=\sqrt{v\left(\hat{θ}\right).}$

The magnitude of the sampling errors can be controlled (reduced) by increasing the sample size and choosing a more efficient sample design or a superior method of estimation.

2.42 **NON-SAMPLING ERRORS**

Errors other than sampling errors. These are systematic errors that arise both in census and sample surveys but tend to be more pronounced in censuses.

Non-sampling errors can be classified into two groups:

1. Random errors; Random errors are errors whose effects approximately cancel out if a large enough sample is used, leading to increased variability.
2. Systematic errors; are errors that tend to go in the same direction and thus accumulate over the entire sample, leading to a bias in the final results. Unlike sampling variance or random errors, this bias is not reduced by increasing the size of the sample. Systematic errors are the principal cause of concern in terms of a survey’s data quality. Unfortunately, nonsampling errors are often extremely difficult and sometimes impossible to measure.

 Non-sampling errors arise primarily from the following sources:

1. Coverage. Coverage errors consist of omissions, erroneous inclusions, duplications and misclassifications of units in the survey frame.
2. Response, content
3. Measurement; Measurement error is the difference between the recorded response to a question and the ‘true’ value. It can be caused by the respondent, the interviewer, the questionnaire, the data collection method or the measuring tool.
4. Non-response errors; There are two types of nonresponse: item (or partial) nonresponse and total nonresponse. Item nonresponse occurs when information is provided for only some items, such as when the respondent responds to only part of the questionnaire. Total nonresponse occurs when all or almost all data for a sampling unit are missing.
5. Processing errors. Errors committed during data processing e.g during editing, coding, programming, data entry and tabulation errors.

2.43 **CAUSES OF NON-SAMPLING ERRORS AND THE REMEDIES**

The causes include;

* Conceptual and definitional problems e.g some concepts keep on changing.
* Defects in the sampling frames due to a number of factors including improper definition of sampling units, incomplete listing or mapping of area units.
* Use of unmotivated and unconscious field staff
* Inadequate training of field staff
* Lack of adequate supervision
* Improper survey instruments including badly designed questionnaires
* Measurement problems
* Respondent’s memory lapse and faulty or inappropriate methods.
* Failure to obtain data from selected units
* Poor operational controls leading to loss of questionnaires
* Poorly designed data entry and tabulation programmers.
* Poor scrutiny and editing
* Using inaccurate or inappropriate estimation procedures
* Language barrier

The remedies include;

* Testing of procedures and materials
* Extensive geographical planning and field work as well as pre-enumeration of all areas
* Better definition of concepts
* Extensive education and publicity campaigns through the national press
* Greater care in recruitment of field and other staff
* Intensive training of field staff
* Better design of survey instruments
* Careful editing of questionnaires
* Better design of data processing procedures

When an estimator has a small sampling error measured by its standard error, it is said to be precise and when an estimator has a small total error, it is declared to be accurate.

**CHAPTER THREE**

1. **SAMPLING PROCEDURES**

 The basic sampling procedures commonly used for selecting a sample are:

* Simple random sampling
* Probability proportional to size sampling
* Stratified random sampling
* Systematic sampling
* Cluster sampling
* Multi-stage sampling

**3.1 SIMPLE RANDOM SAMPLING (SRS)**

This is a sampling design in which every unit in the population has the same probability (1/N) of being selected at each draw. It consists of selecting a sample, unit by unit until the desired number of units (n) is obtained. There are two types of simple random sampling ;

1. Simple random sampling with replacement (SRSWR): If the unit of the population selected in the first draw is replaced back into the population (after writing down the value of the character being measured) before making the second and subsequent draws. The probability of drawing a unit into the sample is the same at all levels of selection and a population unit can repeat itself in a sample and the order in which units appear in the sample does not counts e.g samples AB and BA are treated as different samples. The number of possible samples selected is given as Nn and the probability of each sample is $\frac{1}{N^{n}}.$
2. Simple random sampling without replacement (SRSWOR): if the unit once drawn is not replaced in the sample and subsequent units in the sample are selected from the remaining units of the population at each draw. The probability of selection is $\frac{1}{N\_{C\_{n}}}$ and the number of samples selected is $N\_{C\_{n}}=\frac{N!}{n!(N-n)!}.$ the units do not repeat in the sample and the order of appearance in the sample matters.

SRS has a number of **advantages** over other probability sampling techniques, including:

* It is the simplest sampling technique.
* It requires no additional (auxiliary) information on the frame in order to draw the sample.
* The only information that is required is a complete list of the survey population and contact information.
* It needs no technical development.
* The theory behind SRS is well established, so that standard formulas exist to determine the sample size, population estimates and variance estimates and these formulas are easy to use.

The **disadvantages** of SRS are:

* It makes no use of auxiliary information even if such information exists on the survey frame. This can result in estimates being less statistically efficient than if another sample design had been used.
* It can be expensive if personal interviews are used, since the sample may be widely spread out geographically.
* It is possible to draw a ‘bad’ SRS sample. Since all samples of size n have an equal chance of being included in the sample, it is possible to draw a sample that is not well dispersed and that poorly represents the population.

EXAMPLE 1:

Given a population of size 4 as {E,S,J,R}, select a sample of size 2 using;

1. SRSWR
* Number of possible samples is Nn= 42=16

|  |  |
| --- | --- |
| Unit | Possible samples |
| E | EE | ES | EJ | ER |
| S | SE | SS | SJ | SR |
| J | JE | JS | JJ | JR |
| R | RE | RS | RJ | RR |

1. SRSWOR
* Number of possible samples is $N\_{C\_{n}}=\frac{N!}{n!(N-n)!}=6.$

|  |  |
| --- | --- |
| Unit | Possible samples |
| E | ES | EJ | ER |
| S | SJ | SR |  |
| J | JR |  |  |
| R |  |  |  |

3.11 **ESTIMATION**

**Under**  SRS the following population parameters are used ;

1. Population total (Y) =Σyi OR $Y=N\overbar{y}$
2. Population mean $\overbar{Y}=\frac{1}{N}\sum\_{}^{}y\_{i}$
3. population proportion P = $\frac{N^{,}}{N}$, where N, is number of population units with an attribute of interest.
4. Population ratio (R) between variables y and x is; $R=\frac{\sum\_{}^{}y\_{i}}{\sum\_{}^{}x\_{i}}$
5. Population variance (S2) $=\frac{1}{N-1}\sum\_{}^{}(y\_{i}-\overbar{Y)}^{2}$

The estimators for the above parameters are;

1. Estimator for the mean $\overbar{y}\_{n}=\frac{1}{n}\sum\_{}^{}y\_{i}$ which is an unbiased estimator for the population mean $\overbar{Y.}$ that is ; $E\left(\overbar{y}\right)=\frac{1}{n}\sum\_{}^{}\overbar{y}\_{n}=\overbar{Y}\_{N}.$
2. Variance of the estimator is $ V\left(\overbar{y}\_{n}\right)=\left(\frac{1}{n}-\frac{1}{N}\right)S^{2}=\left(1-f\right)\frac{S^{2}}{n}(for SRSWOR)$ , where S2 is the population variance and f=n/N is the sampling fraction (rate). This variance reduces to S2/n (for SRSWR) as f tends to zero.
3. The standard error (SE) of the sample mean is $SE\left(\overbar{y}\_{n}\right)=+\sqrt{ v\left(\overbar{y}\_{n}\right)}$
4. Estimator of the population total is the sample total given as $y=n\overbar{y}\_{n}=\sum\_{}^{}y\_{i}$ and is unbiased.
5. Sample variance $s^{2}=\frac{1}{n-1}\sum\_{}^{}(y\_{i}-\overbar{y}\_{n})^{2}$
6. Variance of the estimator $\overbar{y}\_{n}$ is $\hat{v}\left(\overbar{y}\_{n}\right)=\frac{s^{2}}{n} OR (\frac{1}{n}-\frac{1}{N})s^{2}$
7. Relative efficiency of SRSWOR to SRSWR = variance under SRSWR/variance under SRSWOR $=\frac{N-1}{N-n}$. This result is greater than one for n larger than one implying that variance under SRSWR is larger than variance under SRSWOR and hence SRSWOR is more efficient than SRSWR.

**EXAMPLE II**

A random sample of size two was drawn from a small village of households having monthly incomes as follows;

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Household | 1 | 2 | 3 | 4 | 5 |
| Income(000) | 250 | 300 | 240 | 150 | 160 |

Calculate;

1. Population mean $\overbar{Y.}, varince δ^{2} and mean square error S^{2}.$
2. Enumerate all possible samples of size 2 by SRSWR and show that ;
3. The sample mean gives an unbiased estimator of the population mean.
4. Find the sampling variance ($v(\overbar{y})$)
5. Show that the sample variance s2 gives an unbiased estimator of the population variance δ2.
6. Enumerate all possible samples of size 2 by SRSWOR and show that;
7. The sample mean gives an unbiased estimator of the population mean and find its sampling variance.
8. The sample mean square (s2) gives an unbiased estimate of the population mean square S2.

SOLUTION

1. Population mean $\overbar{Y}$=$\frac{1}{N}\sum\_{}^{}Y\_{i}=1100/5=$220.
2. Variance $δ^{2}=\frac{1}{N}\sum\_{}^{}(Y\_{i}-\overbar{Y})^{2}=\frac{1}{5}\left(16200\right)=3240.$
3. Mean square error $S^{2}=\frac{1}{N-1}\sum\_{}^{}(Y\_{i}-\overbar{Y})^{2}=\frac{1}{4}\left(16200\right)=4050$

EXERCISE: TRY QUESTIONS (ii) and (iii).

* 1. **CONFIDENCE LIMITS FOR THE POPULATION MEAN** $\overbar{Y.}$

The confidence interval for the mean given that the variance is known is given as;

 $P\left(\overbar{y}-Z\_{\frac{∝}{2}}\sqrt{v\left(\overbar{y}\right) }\leq \overbar{Y} \leq \overbar{y}+Z\_{\frac{∝}{2}}\sqrt{v\left(\overbar{y}\right) }\right)=1-∝, $ where $v\left(\overbar{y}\right)=(1-f)\frac{S^{2}}{n}$ for SRSWOR.

But usually the variance S2 is unknown and the estimate s2 for the sample is used. This approximates to a t- distribution with n-1 degrees of freedom and the confidence interval is given as;

$$P\left(\overbar{y}-t\_{\frac{∝, n-1}{n2}}\sqrt{\hat{v}\left(\overbar{y}\right) }\leq \overbar{Y} \leq \overbar{y}+t\_{\frac{∝, n-1}{2}}\sqrt{\hat{v}\left(\overbar{y}\right) }\right)=1-∝,$$

Under SRSWR, the confidence limits are: $P\left(\overbar{y}-t\_{\frac{∝, n-1}{n2}} \frac{s}{\sqrt{n}}\leq \overbar{Y} \leq \overbar{y}+t\_{\frac{∝, n-1}{2}} \frac{s}{\sqrt{n}}\right)=1-∝$

EXAMPLE III: construct a 95% confidence interval using example II.

3.13 **SAMPLING PROPORTION (P).**

In many situations we need to find out the proportion of units in a population that belong to a given class or posses a certain attribute. The population proportion is $P=\frac{N\_{1}}{N}$ and the sample proportion which is an unbiased estimator of the population proportion P is given as $p=\frac{n\_{1}}{n}.$

* The variance of P under SRSWR is; $v\left(P\right)=v\left(\overbar{Y}\right)=\frac{δ^{2}}{n}$ but δ2 = PQ which implies that $v\left(P\right)=\frac{PQ}{n}.$ the unbiased estimator$\hat{v}\left(p\right)=\frac{s^{2}}{n}$ where the estimator $s^{2}=\frac{npq}{n-1}$
* The variance of P under SRSWOR is given as; $v\left(P\right)=\frac{(N-n)PQ}{n(N-1)}$ and the unbiased estimator of the estimator P is $\hat{v}\left(p\right)=\frac{pq(N-n)}{N(n-1)}.$

The estimated variance is generally given as; $v\left(p\right)=(1-f)\frac{pq}{n-1}$ but if N is large in relation to n, the unbiased estimate for the variance is $\hat{v}\left(p\right)=\frac{pq}{n-1}$.

* The estimate of the total is $\hat{N}\_{1}=NP$ and the variance of the total estimator is $v\left(\hat{N}\_{1}\right)=v\left(NP\right)=N^{2}v\left(p\right)=\frac{N^{2}PQ(N-n)}{n(N-1)})$.
* The estimate of the variance of the total estimator is $\hat{v}\left(\hat{N}\_{i}\right)=\hat{v}\left(NP\right)=\frac{N^{2}(1-f)pq}{n-1}$ and if f is ignored, then $\hat{v}\left(\hat{N}\_{i}\right)=\hat{v}\left(NP\right)=\frac{N^{2}pq}{n-1}$.

**EXAMPLE**

**From** a list of 3042 names and addresses, a simple random sample of 200 names showed on investigation that 38 wrong addresses needed correction.

1. Estimate the total number of addresses needing correction in the list.

For N =3042, n= 200, P=n1/n= 38/200

$$\hat{N}\_{1}=NP=3042\*\frac{38}{200}=577.98.$$

1. Find the standard error of the above estimate

$$v(\hat{N}\_{1})=\sqrt{v\left(\hat{N}\_{1}\right)}=\sqrt{\frac{N^{2}pq}{n-1}}=\sqrt{\frac{3042^{2}\*\frac{38}{200}\*\frac{81}{100}}{199}}=84.596.$$

* 1. **ESTIMATION O F RATIOS**

 Let Y be the study variable and X the auxiliary variable which is correlated with Y. data on variable X are either readily available or can be easily collected for all units in the population.

1. The ratio estimator of the population mean $\overbar{Y}\_{R}=\frac{\overbar{y}\_{n}}{\overbar{x}\_{n}}\*\overbar{X}\_{N}$ where $\overbar{x}\_{n} and \overbar{y}\_{n}\_{n}$ are sample means for the variables x and y.
2. The estimator for the population total is $\hat{Y}\_{R}=N\overbar{Y}\_{R}$. These estimators of the population mean and total are biased.
3. The ratio of the sample means is $ \hat{R}=\frac{\overbar{y}}{\overbar{x}}$ and the ratio of the population means is $R=\frac{\overbar{Y}}{\overbar{X}}$.
4. Estimator of the variance of the ratio estimator for the mean $v(\overbar{y}\_{R})=\frac{1-f}{n(n-1)}\sum\_{}^{}(y\_{i}-rx\_{i})^{2}=\frac{1-f}{n\left(n-1\right)}(y\_{i}^{2}-2r\sum\_{}^{}y\_{i}x\_{i}+r^{2}\sum\_{}^{}x\_{i}^{2}$. where f = n/N and $r=\frac{\overbar{y}\_{n}}{\overbar{x}\_{n}} is the sample ratio.$
5. The coefficient of variation for variable x is defined as $C\_{x}=\frac{S\_{x}}{\overbar{X}\_{n}}.$

EXAMPLE: given the data below for variables x and y from a population of size 30 and the total of X is assumed to be known and given as 195, determine the ratio estimator for the total, the variance and standard error of this estimate.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 3 | 11 | 5 | 4 | 10 | 7 | 2 | 3 | 24 |
| Y | 1 | 6 | 8 | 6 | 10 | 8 | 9 | 8 | 27 |

**SOLUTION**

N =30, n= 9

1. Ratio estimator for the total $\hat{Y}\_{R}=N\overbar{Y}\_{R}=N\frac{\overbar{y}\_{n}}{\overbar{x}\_{n}}\*\overbar{X}\_{N}=\left(30\right)\left(\frac{9.22}{7.67}\right)\left(6.5\right)=234.41.$
2. Variance of the ratio estimator for the total $v\left(\hat{Y}\_{R}\right)=\frac{N\left(N-n\right)}{n\left(n-1\right)}(\sum\_{}^{}y\_{i}^{2}-2r\sum\_{}^{}x\_{i}y\_{i}+r^{2}\sum\_{}^{}x\_{i}^{2})$ =$\frac{30(30-9)}{10(10-1)}\left[1175-\left(2\*1.202\*979\right)+\left(1.202^{2}\*909\right)\right]=\frac{630}{90}\left[134.811\right]=943.677$
3. Standard error of the estimate $se(\hat{y}\_{R})=\sqrt{943.677}=30.72.$
	1. **REGRESSION ESTIMATOR**

If the relationship between variables x and y is linear but does not go through the origin, the population mean is estimated using a linear regression estimator as $y\_{lr}=\overbar{y}\_{n}+b(\overbar{X}\_{N}-\overbar{x}\_{n})$ where b is a regression coefficient given as $=\frac{\sum\_{}^{}\left(x-\overbar{x}\right)(y-\overbar{y})}{\sum\_{}^{}(x-\overbar{x})^{2}}=\frac{S\_{xy}}{S\_{x}^{2}}$ .

The estimated variance of the above regression estimator is $v\left(\overbar{y}\_{lr}\right)=\left(\frac{1}{n}-\frac{1}{N}\right)S\_{y}^{2}(1-r\_{xy}^{2})$ where r2xy is the sample correlation coefficient between y and x.

**EXAMPLE: using the example for 3.3 determine the population mean and its variance.**

**SOLUTION**

* Population mean is $\overbar{y}\_{lr}=\overbar{y}\_{n}+b(\overbar{X}\_{N}-\overbar{x}\_{n})$, where $\overbar{y}\_{n }=9.22, \overbar{X}\_{N}=6.5 and \overbar{x}\_{n}=7.67$ and the coefficient ; $b=\frac{n\sum\_{}^{}xy-\sum\_{}^{}x\sum\_{}^{}y}{n\sum\_{}^{}x^{2}- (\sum\_{}^{}x)^{2}}=\frac{\left(9\*979\right)-(69\*83)}{\left(9\*909\right)-69^{2}}=\frac{3084}{3420}=0.902.$ therefore $\overbar{y}\_{lr}=9.22+\left(0.902\right)\left(6.5-7.67\right)=8.165.$
* Estimated variance $v\left(\overbar{y}\_{lr}\right)=\left(\frac{1}{n}-\frac{1}{N}\right)S\_{y}^{2}(1-r\_{xy}^{2})$ where $s\_{y}^{2}=\frac{\sum\_{}^{}y^{2}-n\overbar{y}^{2}}{n-1}=\frac{1175-(9\*9.22^{2})}{8}=51.24$ and $r\_{xy}^{2}=\frac{n\sum\_{}^{}xy-\sum\_{}^{}x\sum\_{}^{}y}{\sqrt{\left[n\sum\_{}^{}x^{2}-(\sum\_{}^{}x)^{2}\right] [n\sum\_{}^{}y-(\sum\_{}^{}y)^{2}]}}=\frac{\left(9\*979\right)-(69\*83)}{\sqrt{\left[\left(9\*909\right)-69^{2}\right][\left(9\*1175\right)-83^{2}}}=\frac{3084}{\sqrt{3420\*3686}}=0.87.substituting v\left(\overbar{y}\_{lr}\right)=\left(\frac{1}{9}-\frac{1}{30}\right)\left(51.24\right)\left(1-0.87\right)=0.518.$
	1. **SAMPLE SIZE DETERMINATION**

Using a too large sample wastes resources and a too small sample leads to imprecise estimates. The statistical considerations for sample size determination include;

* Degree of precision aimed at.
* Variability in the population
* Number of characteristics for which specified precision is required
* Population subdivisions for which separate estimates a given precision are desired.

When sampling is done with replacement, the sample is determined using the formula; $n=(\frac{Z\_{\frac{∝}{2}S}}{E} )^{2}$ where Z is the level of confidence, S is the standard deviation and E is the tolerated maximum value of the sampling error.

If sampling is done without replacement, the sample size is obtained as; $n=\frac{NZ^{2}S^{2}}{NE^{2}+z^{2}S^{2}}$ where N is the population size and Z is the tabulated value.

**EXAMPLE**

**G**iven a population of size 750 and variance 32, estimate the sample size at 95% confidence interval if the maximum tolerable error is 2.5.

**SOLUTION**

1. If sampling is with replacement $; n=(\frac{Z\_{\frac{∝}{2}S}}{E} )^{2}=(\frac{1.96\*\sqrt{32 }}{2.5})^{2}=4.44^{2}=19.7\~20.$
2. Sampling without replacement; $n=\frac{NZ^{2}S^{2}}{NE^{2}+z^{2}S^{2}}=\frac{750\*1.96^{2}\*32}{750\*2.5^{2}+1.96^{2}\*32}=\frac{92198.4}{4810.4}=19.$
	1. **LIMITATIONS OF SIMPLE RANDOM SAMPLING DESIGN**
* The design does not take into account the size of the units in the population or the structure of the population. Failure to observe the above can lead to a random sample that is not a representative of different population subgroups.
* Use of simple random sampling when the sampling units are highly variable in size leads to less precise estimates.

**3.2 STRATIFIED RANDOM SAMPLING**

With stratified sampling, the population is divided into homogeneous, mutually exclusive groups called strata, and then independent samples are selected from each stratum.

There are three main reasons for stratification. The first is to make the sampling strategy more efficient than SRS or SYS. The second is to ensure adequate sample sizes for specific domains of interest for which analysis is to be performed. The third is to protect against drawing a ‘bad’ sample.

In order to improve the statistical efficiency of a sampling strategy with respect to SRS, there must be strong homogeneity within a stratum (i.e., units within a stratum should be similar with respect to the variable of interest) and the strata themselves must be as different as possible (with respect to the same variable of interest). Generally, this is achieved if the stratification variables are correlated with the survey variable of interest.

The **advantages** of stratified sampling are:

i. It can increase the precision of overall population estimates, resulting in a more efficient sampling strategy. A smaller sample can save a considerable amount on the survey, particularly data collection.

ii. It can guarantee that important subgroups, when defined as strata, are well represented in the sample, resulting in statistically efficient domain estimators.

iii. It can be operationally or administratively convenient.

iv. It can protect against selecting a ‘bad’ sample.

v. It allows different sampling frames and procedures to be applied to different strata (e.g., SRS in one stratum, PPS in another).

The **disadvantages** of stratified sampling are:

i. It requires that the sampling frame contain high quality auxiliary information for all units on the frame, not just those in the sample, that can be used for stratification.

ii. Frame creation is more costly and complex than for SRS or SYS, since the frame requires good auxiliary information.

iii. It can result in a sampling strategy that is less statistically efficient than SRS for survey variables that are not correlated to the stratification variables.

iv. Estimation is slightly more complex than for SRS or SYS.

* 1. **SYSTEMATIC SAMPLING (SYS)**

SYS has a number of **advantages** depending on the circumstances and objective of the survey:

* It is a proxy for SRS when there is no frame.
* It does not require auxiliary frame information, like SRS.
* It can result in a sample that is better dispersed than SRS (depending on the sampling interval and how the list is sorted).
* It has a well-established theory, just like SRS, and so estimates can be easily calculated.
* It is simpler than SRS since only one random number is required.

The **disadvantages** of SYS are:

* It can result in a ‘bad’ sample if the sampling interval matches some periodicity in the population.
* Like SRS, it does not use any auxiliary information that might be available on the frame, and thus it can result in an inefficient sampling strategy.
* The final sample size is not known in advance when a conceptual frame is used.
* It does not have an unbiased estimator of the sampling variance. In order to do variance estimation, the systematic sample is often treated as if it were a simple random sample. This is only appropriate when the list is sorted randomly.
* It can lead to a variable sample size if the population size, N, cannot be evenly divided by the desired sample size, n (but this can be avoided using circular SYS).

**3.4 Probability-Proportional-to-Size (PPS) Sampling**

Probability-proportional-to-size (PPS) sampling is one technique that uses auxiliary data and yields unequal probabilities of inclusion. If population units vary in size and these sizes are known, such information can be used during sampling to increase the statistical efficiency.

PPS can yield dramatic increases in precision if the size measures are accurate and the variables of interest are correlated with the size of the unit. For less accurate size measures, it is better to create size groupings and perform stratified sampling.

The main **advantage** of PPS sampling is that it can improve the statistical efficiency of the sampling strategy by using auxiliary information. This can result in a dramatic reduction in the sampling variance compared with SRS or even stratified sampling.

The **disadvantages** of PPS sampling are:

* It requires a survey frame that contains good quality, up-to-date auxiliary information for all units on the frame that can be used as size measures.
* It is inappropriate if the size measures are not accurate or stable. In such circumstances, it is better to create size groupings and perform stratified sampling.
* It is not always applicable, since not every population has a stable size measure that is correlated with the main survey variables.
* It can result in a sampling strategy that is less statistically efficient than SRS for survey variables that are not correlated with the size variables.
* Estimation of the sampling variance of an estimate is more complex.
* Frame creation is more costly and complex than SRS or SYS, since the size of each unit in the population needs to be measured and stored.

**Methods of PPS Sampling**

How is a PPS sample drawn? There are many PPS sampling schemes, however, three commonly used techniques are the random method, the systematic method and the randomised systematic method. (The following assumes that the size measures are integer values.)

i. The random method for PPS sampling

- for each unit in the population, cumulate the size measures for units up to and including itself.

- determine the range corresponding to each unit in the population, that is, from (but not including) the cumulative sum for the previous unit to the cumulative sum for the current unit.

- select a random number between 0 (if dealing with non-integer size measures) or 1 (for integer size measures) and the total cumulative size and select the unit whose range contains the random number.

- repeat previous step until *n* units have been selected.

To illustrate using the farm example:

**Table 5: PPS Sampling using the Random Method**

|  |
| --- |
| **Farm Size Cumulative Size Range** |
| 1 50 50 1-50 |
| 2 1000 1050 51-1050 |
| 3 125 1175 1051-1175 |
| 4 300 1475 1176-1475 |
| 5 500 1975 1476-1975 |
| 6 25 2000 1976-2000 |

For a sample containing three units, three random numbers between 1 and 2000 are selected. Suppose these numbers are: 1697, 624 and 1109. Then the farms selected are: farm 5, farm 2 and farm 3.

In the case of the random method for PPS sampling without replacement, if more than one unit is selected, complications arise both in attempting to keep probabilities directly proportional to size and in estimating the sampling variances of survey estimates. This becomes even more complicated when more

than two or three units are selected with PPS without replacement.

ii. The systematic method

this involves the following steps;

- for each unit in the population, cumulate the size measures for units up to and including itself.

- determine the range corresponding to each unit in the population, that is, from (but not including) the cumulative sum for the previous unit to the cumulative sum for the current unit.

- determine the sampling interval, *k*=(total cumulative size)/n.

- determine a random start, *r*, between 0 (if dealing with non-integer size measures) or 1 (for integer size measures) and *k.*

- select those units whose range contains the random numbers *r, r+k, r+2k, ... r+(n-1)k*.

3.5  **Cluster Sampling**

Cluster sampling is the process of randomly selecting complete groups (clusters) of population units from the survey frame. It is usually a less statistically efficient sampling strategy than SRS and is performed for several reasons. The first reason is that sampling clusters can greatly reduce the cost of collection, particularly if the population is spread out and personal interviews are conducted. The second reason is that it is not always practical to sample individual units from the population. Sometimes, sampling groups of the population units is much easier (e.g., entire households). Finally, it allows the production of estimates for the clusters themselves (e.g., average revenue per household).

Cluster sampling is a two-step process. First, the population is grouped into clusters (this may consist of natural clustering, e.g., households, schools). The second step is to select a sample of clusters and interview all units within the selected clusters.

There are a number of considerations to bear in mind for cluster sampling. In order for estimates to be statistically efficient, the units within a cluster should be as different as possible. Otherwise, if the units within a cluster are similar, they all provide similar information and interviewing one unit would be sufficient.

Unfortunately, units within a cluster frequently have similar characteristics and therefore are more homogeneous than units randomly selected from the general population. This results in a sampling procedure that is less efficient than SRS.

The statistical efficiency of cluster sampling depends on how homogeneous the units within the clusters are, how many population units are in each cluster and the number of clusters sampled. When neighbouring units are similar, it is more statistically efficient to select many small clusters rather than a few, larger clusters. However, when personal interviews are conducted, the more dispersed the sample is, the more expensive the survey. The statistical agency must strike a balance between the optimal number and size of clusters, and the cost.

The **advantages** of cluster sampling are:

* It can greatly reduce the cost of collection by having a less dispersed sample than SRS. This is particularly important when the population is spread out and personal interviews are conducted, since savings can be achieved by reducing the travel time of interviewers, especially for rural populations.
* It is easier to apply than SRS or SYS to populations that are naturally clustered (e.g., households, schools) and for certain conceptual populations, such as people crossing a border during a specific time interval. For such populations, it may be difficult, expensive or impossible to construct a list of all individual units of the population, required by SRS.
* It allows the production of estimates for the clusters themselves. For example, estimates of the average number of teachers per school (where schools are clusters).
* It can be more statistically efficient than SRS if the units within the clusters are heterogeneous (different) with respect to the study variables and the clusters are homogeneous (similar). But in practice this is usually not the case.

The **disadvantages** of cluster sampling are:

* It can be less statistically efficient than SRS if the units within the clusters are homogeneous with respect to the study variables. This is frequently the case, since units within a cluster tend to have similar characteristics. However, to offset this loss in statistical efficiency, the number of clusters selected can be increased.
* Its final sample size is not usually known in advance, since it is not usually known how many units are within a cluster until after the survey has been conducted.
* Its survey organisation can be more complex than for other methods.
* Its variance estimation will be more complex than for SRS if clusters are sampled without replacement.

**3.6 Multi-Stage Sampling**

Multi-stage sampling is the process of selecting a sample in two or more successive stages. The units selected at the first stage are called primary sampling units (PSU’s), units selected at the second stage are called second stage units (SSU’s), etc. The units at each stage are different in structure and are hierarchical (for example, people live in dwellings, dwellings make up a city block, city blocks make up a city, etc.). In two-stage sampling, the SSU’s are often the individual units of the population.

A common multi-stage sample design involves two-stage cluster sampling using an area frame at the first stage to select regions (the PSU’s) and then a systematic sample of dwellings (the SSU’s) within a region at the second stage. With the one-stage cluster sampling presented earlier, every unit within a sampled cluster is included in the sample. In two-stage sampling, only some of the units within each selected PSU are subsampled.

Each stage of a multi-stage sample can be conducted using any sampling technique. Consequently, one of the chief advantages of a multi-stage sample is its flexibility. For example, within one PSU drawn at the first stage, an SRS sample may be drawn. For another PSU, there may be a measure of size that is correlated with the key survey variables, so PPS may be used within this PSU.

The **advantages** of multi-stage sampling are:

i. It can result in a more statistically efficient sampling strategy than a one-stage cluster design when clusters are homogeneous with respect to the variables of interest (i.e., a sample size reduction).

ii. It can greatly reduce the travel time and cost of personal interviews as a result of the sample being less dispersed than for other forms of sampling, such as SRS.

iii. It is not necessary to have a list frame for the entire population. All that is needed is a good frame at each stage of sample selection.

The **disadvantages** of multi-stage sampling are:

i. It is usually not as statistically efficient as SRS (although it can be more efficient than a one-stage cluster strategy).

ii. The final sample size is not always known in advance, since it is not usually known how many units are within a cluster until after the survey has been conducted. (The sample size can be controlled, however, if a fixed number of units are selected per cluster.)

iii. Its survey organisation is more complex than for one-stage cluster sampling.

iv. Its formulas for calculating estimates and sampling variance can be complex.