Lesson 2. Collection of Data and Sampling Methods.

Collection of data constitutes the first step in statistical inquiry. To collect data means to gather data from relevant sources of data. The data needed for statistical analysis either are readily available or must be collected for the first time by the investigator. Data that are available are known as **secondary data** and data that must be collected for the first time are known as **primary data.**

Primary data are data that are collected for the first time by a particular person or an organization and are used for the specific purpose for which they were collected.

Secondary data refers to data that has already been collected, processed and used by someone else, and is being reused for another study.

Sources of Data

Sources of data are classified as:

i) Primary sources of data. Data collected from primary sources are known as primary data.

ii) Secondary sources of data. Secondary sources of data are also called available data sources. Data obtained from secondary sources are known as secondary data.

Secondary data refers to those data which have been collected earlier for some purpose other than the analysis being undertaken. Sources of secondary data can be classified as:

(i) Internal data sources – data obtained from internal sources are called internal secondary data.

(ii) External secondary data sources – data obtained from external sources are called external secondary data.

Sources of external secondary data

External secondary data refer to data that were generated from outside your organization by an outside person or organization that is not part of your organization. Sources of external secondary data include:

1 Data generated by government agencies such as Central Bank, Ministry of labour, Kenya Bureau of Statistics. e.g. Economic and demographic data collected by KBS.

2 Data obtained from Commercial sources of data: Data collected by companies specializing in collecting and maintaining data are sold to other companies in need of that information.

3 Publications of various industrial and trade associations. Trade associations collect and disseminate data that are related to their industry e.g. Kenya Chamber of Commerce and Industry, Federation of Kenyan Employers.

4 Data obtained from periodicals such as journals found in a library or websites. E.g Data can be obtained from research papers published by university departments and research bureaus.

5 Data can be obtained from media sources. Data on a broad range of subjects is available from broadcast and print media.

6 Publications of International organizations which include:
   - The international labour organisation – publishes data on employment, unemployment, workforce, salaries and wages.
   - The organisation for Economic Cooperation and Development (OECD) which publishes data on foreign trade, industry, food security, transport, science and technology.
   - The International Monetary Fund which publishes reports on national and international foreign exchange transaction, foreign trade, and economic development.

Internal Secondary data

These are the data generated within an organization in the process of routine business activities. Financial accounts, production records, quality control records and sales records are examples of such data, data on employees' salaries, age and years of experience. Data on sales, advertising expenditure, distribution costs, inventory levels and production quantities. Data on customer complaints and service records.

Methods of collecting Primary data

The most popular methods for collecting primary data are:

(i) Survey
(ii) Observation
(iii) Experimentation

Survey

A survey involves the collection of data from individuals about themselves or about the social units to which they belong. The persons from whom data are collected are known as informants or respondents. The responses of questions put to people constitute the major sources of data.

Surveys are an excellent vehicle for collecting a wide variety of unobservable data, such as people's preferences (e.g., political orientation), traits (e.g., self-esteem), attitudes (e.g., attitude toward immigrants), beliefs (e.g., about a new law), behaviors (e.g. smoking or drinking behavior), or factual information (e.g., income).

Survey can be done by using a variety of methods. The most common methods are classified as:

(i) Questionnaires
(ii) Interviews.

Questionnaires

A questionnaire is a set of questions for collecting desired data from the target respondents. A questionnaire can be:

(i) Administered personally by researcher – the researcher assists the respondent to fill the questionnaire.
(ii) Mailed to the respondent – delivered and returned by postal service.
(iii) Drop and Pick questionnaire –Delivered by hand to each respondent and collected later.
(iv) Electronically distributed questionnaire – web-based questionnaires are delivered and collected via the internet.

Interviews

An interview is a verbal interaction between the researcher and the respondents. This method involves presentation of verbal questions orally and recording oral verbal responses.

In this method, the interviewer asks questions in a face-to-face contact or by telephone to the interviewee. Therefore, there are two types of interviews: Personal interview and Telephone interview.

Formats of Questions

Two types of question formats are (1) open-ended questions, (2) closed-ended questions

In a closed-ended question, respondents are offered a set of answers from which they are asked to choose the one that most closely represents their views

Open-ended questions are not followed by any kind of specified responses, and the respondents' responses are recorded in full.

Observation Method. The investigator observes the objects he/she is interested in and records the desired data without asking any questions. For example, an investigator can observe and record the number of people entering a supermarket at a given time. Observation may be participative observation or non-participative observation. Participative observation means that the investigator participates in the activities or the group he is set to investigate. His cover is so complete that as far as other participants are concerned, he is simply one of the group members. A non-participative observer on the other hand stands aloof from the group activities he is observing.

EXPERIMENTATION METHOD

An **experiment** refers to an investigation in which a factor or variable under investigation is manipulated and its effect on another variable called the dependent variable measured. In experiments, the purpose is to identify cause and effect relationships between variables. There are two key variables in an experiment (1) the independent variable, or treatment, and (2) the dependent variable. Persons or objects receiving a treatment are said to be in an experimental group, while those not exposed are in the control group.

CENSUS METHOD VS SAMPLE METHOD

Data can be collected from every member of the population (i.e. use census method/conduct census inquiry) or from only a part of the population (i.e. use sample method/conduct sample inquiry). A census inquiry is an examination of every member of the population i.e. data are collected from every member of the population. Sample inquiry is an examination of a small group of elements selected from the population to represent the entire population.

Why a sample may be preferred to census

- When the population is very large or infinite. A census is not possible if the population size is very large or infinite e.g. an assembly line can keep producing items, a doctor can keep seeing more and more patients.
- If data collection involves destructive testing, then a complete census is out of the question because every item would be destroyed e.g. vehicle crash tests.
- A sample may be preferred when results of the study are needed urgently. Sampling may yield more timely results than a complete census.
- A sample may be preferred when a high degree of accuracy is required. In practice, a sample can be more accurate than a census. Sampling can contribute to accuracy by reducing non-sampling errors.
- A sample is preferred when the cost of conducting a census is prohibitive i.e even if a census is feasible, the cost may exceed our budget.
- A sample is preferred when a researcher wants to get more detailed information about elements of a population.

Methods of Sampling

The term sample refers to a small group selected from the population to represent the entire population. Sampling is the process of selecting a sample from the population.

There are two main categories of sampling methods.

(i) Random Sampling Methods – also called Probability Sampling methods.

(ii) Non-random Sampling Methods – Also called Non-Probability Sampling methods.

Random Sampling Methods

In Probability Sampling, each element in the population has a known or calculable chance of being included in the sample. However, every person or element may not have an equal chance for inclusion.

In probability sampling, elements are chosen by a chance procedure so as to produce a sample that is representative of the population. Such a sample is a random sample and there is a high chance that it is representative of the entire population.

The first step in selecting a random sample from a finite population is to establish a **sampling frame**. A sampling frame is a list of all members of the population. Each member of the population is given a number, then some random method is used to select numbers and the sample members are thus identified.

There are four random sampling techniques:

(i) Simple random sampling

(ii) Systematic random sampling

(iii) Stratified random sampling

(iv) Cluster random sampling.

(i) Simple random sampling

In Simple Random Sampling, a sample is chosen from the population so that every element in the population has the same chance of being included in the sample.

Methods used to select a random sample

a) Lottery method

b) Use of random numbers.

(ii) Systematic Random Sampling

Systematic Sampling involves the selection of every Kth element from a list, after the first element is randomly selected from among the first K elements in the list. K is the ratio of population size to sample size and is called sampling interval or periodicity e.g. choosing every 6^{th} house on a street or selecting every 50^{th} item from assembly line to test for defects.

(iii) Stratified Random Sampling

In stratified sampling, the population is divided into a number of homogeneous and non-overlapping subgroups, called strata, and then a simple random sample is drawn from each subgroup/stratum. Stratified random sampling allows the comparison of population subgroups. Members of one particular stratum have similar characteristics. There are substantial differences between members of different strata. Stratification improves the representativeness of a sample.

(iv) Cluster Sampling

A cluster is defined as a group of elements located near one another, such as residents of the same neighbourhood. Clusters are defined as mutually exclusive groups, each of which is representative of the population. A cluster sample is obtained by separating the population into subpopulations called clusters and then selecting clusters by simple random sampling. After the clusters have been selected we can include every element in each selected cluster or use simple random sampling to select elements from the clusters.

Non probability sampling

Also called non-random sampling

In non-probability sampling, not every element in the population has a chance of being included in the sample, and the process involves at least some degree of personal subjectivity instead of following predetermined probabilistic rules for selection. Non-random Sampling Methods include Quota Sampling, Convenience Sampling, Judgement Sampling and snowball sampling.

Convenience sampling

Convenience sampling involves selection of members of the population who are readily accessible. Sample members are participants who are easiest to reach, available, or willing to take part in the study. For example, selecting the first 30 customers to enter a supermarket would comprise a convenience sample.

Judgemental sampling – also called purposive sampling

Judgmental sampling is a non-probability sampling method where the researcher selects the sample members based on his/her knowledge, expertise, or judgment about which individuals or items will provide the most useful data. Judgemental sampling involves the selection of elements that, according to the judgement of the sampler, accurately reflect the population. In judgemental sampling, personal knowledge and opinion are used to identify the items from the population that are to be included in the sample. For example, choosing experienced doctors instead of the general public when investigating common diseases in a given region.

Quota sampling

This type of sampling ensures that certain groups in the population are adequately represented in the sample through the assignment of a quota for each group. Generally, the quota fixed for each group is based on the total number of each group in the population. The sampler arbitrarily selects a predetermined number of elements from each segment of the population. For example, in a survey of 80 people, a researcher might want 46 women and 34 men in the sample to reflect the population but the actual selection is not performed in a random manner.

Snowball sampling

Snowball sampling is a non-probability sampling method where existing participants help the researcher recruit more participants from among their acquaintances. The sample size "snowballs" (grows) as each new participant refers others. So, the researcher starts with a few participants and uses their networks to recruit others. For example, for hard-to-reach populations such as drug users, the researcher may identify one or two respondents who introduces the researcher to other drug users.

Errors in Survey Research

There are two types of survey errors: sampling error and non-sampling errors.

Sampling error

Sampling error occurs because a sample has been taken instead of a complete census of the population. Sampling error is intrinsic to all sample surveys. Sampling error arises from estimating a population characteristic by measuring only a portion of the population rather than the entire population. Sampling error is a random error. It can also be described as non-directional or non-systematic, because measurements exhibiting random error are just as likely to be too high as they are to be too low. A census has no sampling error since all members of the population are enumerated.

Non-sampling errors.
Include response error, nonresponse error and processing errors.
Response errors. Some respondents may distort the truth when answering a question. They may exaggerate their income, understate their age, or provide answers they think are acceptable.

Nonresponse error

Not everyone in the sample will cooperate in returning the questionnaire or in answering an interviewer's questions. This would not be a problem, except that those who respond may be different from those who don't.