DATA COLLECTION
TYPES, SAMPLING AND ERRORS

manupatra

Data collection refers to the process whereby the researcher collects various information required to analyze the research problem. This stage of data collection begins after the researcher has defined the research problem and has also decided the research design.

The method for collection of data depends on various factors such as:

- The availability of resources
- Credibility of the source from where the information is derived
- Analysis and reporting
- Researcher's skills

3.1 Types of Data

The type of data collected for research is generally categorized in two types i.e. Primary and Secondary data. The table below briefly describes and compares both types of data.

	Primary Data	Secondary Data
Meaning	It refers to original data collected by the researcher himself to undertake the research work.	It refers to the data collected by someone else earlier.
Data	It is real time data	It is past data
Process	Involved and Time Taking	Easy and Quick
Source	Includes Surveys, Experiments, Questionnaires, Interview etc.	Involves Government publications, websites, Reports, Journals, Internal Records etc.
Cost	Expensive	Economical
Accuracy	Highly accurate as it is specifically collected for specific research problem	Less accurate as it is past data collected by someone else
Reliability	High	Low
Form	Available in raw/unorganized form	Refined and Organized
Focus	Always focused on the specific needs of the research	May not be specific as such data is not specific to the researcher's need

3.1.1 Important Methods of Primary Data Collection

As discussed above, primary data forms the most accurate and reliable form of data in research and is generally preferred by the researcher where there are no cost and time limitations. There are a few important methods of such data collection such as:

- **Observation method**

Observation method refers to the method wherein the researcher collected the data with the help of observations by personally doing the fieldwork. It can be understood as a systematic viewing coupled with a consideration of seen phenomenon. Example: Observations being conducted in a particular place and recording it via video and photographs. This method helps in producing large quantities of data and is generally very economical. However, it is often difficult to extract a lot of information through this method as compared to other methods.

	Type	Definition
Classification - 1	Structured Observation	When there is a careful definition of units/people observed, the method of recording and standardized conditions of observation.
	Unstructured Observation	Where the researcher does not prescribe any standardization of observations or definition of unit/people being observed.
Classification - 2	Participant Observation	When the researcher is a part/member of the group he is observing.
	Non-Participant Observation	When the researcher observed people without any information to them.
Classification - 3	Un-Controlled Observation	When observations takes place in natural conditions to get spontaneous responses, pictures etc
	Controlled Observation	Observations take place as per pre-arranged plans with certain experimental procedures under controlled conditions.

- ## **Questionnaires**

In this method, a questionnaire is sent to the person (in person or online or via post/mail) requesting him to answer questions in the questionnaire and return the same to the researcher. It consists of various questions in a specific and definite order.

The characteristics of a good questionnaire are:

- It must be short and simple
- The flow of questions must be from easy to difficult
- Use of technical expressions must be avoided or must be minimum
- The space provided to answer the questions must be appropriate and sufficient
- Sequence of questions must be clear and logical

	Type	Definition
Classification - 1	Open Ended Questions	Such questions provide the person the freedom and ability to respond to the question in his own words and manner he would choose.
	Close Ended Questions	Such questions provide choice to the person to answer the questions from the pre-decided alternatives. Such alternatives can also be referred to as MCQs.

- ## **Schedules**

This method is quite similar to Questionnaire method of data collection. However, in case of schedule, the questions are filled by the enumerator, who is appointed by the person/agency for the very purpose.

In this method, the enumerator goes to people, asks the questions in the questionnaire and records the answers/responses by himself. He is a trained professional appointed specifically for administering the schedule.

Example: Census conducted by the Government. A government agent or Enumerator asks questions and fills in the responses based on the details provided by people.

- **Interview Method**

This method of data collection involves presentation of oral/verbal replies in terms of oral/verbal responses. In this process, the interviewer asks questions aimed at getting information required for the researcher

This method is particularly helpful in getting more information at greater depth as the interviewer is often skillful and can overcome the resistance from the person being interviewed. However, the method can be expensive and also often suffers from various interviewer and respondent biases

Type	Definition
Structured	Consists of pre-decided set of questions
Unstructured	Interviewer does not follow the system of pre-determined set of questions
Group	When a group of individuals is interviewed
Individual	When interviewer meets an individual and interviews him
Selection	Interviews done for selection of people for a job.

3.2. Sampling

Sampling can be defined as process wherein a predetermined number of observations or responses are recorded from a larger population. Sampling is very important for various reasons, such as:

- It is economical and saves time
- It helps in gaining accuracy and controlling the margin of error in data collection
- It is practical and easily approachable

There are various terms associated with the process of sampling. The table below highlights some of the important terms.

Term	Definition
Population	It refers to the entire group which is being studied via the research.
Parameter	It measures and describes the whole population.
Sample	It is the group of people selected from the entire population. The data representing the entire population will be collected from this sample.
Sampling Frame	It is the group of people selected from the entire population. The data representing the entire population will be collected from this sample.
Sampling Design	It is the source material from which the sample is drawn. It is the list of all those who can be sampled and may include categories such as people, institutions, households etc.

It is the group of people selected from the entire population. The data representing the entire population will be collected from this sample.

3.2.1 Probability Sampling Method

In this sampling method, every unit in the entire population has equal chance of being included in the sample. This method is less prone to errors and is more accurate. The various types of sampling under this method are:

• Simple Random Sampling

In this method, the units are selected at the same time and independently of each other. The same is done by assigning a random number to each unit and selection is done by choosing a random number. This helps in reducing subjective bias of the researcher.

- ## Stratified Random Sampling

  In this method, the population is divided into smaller groups and samples are drawn from each group separately. The groups created are a mutually exclusive stratum which ensures that groups don't overlap but together they represent the entire population.

  Generally, the groups are classified based on gender, age, ethnicity, profession, income etc. Members of each group must be distinct and should have an equal chance of being selected for sampling using simple random sampling.

- ## Systematic Sampling

  In this method, researcher makes a pre-determined sequential choice from the population to be part of the sample.

  Example: It is decided that every 5th person in the population will be selected for the sample

- ## Cluster Sampling

  In this method, the random selection of the sample from the population is done when the units are naturally spread out in the population.

  Example: If the population consists of all students from the university, the naturally spread groups can be formed based on the academic course they are pursuing (Law, Engineering, MBA etc.)

- ## Multistage Sampling

  Also known as Multi-stage cluster sampling, it is a more complex form of cluster sampling which contains multiple stages in sample selection. In this process, the sample size is smaller at every subsequent stage. Initially, a primary sample is selected and subsequently, a secondary sample is selected from the primary sample. The last sample collected will be used in the research and analysis.

  Example: A researcher divided the population into 50 groups and 100 samples from each group were selected. In subsequent sample, the researcher selected 50 samples from each group from the 100 originally selected in previous stage.

manupatra

3.2.2 Non-Probability Sampling Method

In this method, every unit in the population does not have equal chances and one cannot estimate the chances of a unit being included in the sampling process. The selection is done on non-random criteria and therefore not every unit has a chance of being selected in the sample.

It is an easier and more economical approach but often fails to make valid statistical inferences about the entire population.

- ## Snowball Sampling

When the population is hard to access, this method can be used to select participants via other participants to get in contact with more and more people. One participant gives contact of a new person and so on.

Example: Alike the concept shown in the bollywood movie – Jai Ho, where the actor says that instead of saying Thank You, pass the help to three people, thereby forming a long chain of people.

- ## Judgment Sampling

In this sampling, the researcher uses his judgment to select the sample which is according to him most appropriate and useful to the research. This type of sampling is often done in quantitative research where the research requires detailed knowledge about a particular issue. It is also known as deliberate/purposive sampling.

Example: In a research to reflect the extent of misuse of a particular law, the researcher chooses various categories of participants such as lawyers, prosecutors, judges and notable academicians to understand the issue rather than recording the sample of general public.

- ## Convenience Sampling

Also known as Accidental Sampling, it includes the individuals who are most accessible to the researcher. This is often done to obtain a large number of samples quickly and economically, but there is no link to establish if this sample is representative of the entire population and therefore fails to generalize the results obtained.

- ## Voluntary Response Sampling

Similar to Convenience Sampling, in this method instead of researcher choosing individuals and directly contacting them, individuals themselves volunteer and respond with the data.

Example: online public survey

- ## Quota Sampling

In this method, the researcher creates a sample involving individuals who fit within certain criteria (for example: Age, Gender, Income etc). This type of sampling ensures that samples satisfy all the characteristics in the market being researched and therefore are representative of the entire population.

Example: An alcohol manufacturing company can undertake study to research what brand of alcohol is preferred in particular city by particular age groups. Here using quota sampling age groups can be formed such as 21-30, 31-40, 41-50 and 50+. This will help the company in analyzing what age group prefers which brand in that particular city.

- ## Dimensional Sampling

This method is similar to and an extension of Quota Sampling, where researcher takes into consideration various characteristics of the population (Age, Income, Gender, Education etc) and have to ensure that the sample collected represents people from each of those characteristics or group.

3.3 Errors

Error is an action in the sampling which is inaccurate or incorrect which occurs during collection of data or analysis.

There are two types of error, **i.e. Sampling Errors and Non-Sampling Errors.** Sampling error refers to statistical errors that occur when the researcher select such a sample which does not represent the entire population and therefore, the results found or obtained through the sample collected does not truly represent the population regarding which the research was undertaken.

On the other hand, a Non-Sampling error is a human error which occurs due to several aspects, most common of which is the error in the data entry. Other such errors include issues in the questions, false information, wrong analysis, mathematical mistake etc.

- Voluntary Response Sampling

In addition to it, there is another category of errors i.e. Response and Non-Response Errors. Response error occurs when the respondent or participant provides inaccurate information or if the information recorded is wrong.

A non-response error generally occurs when people are provided with an option to participate but decide not to, therefore, the results are not included in the data. It is the difference between the people who choose to participate in comparison to the people who did not participate.

NOTE: ERROR v. SAMPLE SIZE

Error and Sample size have an inverse relationship and therefore as the size of the sample increases, the margin of error decreases. It means that the more information is collected by the researcher, the chances of getting more accurate results increases.