UNIT II - DATA COLLECTION AND SOURCES

What is data collection?

Data collection is a systematic method of collecting and measuring data gathered from different sources of information in order to provide answers to relevant questions. An accurate evaluation of collected data can help researchers predict future phenomenon and trends.

Data collection can be classified into two, namely: primary and secondary data. Primary data are raw data i.e. fresh and are collected for the first time. Secondary data, on the other hand, are data that were previously collected and tested.

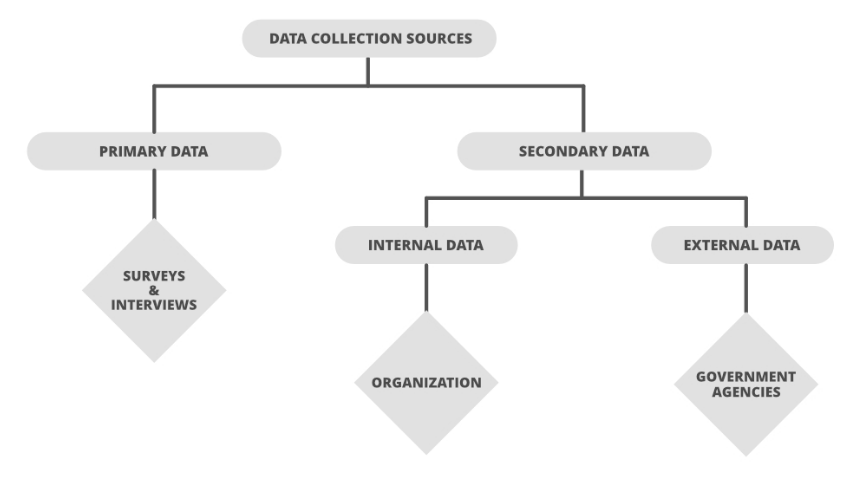
Therefore, data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities.

Data collection is a standout amongst the most essential stages in carrying on a research. You can have the best research plan in the world, however, in the event that you can’t gather the necessary data you will not have the capacity to complete your venture. Data collection is an extremely challenging work which needs exhaustive planning, diligent work, understanding, determination and more to have the capacity to complete the assignment effectively. Data collection begins with figuring out what sort of data is needed, followed by the collection of a sample from a certain section of the population.

Methods of data collection

The system of data collection is based on the type of study being conducted. Depending on the researcher’s research plan and design, there are several ways data can be collected.

The sources of data can be classified into two types: statistical and non-statistical. Statistical sources refer to data that is gathered for some official purposes, incorporate censuses, and officially administered surveys. Non-statistical sources refer to the collection of data for other administrative purposes or for the private sector.



What are the different sources of data?

The following are the two sources of data:

1. **Internal sources**

* When data is collected from reports and records of the organization itself, they are known as the internal sources.
* For example, a company publishes its annual report’ on profit and loss, total sales, loans, wages, etc.

1. **External sources**

* When data is collected from sources outside the organization, they are known as the external sources. For example, if a tour and travel company obtains information on Karnataka tourism from Karnataka Transport Corporation, it would be known as an external source of data.

**Types of Data**

**A) Primary data**

* Primary data means first-hand information collected by an investigator.
* It is collected for the first time.
* It is original and more reliable.
* For example, the population census conducted by the government of India after every ten years is primary data.

Primary Data Collection

* Interviews.

The researcher asks questions of a large sampling of people, either by direct interviews or means of mass communication such as by phone or mail. This method is by far the most common means of data gathering.

* Projective Technique.

Projective data gathering is an indirect interview, used when potential respondents know why they're being asked questions and hesitate to answer. For instance, someone may be reluctant to answer questions about their phone service if a cell phone carrier representative poses the questions. With projective data gathering, the interviewees get an incomplete question, and they must fill in the rest, using their opinions, feelings, and attitudes.

* Delphi Technique.

The Oracle at Delphi, according to Greek mythology, was the high priestess of Apollo’s temple, who gave advice, prophecies, and counsel. In the realm of data collection, researchers use the Delphi technique by gathering information from a panel of experts. Each expert answers questions in their field of specialty, and the replies are consolidated into a single opinion.

* Focus Groups.

Focus groups, like interviews, are a commonly used technique. The group consists of anywhere from a half-dozen to a dozen people, led by a moderator, brought together to discuss the issue.

* Questionnaires.

Questionnaires are a simple, straightforward data collection method. Respondents get a series of questions, either open or close-ended, related to the matter at hand.

**B) Secondary data**

* Secondary data refers to second-hand information.
* It is not originally collected and rather obtained from already published or unpublished sources.
* For example, the address of a person taken from the telephone directory or the phone number of a company taken from Just Dial are secondary data.

Secondary Data Collection

Unlike primary data collection, there are no specific collection methods. Instead, since the information has already been collected, the researcher consults various data sources, such as:

* Financial Statements
* Sales Reports
* Retailer/Distributor/Deal Feedback
* Customer Personal Information (e.g., name, address, age, contact info)
* Business Journals
* Government Records (e.g., census, tax records, Social Security info)
* Trade/Business Magazines
* The internet

**Data Collection Tools**

Now that we’ve explained the various techniques, let’s narrow our focus even further by looking at some specific tools. For example, we mentioned interviews as a technique, but we can further break that down into different interview types (or “tools”).

* Word Association.

The researcher gives the respondent a set of words and asks them what comes to mind when they hear each word.

* Sentence Completion.

Researchers use sentence completion to understand what kind of ideas the respondent has. This tool involves giving an incomplete sentence and seeing how the interviewee finishes it.

* Role-Playing.

Respondents are presented with an imaginary situation and asked how they would act or react if it was real.

* In-Person Surveys.

The researcher asks questions in person.

* Online/Web Surveys.

These surveys are easy to accomplish, but some users may be unwilling to answer truthfully, if at all.

* Mobile Surveys.

These surveys take advantage of the increasing proliferation of mobile technology. Mobile collection surveys rely on mobile devices like tablets or smartphones to conduct surveys via SMS or mobile apps.

* Phone Surveys.

No researcher can call thousands of people at once, so they need a third party to handle the chore. However, many people have call screening and won’t answer.

* Observation.

Sometimes, the simplest method is the best. Researchers who make direct observations collect data quickly and easily, with little intrusion or third-party bias. Naturally, it’s only effective in small-scale situations.

Methods of Collecting Primary Data

1. Direct personal investigation
2. Indirect oral investigation
3. Information through correspondents
4. Telephonic interview
5. Mailed questionnaire
6. The questionnaire filled by enumerators

**Measurements**

Measurement is the process of observing and recording the observations that are collected as part of a research effort. There are two major issues that will be considered here.

First, you have to understand the fundamental ideas involved in measuring. Here we consider two of major measurement concepts. In Levels of Measurement, I explain the meaning of the four major levels of measurement: nominal, ordinal, interval and ratio. Then we move on to the reliability of measurement, including consideration of true score theory and a variety of reliability estimators.

Second, you have to understand the different types of measures that you might use in social research. We consider four broad categories of measurements. Survey research includes the design and implementation of interviews and questionnaires. Scaling involves consideration of the major methods of developing and implementing a scale. Qualitative research provides an overview of the broad range of non-numerical measurement approaches. And unobtrusive measures presents a variety of measurement methods that don’t intrude on or interfere with the context of the research.

Measurement and Scale in Research Methodology. Measurement is the process of describing some property of a phenomenon under study and assigning a numerical value to it. Measurement is considered as the foundation of scientific inquiry. In our daily life, many things are measured continuously in different ways for different purposes.

We can not only measure physical objects but abstract objects also, that means we can measure quantitatively and qualitatively. We can measure height, weight, length, width, income etc., (quantitative measurement) and at the same time, we can measure attitude, personality, perception, intelligence, preference (qualitative measurement) etc. A measurement can give us different kinds of information about a theoretical concept under study.

A more contemporary definition of measurement as “the estimation or the discovery of the ratio of some magnitude of a quantitative attribute to a unit of the same attribute” (Michell, 1997).

According to Warren S Torgerson “The assignment of numbers to objects to represent amounts or degrees of a property possessed by all of the objects.” To understand the nature of the data, we must know at which level the data is measured. So the measurement can occur at different levels, and the relationship among the values assigned determines the level of measurement. There are four hierarchical levels of measurement identified by Stevens (1946); they are nominal, ordinal, interval, and ratio.

Nominal Scale

This is a method of measuring the objects or events into a discrete category.   This is regarded as the most basic form of measurement.  Here we assign a number to an object only for the identification of the object.   So it is a categorical data or qualitative data.  Here the numbers are only used for labeling the object, and there is no quantitative value at all.  This is used to categories the data into different groups.  In a survey, all the respondence are divided into different categories, which should be mutually exclusive and collectively exhaustive. Here the categories have no predefined order.

Examples of nominal scale data connection using a questionnaire.

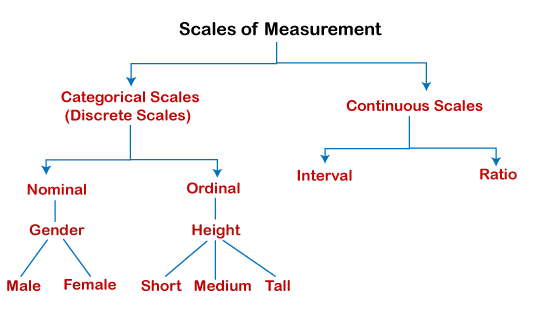
1. **Specify your gender**
   * + **Male**
     + **Female**
2. **Are you married?**
   * + **Yes**
     + **No**
3. **You are from**
   * + **Urban**
     + **Rural**
4. **Specify your working department**
   * + **Marketing**
     + **HR**
     + **Finance**
     + **Sales**
     + **Production**
     + **Operations**
5. **Specify your food habit**

* Vegetarian
* No-Vegetarian

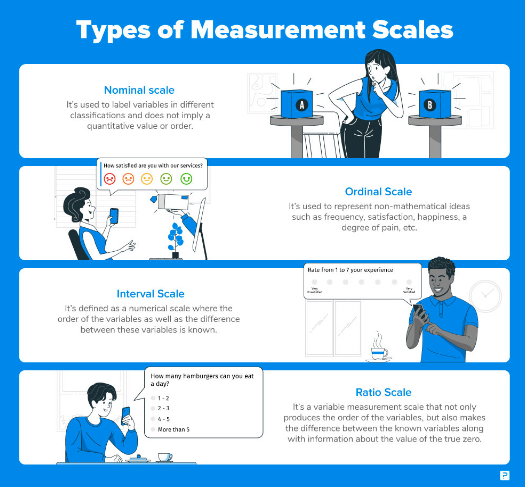
Here we can assign number to each option like 1 to Male and 2 to female, and 1 to Yes, and 2 to No, 1 to Urban, 2 to Rural, 1 to Marketing, 2 to HR, 3 to Finance etc.

Here these numbers have no quantitative values; they only represent the category.  So we cannot apply any arithmetic operations in this type of sale.  We can only count the number of items in each category.

Here we can prepare a frequency distribution table for representing this nominal data.



1. **Nominal Scales**: Nominal scale is the simplest form of measurement. A variable measured on a nominal is one which is divided into two or more categories, for example, gender is categorized as male or female, a question as to whether a family owns a iPhone can be answered ‘Yes’ or ‘No’. It is simply a sorting operation in which all individuals or units or answers can be placed in one category or another (i.e. the categories are exhaustive). The essential characteristic of a nominal scale is that in terms of a given variable, one individual is different from another and the categories are discriminate (i.e. the categories are mutually exclusive). This characteristic of classification if fundamental to all scales of measurement. Nominal scales that consist only two categories such as female-male, agree-disagree,aware-unaware, yes-no, are unique and are called dichotomous scales. Such dichotomous nominal scales are important to researchers because the numerical labels for the two scale categories can be treated as though they are of interval scale value.
2. **Ordinal Scales**: Ordinal scales have all the properties of a nominal scale, but, in addition, categories can be ordered along a continuum, in terms of a given criterion. Given three categories A, B and C, on an ordinal scale, one might be able to say, for e.g., that A is greater than B and B is greater than C. If numerals are assigned to ordinal scale categories, the numerals serve only as ranks for ordering observations from least to most in terms of the characteristic measured and they do not indicate the distance between scale that organizes observations in terms of categories such as high, medium and low or strongly agree, agree, not sure, disagree, and strong disagree.
3. **Interval Scales**: Interval scales incorporate all the properties of nominal and ordinal scales and in addition, indicate the distance or interval between the categories. In formal terms one can say not only that A is greater than B and B is greater than C but also that (A-B)=(B-C) or (A-C)=(A-B)+(B-C). Examples of interval scale include age, income and investments. However, an interval scale is one where there is no absolute zero point. It can be placed anywhere along a continuum e.g., the age can be between 20 to 60 years and need not necessarily start from 0 years. This makes ratio comparison, that A is twice that of B or so wrong.
4. **Ratio Scales**: A special form of interval scale is the ratio scale which differs in that it has a true zero point or a point at which the characteristic that is measured is presumed to be absent. Examples of ratio scales include, weight, length, income, expenditure and others. In each there is a concept of zero income, zero weight, etc. Since ratio scales represent a refinement of interval scales, generally these scales are not distinguished and both the terms are used inter-changeably.



What is a Questionnaire?

A questionnaire is a research instrument that consists of a set of questions or other types of prompts that aims to collect information from a respondent. A research questionnaire is typically a mix of close-ended questions and open-ended questions.

Open-ended, long-form questions offer the respondent the ability to elaborate on their thoughts. Research questionnaires were developed in 1838 by the Statistical Society of London.

The data collected from a data collection questionnaire can be both qualitative as well as quantitative in nature. A questionnaire may or may not be delivered in the form of a survey, but a survey always consists of a questionnaire.

Questionnaires can be classified as both, quantitative and qualitative method depending on the nature of questions. Specifically, answers obtained through closed-ended questions (also called restricted questions) with multiple choice answer options are analyzed using quantitative methods. Research findings in this case can be illustrated using tabulations, pie-charts, bar-charts and percentages.

Answers obtained to open-ended questionnaire questions (also known as unrestricted questions), on the other hand, are analyzed using qualitative methods. Primary data collected using open-ended questionnaires involve discussions and critical analyses without use of numbers and calculations.

There are following types of questionnaires:

Computer questionnaire. Respondents are asked to answer the questionnaire which is sent by mail. The advantages of the computer questionnaires include their inexpensive price, time-efficiency, and respondents do not feel pressured, therefore can answer when they have time, giving more accurate answers. However, the main shortcoming of the mail questionnaires is that sometimes respondents do not bother answering them and they can just ignore the questionnaire.

1. Telephone questionnaire. Researcher may choose to call potential respondents with the aim of getting them to answer the questionnaire. The advantage of the telephone questionnaire is that, it can be completed during the short amount of time. The main disadvantage of the phone questionnaire is that it is expensive most of the time. Moreover, most people do not feel comfortable to answer many questions asked through the phone and it is difficult to get sample group to answer questionnaire over the phone.
2. In-house survey. This type of questionnaire involves the researcher visiting respondents in their houses or workplaces. The advantage of in-house survey is that more focus towards the questions can be gained from respondents. However, in-house surveys also have a range of disadvantages which include being time consuming, more expensive and respondents may not wish to have the researcher in their houses or workplaces for various reasons.
3. Mail Questionnaire. This sort of questionnaires involve the researcher to send the questionnaire list to respondents through post, often attaching pre-paid envelope. Mail questionnaires have an advantage of providing more accurate answer, because respondents can answer the questionnaire in their spare time. The disadvantages associated with mail questionnaires include them being expensive, time consuming and sometimes they end up in the bin put by respondents.
4. Open question questionnaires. Open questions differ from other types of questions used in questionnaires in a way that open questions may produce unexpected results, which can make the research more original and valuable. However, it is difficult to analyze the results of the findings when the data is obtained through the questionnaire with open questions.
5. Multiple choice questions. Respondents are offered a set of answers they have to choose from. The downsize of questionnaire with multiple choice questions is that, if there are too many answers to choose from, it makes the questionnaire, confusing and boring, and discourages the respondent to answer the questionnaire.
6. Dichotomous Questions. These type of questions gives two options to respondents – yes or no, to choose from. It is the easiest form of questionnaire for the respondent in terms of responding it.
7. Scaling Questions. Also referred to as ranking questions, they present an option for respondents to rank the available answers to questions on the scale of given range of values (for example from 1 to 10).

Questionnaires as primary data collection method offer the following advantages:

* Uniformity: all respondents are asked exactly the same questions
* Cost-effectiveness
* Possibility to collect the primary data in shorter period of time
* Minimum or no bias from the researcher during the data collection process
* Usually enough time for respondents to think before answering questions, as opposed to interviews
* Possibility to reach respondents in distant areas through online questionnaire

At the same time, the use of questionnaires as primary data collection method is associated with the following shortcomings:

* Random answer choices by respondents without properly reading the question.
* In closed-ended questionnaires no possibility for respondents to express their additional thoughts about the matter due to the absence of a relevant question.
* Collecting incomplete or inaccurate information because respondents may not be able to understand questions correctly.
* High rate of non-response

Advantages of a good design

* With a survey questionnaire, you can gather a lot of data in less time.
* There is less chance of any bias creeping if you have a standard set of questions to be used to your target audience. You can apply logic to questions based on the respondents’ answers, but the questionnaire will remain standard for a group of respondents that fall in the same segment.
* Surveying online survey software is quick and cost-effective. It offers you a rich set of features to design, distribute, and analyze the response data.
* It can be customized to reflect your brand voice. Thus, it can be used to reinforce your brand image.
* The responses can be compared with the historical data and understand the shift in respondents’ choices and experiences.
* Respondents can answer the questionnaire without revealing their identity. Also, many survey software complies with significant data security and privacy regulations.

Steps Involved in Questionnaire Design

1. Identify the scope of your research:

Think about what your questionnaire is going to include before you start designing the look of it. The clarity of the topic is of utmost importance as this is the primary step in creating the questionnaire. Once you are clear on the purpose of the questionnaire, you can begin the design process.

2. Keep it simple:

The words or phrases you use while writing the questionnaire must be easy to understand. If the questions are unclear, the respondents may simply choose any answer and skew the data you collect.

3. Ask only one question at a time:

At times, a researcher may be tempted to add two similar questions. This might seem like an excellent way to consolidate answers to related issues, but it can confuse your respondents or lead to inaccurate data. If any of your questions contain the word “and,” take another look. This question likely has two parts, which can affect the quality of your data.

4. be flexible with your options:

While designing, the survey creator needs to be flexible in terms of “option choice” for the respondents. Sometimes the respondents may not necessarily want to choose from the answer options provided by the survey creator. An “other” option often helps keep respondents engaged in the survey.

5. The open-ended or closed-ended question is a tough choice:

The survey creator might end up in a situation where they need to make distinct choices between open or close-ended questions. The question type should be carefully chosen as it defines the tone and importance of asking the question in the first place.

If the questionnaire requires the respondents to elaborate on their thoughts, an open-ended question is the best choice. If the surveyor wants a specific response, then close-ended questions should be their primary choice. The key to asking closed-ended questions is to generate data that is easy to analyze and spot trends.

6. It is essential to know your audience:

A researcher should know their target audience. For example, if the target audience speaks mostly Spanish, sending the questionnaire in any other language would lower the response rate and accuracy of data. Something that may seem clear to you may be confusing to your respondents. Use simple language and terminology that your respondents will understand, and avoid technical jargon and industry-specific language that might confuse your respondents.

For efficient market research, researchers need a representative sample collected using one of the many sampling techniques, such as a sample questionnaire. It is imperative to plan and define these target respondents based on the demographics required.

7. Choosing the right tool is essential:

Question Pro is a simple yet advanced survey software platform that the surveyors can use to create a questionnaire or choose from the already existing 300+ questionnaire templates.

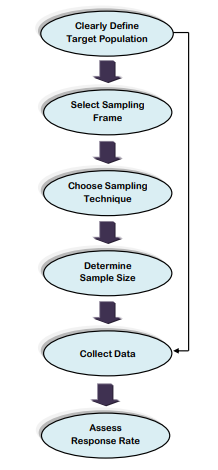
Always save personal questions for last. Sensitive questions may cause respondents to drop off before completing. If these questions are at the end, the respondent has had time to become more comfortable with the interview and are more likely to answer personal or demographic questions.

Differences between a Questionnaire and a Survey

|  |  |  |
| --- | --- | --- |
|  | Questionnaire | Survey |
| Meaning | A questionnaire can is a research instrument that consists of a set of questions to collect information from a respondent. | A survey is a research method used for collecting data from a pre-defined group of respondents to gain information and insights on various topics of interest. |
| What is it? | The instrument of data collection | Process of collecting and analyzing that data |
| Characteristic | Subset of survey | Consists of questionnaire and survey design, logic and data collection |
| Time and Cost | Fast and cost-effective | Much slower and expensive |
| Use | Conducted on the target audience | Distributed or conducted on respondents |
| Questions | Close-ended and very rarely open-ended | Close-ended and open-ended |
| Answers | Objective | Subjective or objective |

**Sampling and methods**

In order to answer the research questions, it is doubtful that researcher should be able to collect data from all cases. Thus, there is a need to select a sample. The entire set of cases from which researcher sample is drawn in called the population. Since, researchers neither have time nor the resources to analysis the entire population so they apply sampling technique to reduce the number of cases. Figure 1 illustrates the stages that are likely to go through when conducting sampling.



. Stage 1: Clearly Define Target Population The first stage in the sampling process is to clearly define target population. Population is commonly related to the number of people living in a particular country.

Stage2: Select Sampling Frame A sampling frame is a list of the actual cases from which sample will be drawn. The sampling frame must be representative of the population.

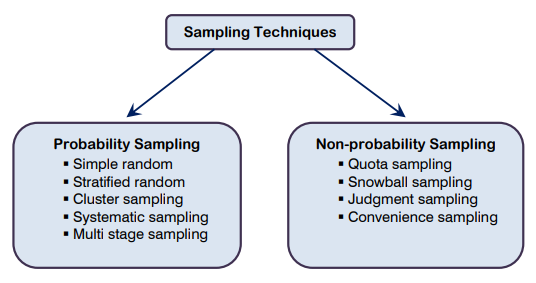
Stage 3: Choose Sampling Technique Prior to examining the various types of sampling method, it is worth noting what is meant by sampling, along with reasons why researchers are likely to select a sample. Taking a subset from chosen sampling frame or entire population is called sampling. Sampling can be used to make inference about a population or to make generalization in relation to existing theory. In essence, this depends on choice of sampling technique.

In general, sampling techniques can be divided into two types:

♣ Probability or random sampling

♣ Non- probability or non- random sampling

Before choosing specific type of sampling technique, it is needed to decide broad sampling technique. Figure 2 shows the various types of sampling techniques.



1. Probability Sampling Probability sampling means that every item in the population has an equal chance of being included in sample. One way to undertake random sampling would be if researcher was to construct a sampling frame first and then used a random number generation computer program to pick a sample from the sampling frame (Zikmund, 2002). Probability or random sampling has the greatest freedom from bias but may represent the most costly sample in terms of time and energy for a given level of sampling error (Brown, 1947).

* 1. Simple random sampling

The simple random sample means that every case of the population has an equal probability of inclusion in sample. Disadvantages associated with simple random sampling include (Ghauri and Gronhaug, 2005):

* 1. A complete frame ( a list of all units in the whole population) is needed;
  2. In some studies, such as surveys by personal interviews, the costs of obtaining the sample can be high if the units are geographically widely scattered;
  3. The standard errors of estimators can be high.
  4. Systematic sampling

Systematic sampling is where every nth case after a random start is selected. For example, if surveying a sample of consumers, every fifth consumer may be selected from your sample. The advantage of this sampling technique is its simplicity.

* 1. Stratified random sampling

Stratified sampling is where the population is divided into strata (or subgroups) and a random sample is taken from each subgroup. A subgroup is a natural set of items. Subgroups might be based on company size, gender or occupation (to name but a few). Stratified sampling is often used where there is a great deal of variation within a population. Its purpose is to ensure that every stratum is adequately represented (Ackoff, 1953).

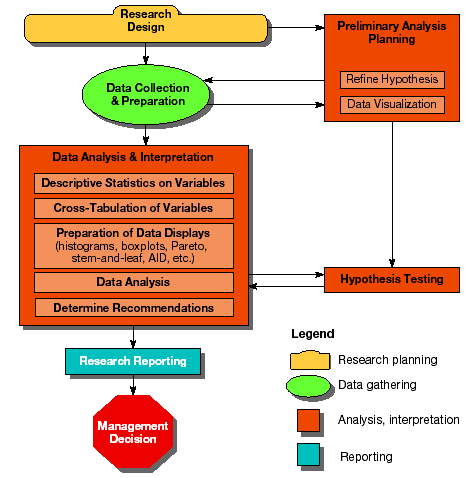
* 1. Cluster sampling Cluster sampling is where the whole population is divided into clusters or groups. Subsequently, a random sample is taken from these clusters, all of which are used in the final sample (Wilson, 2010). Cluster sampling is advantageous for those researchers whose subjects are fragmented over large geographical areas as it saves time and money (Davis, 2005). The stages to cluster sampling can be summarized as follows:
* Choose cluster grouping for sampling frame, such as type of company or geographical region
* Number each of the clusters
* Select sample using random sampling
  1. Multi-stage sampling Multi-stage sampling is a process of moving from a broad to a narrow sample, using a step by step process (Ackoff, 1953). If, for example, a Malaysian publisher of an automobile magazine were to conduct a survey, it could simply take a random sample of automobile owners within the entire Malaysian population. Obviously, this is both expensive and time consuming. A cheaper alternative would be to use multi-stage sampling. In essence, this would involve dividing Malaysia into a number of geographical regions. Subsequently, some of these regions are chosen at random, and then subdivisions are made, perhaps based on local authority areas. Next, some of these are again chosen at random and then divided into smaller areas, such as towns or cities. The main purpose of multi-stage sampling is to select samples which are concentrated in a few geographical regions. Once again, this saves time and money.

1. Non probability Sampling

Non probability sampling is often associated with case study research design and qualitative research. With regards to the latter, case studies tend to focus on small samples and are intended to examine a real life phenomenon, not to make statistical inferences in relation to the wider population (Yin, 2003). A sample of participants or cases does not need to be representative, or random, but a clear rationale is needed for the inclusion of some cases or individuals rather than others.

* 1. Quota sampling Quota sampling is a non random sampling technique in which participants are chosen on the basis of predetermined characteristics so that the total sample will have the same distribution of characteristics as the wider population (Davis, 2005).
  2. 2.2. Snowball sampling Snowball sampling is a non random sampling method that uses a few cases to help encourage other cases to take part in the study, thereby increasing sample size. This approach is most applicable in small populations that are difficult to access due to their closed nature, e.g. secret societies and inaccessible professions (Breweton and Millward, 2001).
  3. Convenience sampling Convenience sampling is selecting participants because they are often readily and easily available. Typically, convenience sampling tends to be a favored sampling technique among students as it is inexpensive and an easy option compared to other sampling techniques (Ackoff, 1953). Convenience sampling often helps to overcome many of the limitations associated with research. For example, using friends or family as part of sample is easier than targeting unknown individuals.
  4. Purposive or judgmental sampling Purposive or judgmental sampling is a strategy in which particular settings persons or events are selected deliberately in order to provide important information that cannot be obtained from other choices (Maxwell, 1996). It is where the researcher includes cases or participants in the sample because they believe that they warrant inclusion.

**Data - Preparing, Exploring, examining and displaying**



Data preparation is the process of gathering, combining, structuring and organizing data so it can be used in business intelligence (BI), analytics and data visualization applications. The components of data preparation include data preprocessing, profiling, cleansing, validation and transformation; it often also involves pulling together data from different internal systems and external sources.

Data preparation work is done by information technology (IT), BI and data management teams as they integrate data sets to load into a data warehouse, NoSQL database or data lake repository, and then when new analytics applications are developed with those data sets. In addition, data scientists, data engineers, other data analysts and business users increasingly use self-service data preparation tools to collect and prepare data themselves.

Data preparation is often referred to informally as data prep. It's also known as data wrangling, although some practitioners use that term in a narrower sense to refer to cleansing, structuring and transforming data; that usage distinguishes data wrangling from the data preprocessing stage.

Purposes of data preparation

One of the primary purposes of data preparation is to ensure that raw data being readied for processing and analysis is accurate and consistent so the results of BI and analytics applications will be valid. Data is commonly created with missing values, inaccuracies or other errors, and separate data sets often have different formats that need to be reconciled when they're combined. Correcting data errors, validating data quality and consolidating data sets are big parts of data preparation projects.

Data preparation also involves finding relevant data to ensure that analytics applications deliver meaningful information and actionable insights for business decision-making. The data often is enriched and optimized to make it more informative and useful -- for example, by blending internal and external data sets, creating new data fields, eliminating outlier values and addressing imbalanced data sets that could skew analytics results.

In addition, BI and data management teams use the data preparation process to curate data sets for business users to analyze. Doing so helps streamline and guide self-service BI applications for business analysts, executives and workers.

What are the benefits of data preparation?

Data scientists often complain that they spend most of their time gathering, cleansing and structuring data instead of analyzing it. A big benefit of an effective data preparation process is that they and other end users can focus more on data mining and data analysis -- the parts of their job that generate business value. For example, data preparation can be done more quickly, and prepared data can automatically be fed to users for recurring analytics applications.

Done properly, data preparation also helps an organization do the following:

* ensure the data used in analytics applications produces reliable results;
* identify and fix data issues that otherwise might not be detected;
* enable more informed decision-making by business executives and operational workers;
* reduce data management and analytics costs;
* avoid duplication of effort in preparing data for use in multiple applications; and
* Get a higher ROI from BI and analytics initiatives.

Steps in the data preparation process

Data preparation is done in a series of steps. There's some variation in the data preparation steps listed by different data professionals and software vendors, but the process typically involves the following tasks:

Data collection. Relevant data is gathered from operational systems, data warehouses, data lakes and other data sources. During this step, data scientists, members of the BI team, other data professionals and end users who collect data should confirm that it's a good fit for the objectives of the planned analytics applications.

Data discovery and profiling. The next step is to explore the collected data to better understand what it contains and what needs to be done to prepare it for the intended uses. To help with that, data profiling identifies patterns, relationships and other attributes in the data, as well as inconsistencies, anomalies, missing values and other issues so they can be addressed.

Data cleansing. Next, the identified data errors and issues are corrected to create complete and accurate data sets. For example, as part of cleansing data sets, faulty data is removed or fixed, missing values are filled in and inconsistent entries are harmonized.

Data structuring. At this point, the data needs to be modeled and organized to meet the analytics requirements. For example, data stored in comma-separated values (CSV) files or other file formats has to be converted into tables to make it accessible to BI and analytics tools.

Data transformation and enrichment. In addition to being structured, the data typically must be transformed into a unified and usable format. For example, data transformation may involve creating new fields or columns that aggregate values from existing ones. Data enrichment further enhances and optimizes data sets as needed, through measures such as augmenting and adding data.

Data validation and publishing. In this last step, automated routines are run against the data to validate its consistency, completeness and accuracy. The prepared data is then stored in a data warehouse, a data lake or another repository and either used directly by whoever prepared it or made available for other users to access.

**Exploratory research: Definition**

E\xploratory research is defined as a research used to investigate a problem which is not clearly defined. It is conducted to have a better understanding of the existing problem, but will not provide conclusive results. For such a research, a researcher starts with a general idea and uses this research as a medium to identify issues that can be the focus for future research. An important aspect here is that the researcher should be willing to change his/her direction subject to the revelation of new data or insight. Such a research is usually carried out when the problem is at a preliminary stage. It is often referred to as grounded theory approach or interpretive research as it used to answer questions like what, why and how.

For example: Consider a scenario where a juice bar owner feels that increasing the variety of juices will enable increase in customers, however he is not sure and needs more information. The owner intends to carry out an exploratory research to find out and hence decides to do an exploratory research to find out if expanding their juices selection will enable him to get more customers of if there is a better idea.

Another example of exploratory research is a podcast survey template that can be used to collect feedback about the podcast consumption metrics both from existing listeners as well as other podcast listeners that are currently not subscribed to this channel. This helps the author of the podcast create curated content that will gain a larger audience.

**Types and methodologies of exploratory research**

While it may sound a little difficult to research something that has very little information about it, there are several methods which can help a researcher figure out the best research design, data collection methods and choice of subjects. There are two ways in which research can be conducted namely primary and secondary.. Under these two types, there are multiple methods which can used by a researcher. The data gathered from these research can be qualitative or quantitative. Some of the most widely used research designs include the following:

**Primary research methods**

Primary research is information gathered directly from the subject.  It can be through a group of people or even an individual. Such a research can be carried out directly by the researcher himself or can employ a third party to conduct it on their behalf. Primary research is specifically carried out to explore a certain problem which requires an in-depth study.

* **Surveys/polls**: Surveys/polls are used to gather information from a predefined group of respondents. It is one of the most important quantitative method. Various types of surveys  or polls can be used to explore opinions, trends, etc. With the advancement in technology, surveys can now be sent online and can be very easy to access. For instance, use of a survey app through tablets, laptops or even mobile phones. This information is also available to the researcher in real time as well. Nowadays, most organizations offer short length surveys and rewards to respondents, in order to achieve higher response rates.

For example: A survey is sent to a given set of audience to understand their opinions about the size of mobile phones when they purchase one. Based on such information organization can dig deeper into the topic and make business related decision.

* **Interviews**: While you may get a lot of information from public sources, but sometimes an in person interview can give in-depth information on the subject being studied. Such a research is a qualitative research method. An interview with a subject matter expert can give you meaningful insights that a generalized public source won’t be able to provide. Interviews are carried out in person or on telephone which have open-ended questions to get meaningful information about the topic.

For example: An interview with an employee can give you more insights to find out the degree of job satisfaction, or an interview with a subject matter expert of quantum theory can give you in-depth information on that topic.

* **Focus groups**: Focus group is yet another widely used method in exploratory research. In such a method a group of people is chosen and are allowed to express their insights on the topic that is being studied. Although, it is important to make sure that while choosing the individuals in a focus group they should have a common background and have comparable experiences.

For example: A focus group helps a research identify the opinions of consumers if they were to buy a phone. Such a research can help the researcher understand what the consumer value while buying a phone. It may be screen size, brand value or even the dimensions. Based on which the organization can understand what are consumer buying attitudes, consumer opinions, etc.

* **Observations**: Observation research can be qualitative observation or quantitative observation. Such a research is done to observe a person and draw the finding from their reaction to certain parameters. In such a research, there is no direct interaction with the subject.

For example: An FMCG company wants to know how its consumer react to the new shape of their product. The researcher observes the customers first reaction and collects the data, which is then used to draw inferences from the collective information.

**Secondary research methods**

Secondary research is gathering information from previously published primary research. In such a research you gather information from sources likes case studies, magazines, newspapers, books, etc.

* **Online research**: In today’s world, this is one of the fastest way to gather information on any topic. A lot of data is readily available on the internet and the researcher can download it whenever he needs it. An important aspect to be noted for such a research is the genuineness and authenticity of the source websites that the researcher is gathering the information from.

For example: A researcher needs to find out what is the percentage of people that prefer a specific brand phone. The researcher just enters the information he needs in a search engine and gets multiple links with related information and statistics.

* **Literature research**: Literature research is one of the most inexpensive method used for discovering a hypothesis. There is tremendous amount of information available in libraries, online sources, or even commercial databases. Sources can include newspapers, magazines, books from library, documents from government agencies, specific topic related articles, literature, Annual reports, published statistics from research organizations and so on.

However, a few things have to be kept in mind while researching from these sources. Government agencies have authentic information but sometimes may come with a nominal cost. Also, research from educational institutions is generally overlooked, but in fact educational institutions carry out more number of research than any other entities.

Furthermore, commercial sources provide information on major topics like political agendas, demographics, financial information, market trends and information, etc.

For example: A company has low sales. It can be easily explored from available statistics and market literature if the problem is market related or organization related or if the topic being studied is regarding financial situation of the country, then research data can be accessed through government documents or commercial sources.

* **Case study research**: Case study research can help a researcher with finding more information through carefully analyzing existing cases which have gone through a similar problem. Such analysis are very important and critical especially in today’s business world. The researcher just needs to make sure he analyses the case carefully in regards to all the[variables](https://www.questionpro.com/blog/nominal-ordinal-interval-ratio/) present in the previous case against his own case. It is very commonly used by business organizations or social sciences sector or even in the health sector.

For example: A particular orthopedic surgeon has the highest success rate for performing knee surgeries. A lot of other hospitals or doctors have taken up this case to understand and benchmark the method in which this surgeon does the procedure to increase their success rate.

**Exploratory research: Steps to conduct a research**

* **Identify the problem:** A researcher identifies the subject of research and the problem is addressed by carrying out multiple methods to answer the questions.
* **Create the hypothesis:** When the researcher has found out that there are no prior studies and the problem is not precisely resolved, the researcher will create a hypothesis based on the questions obtained while identifying the problem.
* **Further research:** Once the data has been obtained, the researcher will continue his study through descriptive investigation. Qualitative methods are used to further study the subject in detail and find out if the information is true or not.

**Characteristics of Exploratory research**

* They are not structured studies
* It is usually low cost, interactive and open ended.
* It will enable a researcher answer questions like what is the problem. What is the purpose of the study? And what topics could be studied?
* To carry out exploratory research, generally there is no prior research done or the existing ones do not answer the problem precisely enough.
* It is a time consuming research and it needs patience and has risks associated with it.
* The researcher will have to go through all the information available for the particular study he is doing.
* There are no set of rules to carry out the research per se, as they are flexible, broad and scattered.
* The research needs to have importance or value. If the problem is not important in the industry the research carried out is ineffective.
* The research should also have a few theories which can support its findings as that will make it easier for the researcher to assess it and move ahead in his study
* Such a research usually produces qualitative data, however in certain cases quantitative data can be generalized for a larger sample through use of surveys and experiments.

**Advantages of Exploratory research**

* The researcher has a lot of flexibility and can adapt to changes as the research progresses.
* It is usually low cost.
* It helps lay the foundation of a research, which can lead to further research.
* It enables the researcher understand at an early stage, if the topic is worth investing the time and resources and if it is worth pursuing.
* It can assist other researchers to find out possible causes for the problem, which can be further studied in detail to find out, which of them is the most likely cause for the problem.

**Disadvantages of Exploratory research**

* Even though it can point you in the right direction towards what is the answer, it is usually inconclusive.
* The main disadvantage of exploratory research is that they provide qualitative data. Interpretation of such information can be judgmental and biased.
* Most of the times, exploratory research involves a smaller sample, hence the results cannot be accurately interpreted for a generalized population.
* Many a times, if the data is being collected through secondary research, then there is a chance of that data being old and is not updated.

**Importance of Exploratory research**

Exploratory research is carried out when a topic needs to be understood in depth, especially if it hasn’t been done before. The goal of such a research is to explore the problem and around it and not actually derive a conclusion from it. Such kind of research will enable a researcher to set a strong foundation for exploring his ideas, choosing the right research design and finding variables that actually are important for the analysis. Most importantly, such a research can help organizations or researchers save up a lot of time and resources, as it will enable the researcher to know if it worth pursuing.