1. Define statistics.

Definitions:

- It is a science which helps us to **collect, analyze** and **present data** systematically.
- It is the process of **collecting, processing, summarizing, presenting, analysing** and **interpreting of data in order to study** and **describe a given problem.**
- Statistics is the **art of learning from data**.
- Statistics may be regarded as (i)the study of populations, (ii) the study of variation, and (iii) the study of methods of the reduction of data.

Definitions:

- The science of Statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates.
- Statistics is the science which deals with collection, classification and tabulation of numerical facts as the basis for explanation, description and comparison of phenomenon.

Importance of Statistics:

- It simplifies mass of data (condensation);
- Helps to get concrete information about any problem;
- Helps for reliable and objective decision making;
- It presents facts in a precise & definite form;
- It facilitates comparison(Measures of central tendency and measures of dispersion);
- It facilitates Predictions (Time series and regression analysis are the most commonly used methods towards prediction.);
- It helps in formulation of suitable policies;

Limitation of statistics:

1. Statistics does not deal with **individual items;**
2. Statistics deals only with **quantitatively expressed items**, it does not study qualitative phenomena;
3. Statistical results are **not universally true**;
   - Statistical laws are only approximations and not exact. Of
   - in terms of probability and chance
   - Eg. It has been found that 20 % of-a certain surgical operations by a particular doctor are successful."
4. Statistics is **liable/responsible/ to be misused**.
   - can be moulded and manipulated in any manner to support one's way of argument and reasoning.

2. Give functions of Statistics.

- Someone has jokingly said :

"Since statistics is the science of averages, so if head is kept in boiler and leg freeze, then the temperature of the stomach is statistics."

1.1.2 Functions of Statistics

Three main functions of statistics are :

(1) Collection of data

Following are methods of collection of data :

(a) Direct personal enquiry method.

(b) Indirect oral investigation.

(c) By fitting of schedules

(d) By mailed questionnaires.

(e) Information from local agents and correspondents.

(f) By old records.

(g) By direct observational methods.

(2) Presentation of data

There are two kinds of statistical data, they are :

(a) Primary data and

(b) Secondary data

(3) Analysis of data

The requisites are :

(a) It should be complete.

(b) It should be consistent.

(c) It should be accurate.

(d) It should be homogeneous in respect of unit of information.

3. Write any five applications of Statistics

Application areas of statistics

➢ Engineering:

Improving product design, testing product performance, determining reliability and maintainability, working out safer systems of flight control for airports, etc.

➢ Business:

Estimating the volume of retail sales, designing optimum inventory control system, producing auditing and accounting procedures, improving working conditions in industrial plants, assessing the market for new products.

➢ Quality Control:

Determining techniques for evaluation of quality through adequate sampling, in process control, consumer survey and experimental design in product development etc.

Realizing its importance, large organizations are maintaining their own **Statistical Quality Control Department.**

➢ Economics:

Measuring indicators such as volume of trade, size of labor force, and standard of living, analyzing consumer behavior, computation of national income accounts, formulation of economic laws, etc.

Particularly, Regression analysis extensively used in the field of Economics.

- **Health and Medicine:**

  Developing and testing new drugs, delivering improved medical care, preventing diagnosing, and treating disease, etc. Specifically, inferential Statistics has a tremendous application in the fields of health and medicine.

- **Biology:**

  Exploring the interactions of species with their environment, creating theoretical models of the nervous system, studying genetically evolution, etc.

- **Psychology:**

  Measuring learning ability, intelligence, and personality characteristics, creating psychological scales and abnormal behavior, etc.

- **Sociology:**

  Testing theories about social systems, designing and conducting sample surveys to study social attitudes, exploring cross-cultural differences, studying the growth of human population, etc.

4. Define Classification. Explain types of Classification.

We mention below some definitions of classification.

"Classification is the process of arranging data into sequences and groups according to their common characteristics, or separating them into different but related parts".

- Secrist.

"A classification is a scheme for breaking a category into a set of parts, called classes, according to some precisely defined differing characteristics possessed by all the elements of the category".

- Tuttle A.M.

- Thus 'classification' is the arrangement of the data into different classes, which are to be determined depending upon the nature, objectives and scope of the enquiry.
- For example, the number of students registered in Pune University during the academic year 2020-21 may be classified on the basis of the following criterion :

  (i) Sex, (ii) Age, (iii) Religion, (iv) The state to which they belong, (v) Different faculties : Engineering, Medical, Arts, Science, Law, Commerce etc. (vi) Heights or weights, (vii) Institution / College and so on.
- The same data can be classified into different groups in a number of ways. That is based on physical, mental or social characteristics.
- The data in one class will be different from those of another class with respect to some characteristic called the **basis** or **criterion of classification.**

1.4.3 Types of Data Classification

(1) Data classification often involves a multitude of tags and labels that define the type of data, its confidentiality and its integrity.

(2) Availability may also be taken into consideration in data classification processes.

(3) Data's level of sensitivity is often classified and it is based on varying levels of importance or confidentiality, which then correlates to the security measures put in place to protect each classification level.

1.4.4 Three Main Types of Data Classification

These are considered as industry standards :

(i) **Content :** Content-based classification inspects and interprets files looking for sensitive information.

(ii) **Context :** Context-based classification looks at application, location or creator among other variables as indirect indicators of sensitive information.

(iii) **User :** User-based classification depends on a manual, end-user selection of each document. User-based classification relies on user knowledge and discretion at creation, edit, review or dissemination to flag sensitive documents.

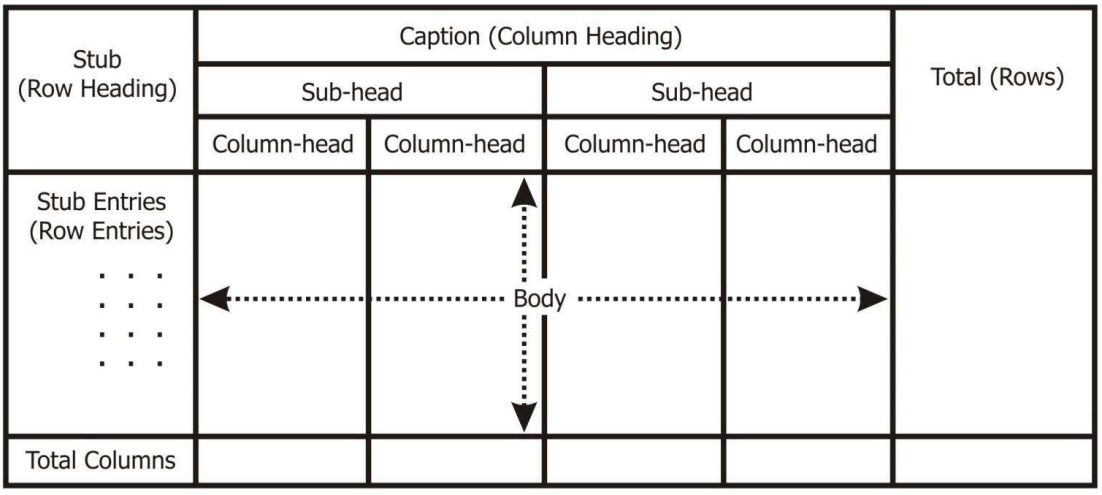
Content-context and user-based approaches can be both right or wrong depending on the business-need and data-type.

5. Which are the different parts of the Table? Explain with suitable example

1. Tabular presentation of data:

- The collected raw data should be put into an ordered array in either ascending or descending order so that it can be organized in to a Frequency Distribution (FD)
- Numerical data arranged in order of magnitude along with the corresponding frequency is called frequency distribution (FD).
- FD is **of two kinds namely ungrouped /and grouped frequency distribution**.

Table Number: .....
Title: .....
(Head Note, if any)



Source Note:
Footnote:

1.5.1 The Parts of a Table

- The parts of a table vary from problem to problem. And that depends upon the nature of the data and purpose of the investigation.
- The following points are important in a good statistical table :

(i)	Table number	(ii)	Title
(iii)	Head (main) notes	(iv)	Captions and stubs
(v)	Body of the table	(vi)	Foot-note
(vii)	Source-note		

We discuss these in brief.

(i) Table number

If the data contains more than one table, then all the tables should be numbered in a logical sequence for proper identification and easily accessible for further reference.

(ii) Title

- Every table must be given a suitable title. A title must be self-explanatory.
- It must describe in brief and concise form the contents of the table.
- It should describe the nature of the data, the place (i.e. geographical region or area to which the data release), the time (i.e. period to which the data relate) and the source of data.
- The title should be brief but not incomplete one and not at cost of clarity. Sometimes, it is necessary to use long title for the sake of clarity.
- In such a case a **'brief note'** may be given above the main title.

(iii) Head notes

- It so required, head not is given just below the title in a prominent type usually centred and enclosed in brackets for further description of the contents of the table.
- It provides an explanation concerning the entire table or its major parts.

(iv) Captions and stubs

- Captions are the headings or designations for vertical columns and stubs are the headings or designations for the horizontal rows.
- They should be brief and self-explanatory Captions are written in the middle of the columns in small letters.

- Each column and row must be given a number for reference.
- If two or more columns or rows correspond to similar classifications (or with the same headings) then they may be grouped together under a common heading to avoid repetitions.

(v) Body of the table

- The arrangement of the data according to the descriptions given in the columns (captions) and rows (stubs) forms the body of the table.
- Numerical information forms the most important part of the table.
- To increase the usefulness of the table, totals must be given for each separate class below the columns or against the rows.

(vi) Foot note

- Foot notes are to be used, when some characteristic or feature of the item of the table needs elaborate explanation.
- Foot-notes, if any are placed below the body of the table.
- Foot notes are identified by the symbols *, **, ***, @ etc.

(vii) Source note

- The source note is required if the secondary data are used.
- Source note is to be given at the bottom of the table. If the data are taken from a research journal or periodical, then the source note should give the name of the journal or periodical along with the date of publication, its volume number, page number, so that anybody who uses this data may verify the accuracy of the figures by referring to the original source.
- A table should have an attractive get up which is appealing to the eye and the mind so that the reader may grasp it without any strain.
- Hence special attention to the size of the table and proper spacings of rows and columns must be given.

1.5.2 Types of Tabulation

- Statistical tables are generally constructed in the following ways :

  (i) Objectives and scope of the enquiry.

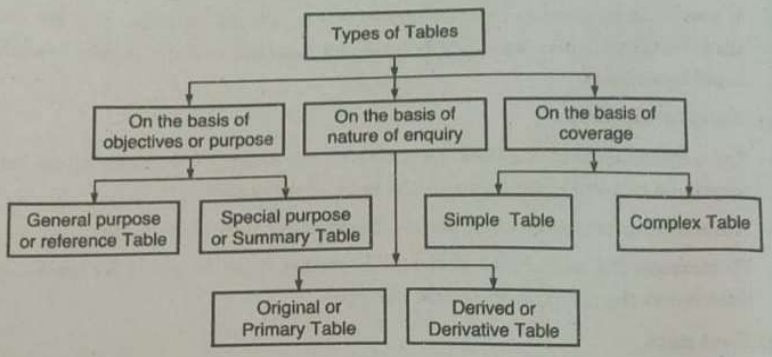
  (ii) Nature of the enquiry (primary or secondary).

  (iii) Extent of coverage given in the enquiry.

- We mention below diagrammatic scheme that displays the various forms of tables commonly used in practice.

6. Give different types of Tables

1.5.3 Types of Table



1.5.4 Solved Examples on Preparation of Tables

Ex. 1.5.1 : Present the following information in a suitable tabular form, supplying the figures.

In 1995, out of total 2000 workers in a factory, 1550 were members at a trade union. The number of women workers employed was 250, out of which 200 did not belong to any trade union.

In 2000, the number of union workers was 1725 of which 1600 were men. The number of non-union, workers was 380 among which 155 were women.

☑ **Soln. :**

Comparative study of the membership of trade union in a factory in 1995 and 2000.

Year →	1995			2000		
Trade union ↓	Males	Female	Total	Males	Females	Total
Members	1550 – 50 = 1500	250 – 200 = 50	1550	1600	1725 – 1600 = 125	1725
Non-members	1750 – 1500 = 250	200	2000 – 1550 = 450	380 – 155 = 225	155	380
Total	2000 – 250 = 1750	250	2000	1600 + 225 = 1825	125 + 155 = 280	1725 + 380 = 2105

Here, we have presented the comparative study of the membership of trade union in a factory in **1995 and 2000.**

7. Define Statistics & state its Limitations .

Examples on

8. Cumulative Frequency Distribution

2. Cumulative Frequency Distribution (CFD):

- It is applicable when we want to know how many observations lie **below or above a certain value/class boundary.**

- **CFD** is of two types, **LCFD and MCFD:**
  - ✓ **Less than Cumulative Frequency Distribution (LCFD):** shows the collection of cases lying **below the upper class boundaries of each class.**
  - ✓ **More than Cumulative Frequency Distribution (MCFD):** shows the collection of cases lying **above the lower class boundaries of each class.**

Test score	CF
Less than 37.5	0
Less than47.5	4
Less than 57.5	12
Less than 67.5	25
Less than 77.5	35
Less than 87.5	38
Less than 97.5	40

Test score	CF
more than 37.5	40
more than 47.5	36
more than 57.5	28
more than 67.5	15
more than 77.5	5
more than 87.5	2
more than 97.5	0

9. By considering given data, construct Continuous Frequency Distribution table. Calculate Mean, Mode, Median.

B. Grouped (continuous) Frequency Distribution (GFD)

- ✓ It is a tabular arrangement of data in order of magnitude by classes together with the corresponding class frequencies.
- ✓ In order to estimate the number of classes, the ff formula is used:

  Number of classes=1+3.322(log N) where N is the Number of observation.

$$\text{The Class size (class width)} = \frac{\text{Range}}{1+3.322(\log N)} \quad \text{(round up)}$$

Exercise:

Construct a GFD of the following aptitude test scores of 40 applicants for accountancy positions in a company with

a. 6 classes b. 8 classes

96	89	58	61	46	59	75	54
41	56	77	49	58	60	63	82
66	64	69	67	62	55	67	70
78	65	52	76	69	86	44	76
57	68	64	52	53	74	68	39

Types of Grouped Frequency Distribution

1. Relative frequency distribution (RFD)
2. Cumulative Frequency Distribution (CFD)
3. Relative Cumulative Frequency Distribution (RCFD)

Types of Grouped Frequency Distribution

1. Relative frequency distribution (RFD):

- A table presenting the ratio of the **frequency of each class to the total frequency of all the classes**.
- Relative frequency generally expressed **as a percentage**, used to show the percent of the total number of observation in each class.

For example

Test score	F	RFD	PFD
37.5-47.5	4	4/40=0.1	10%
47.5-57.5	8	8/40=0.2	20%
57.5-67.5	13	13/40=0.325	32.5%
67.5-77.5	10	10/40=0.25	25%
77.5-87.5	3	3/40=0.075	7.5%
87.5-97.5	2	2/40=0.05	5%

3. Relative Cumulative Frequency Distribution (RCFD)

It is used to determine the ratio or the percentage of observations that lie below or above a certain value/class boundary, to the total frequency of all the classes. These are of two types: The LRCFD and MRCFD.

- **Less than Relative Cumulative Frequency Distribution (LRCFD):** A table presenting the ratio of the cumulative frequency **less than upper class boundary of each class to the total frequency of all the classes**
- **More than Relative Cumulative Frequency Distribution (MRCFD):** A table presenting the ratio of the cumulative frequency **more than lower class boundary of each class to the total frequency of all the classes.**

LRCFD

Test score	LCF	LRCF	LPCF
Less than 37.5	0	0/40=0	0%
Less than 47.5	4	4/40=0.1	10%
Less than 57.5	12	12/40=0.3	30%
Less than 67.5	25	25/40=0.625	62.5%
Less than 77.5	35	35/40=0.875	87.5%
Less than 87.5	38	38/40=0.95	95%
Less than 97.5	40	40/40=1	100%

MRCFD

Test score	MCF	MRCF	MPCF
More than 37.5	40	40/40=1	100%
More than 47.5	36	36/40=0.9	90%
More than 57.5	28	28/40=0.7	70%
More than 67.5	15	15/40=0.375	37.5%
More than 77.5	5	5/40=0.125	12.5%
More than 87.5	2	2/40=0.05	5%
More than 97.5	0	0/40=0	0%

10. Construct Frequency Distribution Table for Bivariate data.

Construct a bivariate frequency distribution table of the marks obtained by students in English (X) and statistics (Y).

Marks in statistics(X)	37	20	46	28	35	26	41	48	32	23	20	39	47	33	27	26
Marks in English(Y)	30	32	41	33	29	43	30	21	44	38	47	24	32	21	20	21

Construct a bivariate frequency distribution table for the given data by taking class interval 20 – 30, 30 – 40 etc. for both X and Y. Also find the marginal distribution and conditional frequency distribution of Y where X lies between 30 – 40.

The whole set of cell frequencies will then define a bivariate frequency distribution.

So, the Bivariate frequency distribution table:

$Y \backslash X$	20 – 30	30 – 40	40 – 50	f_y
20 – 30	11 (2)	11 (2)	1 (1)	5
30 – 40	111 (3)	11 (2)	11 (2)	7
40 – 50	11 (2)	1 (1)	1 (1)	4
f_x	7	5	4	16

Marginal frequency distribution of English (X): -

X	20 – 30	30 – 40	40 – 50	Total
F	7	5	4	16

Marginal frequency distribution of Statistics (Y): -

X	20 – 30	30 – 40	40 – 50	Total
F	5	7	4	16

A conditional distribution is a probability distribution for a sub-population. In other words, it shows the probability that a randomly selected item in a sub-population has a characteristic you're interested in.

Conditional frequency distribution of y when x lies between 30 – 40: -

X	20 – 30	30 – 40	40 – 50	Total
F	2	2	1	5

So, all tables are formed.

Note: During calculating the bivariate table for the given data always mind that all the data should

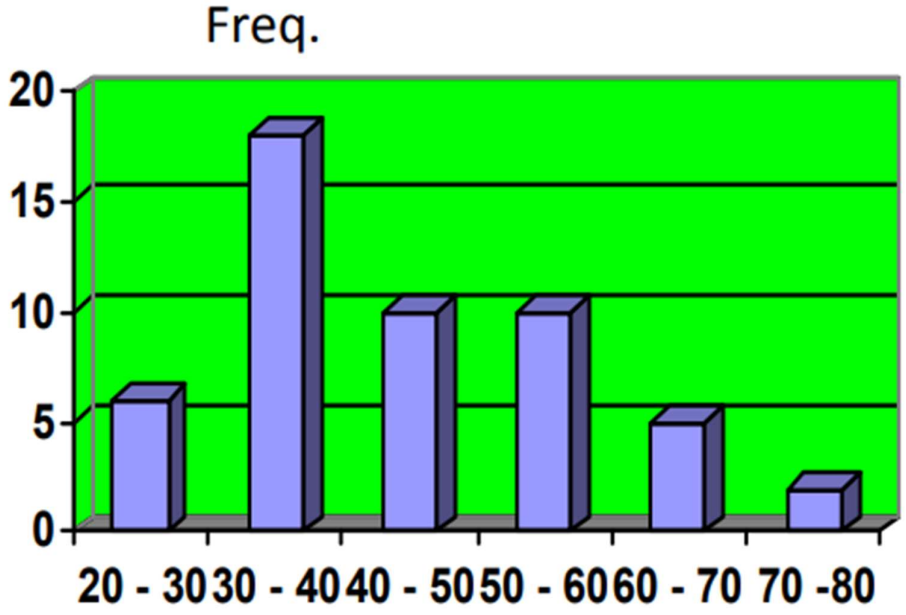
11. Tabulation of Data (3 way complex data) by given conditions.
12. Bar Graph
13. Pie Chart

14. Histogram

1. Histogram:

✓ A graphical presentation of grouped frequency distribution consisting of a series of adjacent rectangles whose bases are the class intervals specified in terms of class boundaries (equal to the class width of the corresponding classes) shown on the x-axis and whose heights are proportional to the corresponding class frequencies shown on the y-axis.

Histogram: E.g.

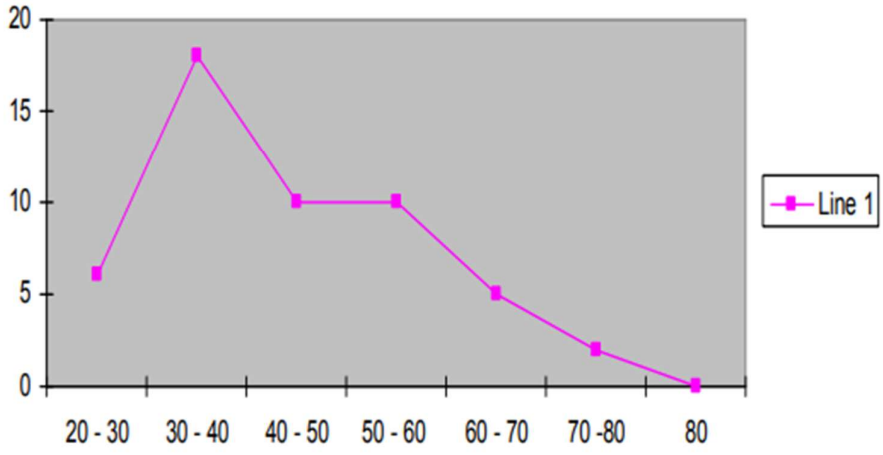


15. Frequency Polygon

2. Frequency Polygon:

It is a line graph of grouped frequency distribution in which the class frequency is plotted against class mark that are subsequently connected by a series of line segments to form line graph including classes with zero frequencies at both ends of the distribution to form a polygon.

'equency Polygon:

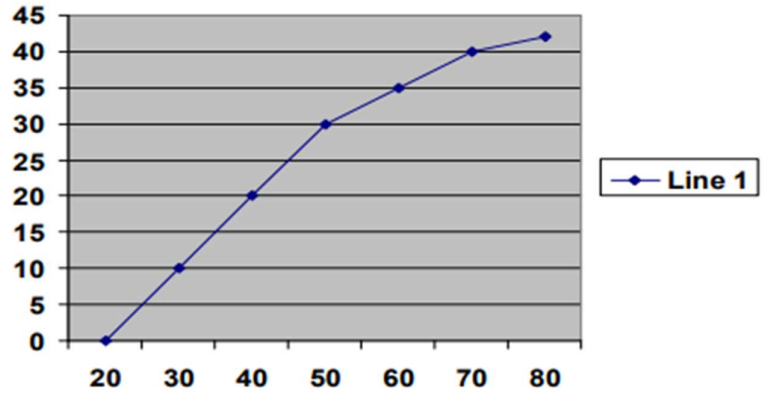


16. Ogive Curve

3. O-GIVE curve (Cumulative Frequency Curve / percentage Cumulative Frequency Curve)

- ✓ It is a line graph presenting the cumulative frequency distribution.
- ✓ O-gives are of two types: The **Less than O-give** and The **More than O-give**.
- ➢ **The Less than O-give** shows the cumulative frequency less than the upper class boundary of each class; and
- ➢ **The More than O-give** shows the cumulative frequency more than the lower class boundary of each class.

Ogive: E.g.



Steps to draw O-gives

i. Mark class boundaries on the x-axis and mark non overlapping intervals of equal length on the y-axis to represent the cumulative frequencies.

ii. For each class boundaries marked on the x-axis, plot a point with height equal to the corresponding cumulative frequencies.

iii. Connect the marked points by a series of line segments where the less than O-give is done by plotting the less than cumulative frequency against the upper class boundaries

17. What is Data? Explain primary data & secondary data. List advantages and disadvantages of primary & secondary data.

Data is a collection of facts, figures, objects, symbols, and events gathered from different sources. Organizations collect data to make better decisions. Without data, it would be difficult for organizations to make appropriate decisions, and so data is collected at

various points in time from different audiences. For instance, before launching a new product, an organization needs to collect data on product demand, customer preferences, competitors, etc. In case data is not collected beforehand, the organization's newly launched product may lead to failure for many reasons, such as less demand and inability to meet customer needs.

Although data is a valuable asset for every organization, it does not serve any purpose until analyzed or processed to get the desired results.

Data collection is a process of collecting information from all the relevant sources to find answers to the research problem, test the hypothesis and evaluate the outcomes.

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 Population census conducted by the government of India after every 10 years.

B) Secondary Data

Secondary data refers to 'Second-hand information'.

These are not originally collected rather obtained from already published or unpublished sources.

For example the Address of a person taken from the Telephone Directory or Phone number of a company taken from 'Just Dial'.

Students can also refer to Meaning and Sources of Secondary Data

Advantages of primary data:

- **Resolve specific research issues**
  Performing your own research allows you to address and resolve issues specific to your own business situation. The collected information is the exact information that the researcher wants to know and he reports it in a way that benefits the specific situation in an organization. Marketers and researchers are asked to find data regarding specific markets instead of finding data for the mass market. This is the main difference from secondary data.
- **Better accuracy**
  Primary data is much more accurate because it is directly collected from a given population.
- **A higher level of control**
  The marketer can control easily the research design and method. In addition, you have a higher level of control over how the information is gathered.
- **Up-to-date information**
  The primary market research is a great source of latest and up-to-date information as you collect it directly from the field in real-time. Usually, secondary data is not so up-to-date and recent.
- **You are the owner of the information**
  Information collected by the researcher is their own and is typically not shared with others. Thus, the information can remain hidden from other current and potential competitors.

Disadvantages:

- **More expensive**
  It could be very expensive to obtain primary data collection methods because the marketer or the research team has to start from the beginning. It means they have to follow the whole study procedure, organizing materials, process and etc.
- **Time-consuming**
  It is a matter of a lot of time to conduct the research from the beginning to the end. Often it is much longer in comparison with the time needed to collect secondary data.
- **Can have a lot of limits**
  Primary data is limited to the specific time, place or number of participants and etc. To compare, secondary data can come from a variety of sources to give more details.
- **Not always possible**
  For example, many researches can be just too large to be performed by your company.

Advantages of Secondary Data:

- **Ease of Access**
  The secondary data sources are very easy to access. The internet world changed how secondary research exists. Nowadays, you have so much information available just by clicking with the mouse in front of the computer.
- **Low Cost or Free**
  The majority of secondary sources are absolutely free for use or at very low costs. It saves not only your money but your efforts. In comparison with primary research where you have to design and conduct a whole primary study process from the beginning, secondary research allows you to gather data without having to put any money on the table.
- **Time-saving**
  As the above advantage suggests, you can perform secondary research in no time. Sometimes it is a matter of a few Google searches to find a credible source of information.
- **Generating new insights and understandings from previous analysis**
  Reanalyzing old data can bring unexpected new understandings and points of view or even new relevant conclusions.
- **Larger sample size**
  Big datasets often use a larger sample than those that can be gathered by primary data collection. Larger samples mean that the final inference becomes much more straightforward.
- **Longitudinal analysis**
  Secondary data allows you to perform a longitudinal analysis which means the studies are performed spanning over a large period of time. This can help you to determine different trends. In addition, you can find secondary data from many years back up to a couple of hours ago. It allows you to compare data over time.
- **Anyone can collect the data**
  Secondary data research can be performed by people that aren't familiar with the different types of quantitative and qualitative research methods. Practically, anyone can collect secondary data.

Disadvantages:

- **Not specific to your needs**
  Here is the main difference with the primary method. Secondary data is not specific to the researcher's need due to the fact that it was collected in the past for another reason. That is why the secondary data might be unreliable and unuseful and in many business and marketing cases. Secondary data sources can give you a huge amount of information, but quantity does not mean appropriateness.
- **Lack of control over data quality**
  You have no control over the data quality at all. In comparison, with primary methods that are largely controlled by the data-driven marketer, secondary data might lack quality. It means the quality of secondary data should be examined in detail since the source of the information may be questionable. As you relying on secondary data for your decision-making process, you must evaluate the reliability of the information by finding out how the information was collected and analyzed.
- **Biasness**
  As the secondary data is collected by someone else than you, typically the data is biased in favor of the person who gathered it. This might not cover your requirements as a researcher or marketer.
- **Not timely**
  Secondary data is collected in the past which means it might be out-of-date. This issue can be crucial in many different situations.
- **Not proprietary Information**
  Generally, secondary data is not collected specifically for your company. Instead, it is available to many companies and people either for free or for a little fee. So this is not exactly an "information advantage" for you and your competitors also have access to the data.

Points	Primary Data	Secondary Data
Meaning	Primary data is collected directly from the first-hand experience. This is the information that you gather for the purpose of a particular research project.	Secondary data is the data that have been already collected for another purpose. The data is collected by someone else instead of the researcher himself.
Main Sources	Interview, surveys, questionnaires, field observation, experiments, action research, case studies and etc.	Previous research, mass media products, Government reports, official statistics, web information, historical data and etc.
Data Time	Real - Time Data	Past Data
Specific to the Researcher Needs	Always specific to the researcher's needs.	Often, it is not specific to the researcher's needs.
Costs	Expensive	Low Cost or Free
Level of the control over data quality	Higher level of control	Lack of control over data quality
Time consuming	More time consuming	Less time consuming
Proprietary Information	You are the owner of the data. Thus, the information can remain hidden from the competitors.	You are not the owner of the data. Your competitors also have access to the data.
Capability	More capable to solve a specific problem	Less capable to solve a specific problem

18. Define Data Collection. Which are the different sources of data?

Data collection is a process of collecting information from all the relevant sources to find answers to the research problem, test the hypothesis and evaluate the outcomes.

SOURCES OF DATA:

Sources of Data can be classified into 2 types. Statistical sources refer to data that are gathered for some official purposes and 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?

Following are the two sources of data:

1. Internal Source

When data are collected from reports and records of the organisation itself, it is known as the internal source.

For example, a company publishes its 'Annual Report' on Profit and Loss, Total Sales, Loans, Wages etc.

2. External Source

When data are collected from outside the organisation, it is known as the external source.

For example, if a Tour and Travels Company obtains information on 'Karnataka Tourism' from Karnataka Transport Corporation, it would be known as external sources of data.

19. Explain methods of collecting Primary data

Methods of Collecting Primary Data

Direct Personal Investigation

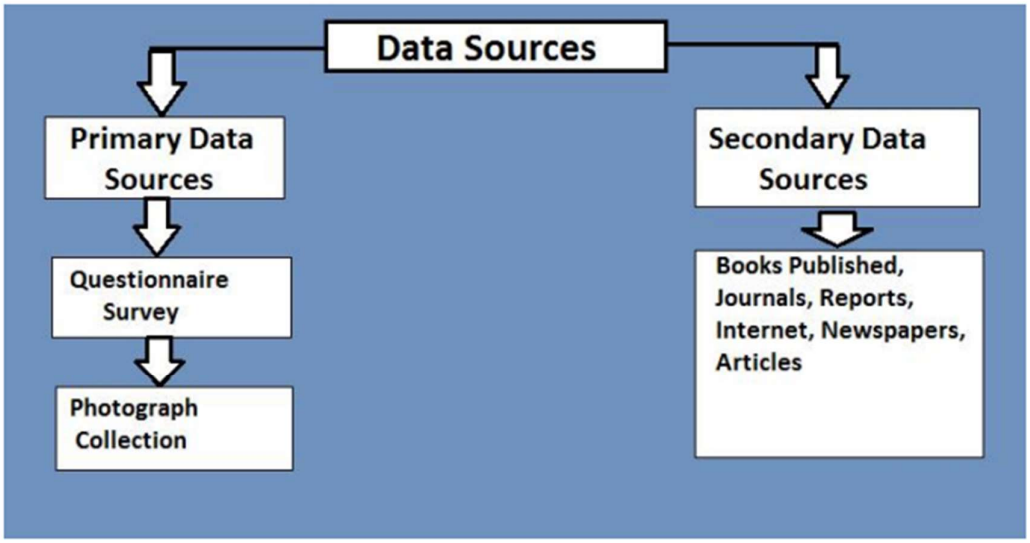
Indirect Oral Investigation

Information Through Correspondents

Telephonic Interview

Mailed Questionnaire

The questionnaire filled by enumerators



Procedure for data collection

. There are various steps commonly involved in the Data Collection process.

2.2.3 We Elaborate the Above-Mentioned Points

(I) Direct Personal Investigation

(a) Under this method, the investigator obtains the first hand information from the respondents themselves.

(b) He personally visits the respondents to collect the data (information).

(c) **Merits of direct personal investigation**

1. The data collected is first-hand and original in nature. So it is more reliable and accurate.
2. In this method, the questions can be modified according to the level of the respondent or other situations.
3. Some additional information can also be obtained along with the required information.
4. This additional information can be used in further investigations.

(d) **Demerits of direct personal investigation**

(a) It is not suitable if the coverage-area is considerably wide.

(b) It is time-consuming method since the investigator personally visits various places and meets different people to collect information.

(c) This method is expensive, particularly when the field of investigation is large.

(d) The data collected in this method is subjected to personal bias.

(ii) Indirect Oral Investigation

(a) The investigator interview several other persons who are directly or indirectly in touch with the informants. Here there is no direct approach to the informants.

(b) Merits of indirect oral investigation

1. Through this method, a large area can be brought under investigation.
2. It is economical in terms of time, money and manpower.

(c) Demerits of indirect oral investigation

1. Since the information is not collected directly from the party, there is a possibility that it may not be completely true.
2. As compared to direct personal investigation the degree of accuracy of the data is likely to be lower.
3. Information collected from different persons for the same party may not be homogeneous and comparable.

4. Respondent / witness can modify the information according to his personal interest.

(iii) Information though Correspondents

(a) For this method, local agents or correspondents are appointed and trained to collect the information from the respondents.

(b) If the field of investigation is large and the information is to be collected from different parts of the country, then this method is useful.

(c) This method is economical and time-saving.

(d) The method is convenient for some special purpose investigations.

(e) It is very useful for collecting information on a regular basis.

(f) **Demerits of information though correspondents :**

1. The information supplied by different correspondents often lacks homogeneity, hence it is not comparable.
2. Data obtained by this method may not be very reliable because of the possibility of personal bias and prejudice of the enumerator.
3. For a high degree of accuracy, this method is not very useful.
4. To collect the information through correspondence, a lot of time and money is spent.

(iv) Telephonic interviews

(a) In this method, data is collected through interviews over the telephone.

(b) **Merits of telephonic interviews**

1. Again this method is quite economical and time-saving.
2. This method is useful where the field of investigation is very wide and the information is to be collected from different parts of the country.
3. The data is reliable since it is obtained directly from the party.

(c) **Demerits of telephonic interviews**

1. The disadvantage of this is limited accessibility to people. One cannot reach to the people, who do not have or own a telephone or mobile.
2. Telephone interviews obstruct visual reactions of the respondents. These reactions become helpful in obtaining information on sensitive issues.

(v) Mailed Questionnaire method

(a) In this method, a questionnaire contains a number of questions related to the investigation. They are prepared thoroughly.

(b) It is then sent by post to informants along with the instructions to fill.

(c) The informants, after filling up the questionnaire, send it back to the investigator.

(d) **Merits of mailed questionnaire method**

1. Again, this method is useful where the field of investigation is very large and the information is to be collected from different parts of the country.
2. This method is economical as it requires less money and labour.
3. Since the informants are directly involved in the collection of data, the data is very much original.
4. Every question is interpreted by the respondent in his own way. Hence, it is free from the personal bias of the investigator.
5. The method is very convenient for sensitive questions and maintains the anonymity of respondents.

(e) Demerits of the mailed questionnaire method :

1. This method is applicable only where the respondents have good knowledge about the questions that are sent.
2. Many of the informants do not return the questionnaire.
3. The informants are least interested in the investigation, hence there is a lack of response from their side.
4. Informants may fail to understand the correct sense of some questions, and may not answer them. Sometimes, informants may provide vague and ambiguous answers.
5. The process is time-consuming, particularly when the information is to be obtained by post.

(vi) Questionnaire and its Qualities

(a) A questionnaire is a list or set of printed questions, which is filled by the informants. If it is filled by enumerate, then it is known as a schedule.

(b) **Characteristics of a good questionnaire**

1. Questions should be short, simple and straightforward.
2. The number of questions should be limited and they should be in a logical order.
3. To assist the importance, clear instructions should be given.
4. To know the shortcomings of a questionnaire, it should be tried on a small selected group.

5. Questions containing mathematical calculations should be completely avoided.
6. Personal questions affecting sentiments and controversial questions related to religion, politics, etc. should be avoided.
7. Respondents should be given assurance that their response will not be shared with anyone.
8. To convey the purpose of how it will help the parties involved, a precise cover letter should be enclosed.

(c) **Method of filling Questionnaire**

Under this method, an enumerator personally visits informants along with a questionnaire, asks questions, and note down their response in the questionnaire in his own language.

(d) **Merits of questionnaires**

1. Since the investigator has direct contact with the respondents, he can have accurate and reliable information.
2. The presence of enumerator may induce the respondents to give information.

   Hence the chances of **no response** in questionnaire method are very less. But in mailed questionnaire, there is possibility of **no response.**
3. This method can be used even though the respondents are not educated. But in case of mailed questionnaire, this is not possible.

(e) **Demerits of Questionnaire**

1. This method is expensive as expenditure on training, renumeration and conveyance is to be borne by the investigator.
2. This method is very time consuming as the enumerator has to visit the informants personally.
3. If the enumerators are not properly trained or of biased views, then they become inefficient and unable to carry out the enquiry properly. This affects adversely on the results of the enquiry.

20. Write short note on: Tools for Data Collection

TOOLS OF DATA COLLECTION / INSTUMRNTS FOR DATA COLLECTION

Data collection is an important step in the research process. The instrument you choose to collect the data will depend on the type of data you plan on collecting (qualitative or quantitative) and how you plan to collect it.

A number of common data-collecting instruments are used in construction research:

1. Questionnaires
2. Interviews
3. Observations
4. Archival documents and government sources
5. Laboratory experiments
6. Quasi experiment

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 Data Gathering**

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.

21. Which are the qualitative methods for Primary Data Collection?

Qualitative Methods:

Qualitative methods are especially useful in situations when historical data is not available. Or there is no need of numbers or mathematical calculations. Qualitative research is closely associated with words, sounds, feeling, emotions, colors, and other elements that are non-quantifiable. These techniques are based on experience, judgment, intuition, conjecture, emotion, etc.

Quantitative methods do not provide the motive behind participants' responses, often don't reach underrepresented populations, and span long periods to collect the data. Hence, it is best to combine quantitative methods with qualitative methods.

Surveys

Surveys are used to collect data from the target audience and gather insights into their preferences, opinions, choices, and feedback related to their products and services. Most survey maker software often a wide range of question types to select.

You can also use a ready-made survey template to save on time and effort. Online surveys can be customized as per the business's brand by changing the theme, logo, etc. They can be distributed through several distribution channels such as email, website, offline app, QR code, social media, etc. Depending on the type and source of your audience, you can select the channel.

Once the data is collected, survey software can generate various reports and run analytics algorithms to discover hidden insights. A survey dashboard can give you the statistics related to response rate, completion rate, filters based on demographics, export and sharing options, etc. You can maximize the effort spent on online data collection by integrating survey builder with third-party apps.

Polls

Polls comprise of one single or multiple choice question. When it is required to have a quick pulse of the audience's sentiments, you can go for polls. Because they are short in length, it is easier to get responses from the people.

Similar to surveys, online polls, too, can be embedded into various platforms. Once the respondents answer the question, they can also be shown how they stand compared to others' responses.

Interviews

In this method, the interviewer asks questions either face-to-face or through telephone to the respondents. In face-to-face interviews, the interviewer asks a series of questions to the interviewee in person and notes down responses. In case it is not feasible to meet the person, the interviewer can go for a telephonic interview. This form of data collection is suitable when there are only a few respondents. It is too time-consuming and tedious to repeat the same process if there are many participants.

Delphi Technique

In this method, market experts are provided with the estimates and assumptions of forecasts made by other experts in the industry. Experts may reconsider and revise their estimates and assumptions based on the information provided by other experts. The consensus of all experts on demand forecasts constitutes the final demand forecast.

Focus Groups

A small group of people, around 8-10 members, discuss the common areas of the problem. Each individual provides his insights on the issue concerned. A moderator regulates the discussion among the group members. At the end of the discussion, the group reaches a consensus.

Questionnaire

A questionnaire is a printed set of questions, either open-ended or closed-ended. The respondents are required to answer based on their knowledge and experience with the issue concerned. The questionnaire is a part of the survey, whereas the questionnaire's end-goal may or may not be a survey.

Secondary Data Collection Methods

Secondary data is the data that has been used in the past. The researcher can obtain data from the sources, both internal and external, to the organization.

22. Which are the quantitative method for Primary Data Collection?

Primary Data Collection Methods

Primary data is collected from the first-hand experience and is not used in the past. The data gathered by primary data collection methods are specific to the research's motive and highly accurate.

Primary data collection methods can be divided into two categories: quantitative methods and qualitative methods.

Quantitative Methods:

Quantitative techniques for market research and demand forecasting usually make use of statistical tools. In these techniques, demand is forecast based on historical data. These methods of primary data collection are generally used to make long-term forecasts. Statistical methods are highly reliable as the element of subjectivity is minimum in these methods.

Time Series Analysis

The term time series refers to a sequential order of values of a variable, known as a trend, at equal time intervals. Using patterns, an organization can predict the demand for its products and services for the projected time.

Smoothing Techniques

In cases where the time series lacks significant trends, smoothing techniques can be used. They eliminate a random variation from the historical demand. It helps in identifying patterns and demand levels to estimate future demand. The most common methods used in smoothing demand forecasting techniques are the simple moving average method and the weighted moving average method.

Barometric Method

Also known as the leading indicators approach, researchers use this method to speculate future trends based on current developments. When the past events are considered to predict future events, they act as leading indicators.

23. Explain Fundamental types of interviews in research

Interview Method of Data Collection

Types of interviews

An interview is generally a qualitative research technique which involves asking open-ended questions to converse with respondents and collect elicit data about a subject. The interviewer in most cases is the subject matter expert who intends to understand respondent opinions in a well-planned and executed series of questions and answers. Interviews are similar to focus groups and surveys when it comes to garnering information from the target market but are entirely different in their operation – focus groups are restricted to a small group of 6-10 individuals whereas surveys are quantitative in nature. Interviews are conducted with a sample from a population and the key characteristic they exhibit is their conversational tone.

Fundamental Types of Interviews in Research

A researcher has to conduct interviews with a group of participants at a juncture in the research where information can only be obtained by meeting and personally connecting with a section of their target audience. Interviews offer the researchers with a platform to prompt their participants and obtain inputs in the desired detail. There are three fundamental types of interviews in research:

Structured Interviews:

Structured interviews are defined as research tools that are extremely rigid in their operations are allows very little or no scope of prompting the participants to obtain and analyze results. It is thus also known as a standardized interview and is significantly

quantitative in its approach. Questions in this interview are pre-decided according to the required detail of information.

Structured interviews are excessively used in survey research with the intention of maintaining uniformity throughout all the interview sessions.

They can be closed-ended as well as open-ended – according to the type of target population. Closed-ended questions can be included to understand user preferences from a collection of answer options whereas open-ended can be included to gain details about a particular section in the interview.

Advantages of structured interviews:

Structured interviews focus on the accuracy of different responses due to which extremely organized data can be collected. Different respondents have different type of answers to the same structure of questions – answers obtained can be collectively analyzed.

1. They can be used to get in touch with a large sample of the target population.
2. The interview procedure is made easy due to the standardization offered by structured interviews.
3. Replication across multiple samples becomes easy due to the same structure of interview.
4. As the scope of detail is already considered while designing the interview, better information can be obtained and the researcher can analyze the research problem in a comprehensive manner by asking accurate research questions.
5. Since the structure of the interview is fixed, it often generates reliable results and is quick to execute.
6. The relationship between the researcher and the respondent is not formal due to which the researcher can clearly understand the margin of error in case the

respondent either degrees to be a part of the survey or is just not interested in providing the right information.

Disadvantages of structured interviews:

1. Limited scope of assessment of obtained results.
2. The accuracy of information overpowers the detail of information.
3. Respondents are forced to select from the provided answer options.
4. The researcher is expected to always adhere to the list of decided questions irrespective of how interesting the conversation is turning out to be with the participants.
5. A significant amount of time is required for a structured interview.
6. Learn more: Market Research

Semi-Structured Interviews:

Semi-structured interviews offer a considerable amount of leeway to the researcher to probe the respondents along with maintaining basic interview structure. Even if it is a guided conversation between researchers and interviewees – an appreciable flexibility is offered to the researchers. A researcher can be assured that multiple interview rounds will not be required in the presence of structure in this type of research interview.

Keeping the structure in mind, the researcher can follow any idea or take creative advantage of the entire interview. Additional respondent probing is always necessary to garner information for a research study. The best application of semi-structured interview is when the researcher doesn't have time to conduct research and requires detailed information about the topic.

Advantages of semi-structured interviews:

Questions of semi-structured interviews are prepared before the scheduled interview which provides the researcher with time to prepare and analyze the questions.

1. It is flexible to an extent while maintaining the research guidelines.
2. Researchers can express the interview questions in the format they prefer, unlike the structured interview.
3. Reliable qualitative data can be collected via these interviews.
4. Flexible structure of the interview.

Disadvantages of semi-structured interviews:

Participants may question the reliability factor of these interviews due to the flexibility offered.

Comparing two different answers becomes difficult as the guideline for conducting interviews is not entirely followed. No two questions will have the exact same structure and the result will be an inability to compare are infer results.

Unstructured Interviews:

Also called as in-depth interviews, unstructured interviews are usually described as conversations held with a purpose in mind – to gather data about the research study. These interviews have the least number of questions as they lean more towards a normal conversation but with an underlying subject.

The main objective of most researchers using unstructured interviews is to build a bond with the respondents due to which there are high chances that the respondents will be 100% truthful with their answers. There are no guidelines for the researchers to follow and so, they can approach the participants in any ethical manner to gain as much information as they possibly can for their research topic.

Since there are no guidelines for these interviews, a researcher is expected to keep their approach in check so that the respondents do not sway away from the main research motive. For a researcher to obtain the desired outcome, he/she must keep the following factors in mind:

Intent of the interview.

The interview should primarily take into consideration the participant's interest and skills.

All the conversations should be conducted within permissible limits of research and the researcher should try and stick by these limits.

The skills and knowledge of the researcher should match the purpose of the interview.

Researchers should understand the do's and don'ts of unstructured interviews.

Advantages of Unstructured Interviews:

Due to the informal nature of unstructured interviews – it becomes extremely easy for researchers to try and develop a friendly rapport with the participants. This leads to gaining insights in extreme detail without much conscious effort.

The participants can clarify all their doubts about the questions and the researcher can take each opportunity to explain his/her intention for better answers.

There are no questions which the researcher has to abide by and this usually increases the flexibility of the entire research process.

Disadvantages of Unstructured Interviews:

As there is no structure to the interview process, researchers take time to execute these interviews.

The absence of a standardized set of questions and guidelines indicates that the reliability of unstructured interviews is questionable.

In many cases, the ethics involved in these interviews are considered borderline upsetting.

Learn more: Qualitative Market Research

24. Write short note on: Questionnaire in Data Collection

Questionnaire method of data collection

A questionnaire is a research instrument consisting of a series of questions for the purpose of gathering information from respondents. Questionnaires can be thought of as a kind of written interview. They can be carried out face to face, by telephone, computer or post. Questionnaires provide a relatively cheap, quick and efficient way of obtaining large amounts of information from a large sample of people. Questionnaire is as an instrument for research, which consists of a list of questions, along with the choice of answers, printed or typed in a sequence on a form used for acquiring specific information from the respondents. The questionnaire is prepared in such a way that it translates the required information into a series of questions, that informants can and will answer.

Characteristics of a Good Questionnaire

The following are characteristics of good questionnaires:

1. It should consist of a well-written list of questions.
2. The questionnaire should deal with an important or significant topic to create interest among respondents.
3. It should seek only that data which cannot be obtained from other sources.
4. It should be as short as possible but should be comprehensive.
5. It should be attractive.
6. Directions should be clear and complete.
7. It should be represented in good psychological order proceeding from general to more specific responses.
8. Double negatives in questions should be avoided.
9. Putting two questions in one question also should be avoided. Every question should seek to obtain only one specific information.
10. It should be designed to collect information which can be used subsequently as data for analysis.

Format of Questions in Questionnaires

The questions asked can take two forms:

Restricted questions, also called closed-ended, ask the respondent to make choices — yes or no, check items on a list, or select from multiple choice answers.Restricted questions are easy to tabulate and compile.

Unrestricted questions are open-ended and allow respondents to share feelings and opinions that are important to them about the matter at hand.

Unrestricted questions are not easy to tabulate and compile, but they allow respondents to reveal the depth of their emotions.

If the objective is to compile data from all respondents, then sticking with restricted questions that are easily quantified is better.

If degrees of emotions or depth of sentiment is to be studied, then develop a scale to quantify those feelings.

Advantages of Questionnaire

1. One of the greatest benefits of questionnaires lies in their uniformity — all respondents see exactly the same questions.
2. It is an inexpensive method, regardless of the size of the universe.
3. Free from the bias of the interviewer, as the respondents answer the questions in his own words.
4. Respondents have enough time to think and answer.
5. Due to its large coverage, respondents living in distant areas can also be reached conveniently.
6. Comparability

Limitations of Questionnaire

The risk of collection of inaccurate and incomplete information is high in the questionnaire, as it might happen that people may not be able to understand the question correctly.

The main demerits of this system can also be listed here:

1. Low rate of return of the duly filled in questionnaires; bias due to no-response is often Indeterminate.
2. It can be used only when respondents are educated and cooperating.
3. The control over questionnaire may be lost once it is sent.

4. There is inbuilt inflexibility because of the difficulty of amending the approach once
5. questionnaires have been dispatched.
6. There is also the possibility of ambiguous replies or omission of replies altogether to certain
7. Questions; interpretation of omissions is difficult.
8. It is difficult to know whether willing respondents are truly representative.
9. This method is likely to be the slowest of all.

25. What is Interview Schedule? Explain following terms with respect to it: Investigator, Enumerator, Informant, Closed ended questions, Open ended questions What is Census Method? Give merits and limitations of it

Interview Schedules

Data Collection through Schedules – Very similar to the Questionnaire method. The main difference is that a schedule is filled by the trained enumerator who is specially appointed for the purpose. Enumerator goes to the respondents, asks them the questions from the Performa in the order listed, and records the responses in the space provided.

Open-ended – Questions in which the respondent answers in his own words.

Closed-ended (or Fixed Alternative) – Question in which respondent selects one or more options from pre-determined set of responses.

Simple dichotomy → Closed ended question with only two response alternatives

Multiple Choice → Closed ended question with more than two response alternatives.

Determinant choice – Multiple choice question in which respondent must select only one of the response alternatives.

Checklist question - Multiple choice question in which respondent can select more than one of the response alternatives.

Investigator - One who conducts the investigation i.e. statistical enquiry and seeks information is known as Investigator. It can be an individual person or an organization.

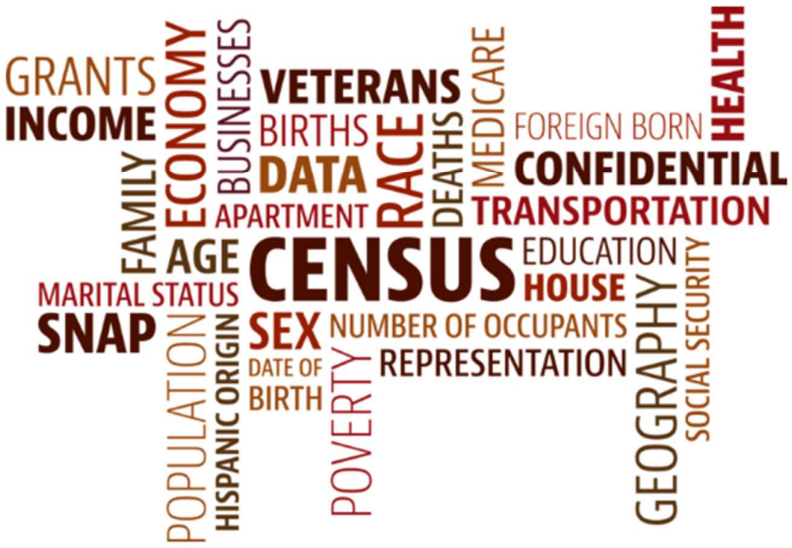
Enumerators- Enumerators are the persons who help the Investigators in the collection of data.

Informant - Informants are the respondents who supply the information to the investigator or enumerators.

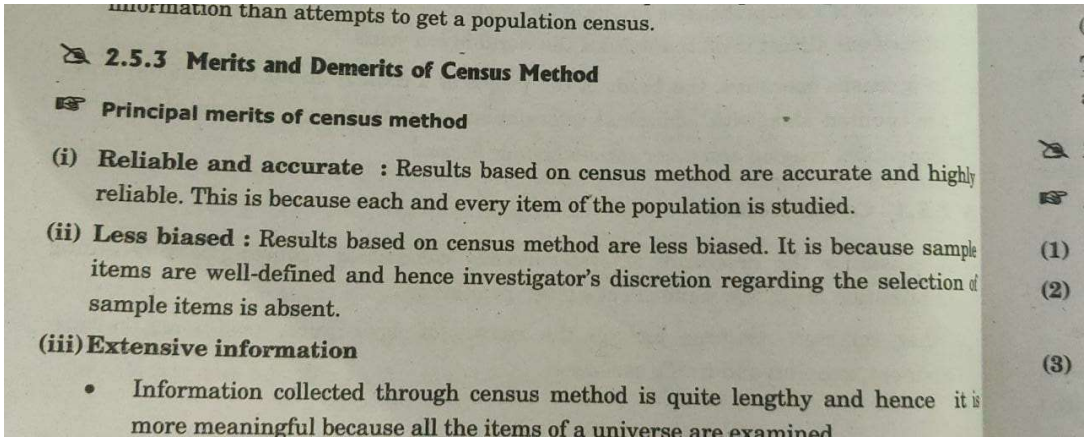
Census Definition

Under the census or complete enumeration method, the statistician collects the data for each and every unit of the population or universe. This universe is a complete set of items which are of interest in any situation.

To give you an example, if you record the marks of all students of B.Com of the Mumbai University for analysis, it is a census investigation.



The population census is another example of a census investigation. Usually, this method is recommended in cases where the area of investigation is limited and requires intensive examination of the population.



information than attempts to get a population census.

2.5.3 Merits and Demerits of Census Method

Principal merits of census method

(i) **Reliable and accurate** : Results based on census method are accurate and highly reliable. This is because each and every item of the population is studied. (1)

(ii) **Less biased :** Results based on census method are less biased. It is because sample items are well-defined and hence investigator's discretion regarding the selection of sample items is absent. (2)

(iii) **Extensive information** (3)

- Information collected through census method is quite lengthy and hence it is more meaningful because all the items of a universe are examined

- For example, population census in India gives exhaustive information relating to the number of people in different parts of the country, their age and sex composition, education, status, occupation, and the like.

(iv) **Diverse Characteristics :** By using census method, one can study diverse characteristics of the universe.

(v) **Complex investigation :** When items in a universe are of complex nature and if it is required to study each of them, only census method can produce the desired results. Data on country's population are collected by this method.

(vi) **Indirect investigation :** Census method can be successfully used in indirect investigations, relating to unemployment, property, corruption, etc.

☞ **Demerits of census method**

There are certain demerits of census method :

(1) **Costly :** Census method is very costly and hence it is generally not used for ordinary investigations. Only the government or some big institutions can afford to use this method, and that too for specific purposes only.

(2) **Lane manpower :** Census method requires a lot of manpower. Training of a large number of enumerators becomes essential but is a very difficult process.

(3) **Not suitable for large investigations :** If the universe comprises large number o items, it may not be possible to cover each and every item : census method become practically inoperative in such situations. Census method is suitable when :

(i) Area of investigation is limited.

(ii) The units are of different qualitative phenomenon.

(iii) More accuracy is desired.

The use of census method depends on the area of research, its purpose and t available resources : time, money and energy.

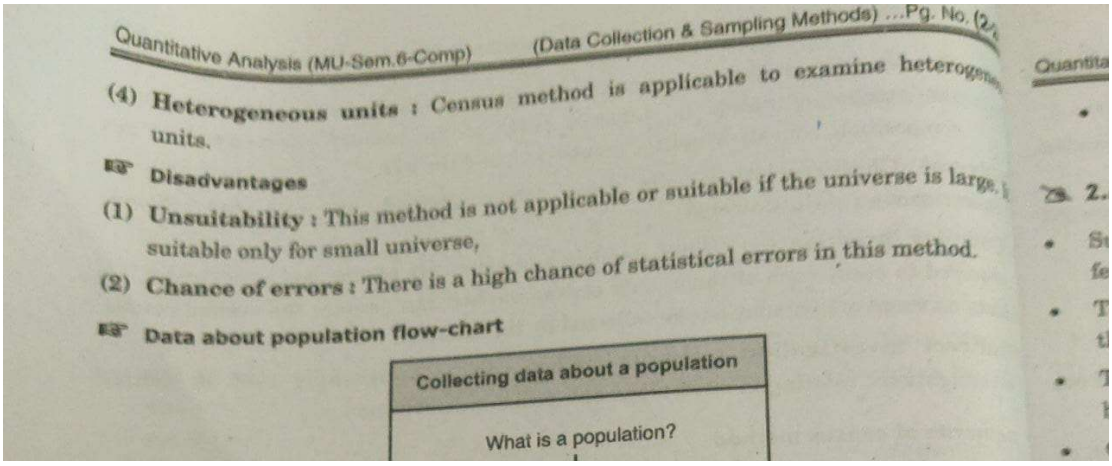
2.5.4 Advantages and Disadvantages of Census Method

☞ **Advantages**

(1) **Suitability :** This method is effective if the universe is small.

(2) **Intensive study :** Census method completely examines each unit and gath important data for intensive study.

(3) **Indispensable :** Census method is indispensable in certain cases where ot methods cannot provide reliable and accurate result.



26. Define Sampling. Give classification of different sampling methods Explain restricted random sampling methods

Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population.

Different sampling methods are widely used by researchers in market research so that they do not need to research the entire population to collect actionable insights. It is also a time-convenient and a cost-effective method and hence forms the basis of any research design. Sampling techniques can be used in a research survey software for optimum derivation.

For example, if a drug manufacturer would like to research the adverse side effects of a drug on the country's population, it is almost impossible to conduct a research study that involves everyone. In this case, the researcher decides a sample of people from each demographic and then researches them, giving him/her indicative feedback on the drug's behavior.

Types of sampling: sampling methods

Sampling in market research is of two types –

1. Probability sampling and non-probability sampling. Let's take a closer look at these two methods of sampling.

Probability sampling: Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.

Non-probability sampling: In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.

In this blog, we discuss the various probability and non-probability sampling methods that you can implement in any market research study.

Types of probability sampling

Probability sampling is a sampling technique in which researchers choose samples from a larger population using a method based on the theory of probability. This sampling method considers every member of the population and forms samples based on a fixed process.

For example, in a population of 1000 members, every member will have a 1/1000 chance of being selected to be a part of a sample. Probability sampling eliminates bias in the population and gives all members a fair chance to be included in the sample.

There are four types of probability sampling techniques:

Simple random sampling: One of the best probability sampling techniques that helps in saving time and resources, is the Simple Random Sampling method. It is a reliable method of obtaining information where every single member of a population is chosen randomly, merely by chance. Each individual has the same probability of being chosen to be a part of a sample.

For example, in an organization of 500 employees, if the HR team decides on conducting team building activities, it is highly likely that they would prefer picking chits out of a bowl. In this case, each of the 500 employees has an equal opportunity of being selected.

Cluster sampling: Cluster sampling is a method where the researchers divide the entire population into sections or clusters that represent a population. Clusters are identified and included in a sample based on demographic parameters like age, sex, location, etc. This makes it very simple for a survey creator to derive effective inference from the feedback.

For example, if the United States government wishes to evaluate the number of immigrants living in the Mainland US, they can divide it into clusters based on states such as California, Texas, Florida, Massachusetts, Colorado, Hawaii, etc. This way of conducting a survey will be more effective as the results will be organized into states and provide insightful immigration data.

Systematic sampling: Researchers use the systematic sampling method to choose the sample members of a population at regular intervals. It requires the selection of a starting point for the sample and sample size that can be repeated at regular intervals. This type of sampling method has a predefined range, and hence this sampling technique is the least time-consuming.

Stratified random sampling: Stratified random sampling is a method in which the researcher divides the population into smaller groups that don't overlap but represent the entire population. While sampling, these groups can be organized and then draw a sample from each group separately.

For example, a researcher looking to analyze the characteristics of people belonging to different annual income divisions will create strata (groups) according to the annual family income.. By doing this, the researcher concludes the characteristics of people belonging to different income groups. Marketers can analyze which income groups to target and which ones to eliminate to create a roadmap that would bear fruitful results.

Types of non-probability sampling

The non-probability method is a sampling method that involves a collection of feedback based on a researcher or statistician's sample selection capabilities and not on a fixed selection process. In most situations, the output of a survey conducted with a non-probable sample leads to skewed results, which may not represent the desired target population. But, there are situations such as the preliminary stages of research or cost constraints for conducting research, where non-probability sampling will be much more useful than the other type.

Four types of non-probability sampling explain the purpose of this sampling method in a better manner:

Convenience sampling: This method is dependent on the ease of access to subjects such as surveying customers at a mall or passers-by on a busy street. It is usually termed as convenience sampling, because of the researcher's ease of carrying it out and getting in touch with the subjects. Researchers have nearly no authority to select the sample elements, and it's purely done based on proximity and not representativeness. This non-probability sampling method is used when there are time and cost limitations in collecting feedback. In situations where there are resource limitations such as the initial stages of research, convenience sampling is used.

For example, startups and NGOs usually conduct convenience sampling at a mall to distribute leaflets of upcoming events or promotion of a cause – they do that by standing at the mall entrance and giving out pamphlets randomly.

Judgmental or purposive sampling: Judgmental or purposive samples are formed by the discretion of the researcher. Researchers purely consider the purpose of the study, along with the understanding of the target audience. For instance, when researchers want to understand the thought process of people interested in studying for their master's

degree. The selection criteria will be: “Are you interested in doing your masters in …?” and those who respond with a “No” are excluded from the sample.

Snowball sampling: Snowball sampling is a sampling method that researchers apply when the subjects are difficult to trace. For example, it will be extremely challenging to survey shelter Less people or illegal immigrants. In such cases, using the snowball theory, researchers can track a few categories to interview and derive results. Researchers also implement this sampling method in situations where the topic is highly sensitive and not openly discussed—for example, surveys to gather information about HIV Aids. Not many victims will readily respond to the questions. Still, researchers can contact people they might know or volunteers associated with the cause to get in touch with the victims and collect information.

Quota sampling: In Quota sampling, the selection of members in this sampling technique happens based on a pre-set standard. In this case, as a sample is formed based on specific attributes, the created sample will have the same qualities found in the total population. It is a rapid method of collecting samples.

Q.1 Explain the different types of random sampling. List the methods covered under each category.

Answer:

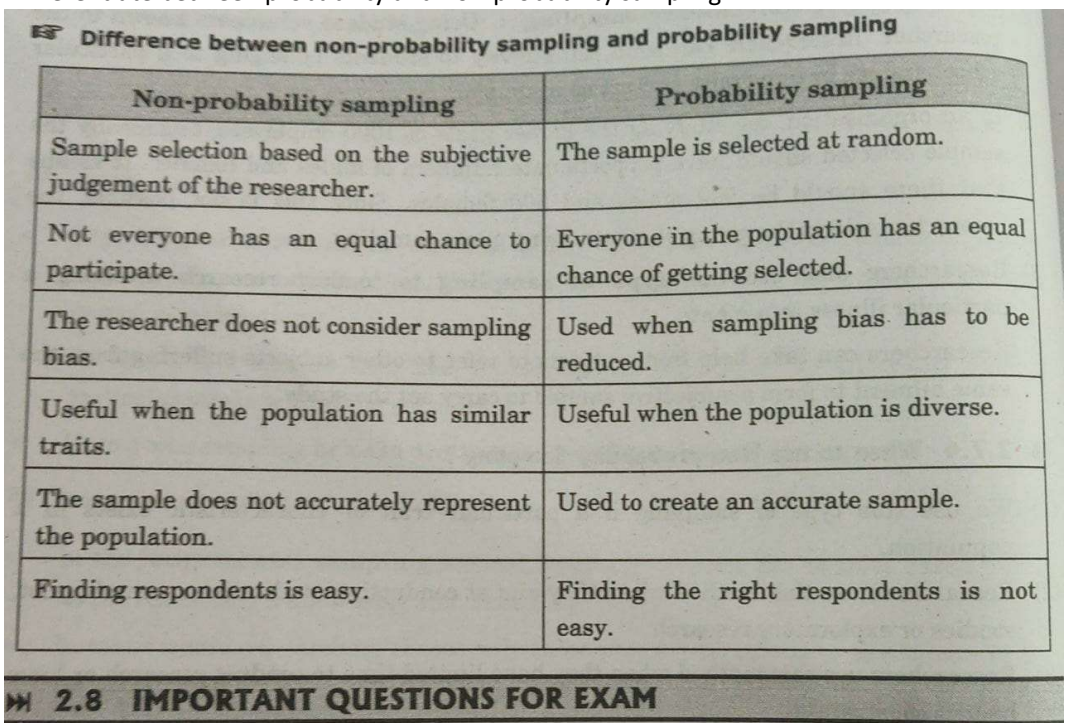
There are two types of random sampling.

1. Simple or unrestricted random sampling
2. Restricted random sampling

(A) Simple random sampling **(Unrestricted random sampling)**	• A **simple random sampling** is one in which every item of the population has an equal chance of being selected. • This method is also known as **unrestricted random sampling.** • The process used decides the chances of selection of an item, not an investigator. • Under this type of random sampling, the samples are selected by using the following two methods: 1. Lottery method 2. Table of random numbers
(B) Restricted random sampling	• In the case of the heterogeneous population, when samples are selected randomly but under certain restrictions, it is termed as restricted random sampling. • It involves the personal attention of the investigator while selecting a sample. • It is not purely random. • Important methods under this category are as follows: i. Stratified random sampling ii. Systematic sampling iii. Cluster or multistage sampling

27. Explain types of non probability random sampling methods

28. Differentiate between probability and non-probability sampling



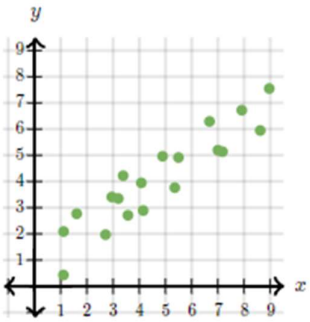
☞ **Difference between non-probability sampling and probability sampling**

Non-probability sampling	Probability sampling
Sample selection based on the subjective judgement of the researcher.	The sample is selected at random.
Not everyone has an equal chance to participate.	Everyone in the population has an equal chance of getting selected.
The researcher does not consider sampling bias.	Used when sampling bias has to be reduced.
Useful when the population has similar traits.	Useful when the population is diverse.
The sample does not accurately represent the population.	Used to create an accurate sample.
Finding respondents is easy.	Finding the right respondents is not easy.

⏭ 2.8 IMPORTANT QUESTIONS FOR EXAM

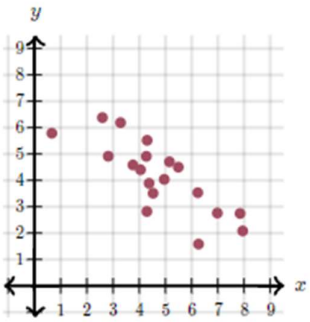
29. Examples on Stratified Random Sample & Cluster Sampling
30. What is correlation? Sketch the scatter plots for positive correlation, negative correlation, strong correlation, no correlation.

Positive correlation



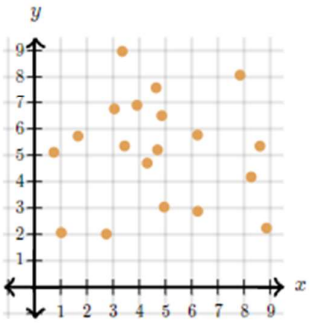
When the y variable tends to decrease as the x variable increases, we say there is a **negative correlation** between the variables.

Negative correlation



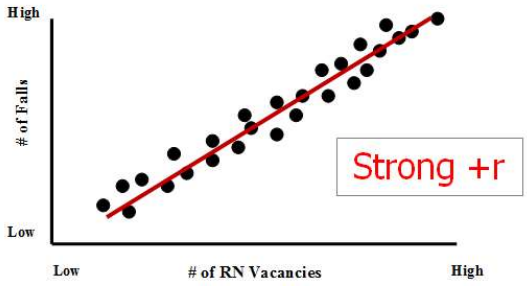
When there is no clear relationship between the two variables, we say there is **no correlation** between the two variables.

No correlation

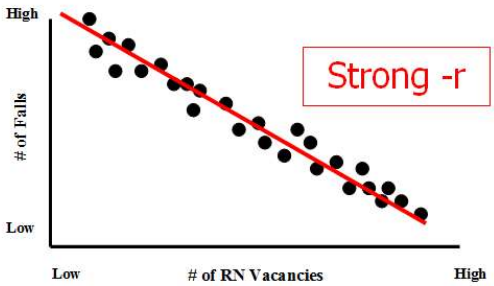


Want to learn more about types of correlation? Check out this video.

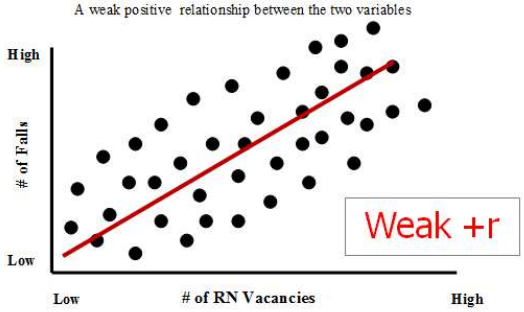
A strong positive relationship between the two variables



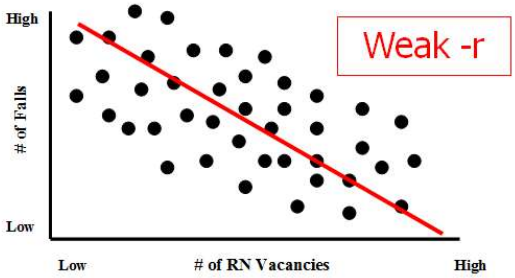
A strong negative relationship between the two variables



A weak positive relationship between the two variables



A weak negative relationship between the two variables



3.1 CORRELATION

- Correlation is the relationship that exists between two or more variables. T variables are said to be **correlated** if a change in one variable affects a change in t other variable. Such a data connecting two variables is called bivariate data.
- Correlation measures the closeness of the relationship between the variables.
- Some examples of a relationship are as follows:
  - Relationship between heights and weights
  - Rainfall and crop-yield correlated.
  - Relationship between age of husband and age of wife

3.1.1 Types of Correlation

Correlation is classified into four types:

(1) Positive correlation and negative correlation

(2) Simple correlation and multiple correlation

(3) Partial correlation and total correlation

(4) Linear correlation and nonlinear correlation

Positive and negative correlations

Positive correlation

- If the value of one variable increases, the value of the other variable also increases, or, if value of one variable decreases, the value of the other variable also decreases. This type of correlation is said to be **positive correlation**.

e.g. The correlation between heights and weights of group of persons

Height (cm)	145	150	160	162	165	175
Weight (kg)	55	60	62	65	67	68

Negative correlation

- If the value of one variable increases, the value of the other variable decreases, or, if value of one variable decreases, the value of the other variable increases. This type of correlation is said to be **negative correlation**.

e.g. The correlation between the price and demand of a commodity

Price (Rs per unit)	15	10	8	7	6	3
Demand (units)	150	200	220	260	300	320

1. Simple and multiple correlations

☞ **Simple correlation**

The relationship between only two variables is described as simple correlation.

e.g. The quantity of money and price level, demand and price

☞ **Multiple correlation**

The relationship between more than two variables is described as multiple correlation.

e.g. Relationship between price, demand and supply of a commodity

2. Partial and total correlations

☞ **Partial correlation**

When more than two variables are studied excluding some other variables, the relationship is termed as partial correlation.

☞ **Total correlation**

When more than two variables are studied without excluding any variables, the relationship is termed as total correlation.

Linear and nonlinear correlations

☞ **Linear correlation**

If the ratio of change between two variables is constant, the correlation is said to be linear. The graph of a linear relationship will be a straight line.

e.g.

Milk (*l*)	5	10	15	20	25	30
Curd (kg)	2	4	6	8	10	12

☞ **Nonlinear correlation**

If the ratio of change between two variables is not constant, the correlation is said to be nonlinear.

The graph of a nonlinear relationship will be a curve.

31. Give the values of correlation coefficient for perfectly positive correlation, perfectly negative correlation and zero correlation. Hence, give the range of correlation coefficient. Write the equations for lines of regression of y on x and x on y respectively.

e.g.

Price (Rs per unit)	15	10	8	7	6	3
Demand (units)	150	200	220	260	300	320

3.1.2 Scatter Diagram

There are various relationship between two variables represented by the following scatter diagrams.

- **Perfect positive correlation :** If all the plotted points lie on a straight line rising from the lower hand corner to the upper righthand corner, the correlation is said to be perfect positive correlation.

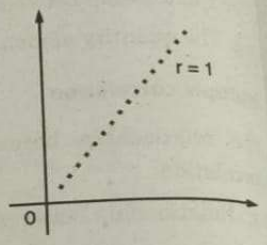


Fig. 3.1.1 : Perfect positive correlati

- **Perfect negative correlation :** If all the plotted points lie on a straight line from the upper left-hand corner to the lower right-hand corner, the correlation is said to be perfect negative correlation.

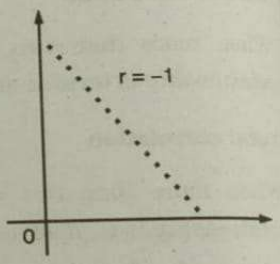


Fig. 3.1.2 : Perfect negative correlati

- **High degree of positive correlation :** If all the plotted points lie in the narrow strip, rising from the lower left-hand corner to the upper right -hand corner, it indicates a high degree of positive correlation.

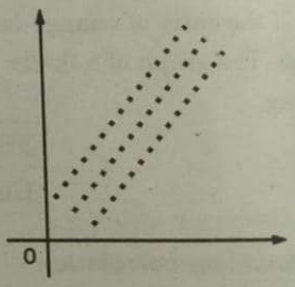


Fig. 3.1.3 : High degree of positive correlation

- **High degree of negative correlation :** If all the plotted points lie in a narrow strip, falling from the upper left-hand corner to the lower right-hand corner, it indicates the existence of a high degree of negative correlation. degree of negative correlation.

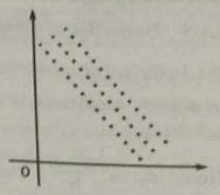


Fig. 3.1.4 : High degree of negative correlation

- **No correlation :** If all the plotted points lie on a straight line parallel to the x-axis or y-axis, it indicates the absence of any relationship between the variables.

3.1.3 Karl Pearson's Coefficient of Correlation

The coefficient of correlation is the measure of correlation between two random variables X and Y, is denoted by r.

$$r = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \qquad ...(1)$$

Where,

$\text{cov}(X, Y)$ = the covariance of variables X and Y

$$= \frac{1}{n}\sum (x-\bar{x})(y-\bar{y})$$

σ_X = the standard deviation of variable X

$$= \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

σ_Y = the standard deviation of variable Y

$$= \sqrt{\frac{\sum (y-\bar{y})^2}{n}}$$

So,
$$r = \frac{\frac{1}{n}\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\sum (x-\bar{x})^2}{n}}\sqrt{\frac{\sum (y-\bar{y})^2}{n}}}$$

By simplifying,
$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}}$$

The above expression r is called Karl Pearson's coefficient of correlation.

3.1.4 Properties of Coefficient of Correlation

(1) The coefficient of correlation lies between – 1and 1. i.e. $-1 \le r \le 1$.

(2) Correlation coefficient is independent of change of origin and change of scale. i.e. $r_{xy} = r_{d_x d_y}$

Here, $d_x = \frac{x-a}{h}$ and $d_y = \frac{y-b}{k}$

i.e. $$r = \frac{n\sum d_x d_y - \sum d_x \sum d_y}{\sqrt{n\sum d_x^2 - (\sum d_x)^2}\sqrt{n\sum d_y^2 - (\sum d_y)^2}}$$

(3) Two independent variables are uncorrelated. i.e. $r = 0$.

Ex. 3.1.1 : Calculate the coefficient of correlation for the following data.

x	9	8	7	6	5	4	3	2	1
y	15	16	14	13	11	12	10	8	9

☑ **Soln. :**

Here, n = 9

x	y	x^2	y^2	xy
9	15	81	225	135
8	16	64	256	128
7	14	49	196	98
6	13	36	169	78
5	11	25	121	55
4	12	16	144	48
3	10	9	100	30
2	8	4	64	16
1	9	1	81	9
$\sum x = 45$	$\sum y = 108$	$\sum x^2 = 285$	$\sum y^2 = 1356$	$\sum xy = 597$

The coefficient of correlation is

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{9(597) - (45)(108)}{\sqrt{9(285) - 45^2}\sqrt{9(1356) - 108^2}}$$

$$r = 0.95$$

This allows the researcher to estimate the **conditional expectation** (or population average value) of the dependent variable when the independent variables take on a given set of values.

3.2.2 Lines of Regression

Definition

(1) Line of regression of y on x is the line which gives the best estimate for the value of y for any specified value of x.

We have already seen that the equation of line of regression of y on x is

$$y - \bar{y} = r\left(\frac{\sigma_y}{\sigma_x}\right)(x - \bar{x}) \quad \ldots(i)$$

(2) The line of regression of x on y is the line which gives the best estimate of x for any given value of y. And the equation of line of regression of x on y is

$$x - \bar{x} = r\left(\frac{\sigma_x}{\sigma_y}\right)(y - \bar{y}) \quad \ldots(ii)$$

Angle between regression lines

Whenever two lines intersect, there are two angles bet...

Lines of Regression

- Line of Regression of X on Y

$$X - \bar{X} = r\frac{\sigma_x}{\sigma_y}(Y - \bar{Y}) \qquad b_{xy} = r\frac{\sigma_x}{\sigma_y}$$ Regression Coefficient

- Line of Regression of Y on X

$$Y - \bar{Y} = r\frac{\sigma_y}{\sigma_x}(X - \bar{X}) \qquad b_{yx} = r\frac{\sigma_y}{\sigma_x}$$ Regression Coefficient

32. What is Linear Regression

Regression Model

- It is easier to use mathematical tools on linear functions than on nonlinear functions.
- So linear regression model is more preferred over non-linear regression model.
- Linear function: dependent variable – *y*, independent variable – *x*

$$y = \alpha + \beta x$$

- *A- intercept term* β: slope or rate of change in *y* wrt *x*
- Relationship between *y* and *x* is linear

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Linear regression is a statistical method used to analyze the relationship between a dependent variable (also called the response or outcome variable) and one or more independent variables (also called the predictor or explanatory variables).

The goal of linear regression is to find the best linear relationship between the dependent variable and the independent variable(s) that can be used to make predictions about the dependent variable. This is achieved by fitting a straight line (or plane, in the case of multiple independent variables) to the data that minimizes the distance between the line and the actual data points.

Linear regression can be used for both simple and multiple regression analysis. Simple linear regression involves only one independent variable, while multiple linear regression involves two or more independent variables.

Linear regression is widely used in many fields such as economics, finance, psychology, and engineering for modeling relationships between variables and making predictions about future outcomes.

☞ **Linear regression**

In linear regression, the relationship between the variables, is linear and is represented by straight line, known as a regression line or the line of average relationship or prediction equation.

☞ **Regression line of y on x**

Suppose in the study of relationship between two variables x and y if y is dependent on x, then the simple linear equation $y = a_0 + a_1 x$ is known as regression line of y on x.

Similarly, if x depends on y, then $x = b_0 + b_1 y$

is known as regression line of x on y.

In multiple regression the equation is, $y = f(x_1, x_2, \ldots x_n)$

In multiple linear regression, f is linear,

i.e. $y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$

In multiple non-linear regression, f is non-linear, for example,

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2 + b_4 x_1^2 + b_5 x_2^2$$

3.4.3 Some Results

33. Explain the mathematical equation of Linear Regression

In linear regression, we try to fit a straight line (in case of simple linear regression) or a plane (in case of multiple linear regression) to the data that can best explain the relationship between the independent and dependent variables. This straight line or plane can be represented by the following mathematical equation:

$y = b0 + b1*x1 + \varepsilon$

where:

- y is the dependent variable (also called the response or outcome variable)
- x1 is the independent variable (also called the predictor or explanatory variable)
- b0 is the intercept (the value of y when x1 is equal to 0)
- b1 is the slope (the change in y for every unit increase in x1)
- ε is the error term (the difference between the predicted value of y and the actual value of y)

The equation shows that the value of the dependent variable y is a function of the value of the independent variable x1, and the intercept and slope coefficients (b0 and b1) that we estimate from the data. The error term represents the variability or randomness in the data that cannot be explained by the model.

The objective of linear regression is to estimate the values of the intercept and slope coefficients (b0 and b1) that minimize the sum of the squared errors between the predicted values of y and the actual values of y for all the data points. This is typically done using a method called Ordinary Least Squares (OLS) regression. Once we have estimated the coefficients, we can use the equation to make predictions about the value of y for any given value of x1.

34. What is the significance of slope and intercept with respect to Linear regression

Regression Model

- It is easier to use mathematical tools on linear functions than on nonlinear functions.
- So linear regression model is more preferred over non-linear regression model.
- Linear function: dependent variable – y, independent variable – x

$$y=\alpha+\beta x$$

- A- *intercept term* β: slope or rate of change in y wrt x
- Relationship between y and x is linear

Model

- In practice, exact relationship may not hold due to several reasons.
- Introduce a random error component (u)

$$y=\alpha+\beta x+u$$

- u reflects the effect of
  - -Total effect of all variables/ relevant factors left
  - - randomness in human behaviour/ responses
  - - qualitative variables etc

The slope and intercept are two important parameters in linear regression that help to describe the relationship between the independent variable(s) and dependent variable. Here's how they are significant:

1. Slope:
   The slope (represented by the coefficient b1) is the change in the dependent variable (y) for each unit change in the independent variable (x). In other words, it tells us the rate of change of y with respect to x. If the slope is positive, then an increase in x leads to an increase in y. If the slope is negative, then an increase in x leads to a decrease in y. The magnitude of the slope tells us how strong the relationship between x and y is. A larger magnitude slope indicates a stronger relationship between the variables.
2. Intercept:
   The intercept (represented by the coefficient b0) is the value of the dependent variable (y) when the independent variable(s) (x) is zero. In other words, it is the starting point of the line or plane that we are fitting to the data. The intercept can give us important information about the data that is not apparent from the slope. For example, if the intercept is not zero, it suggests that there is a non-zero value of y when x is equal to zero.

Together, the slope and intercept provide a simple and interpretable way to describe the relationship between the variables in linear regression. They can help us to understand the direction, strength, and starting point of the relationship between the independent and dependent variables, and to make predictions about the value of the dependent variable for any given value of the independent variable.