UNIT 8: TECHNIQUES OF DATA COLLECTION

UNIT STRUCTURE

8.1 Learning objectives

8.2 Introduction

8.3 Sources of Data Collection

- 8.3.1 Observation Method
- 8.3.2 Interview Method

8.4 Meaning of Sampling

- 8.4.1 Sampling Design
- 8.4.2 Characteristics of a good sampling design
- 8.4.3 Advantages of Sampling
- 8.4.4 Types of Sampling

8.5 Questionnaire

- 8.5.1 Preparation of a questionnaire
- 8.5.2 Features of a good questionnaire
- 8.5.3 Merits of Questionnaire
- 8.5.4 Demerits of Questionnaire

8.6 Let Us Sum Up

8.7 Further reading

8.8 Answers to your progress

8.9 Possible Questions

8.1 LEARNING OBJECTIVES:

After going through this unit, you will be able to –

- understand the meaning of primary and secondary data
- identify the different methods of data collection like interview, observation
- explain sampling as a technique that simplifies data collection
- discuss the meaning, features, advantages and disadvantages of a questionnaire and also the preparation of a questionnaire.

8.2 INTRODUCTION

Data collection can be considered the heart of research because data is the main core of any research. Without proper and adequate data we cannot derive any acceptable and satisfactory findings. Data collection is, therefore, one of the most vital aspects of research. Once the topic for research or the research problem is selected, the researcher needs to decide on the methods he/she would use to collect data. The methods that the researcher would use must complement the research design and only then can he/she arrive at a valid conclusion. Thus, a lot of effort, caution and hard work are necessary on the part of the researcher while deciding on the methods of data collection and collecting data. In the previous unit we have discusses the fundamental concepts or esearch and the different types of research. In this unit we would discuss the meaning and concept of two types or sources of data i.e. Primary data and Secondary data. After having an understanding about them, we would learn the different methods of collecting primary or secondary data like Observation, Interview etc... Further we would also try to understand Sampling as a technique that helps in collecting data. Finally, we will have a discussion on Questionnaire, way of preparing a questionnaire, essential features of a good questionnaire and the advantages and disadvantages of questionnaire.

8.3 SOURCES OF DATA COLLECTION

Data collection as mentioned before is a very important component of research. There can be no research without data. Thus, immediately after identifying the topic of research, the researcher has to decide on the method and technique of data collection. There are two sources of data- Primary and Secondary.

PRIMARY DATA: Primary data are those data which are collected for the first time and are original in nature. These are first hand data. The researcher should be the first one to collect these data. We can collect it through experiments by conducting experimental research or through surveys, or with the help of observation and interviews. Thus it means that there are

several methods of collecting primary data. The most important methods of gathering primary data are: Interview method, Observation, Focus group discussion, etc.

SECONDARY DATA: Secondary data means the data that are already available i.e. they refer to the data which have already been collected and analysed by someone else. The researcher himself /herself does not collect the data but uses or refers to the data collected by others. Secondary data may be published or unpublished data. Usually published data are available in various publications of the central, state or local governments; various publications of foreign governments or of international bodies and their subsidiary organisations; technical and trade journals; books, magazines and newspapers; reports prepared by research scholars, universities; public records and statistics, historical documents and other sources of published information. The sources of unpublished data are diaries, letters, unpublished biographies and autobiographies etc.

SOURCES OF PRIMARY DATA:

8.3.1 OBSERVATION

It is an important technique of collecting primary data. It implies the researcher being present in the field and observing the activities and events in the field. Observation can be both participant as well as non participant. 31It is a basic method of obtaining information about the world around us. However all observation is not scientific observation. It becomes a scientific tool for the researcher to the extent that it serves a formulated research purpose and it is planned systematically

- **Participant Observation:** Here the researcher tries to make himself/herself a member of the group he or she is studying. The researcher participates in the events and activities in the field and makes an attempt to get the insiders perspective. He/ She also builds up close relationship with the members of the group and tries to experience any occurrence or event like the other members. Since the researcher stays in the field to be a

part of the everyday life activities of the group members, so this type of research is called Participant observation.

- **Merits of Participant Observation:**
  a. The researcher can conveniently witness, understand and the natural behaviour of the group.
  b. Participant observation enables one to get the insider's perspective. The study can be very in-depth in this process.
  c. The researcher can gather much more information which would not have been possible otherwise.
  d. Since the researcher comes in close contact with the respondents, he/she may be able to gain their faith and trust and thereby get their views, ideas or perceptions even on sensitive or difficult issues.
- **Limitations of Participant Observation:**
  a. This method demands from an observer a high level of empathy. The observer is required to recreate in his/her imagination, the thought and feelings which are in the minds of his subjects
  b. This type of research involves a lot of time.
  c. It is not very conducive for all situations. For example if someone is conducting research on a terrorist organisation then it may not be possible for him/her to have participant observation or it will be difficult and unsafe.
  d. The observer may sometime become one with the group and lose his/her objectivity. As a consequence he/she may ignore several occurrences in the group and not record them at all thinking that they are just commonplace and natural
- **Non Participant Observation:** Here the researcher is not an active participant. In non participant observation it is not necessary for the researcher to make himself/herself a member of the group under the study. He/She must be present in the field, but need not take part in the group actions and activities. The researcher observes the events in a detached manner without any active participation in

it. The researcher would observe and analyse the events or activities as an outsider or non member of the group. The advantage of this type of observation is that there is less chance of biasness or loss of objectivity on the part of the researcher.

8.3.2 INTERVIEW

It is probably the most common technique of primary data collection. It implies oral questioning and response or verbal communication. Here the researcher goes to the field and verbally converse with the respondents and gets their views and opinions on the different issues related to his/her topic of study. It can be considered as 'meeting' in which the interviewer puts questions to the interviewer and records his responses. This meeting is always face-to-face where both the parties communicate with each other not only through gestures, mannerisms, facial expression etc. Interview is one of the most effective and extensively used tools of data collection. There are different types of interviews:

- **Unstructured Interviews:** Unstructured interviews are those where issues are raised in a conversational form. The questions are not predetermined or pre-set. The researcher communicates with the respondents keeping in mind the broad framework of the research and tries to gather as much information as possible. The questions are open ended, flexible, and allow the respondent to give his/her opinion, put forward his/her views and arguments in detail. Since the interview is in a conversational form, questions are put in a way that helps in maintaining the flow of the interview. These interviews are most often in-depth and are effective in the context of sensitive issues. Usually qualitative data are collected through this method. This type of interviews are generally used in exploratory studies or in studies pertaining to sensitive subjects.

- **Semi Structured Interview:** In this type of interview ask a number of specific major questions but he is also free to probe beyond the answers to these questions. Interview guides are used in this type of interview so that it allows comparison between the data or information gathered from different interviews. The questions here are both close and open ended. The interview guides help in putting the questions in a proper sequence so that the flow of the interview is well maintained and the questions remain within the broad ambit of the research topic. This type of interview generates both quantitative and qualitative data.
- **Structured Interview:** Structured questionnaires are used in this type of interview. The questions are predetermined and put in a definite sequence to seek opinions and perceptions on the issues pertaining to the topic of study. The questionnaires usually contain number of short questions, the answers of which are coded numerically. All interviews are conducted in the same way to generate reliability using set questions and set response codes. Structured interviews generate a lot of information but they lack in depth. Both quantitative and qualitative data can be gathered with the help of structured interviews.



CHECK YOUR PROGRESS

Q.1: What is meant by Participant Observation?

_____

_____

Q.2: What is meant by Unstructured Interview?

_____

_____

Q.3: Is non-participant observation a source of secondary data?

_____

_____

8.4 MEANING OF SAMPLING

Sampling is one of the most effective and widely used techniques for collecting data. The term sampling implies the selection of a section or a part of a group with the aim of collecting information about the whole group. Sampling is used mainly in survey kind of research or quantitative approach. When the population or the universe of the study is too large it becomes difficult to investigate or study each and every individual member, and therefore the researcher selects a limited number of respondents who can adequately represent the entire group. Thus, he/she takes up sample study method by which he/she selects only a few individuals from the entire group and studies them. The selected respondents or individuals are called the samples and the process of selecting them is called the sampling technique. Sample is, therefore, only that subset which gives us an idea about the total population. The researcher, however, needs to be very careful while selecting the sample because the sample he/she selects must be of adequate size; it must be able to represent the whole group and the findings he/she derives from studying the samples must be applicable and valid for the entire group. Sampling is helpful in the study of a large population or universe. It depicts the scenario for the entire group even if each and every member of the group is not a respondent or not taken as a sample of study. Further, Sampling is helpful because it saves time and cost. One can conduct a vast study in less time and cost using this technique.

8.4.1 SAMPLING DESIGN:

A sample design refers to a specific, explicit plan for selecting or arriving at a sample from a given population or universe. It implies the method or the procedure that would be adopted by the researcher to select his sample. The researcher prepares his sample design on the basis of the nature and extent of his study. The sample design is formulated prior to the collection of data. The selection of samples and the collection of data are all guided by the sample design. The design also depicts the sample size of the study.

8.4.2 Characteristics of a Good Sampling Design:

- Sampling design must emphasise on an adequate Sample size. It must be proportionate to the entire group or universe of study.
- Sample design must result in a representative sample.
- Sample design must be such as to lead to minimum sampling error.
- Sample design must be cost effective.
- Sample design must have a control over or proscribe sampling biasness.
- The results derived from the sample should be applicable to or valid for the entire group.

8.4.3 Advantages of Sampling

- Sampling is cost effective. Since only a section of the entire population is studied in this method, it reduces the cost to be incurred in data collection by the researcher.
- Sampling saves time. It takes less time in sampling technique because here we do not study the whole group but only a few members of the group. It helps in quicker results or findings.

8.4.4 Sampling

On the basis of representation and element selection technique there are different types of sample designs. Sample may be probability sampling and non probability sampling on the representation basis, while on element selection basis; the sample may be either restricted or unrestricted. When each sample element is drawn individually from the population at large, then the sample so drawn is known as ‘unrestricted sample; whereas all other forms of sampling are covered under the term ‘restricted sampling’. Based on the nature of the topic of research, the researcher decides on the type of his/her sample.

Let us now try to understand the meaning and types of Probability non probability sampling

NON PROBABILITY SAMPLING:

In non probability sampling the researcher selects the sample according to his/her convenience or purpose. The samples are not selected randomly but rather they are chosen intentionally or deliberately. The choice and the assessment of the researcher play a vital role in selection of the samples. Here the sample may not represent all the characteristics of the population. Therefore there is little chance for scientific generalisation in such research.

TYPES OF NON PROBABILITY SAMPLING

Convenience Sampling: This type of sampling depends upon the convenience of the researcher. He/She selects the samples according to his convenience. If there are too many constraints in using the other types of sampling, the researcher goes for this type. Lack of time, difficulty in finding respondents, etc. leads the researcher to use convenience sampling.

Purposive Sampling: This type of sample is also known as judgement sample . The purpose of sampling is defined more by the purpose of research. The sample is not selected on random basis but depending on the rationale or purpose of research, the samples are chosen. It is selected purposively for a definite objective. It is very simple to draw. It is often used in exploratory investigations which proceede major survey. It is more representative than the random sample if the size of sample is small.

Quota Sampling: When the population is stratified or heterogeneous and its size is large we go for quota sampling. Here we fix or assign quotas and we fill these quotas from the different strata's in the given population. This sampling reduces cost of preparing sample and fieldwork.

Snowball Sampling: Snowball sampling technique is used when the researcher is not able to or finds difficulty in identifying his sample population. Here the researcher may find one or two person whom

he/she can include as his sample population and then from them he/she gets the information about other persons whom he can include as his sample. This technique is therefore used in such situations or circumstances when the researcher is not able to locate his sample population. For example if a researcher is studying on HIV AIDS he may not know the people who are suffering from AIDS. In this case if he/she can find out one person who is suffering from the disease, that person may give the researcher contacts and information's about other people suffering from the same disease. In this way from one or two person the researcher may get a sufficient number of people who fall into the same category and whom he can include as his/her sample in the study. The sample here may not be representative in nature.

PROBABILITY SAMPLING

Probability Sampling is also known as 'random sampling' or 'chance sampling'. Here every unit of the universe or total population has the equal chance of being selected. Samples are selected on random basis by using lottery or other similar methods. The samples are thus chosen mechanically. There is no purposive or deliberate selection of sample in this sampling design. The results obtained from random sampling can be assured in terms of probability i.e. we can measure the errors of estimation or the significance of results obtained from a random sample. In probability sampling, the sample represents the population and the findings can be generalized to the whole population.

TYPES OF PROBABILITY SAMPLING

Simple Random Sampling: This type of sampling is also known as chance sampling. It is the most basic type of probability sampling. Here samples are selected from the random table and therefore each and every unit in the universe has the equal chance of being included in the sample. This procedure gives each item an equal probability of being selected.

Systematic sampling: An element of randomness is usually introduced into this kind of sampling by using random numbers to pick up the unit with which to start. This procedure is useful when sampling frame is available in the form of a list. In such a design the selection process starts by picking some random point in the list and then every *n*th element is selected until the desired number is secured.

Stratified sampling: Stratified sampling is used when the population which is to be studied does not constitute a homogeneous group. When the population is diverse and has numerous categories of people then we go for stratified sampling. Here we select or pick up samples from each category or group in the given population. For this purpose, the different groups or categories in the population are identified and then samples are selected from each strata randomly.

Multi-stage sampling: This technique is meant for big inquiries extending to a considerably large geographical area like an entire country. Under multi-stage sampling the first stage may be to select large primary sampling units such as states, then districts, then towns and finally certain families within towns. If the technique of random sampling is applied at all stages, then the sampling procedure is described as multi-stage random sampling.

Cluster sampling: Cluster sampling involves grouping the population and then selecting the groups or the clusters rather than individual elements for inclusion in the sample. The clustering approach makes the sampling procedure relatively easier and increase the efficiency of fieldwork, especially in the case of personal interviews.

8.5 QUESTIONNAIRE

There are various tools available for conducting personal interview. Questionnaire is one such research tool used to collect primary data from the interviewee. A questionnaire consists of a number of questions printed

or typed in a definite order or on a form or set of forms. The questionnaire is given to the respondents who are expected to read and understand the questions and write down the reply in the space meant for the questionnaire itself. The respondents have to answer the questions on their own.

8.5.1 PREPARATION OF A QUESTIONNAIRE

There are certain steps and procedures that one needs to follow while preparing a meaningful questionnaire. Some of the important aspects are discussed below.

- **General Form:** Questionnaires are generally structured in which there are definite, concrete and predetermined questions. Each interviewee is asked the same questions. The questions are therefore presented with exactly the same wording and in the same order to all the respondents.
- **Question Sequence:** The researcher must be very careful about the question sequence while preparing a questionnaire. The questionnaire would be meaningful and effective only when the questions are put in a definite order or sequence. The questions must follow a specific pattern; each question following from the above. The questions must be easy; its meaning clear and there should be a flow from one question to the other so that it helps the researcher to easily understand and give his/her views/ answers. The initial or the first few questions should be the easiest to answer and should be able to arouse the interest of the respondents.
- **Question formulation and wording:** The questions put in the questionnaire should be simple and written in a lucid language. The questions should not lead to any misunderstanding and miscommunication. The questions should be such that they are understood easily. Further, it should convey only one thought at a time.

8.5.2 FEATURES OF A GOOD QUESTIONNAIRE

- Questionnaire should be comparatively short and simple

- Questions should proceed in logical sequence moving from easy to more difficult questions
- Personal and intimate questions should be asked towards the end.
- One must try to avoid technical or difficult terms and vague language which may be misunderstood or misinterpreted.
- The researcher must avoid questions which affect or hurt the sentiments of the respondents.
- There should be sufficient space for the respondent to answer the question especially the open ended ones.

8.5.3 Merits of a questionnaire

- It is cost effective, especially in the case of a large population. It saves money that one would have to incur if he/she has to visit and attend to each respondent.
- It can save time of the researcher. Rather than personally meeting and interviewing every respondent in case of a large population, the researcher can simply send the questionnaires to many at the same time and get their responses within a short duration.
- Here there is less chance of the interviewer's biasness or manipulation. Since he/she is not present, the respondent can write the replies on his/her own without being influenced by the researcher or interviewer.
- The respondents get enough time and space to provide their views and opinions.
- Large samples can be made use of and thus the results can be made more dependable and reliable

8.5.4 Demerits of a questionnaire

- Questionnaires are useful only in the case of educated and cooperative respondents.
- Sometimes the respondents may misunderstand the questions and hence the researcher may fail to obtain the desired result.

- There is a possibility of ambiguous replies or omission of replies altogether to certain questions; interpretation of omissions is difficult
- At times it becomes difficult to receive the questionnaires back from the respondents and then it delays the whole process.



CHECK YOUR PROGRESS

O.4: Why is there less expenditure in using questionnaires in research?

_____

_____

O.5: Why is questionnaire usually considered a slow method?

_____

_____



8.6 LET US SUM UP

- After identifying the topic of research, the researcher has to decide on the method and technique of data collection. There are two sources of data- Primary and Secondary.
- Primary data are those data which are collected for the first time and are original in nature.
- Secondary data means data that are already available i.e. they refer to the data which have already been collected and analysed by someone else.
- The most important methods of gathering primary data are: Interview method; Observation, Focus Group discussion etc while journals; books, magazines and newspapers; reports prepared by research scholars, universities; public records and statistics, historical documents and other sources of published information comprise the secondary data.

- Observation is an important technique of collecting primary data. Observation can be both participant as well as non participant.
- In Participant Observation the researcher tries to make himself/herself a member of the group he or she is studying, while in Non Participant Observation the researcher is not an active participant.
- Interview in simple words mean oral/verbal inquiring and response or verbal communication.
- The term sampling refers to the selection of a section or a part of a group with the aim of collecting information about the whole group. There are basically two type of Sampling design-Probability Sampling; Non-Probability Sampling.
- A questionnaire is one which consists of a number of questions printed or typed in a definite order or on a form or set of forms. It is given to the respondents who are expected to read and understand and write down their reply in the space provided.



8.7 FURTHER READING

1. Guthrie Gerard, *Basic Research Methods: An Entry to Social Science Research*. New Delhi: Sage publication. India. 2010.
2. Kothari C.R. *Research Methods and Techniques.* New Delhi: New Age International Publisher. 2004,
3. Young, P.V. *Scientific Social Surveys and Research*. New Delhi: Prentice Hall, 2009



8.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q.No.1: Participant Observation is a method of data collection where the researcher tries to be an active participant and makes himself a member of the group he or she is studying.

Ans to Q.No.2: Unstructured interviews are those where the researcher raises the questions in a conversational form. The questions are not

predetermined or pre-set; rather the questions are picked up from the opinions given by the researcher.

Ans to Q.No.3: No, non participant observation is not a source of secondary data; rather it is a source of primary data.

Ans to Q.No.4: Questionnaire involves less cost because the researcher does not have to visit every respondent, especially those who are at a far off place. It saves money that one would have to incur if he/she has to visit and attend to each respondent.

Ans to Q.No.5: Sometimes it becomes difficult to receive the questionnaires back from the respondents and then it delays the whole process. It takes a long time to collect the questionnaires from the respondents and hence it is considered a slow method.



8.9 MODEL QUESTIONS

A) SHORT QUESTIONS (Answer within 150 words)

Q1: Why is data the most important component of research?

Q2: Distinguish between primary and secondary data?

Q3: What is semi structured interview?

B) LONG QUESTIONS (Answer within 300-500 words)

Q1: Discuss in detail the methods used for collecting primary data.

Q2: What is sampling? How is probability sampling different from non probability sampling?

Q3: 'Data Collection involves different methods and techniques'. Elaborate the Statement.

*** ***** ***

UNIT 9 : THEORY BUILDING

UNIT STRUCTURE

9.1 Learning Objectives

9.2 Introduction

9.3 Meaning of theory

9.4 Methods used in theory building

9.5 Construction of a theory

9.6 Fact versus theory

9.7 Theory Typologies/Paradigm in theory building

9.8 Relation between theory and research

9.9 Let Us Sum Up

9.10 Further reading

9.11 Answers to check your progress

9.12 Possible Questions

9.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- know what is theory
- know the methods used in theory building.
- recognise the difference between theory and facts

9.2 INTRODUCTION

Every research requires a theoretical background which will provide a framework for analysis, interpretation of data and development of the research field. Theory is required to investigate the practical world problems scientifically. Without having a theoretical background in our research, our research will remain a mere fact finding process which will fail to develop concepts and test the relationship between variables. A good theory must have uniqueness, exclusivity and generalizability. It should have all the capacities within itself to describe critically and comprehensively the causal relationship hidden among the variables found in a social phenomenon. It should be internally consistent and reliable, that can be tested empirically.

It must be also a fact supported theory and should not be a mere guess. Apart from having consistency it must have empirical riskiness. This means its validity might be tested repeatedly when situation demands and all these qualities are to be kept in the mind while building a theory in any kind of research.

Good theories exhibit a dynamic relationship between hypothetical-deduction model (alternately referred to by theorists as nomothetic, positivism, post positivism, empirical-analytical, or hierarchialism) and the inductive-synthesis model (alternately referred to as idiographic, grounded theory, constructivism, or interpretive theory).Here what is desirable is on our part is that we must have a combination of both by adopting mixed method research framework, where both the above mentioned model could be used. In this unit we will learn about the theory building process in detail.

9.3 MEANING OF THEORY

A theory must possess the characteristics of having well sustained and continuous explanation of at least some aspect of the real world. The facts used in a theory should be confirmed repeatedly through observation and experiment. It must be supported by facts and should have a structure of interrelated concepts, definitions and propositions which can further explain and predict a given situation by significantly finding out the relations among variables.

Gelso (2006) had concisely stated that a theory is a statement or declaration of the suspected or assumed relationship between and among variables.

Heinen (1985) stated that a theory can be defined as a group of logically, rationally and sensibly organized laws or relationships that constitute explanation and clarification in a discipline.

Similarly, Sutton and Staw (1995) argued that a theory must essentially and fundamentally answer the question why something is occurring in a given situation. Theory describes causal relationships and explains the order and timing of events as well as deciphers reasons why a relationship exist and functions. Thus its broad applicability and its

generalizability are the two most important criteria. The nature of theory should be abstract and should not be content or topic specific. Theory building in Quantitative research is affected by the empiricist paradigm, which means that it is concerned with cause and effect of social phenomena and uses those data which is based on empirical observation and their critical interpretation.

9.4 METHODS IN THEORY BUILDING

In general we can have the following methods in theory building:

Though there are numerous theory taxonomies, here we will concentrate on some basic and fundamental traditions in theory building,

- Inductive-synthesis (alternately referred to as idiographic, grounded theory, constructivism, or interpretive theory);
- Hypothetico-deduction (alternately referred to by theorists as nomothetic, positivism, post positivism, empirical-analytical, or hierarchialism; and
- Critical theory (alternately referred to as radical, neo-Marxist, or social justice theory).

Let us summerise the above mentioned points through the following approaches which are extensively used in theory building.

❍ **Inductive Approach**

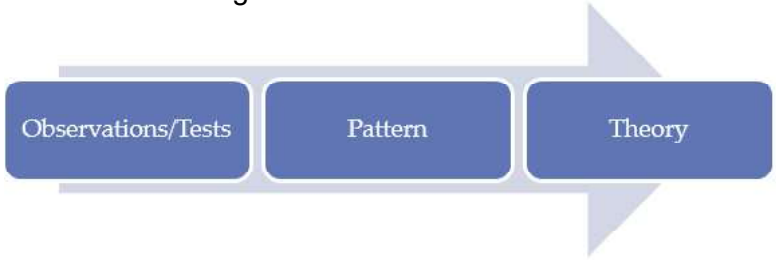
When a research aims at discovering truth or law which is generalizable across populations, then the hypotheses are required to be tested or hypotheses testing perspective can be used. If, however, the goal of the research is to understand what is taking place in a definite surrounding and how different stakeholders make sense of the phenomenal elements under observation, then inductive approach is used to build grounded theory, where the approach is qualitative and not quantitative. Inductive reasoning, by its very nature, is idiographic (i.e., relating to a particular situation), more open-ended and exploratory, especially at the beginning. It is a technique of studying human behaviour by a thorough analysis of individual cases. Here emphasis is more on detailed and thorough description and in depth study than on generalization. It is widely used in qualitative research.

Thus in theory building if we use an inductive approach, the construction of interpretative models is based on observations. It is oriented by the research questions, which are formulated through a literature review or a pre-survey. These observations allow the construction of new concepts, categories of analysis, and theoretical propositions.

Starting from the precise and highly contextualized observations (for a given sociological and geographic space), the inductive approach then generalizes the knowledge to have larger, bigger and more diversified spaces through comparisons with numerous study sites. It proceeds from the specific towards the general. This research approach allows construction of concepts, categories of analysis, and laws or models in a dialectic back-and-forth process between field and theory. It is a characteristic of the constructivist research approach, which is used in theory building in qualitative research.

Inductive theories are based on inductive reasoning which generates more abstract generalizations and ideas because it starts with detailed observations and explanation of the world. When a researcher is following an inductive approach, starting with a topic, he/she is required to develop empirical generalizations and should identify or recognize the first round relationships as he/she progresses with his/her research. No hypotheses can be formulated during the preliminary stages of the research because the researcher is not certain and convinced about the type and nature of the research findings until the study is completed. Thus an inductive approach is closely associated with the generation of new theory from the accumulated data.

As it is described in the figure below, inductive reasoning is often referred to as a “bottom-up” approach where the researcher uses observations to build an abstraction or to describe the picture of the phenomenon that is being studied.



❍ Deductive Approach

Deductive reasoning is narrower and constricted in character and is concerned with testing an existing theory or confirming hypotheses. It is nomothetic in nature (i.e., seeks to explain and clarify a broad class of situations, individuals, or behaviours). Because it is nomothetic in nature, it has better explanatory power. It is typically quantitative in nature and the best example is survey method. Here analysis is directed towards formulation of general principles. In sociology this method helps us to make us aware that individual events are studied not only for their own sake but also for their significance in so far as they contribute towards the formulation of generalization. This approach represents a different approach to convey a connection between theory, data, and analysis. From this perspective, the theory specifies explicitly what kind of data one is required to collect. Here the primary purpose is to discover if hypotheses have observed empirical backing (prediction) or not and if the theory is fallacious using explicit coefficients and other criteria. For example the focus of critical theory is to uncover facts about power relations that are inexplicable or purposely complicated to members of society with the goal of transforming political, social, and economic inequalities. This theoretical perspective has contributed towards exploring the issues of exploitation, supremacy, and subjugation in order to make social actors aware of the existing inequalities and make emancipation of the subjugated class possible.

The construction of the statement of problem in the deductive approach is based on an interpretative model. By functioning logically, this model generates hypotheses, concepts, and indicators for which correspondences must be sought in reality (variables to be informed through observation or survey).In this method, the model and its laws are formulated by the researcher in the form of hypotheses to be proved and verified through statistically representative surveys. This provides statistical proof of their validity or invalidity. In a deductive approach, the formulation of hypotheses must meet the two specific requirements and they are:

- Hypothesis must be empirically verifiable through a survey.
- Hypothesis must be plausible or conceivable.

Generally, theory building using deductive approach follows the following stages:

a. *Deducing* hypothesis from already existing theory

b. *Formulating* hypothesis in operational terms and proposing relationships between two specific variables

c. *Testing* hypothesis with the application of relevant methods

d. *Examining* the outcome of the test, and thus confirming or rejecting the theory.

e. *Modifying* theory or building new theory in instances when hypothesis is not confirmed.

❍ **Mixed method**

This is an approach in research investigation that combines the unique methodology and methods of both quantitative [nomothetic inquiry] and qualitative [idiographic inquiry] epistemologies. By using mixed method approach we can achieve a greater insight to understand the data than when either quantitative or qualitative method is used alone. In determining the appropriateness of which methodology and methods to use (quantitative, qualitative, mixed methods), the obvious starting criteria would be, the purpose of the research, and the type of data being sought to answer the research questions.

In some quantitative research, research questions often require using quantitative methods. Similarly certain qualitative questions are required to be answered by employing qualitative methodologies and methods. But whenever possible the researcher should consider using the mixed method approach by using both qualitative and quantitative approaches. The motive for doing so is that a mixed methods typology has the benefit of capitalizing on the strengths of both quantitative and qualitative approaches while offsetting the weaknesses of both.

By applying mixed method, a researcher can compare the findings derived from qualitative and quantitative data sources. It involves collecting both types of data at roughly the same time, assessing and reviewing

information using parallel constructs for both types of data, separately analyzing both types of data and comparing results through side by side comparison in a discussion, transforming the qualitative data set into quantitative scores, or jointly displaying both forms of data. For example if you are doing a research in health sector and you may want to gather data on the impact of the government health care schemes, then you can gather qualitative data (through in-depth interview) to assess the personal experiences of patients and at the same time you can also gather data by conducting a survey to measure the quality of care. The two types of data can provide validation for each other and also create a solid and firm foundation for drawing conclusions about the intervention.

Some of the essential aspects of the mixed method are briefly discussed below:

Timing

Before you select which design you want to use, make sure you consider the timeline of your study, and when you want your strands of qualitative and quantitative research to occur.

- **Timing**: Before selecting the design of study, one must ponder upon the timeline of the study.
- **Concurrent**: Qualitative and quantitative data collection happens at the same time.
- **Sequential:** The researcher implements the strands (qualitative and quantitative) in two distinct phases.
- **Multiphase:** The researcher is required to do both concurrent and sequential data collection and analysis over a longer period of time.

Mixed Method Designs: Some of the mixed method designs are discussed below:

- **Convergent Parallel:** In this method quantitative data collection and analysis occur at the same time and they are compared after the completion of the study.

- **Explanatory Sequential**: In this design, quantitative data collection and analysis occurs first which is followed by qualitative data collection and analysis.
- **Exploratory Sequential**: In this method qualitative data collection and analysis occurs first, followed by quantitative data collection and analysis.
- **Embedded:** In this method quantitative and qualitative data are collected and analyzed simultaneously within a quantitative design (i.e experimental) or qualitative design (i.e. ethnography).
- **Transformative**: In this method, qualitative and quantitative data collection and analysis can occur concurrently or sequentially and here a transformative design allows the researcher to work within a specific theoretical framework (for example feminist theory).
- **Multiphase:** In this method separate quantitative and qualitative studies are conducted to gather data before a mixed method of study is conducted.

Thus, mixed method research provides more complete and wide ranging evidences for studying a research problem than by either quantitative or qualitative research alone. The mixed method has been used in building the middle range theories.



CHECK YOUR PROGRESS

Q1: a) Mention three methods of theory building.

_____

_____

b) What is meant by mixed method?

_____

_____

9.5 CONSTRUCTION OF A THEORY

Theory building is an essential part of a scientific research. We can see how sociological theories differ from the implicit, inherent or even hidden theories of everyday life. In context of sociology, one of the sharpest dilemma has been whether the focus should be on discovering identical and regular relationships between objective social facts (as argued by Durkheim) or whether it must be concentrate primarily with the understanding of subjective meanings which vary in different cultural and societal contexts (as Weber argued).

In spite of the fact that whether our focus is on objective facts or subjective meanings, the scientific method requires and demands formulating concepts and ideas that has to be established and recognized with sufficient clarity, precision, simplicity, usefulness, utility and accuracy so that they can be used to guide research.

Following are some steps for the construction of a sociological theory.

- **Establishment of a cause effect relationship**

The first step involved in theory construction is the establishment of a cause-effect relationship between independent and dependent variables. And this must be explained logically. The explanation of cause and effect relationship must have the essentials of universality and uniformity with reference to its time and place.

- **Giving definitions of the concepts**

One must give the definition of concepts in a proposition. These concepts must be defined and explained clearly which give precise and clear understanding of the concepts used and must be having meaningful and understandable characteristics.

- **We must establish a relationship between concepts**

We must try to establish a logical relationship between the well-defined concepts of the propositions. The concepts must be explained with logic and reason. This explanation should be understandable and should be clear from any confusion or contradiction. The explanation of the concepts must have a logical relationship with the statement of the research problem.

- **The use of Theoretical concepts**

Theoretical concepts are the changing values or variables while the logical relationship is considered as hypothesis, and are subjected to test. This is very, important component in theory building because the validity of sociological theory is based on the validity of hypothesis to be tested.

- **Testing the Hypothesis**

The fifth step in theory construction is the testing of proposed hypothesis empirically through scientific research methods. We can find the hypothesis as either true or false after testing is done. This process is one of the most important steps considered in theory building.

- **Analysing the Data**

The data which we collect in figure or the numerical data is called statistical data is mostly analysed through the use of statistical data analysis processes or devices. And the descriptive information is called non-statistical analysis. Data collection and its analysis are based on the tested and verified hypothesis. The analysis of data may give us certain findings in favour of our hypothesis or even sometimes it may go against the hypothesis.

- **Generalization**

The collected data is interpreted and generalized to develop a theory. This is very crucial and critical stage of theory construction because the findings may be interpreted in many ways. Here lies the danger of misinterpretation of the findings too. The generalization of the data would be either in favor or against the theory. It is an act of reasoning and analyzing that helps us to draw broad inferences and conclusions from our observation which is widely acknowledged in quantitative research. But its applicability in qualitative research is a bit controversial and questionable.

- **Evaluation of a Theory**

The last step in theory construction is the evaluation of theory construction. It means the process of evaluation of theory in the light of collected facts and specifics after interpretation and generalization. The theory may be evaluated as true or false, partially or fully. Postulates are the first step of theory and its mathematical form is called the law.

9.6 FACT VERSUS THEORY

Facts come before theory. Therefore the most important task is to gather all available data (facts) about the situation before proceeding to build a theory. We must also remember that we cannot ignore a fact simply because it does not agree with our theory.

Facts and theories have different meanings and are used differently in scientific research. Yet there is a common thing that appears among them which is both are used extensively in science. In the scientific world, facts (or scientific facts) are what one can readily observe. But in sociology fact can be gathered from subjective interpretation too. It can pertain to any objective and real phenomenon. It may be the falling of the ball after being thrown upwards or other simple observable occurrences or it can be a collective conscience about a particular religious practice etc. In this regard, the fact in natural science is that the ball will fall. More so, if this test is being done repeatedly under a controlled environment that cancels all unnecessary variables in the phenomenon then it will become obvious and irrefutable and unquestionable fact. It is considered a fact because it will remain as true even after several centuries unless there is a more rigid and precise way of measuring a certain phenomenon.

On the contrary, theories in social science are likened to the explanations to what has been observed. It is relatively gives greater weight to the hypothesis. If a hypothesis is the first base of formulating a scientific law then (generally) theories are placed at the second base. These are the statements that are assumed and understood to be true (because they seem so) even if there is no hundred percent real and concrete evidences. Yet, theories are always presented to be true even if the claims in the said theories are mere speculations or a general agreement between a significant numbers of experts. Moreover, theories are the statements that often undergo a series or sequences of tests to nullify the claims and statements made by those who propose them. Thus we can say that:

- Facts are observations whereas theories are the explanations to those observations.

- Theories are vague truths or unclear facts whereas facts are really fact.

Thus we can say that facts are the subset of truths that have been verified. There are an infinite number of truths waiting to be discovered and verified. But theories are the explanation and clarification of causal relationships between facts to explain and make clear the ordering and timing of events in that relationship as well as reasons why a relationship exists. This is how we distinguish and decide the difference between theories and facts. Within human reality and authenticity, truths are subjective: we do not know, for sure, what is true. Facts are objective: we have verified them. But, within objective reality, there's no difference between truths and facts. All of human reality is provisional and temporary. All of objective reality is a mystery and most of the time suffering from ambiguity.

Scientific theories, in the domain of natural science provide us with the best available explanations and understanding of physical things, phenomena and events. Theories, unless one is speaking of speculation or guess, is mostly associated with science and mathematics. In philosophy and religion, theories are usually the principles, ideology, methods and dogmas that mostly are based upon the discourse of analysis.

9.7 THEORY TYPOLOGIES/ PARADIGM IN THEORY BUILDING

Many researchers use certain typologies and classifications to describe the types of theory. They do it to deal with the diversity and complexity of the definitions and criteria associated with theory. Tarraco's (2002) taxonomy identified five types of theory:

(a) Hypothetico-deductive method,

(b) Inductive grounded theory,

(c) Meta-analytic theory,

(d) Social constructionist theory, and

(e) Case study theory.

According to Gelso (2006), theories have four functions and they are descriptive, delimiting, generative, and integrative. The *descriptive function* explains the why of things (i.e., causal explanation to know why something is happening and what is the cause and consequence associated with it.). The *delimiting function* puts boundaries on what is required to be examined. The *generative function* tends to encourage and support new research (heuristic value) thereby expanding the existing body of knowledge. And finally the *integrative function* seeks to provide a coherent, logical and unified picture of often diverse, varied and at times seemingly disparate as well as different facts.

According to Lynham (2002), Hagerman's three-perspective classification comprise a more parsimonious and practical taxonomy. They are : **empirical-analytical, interpretive, and critical theory** (Lynham 2002). DiMaggio (1995) suggested that there are at least three views regarding what theory should be. They are:

(1) Theory as covering laws,

(2) Theory as enlightenment, and

(3) Theory as narrative.

Finally according to Heinen (1985) there only two kinds of theories: a *concatenated theory* (i.e., inductive-synthesis) and *hierarchical theory* (i.e., hypothetico-deductive).

All the above mentioned views point that we need to distinguish theory from related concepts, such as hypothesis, paradigm, model, and concept. Concepts are the terms which describe the things about which a science tries to make sense or meaning. In other words, a concept is a mental image or constructs that summarizes a set of similar observations. According to Stinchcombe (1968), concepts must have various values, and should be defined in such a way that one can tell by means of observations which value it has in a particular occurrence or happening (i.e., variables). Propositions are the statements that express and articulate the relationship between two or more concepts. According to Gelso (2006), propositions form the basis and foundation of hypotheses, or more

precisely, a hypothesis is a proposition stated in a manner that is empirically testable.

Hypotheses have independent and dependent variables which predict a measurable and assessable observed difference between them. Typically, when a hypotheses is stated in the null form (i.e., no relationship exist between the variables) then a researcher seeks to reject or refuse the null hypothesis. By rejecting the null hypothesis, researchers empirically and scientifically establish and create the degree to which the concepts of a proposition share a co-relationship between variables. Sutton and Staw (1995) argued that hypotheses serve as a crucial bridges or linkages between theory and data. It explains how variables and relationships are operationalized and systematically organized. Thus, a hypothesis is a concise and to the point statement about what is predicted to occur. And theory presents the causal logic of why it is expected to occur. Therefore, theory is a set of interrelated constructs (concepts), definitions and propositions that give us a systematic and methodical view of a phenomenon by specifying and signifying relations among variables. It has the purpose of explaining and predicting the phenomenal changes as well as the existing norms of an occurring behavior.

With regards to models, diagrams, and paradigms, Sutton and Staw (1995) argued that diagrams and models can help to construct and assemble theory or theories and helps to visualize as well as imagine the inter-relatedness of variables. It shows their causal and underlying direction and strength of the relationship.

A theory should be used to elucidate what the diagram shows as opposed to the diagrams explaining the theory. Thus, models or diagrams are often used in research to illustrate relationships within a theory but, by themselves, do not constitute theory. Paradigms are common ways of observing the phenomenal world on the other hand theories, are the systematic sets of interrelated statements intended to explain some aspect and characteristics of social life. Thus, theories flesh out and specify paradigms. In other words, theories attempt to authenticate and validate our paradigms.

9.8 RELATIONSHIP BETWEEN THEORY AND RESEARCH

One of the most important questions in academia is whether research contributes to theory in a discipline or vice versa. The existing literature proposes that research (both quantitative and qualitative) contributes to theory bi-dimensionally:

- Originality [incremental or revelatory] and
- Utility [scientific or practical]

According to Smith, Bekker and Cheater (2011) there exists a tension between theoretical design (theory-building) and pragmatic design (utilitarian practice) .Quantitative contribution to theory may lead to uniqueness or efficacy. At the same time it tends to be more generalizable due to arduous and detailed hypotheses testing. It yields greater explanatory power/control and predictability.

Qualitative contribution to theory is narrower in scope but at the same time it is equally significant, chiefly while dealing with issues that are difficult to quantify. When one tries to make sense of multifaceted and critical social situations, one has to depend on qualitative approach, particularly while responding to the question such as “what is happening here” or while trying to explain how people make sense of their situation. Though idiographic, the resultant theory is grounded in social reality. Thus, qualitative methods have the ability and vision to make vital contributions towards the growth of behavioral and social science. Qualitative methods can elucidate changes in the prevailing constructs, build new relationship among constructs, and also can provide the direction of causation between predictors and outcomes. Thus, there are numerous ways in which research can contribute towards the development of a theory. As stated above, research contributes to theory bi-dimensionally, one is in its originality and the other is based on its utility. Let us briefly discuss these two dimensions.

Firstly, research can add to theory’s explanatory power. It generates research knowledge that may lead to paradigm shift. But often research knowledge tends to contribute in developing a theory in a more incremental

or argumentative manner. It also helps in adding to the lexicon of facts. *Secondly,* theory must have the ability to improve the existing practices and understanding among practitioners. Another major purpose of theory is to establish and generate certain best-practices for the researchers.

It can be summarized from the above discussion that if a theory does not hold true in practice, then that theory holds no value. In other words, the purpose or aim of theory is to bridge the gap between research and practice in order to be applicable in the real world.



CHECK YOUR PROGRESS

Q.2: a) Mention the difference between theory and fact.

_____

_____

b) State the meaning of concept.

_____

_____



9.9 LET US SUM UP

- Theory building has an important place in Sociology because it provides a framework for analysis which facilitates the development of the research field. It is also needed to investigate the practical real world problems.
- In theory building if we use an inductive approach, we can use interpretative models, phenomenological interpretative model, ethnography etc., which are based on observations, oriented by the research questions. They are themselves formulated through a literature review or a pre- survey.
- A deductive approach is concerned with developing a hypothesis (or hypotheses) based on existing theory. After that it aims at

designing a research strategy to test the hypothesis. In deductive approach we go from the general- the theory –to the specific-the observations. It starts out with a general statement or declaration or hypothesis and examines the possibilities to reach a specific, logical conclusion.

- Deduction begins with an expected and predictable pattern which is tested against observations, whereas induction begins with observations and seeks to find a pattern within them.
- Mixed method is an approach of investigation that combines the unique methodology and methods of both quantitative (nomothetic inquiry) and qualitative (ideographic inquiry) epistemologies.
- There are five paradigm / patterns of theory building. They are Hypothetico-deductive method, inductive grounded theory model, meta-analytic theory model, social constructionist theory model, and case study theory.



9.10 FURTHER READING

1. Denzin, N. & Lincoln, Y. (Eds.) (2000). *Handbook of qualitative research* (2nd ed.). Thousand Oaks, C.A.: Sage.
2. Dubin, R. (1978). *Theory building* (Rev. ed.). NY: Free Press.
3. Homans, G. (1964). Contemporary theory in sociology. In R. L. Farris (Ed.), Handbook of modern sociology (pp. 951-977). Chicago: Rand-McNally.
4. Stinchcombe, A. (1968). Constructing social theories. Chicago: University of Chicago Press.



9.11 ANSWER TO CHECK YOUR PROGRESS

Answer to Q1: a) Hypothetico-deduction, Inductive- synthesis, Critical Theory

Answer to Q1: b) Mixed method is an approach to inquiry that combines the unique methodology and methods of both quantitative (nomothetic inquiry) and qualitative (ideographic inquiry) epistemologies.

Answer to Q2: a) Facts are observations whereas theories are the explanations to those observations.

Answer to Q2: b) Concepts are the terms which describe the things about which a science tries to make sense or meaning.



9.12 POSSIBLE QUESTIONS

A) Short Question: (Answer within 150 words)

Q.1: What is meant by fact?

Q.2: What is meant by inductive approach.

B) Long Question: (Answer within 350-500 words)

Q.1: What is the relation between theory and research?

Q.2: Discuss the different paradigms used in theory building.

*** ***** ***

UNIT 10: CLASSIFICATION, TABULATION AND GRAPHICAL PRESENTATION OF DATA

UNIT STRUCTURE

10.1 Learning objective

10.2 Introduction

10.3 Meaning of Classification

- 10.3.1 Types of Classification
- 10.3.2 Methods of Classification

10.4 Meaning of Tabulation

- 10.4.1 Parts of a Table
- 10.4.2 Rules for making a table
- 10.4.3 Types of table

10.5 Graphical Presentation of data

- 10.5.1 Bar Graphs
- 10.5.2 Histogram
- 10.5.3 Frequency polygon and frequency curve
- 10.5.4 Ogive
- 10.5.5 Pie Diagram

10.6 Let us sum up

10.7 Further reading

10.8 Answers to your progress

10.9 Possible Questions

10.1 LEARNING OBJECTIVES

After going through this unit, you will be able to -

- know how to deal with raw data collected from field.
- understand how to categorize data into major groups.
- understand how to organize data and present them in a meaningful manner.
- understand how to illustrate relationship between variables through graphical representation.

10.2 INTRODUCTION

Social sciences in order to have an understanding of the functioning of the society need to look into the various components of the society. In addition to qualitative data, they have to rely a lot on quantitative information to draw inferences about different aspects. Thus, to understand findings it becomes necessary to have an understanding of the elementary quantitative methods. Sociology makes observations about people and their functioning. But the people themselves are not data. Data are the counts, measurements and observations made on people. They have to be generated by observing the aspect that is intended to be studied. Once the data are generated, they can take a number of different forms. This is when classification and tabulation of data becomes necessary. No inferences can be made from raw data. They have to be arranged into categories. This part of statistics is called Descriptive Statistics. The objective of descriptive statistics is to provide precise description of a phenomenon. They consist of standardized procedures through which masses of data can be reduced to manageable proportions. The variables may be grouped into categories, which can then be arranged into frequency distribution or better still presented graphically to illustrate the relationship between variables. In the previous unit we have discussed the different tools and techniques of data collection. From that unit we have learned that there are two types of data, namely primary and secondary data and there are different methods of collecting these data. From the discussions in the previous units we are now aware that data is the most significant aspect of research and there are not only different techniques of collecting data but there are also different ways and procedures of classifying, analyzing and representing data. In this unit we will therefore discuss about the classification of data, tabulation of data and graphical presentation of data.

10.3 MEANING OF CLASSIFICATION

The process of arranging data into categories or groups on the basis of similarities, related facts or common characteristics is called

classification. After classification the data becomes comprehensible and the important aspects of the data become apparent. The similarities and differences may also be understood, which makes comparison possible. Another advantage of classification is that it makes the data fit for statistical treatment. Unless and until classification is done, no statistical treatment can be done. From the nature of data, we can also deduce what statistical tools and technique to be used for analyzing the data.

10.3.1 TYPES OF CLASSIFICATION

Data can be classified mainly into two types. One is Qualitative Data and the other is Quantitative Data.

Qualitative data is based on attributes which are of descriptive nature. Some examples of qualitative data is gender (male/female), marital status (married/unmarried) etc. In other words, the attributes used for classification cannot be measured, but we can find out whether they are present or absent among the population under study.

Quantitative data is based on some characteristic or parameter that can be measured. For example, the height, weight, age, income, etc. Such measurable characters are arranged into groups or classes called class intervals.

When classification is done according to qualitative aspect, then it is called qualitative classification and when it is done according to quantitative data or class intervals then it is called quantitative classification. Two elements are very essential in quantitative classification. One is the variable and the other is the frequency. The term variable refers to the characteristic that varies. The variable may be the height of the population which varies from person to person. Frequency is the number of times the variable occurs.

Quantitative classification may be continuous or discrete depending on the nature of data. A continuous variable is also called a continuous random variable. It is capable of manifesting every conceivable fractional value within the range of possibilities. In other

words, it can assume any interval of measurement. For example, when a person grows in height from 95 cm to 105cm, he passes through all the values within these limits. The discrete variable, on the other hand, is the one which can vary only by finite units and cannot manifest every conceivable fractional value. Each unit of data is separate, complete and not capable of division. For example, the number of books a person has can only take a certain value, i.e. 15 or 20 but can never be 15.5 or 18,8 books. Therefore it can be said that continuous data can be obtained by measuring and discrete data can be obtained by counting. Classification of a series can be said to be continuous when it is described by continuous variables and can be said to be discrete when it is described by certain fixed or discontinuous measurements.

Qualitative classification may be simple or manifold. Classification is said to be simple when only one attribute is used. For example, when we are studying population structure, the population may be divided into male and female for understanding gender distribution. Here only one characteristic or attribute is used. Similarly, the population may be divided according to their educational status, religion, marital status, etc. When we classify the population using only one attribute, then it is called simple classification and when we classify using more than one attribute, then such a classification is called a manifold classification. An example of manifold classification may be that a population may be classified according to what proportion of males and females live in rural and urban areas, or the proportion of males and females living in rural and urban areas by their educational or marital status. In such a classification, we are using more than one attribute at the same time.

10.3.2 METHODS OF CLASSIFICATION

In case of quantitative data, classification may be done according to class interval. The entire data may be divided into

several classes, each of which is known as class interval. In making the classes it should be noted that the interval is not too large as it may conceal important information. Moreover care should be taken to make the classes mutually exclusive. There should be no ambiguity or doubt as to the class in which any variable belongs. For example, if two classes are made at intervals 2.5-3.0 and 3.0-3.5, then the variable 3.0 would fall in both the classes due to overlapping. Such a situation must be avoided. Classes must also be exhaustive i.e. the entire range of data must be reflected by the classes.



CHECK YOUR PROGRESS

Q1: What are the elements essential in quantitative classification?

_____

_____

Q2: What is a Discrete Variable?

_____

_____

Q3: What do you understand by Manifold Classification?

_____

_____

10.4 MEANING OF TABULATION

Classification and tabulation are actually not two distinct processes but they more or less go together. Classification may be said to be the first step in tabulation. Tabulation is a simple device for summarizing data and presenting them in a meaningful manner. It helps in consolidating the primary data. Tabulation may be defined as the systematic and scientific method of arranging data. A table is generally a systematic method of arranging data in columns and rows. Columns are vertical arrangements whereas rows are horizontal ones. Classified data may be displayed in columns and rows

so that their relationship can also be understood. The data after being arranged in a table becomes easy to comprehend and fit for analysis and interpretation.

10.4.1 PARTS OF A TABLE

A table comprises of different parts which are as follows:

- Number and title – Each table should be numbered. The number is accompanied by a suitable title. The title is a short description of the contents of the table. The title should be clear, brief and written in a form different from the contents of the table.
- Caption – The heading of a column is called the caption. The caption tells us what the column represents. A column may also have subheads and more importantly it should indicate the unit of measurement of the data contained in the column.
- Stub – The contents of the rows are indicated from the stub. The stub is the first column in a table and tells us what the data contained in the rows represent. Column is the first horizontal row and stub is the first column in the table.
- Body – The numerical information contained in the columns and rows of the table is called the body of the table. This is the most important and vital part of the table.
- Footnotes – Any explanation regarding any part of the table which is necessary for understanding the table is added as footnotes. The footnotes are placed just below the body of the table, outside the margin of the table. The footnote is added either, to point out exceptions, clarify something about the table or cite source of secondary data. If there is more than one footnote necessary, then they should be numbered or simply explained by adding a star.

10.4.2 RULES FOR MAKING A TABLE

While tabulating the data certain precautions should be taken, which are as follows:

- If the data are too large then separate tables should be made. Each table should portray a particular set of data. This would make the data more comprehensible.
- If any statistical calculation like total, average, percentage, etc. is to be shown, then they should be placed together after the data.
- The caption, stub and body of the table should follow a systematic order. Then only the data becomes readable and allows important items to be emphasized. The data should be arranged alphabetically, chronological, or by following some accepted conventional arrangement depending on the nature of the data.
- Zero should be used to indicate zero quantity and not non availability of information. If data for a particular box is not available then a dash (-) can be used.

The objective of tabulation is to portray clearly and understandably the collected data. Three things that need to be kept in mind while making a table are that the data should be clearly represented; secondly, it should contain only those information which are important for understanding a particular aspect of the total data; Thirdly it has to be assured that the contents of the table are accurately represented.

10.4.3 TYPES OF TABLE

There are mainly two types of tables. One is the simple type where only one characteristic is shown and the other is the Complex tables in which information about two or more inter-related characteristics is shown together. This type of table is commonly used as they facilitate a comparison and consideration of all related facts. A simple table is also known as a one way table. When two characteristics are shown, it is known as a two way table. When four or more characteristics are simultaneously shown then such a type of tabulation is called a manifold tabulation.

Tabulation helps in consolidating and presenting raw data. Tables or frequency distributions are not the end in itself. They are commonly a preface to a more intensive analysis. Tabulation can be said to be the last step in the collection and compilation of data. It leads to analysis and interpretation of data.



CHECK YOUR PROGRESS

Q4: What do you understand by Tabulation?

Q5: What are Footnotes?

Q6: Is Complex Table known as One way table?

10.5 GRAPHICAL PRESENTATION OF DATA

Graphs provide an illustrative method of presenting data. They help in making a concise summary statement of data. The salient features of the data become more apparent through graphical presentation. There are mainly five types of graphical presentation. They are bar graph, histogram, frequency polygon and frequency curve, ogive and pie diagram.

10.5.1 BAR GRAPHS

The bar graph demonstrates the relationship between variables by means of vertical or horizontal rectangles. They are important when the data on one of the variables are measurable on nominal scale. For example if we want to show the proportion of elderly to the total population of different countries, we can show the percentage with a bar graph for each country. Even without looking at the percentage,

we can see which country has a higher proportion than the other from the size of the rectangle.

10.5.2 HISTOGRAM

The histogram is another very helpful graph in understanding basic statistical and probabilistic concepts. It helps in summarizing data and provides a two dimensional map of the frequency distribution table. To draw the histogram for a continuous frequency distribution, the class intervals are shown on the X-axis. For each class interval one rectangle is drawn in proportion to the frequency of each class.

10.5.3 FREQUENCY POLYGON AND FREQUENCY CURVE

To draw the frequency polygon, the midpoints of the class intervals are shown on the X-axis and the frequencies on the Y-axis. The frequencies are then joined by means of straight lines. The frequency curve is obtained from a frequency polygon. The polygon is then smoothened out keeping the area of the curve to that of the histogram.

10.5.4 OGIVE

Ogives are also known as cumulative curves. They are used for expressing graphically the cumulative frequency distribution. It illustrates relations between a class interval and frequency at or below its upper limit. To draw the ogive, the upper limits of the class are shown on the X-axis and the cumulative frequencies on the Y-axis. The points are then joined by free hand smooth curve. The curve of the ogive rises as the variable increases.

10.5.5 PIE DIAGRAM

Pie diagrams are also known as circle graphs. They provide a useful method for illustrating the inter relationships between components of a unitary whole. For example, if we want to show the educational status of a population in a pie diagram, the circle may be divided

into relative proportions depicting the different percentages of achieved educational attainment. Pie diagrams should be simple because too many categories can obscure the overall relationships.



10.6 LET US SUM UP

By now we have come to know that –

- Generated raw data has to be organized into categories so that inferences can be drawn from the data.
- Classification is the first step in organizing data and is the process of arranging data into categories on the basis of similarities.
- Classification may be qualitative or quantitative depending on the nature of data.
- Tabulation is the simple device of summarizing data into a systematic manner in columns and rows.
- A table consists of table number and title, caption, stub, body and footnotes.
- When data are too large, then separate tables should be made.
- Graphical presentation is an illustrative method of presenting a concise summary statement.
- The main types of graphical presentation of data are bar graphs, histogram, frequency polygon and frequency curve, ogive and pie diagrams.



10.7 FURTHER READING

1. C.R. Kothari 2004 Research Methodology Methods and Techniques, 2nd Edition: New Delhi, New Age International Publisher
2. L. Choudhury 2000 An introduction to Statistics, Volume One: Guwahati Kitapghar



10.8 ANSWER TO CHECK YOUR PROGRESS

Answer to Q1: The two essential elements in quantitative classification are Variable and Frequency

Answer to Q2: A discrete variable is the one which can vary only by finite units and cannot manifest every conceivable fractional value

Answer to Q3: When we classify a given population using more than one attribute it is called manifold classification.

Answer to Q4: In simple words tabulation implies systematic arrangement of data into rows and columns

Answer to Q5: Footnotes can be described as the explanations regarding any part of table which are placed at the bottom of the table i.e below the body and outside the margin of the table.

Answer to Q6: No a complex table is not known as one way table. Only simple table is called one way table.



10.9 POSSIBLE QUESTIONS

A) SHORT QUESTIONS: (Answer each question in about 150 words)

Q1: What is classification of data? Describe the types of classification used in statistics.

Q2: Describe the different parts of the table.

B) LONG QUESTIONS: (Answer each question in about 300-500 words)

Q1: Discuss some precautions to be taken while classifying data into class intervals.

Q2: Discuss some of the rules to be followed while tabulating data.

Q3: What are the different ways by which data can be presented graphically?

*** ***** ***

UNIT-11: SCALING TECHNIQUES IN SOCIOLOGICAL RESEARCH

UNIT STRUCTURE

11.1 Learning Objectives

11.2 Introduction

11.3 Meaning of Scale and Scaling

11.4 Types of Scale

11.5 Scales to analyse Social distance

11.6 Importance of Scaling

11.7 Difficulties in Application of Scaling Techniques

11.8 Let us Sum up

11.9 Further Reading

11.10 Answers to Check Your Progress

11.11 Possible Question

11.1 LEARNING OBJECTIVES

After going through this unit you, will be able to –

- know the meaing of scale, scaling and types of scaling
- know the importance of scaling and difficulties of using scaling teaching in sociological research.

11.2 INTRODUCTION

An important trend in sociology during the past few decades has been an increasing emphasis on the development of measuring instruments. Such instruments are commonly referred to as as 'scales'. As time changes the instruments of measurement also changes. New techniques of measurement are being developed improve day by day. Scaling always hypothesizes the existence of a continuum of some kind. Its nature must be inferred from the character of the items selected to make up the scale. In this unit, we shall discuss about scale and scaling techniques, different types of scales and we shall also learn about social

distance scale and sociometry scale and their application in Sociological research.

11.3 MEANING OF SCALE AND SCALING

Scaling is the branch of measurement that involves the construction of an instrument that associates qualitative constructs with quantitative metric unit. Scaling evolved out of efforts in psychology and education to measure “immeasurable” construct like authoritarianism and self-esteem. Scaling describes the procedures of assigning numbers to various degrees of opinion, attitude and other concepts. The process of measuring abstract concepts in concrete terms remains one of the most difficult tasks in empirical social science research. Scaling is a branch of measurement that involves the construction of measures by associating qualitative judgments about unobservable constructs with quantitative, measurable metric units. Stevens (1946) said, “Scaling is the assignment of objects to numbers according to a rule.” In many ways, scaling remains one of the most arcane and misunderstood aspects of social research measurement, and it attempts to do one of the most difficult of research tasks- measure abstract concepts. The problem to which techniques are applied is that of ordering a series of items along some sort of continuum. In other words, they are methods of turning a series of qualitative facts into a quantitative series.

Originally the word ‘scale’ comes from the Latin word ‘scala’, meaning a ladder or flight of steps. A scale represents a series of ordered steps at fixed intervals used as a standard of measurement. Scales are used to rank people’s judgments of objects, events, or other people from low to high or from poor to good. Scales can be unidimensional or multidimensional, based on whether the underlying construct is unidimensional (e.g., weight, wind speed, firm size) or multidimensional (e.g., academic aptitude, intelligence).The three most popular unidimensional scaling methods are: (1) Thurston’s equal-appearing scaling, (2) Likert’s summative scaling, and (3) Guttman’s cumulative scaling. A scale is a continuum from the highest to the lowest points, and has intermediate points in between these two extremes. The scale point is higher

than the third point and so on. A scale is a method of measurement. In social research; attitude, behaviour, and other qualitative characteristic are measured by means of different scales. The scalability of the phenomena depends on its continuum. The various factors of the phenomena must be logically interrelated and should be capable of continued measurement. A scale must be reliable.

Scaling is the method of turning a series of qualitative facts (referred to as attributes) into a quantitative series (referred to as variable). The complexity of this undertaking, and hence it may have pitfalls, can best be seen by considering what is meant by ordering items along a continuum. For example, when one wishes to measure the size of a series of families, no problem arises. Families sare counted as having certain number of members; all families which have seven members also have at least six; all those with three also have at least two, and so on. This is obvious and in the nature of the ordinal character of measurement. It is clear that no man can save his second thousand dollars until the first is accumulated. This is because the amount of money in the bank has a quantitative character. Construction of scale and its application requires a high degree of consciousness of the logical problems surrounding the twin assumptions of (i) the existence of a continuum and (ii) the representativeness of that continuum achieved by the particular scale in mind. The question of representativeness is checked by tests of validity.

11.4 TYPES OF SCALES

Basically there are four types of scales found. They are discussed below:

- **Nominal scale**

Nominal scales, also called categorical scales, measure categorical data. These scales are used for variables or indicators that have mutually exclusive attributes. For examples gender (two values: male or female), industry type (manufacturing, financial, agriculture, etc.), and religious affiliation (Christian, Muslim, Jew, etc.). Nominal scale is simply a system of assigning umber symbols to events in order to label them. Nominal scale provides convenient ways of keeping track of people, objects and events.

One cannot do much with the numbers involved. Nominal scales merely offer *names* or *labels* for different attribute values. The appropriate measure of central tendency of a nominal scale is mode, and neither the mean nor the median can be defined. Nominal terms have no corresponding reality, either in or outside our mind. These are mere words. Nominal scales are only in name. Numbers are assigned to them like labels. These provide identity. The identifying numbers can be switched around. They can be interchanged. This scale consists of two or more named categories into which individuals, objects or responses are classified.In a nominal scale, it is possible to distinguish two or more categories in relation to the specified attribute. Nominal data are marked by their equality. These cannot be placed in a meaningful order, or ranked as higher or lower with regard to each other, except according to some other yardstick. Nominal scale is the least powerful level of measurement. It indicates no order or distance relationship and has no arithmetic origin. It simply describe differences between things by assigning them to categories, for that reason nominal data is counted data. It is a simple method of classification rather than an arrangement along a continuum. The question of dimension is not important in this type of scale. If desired, the different groups may be numbered. Classification of individuals according to religion is an example of nominal scale. The categories of nominal scale must be related to each other. They must lie along the same dimension like landownership, gender, voting behaviour etc.

- **Ordinal Scale**

Ordinal scale is the lowest level of the ordered scale that is commonly used. Ordinal scale places events in order, but there is no attempt to make the intervals of the scale equal in terms of some rule. Ordinal scales generally measures non-numeric concepts like happiness, satisfaction, discomfort etc. Ordinal scales are those that measure rank-order data, such as the ranking of students in a class as first, second, third, and so forth, based on their grade point average or test scores. However, the actual or relative values of attributes or difference in attribute values cannot be assessed. Ordinal scale is easy to remember because it is hearing like order and that's the key to remember with ordinal scales- it is

the order that matters. In addition to the qualities possessed by nominal scales, ordinal scales show transitivity. They possess asymmetrical order or succession. The use of an ordinal scale implies a statement of 'greater than' or 'less than' without our being able to state how much greater or less. The real difference between rank 1 and 2 may be more or less than the difference between rank 5 and 6. The central tendency measure of an ordinal scale can be its median or mode, and means are uninterruptable. Hence, statistical analyses may involve percentiles and non-parametric analysis, but more sophisticated techniques such as correlation, regression, and analysis of variance, are not appropriate. But, as like the nominal scale, ordinal scale does not measure the distance between the positions.

- **Interval Scale**

In this type of scale the distance between two items is known. In interval scale the intervals are adjusted in terms of some rule that have been established as a basis for making the units equal. The units are equal only in so far as one accepts the assumptions on which the rule is based. The exact differences between two values are known in this scale. This scale has equal units of measurement. For that, it is possible to interpret not only the order of scale scores but also the distance between them.Interval scalesare those where the values measured are not only rank-ordered, but are also equidistant from adjacent attributes. For example, the temperature scale (in Fahrenheit or Celsius), where the difference between 30 and 40 degree Fahrenheit is the same as that between 80 and 90 degree Fahrenheit.. The intelligence quotient (IQ) scale is also an interval scale, because the scale is designed such that the difference between IQ scores 100 and 110 is supposed to be the same as between 110 and 120 (although we do not really know whether that is truly the case).Interval scale allows us to examine "how much more" is one attribute when compared to another, which is not possible with nominal or ordinal scales. Allowed central tendency measures include mean, median, or mode, as are measures of dispersion, such as range and standard deviation. Interval scale provides more powerful measurement than ordinal scales because interval scale also incorporates the concept of equality of interval.

- **Ratio Scale**

This scale incorporates the properties of an interval scale together with a fixed origin or zero point. Weight, length, and time are obvious examples. A ratio scale possesses a meaningful zero point.The zero point is not as precise as it was once believed to be. We can conceive of an absolute zero of length and similarly we can conceive of an absolute zero of time. This type of scale represents the actual amount of variables. Measures of physical dimensions such as weight, height, distance, etc. are examples. Researchers in physical science have the advantage to describing variables in ratio scale form but the behavioral sciences are generally limited to describe variables in interval scale form, a less precise type of measurement. This point exists at that point where the scale possesses no quality either this way or that. It is the top level of measurement and is not often available in social research. The factor which clearly defines a ratio scale is that it has a true zero point. A rational zero is a location on an interval scale deliberately chosen for reasons other than the current data. The distinctive feature of a ratio scale is that it has an origin defined by a dominating theory. These use of a ratio scale permits comparisons such as being twice as high, or one-half as much. Reaction time uses a ratio scale of measurement-time. Onthe basis of a ratio scale one can compare both differences in scores and the relative magnitude of scores. Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, and divided.

- **Attitude Scale:**

An attitude scale is a special type of questionnaire designed to produce scores indicating the intensity and direction of a person's feeling about an object or event. In this type of scale, the attitude of an individual towards a matter, thing, object or system can be known from the score of his responses given on a questionnaire. The score will place him on a scale. He simply expresses his like or dislike, agreement or disagreement with the issue involved, as given in the forms of questions. On the basis of his reply, he is assigned a score which indicates his position. Some relevant and indirect statements can also be used to reveal the attitude. In the attitude

scale, most of the scales used are ordinal in nature, though there is attempt to treat the resulting data as internally scaled. This type of scales provide a quantitative measurement of attitudes, opinions or values by summarizing numerical scores given by researchers to people's responses to sets of statements exploring dimensions of an underlying theme.

There are various types of attitude scales. These are discussed below-

- **Point Scale:**

In this scale, at first, a crucial number of words of which the opinion are required is selected. The word may be prayer, dancing, birth control, smoking and so on. The respondent is to cross out every word that is more annoying than pleasing to him. One point is given to each agreement or disagreement, whichever is to be chosen. There is another variety of point scale. In such a scale two sets of words indicating both favourable and unfavourable opinions are given. The unfavourable items may be crossed or favourable item may be left uncrossed.

- **Thurstone Scale:**

Louis Thurstone, one of the earliest and most famous scaling theorists, published a method of equal-appearing intervals in 1925. It is the most widely used method of scaling. Its procedure is used to develop scale items to measure attitudes. It is also known as equal-appearing scale, and judgment scale. It is a uni-dimensional scale, for it measures aggregate attitude towards a specific issue. A Thurston scale is an attitude scale consisting of items with which the respondent has either to agree or disagree. Only those items with which the respondent agrees are second. Each item has a value and the respondent's score on the scale corresponds to the median score of the items with which the respondent agree. In this scale a number of statements, whose position on the scale has been determined by judges, is used. The position is determined by the method of equal-appearing intervals.

Thurstone attempted to devise a method that would represent the attitudes of a group on specific issue in the form of a frequency distribution, the baseline indicating the whole range of attitude gradation from the most favourable at one end to the least favourable at the other with a neutral zero

in between' (Thurstone L. L. and Chave E. J. 1929 cited in Young 1979: 352).

- **Summated (Likert) Scale:**

Likert scaling was developed by Rensis Likert in 1932. This is also referred to as the 'method of summated ratings' because a respondent's score is computed by summing the response values. It is like marks obtained by a student in a test. It is a uni-dimensional scaling method. It is the most widely used scaling technique to assess social attitudes. It is a very popular rating scale for measuring ordinal data in social science research. This scale includes Likert items that are simply-worded statements to which respondents can indicate their extent of agreement or disagreement on a five or seven-point scale ranging from "strongly disagree" to "strongly agree". Unlike the differential scale, the Likert scale uses only the definitely favourable and unfavourable statements. Likert scale excludes the intermediate opinions. This scale consists of a series of statements to which the respondent is to react. Likert scale needs a minimum of two categories, such as 'agree' and 'disagree'. To know the degree of intensity it may further be divided into five or seven point response scale like; 'Strongly approve', 'approve', 'undecided', 'disapprove', 'strongly disapprove' etc. These five points will constitute the scale. Each point of the scale carries a score. "Strongly approve" is given the highest score (5 or +2) and "strongly disapprove" is given the least score (1 or -2). The score values are not indicated in the questionnaire. The following is a Likert type scale:

Statement 1: Prohibition should be made compulsory

1	2	3	4	5
Strongly approve	Approve	Undecided	Disapprove	strongly disapprove
(5 or +2)	(4 or +1)	(3 or 0)	(2 or -1)	(1 or -2)

A key characteristic of a Likert scale is that even though the statements vary in different items or indicators, the anchors ("strongly disagree" to "strongly agree") remain the same. Likert scales are ordinal scales because the anchors are not necessarily equidistant, even though sometimes we treat them like interval scales.Likert scale is considered more reliable because under it respondents answer each statement included in the instruments and it also provide more information and data. It takes less time to construct; it is frequently used by the students of opinion research.

- **Guttman Scale:**

This scale is designed by Louis Guttman and it is also known as the cumulative scaling method. It is based on Emory Bogardus' social distance technique, which assumes that people's willingness to participate in social relations with other people vary in degrees of intensity. It measures that intensity by using a list of items arranged from "least intense" to "most intense". The idea is that people who agree with one item on this list also agree with all previous items. In practice, we seldom find a set of items that matches this cumulative pattern perfectly. A scalogram analysis is used to examine how closely a set of items corresponds to the idea of cumulativeness. Guttman scale uses a series of items arranged in increasing order of intensity of the construct of interest, from least intense to most intense. The Guttman scale is based on the assumption that the various attitude statements in the scale belong to the same dimension. The attainment of a high degree of unidimensionality is the major concern of the Guttman scale. Guttman's scalogram collects response categories as yes/no, agree/disagree for each item. The affirmative responses are given 1 and 0 for negative response. The respondent's score is obtained by summing up the score for all the items. Guttman model is deterministic in nature. It assumes that a person, who responds positively/negatively to one item, must respond positively negatively to a series of others. The model can be made probabilistic rather than deterministic.The strength of the Guttman method is its capacity to identify more than one dimension in the scale. The coefficient of reproducibility is indicative of the extent to which

the material relates to a single dimension. Construction of Guttman scale involves different phases. In the first phase, items in the form of questions are generated based on the concept under study. Each item must be given two options – in affirmative (agree) and negative (disagree). Items must cover the whole range of the underlying attitude continuum. In other words, items must move from simple agreement to overwhelming agreement. A group of judges rate the statements in terms of their favourableness or unfavourableness towards the concept. They then review the items and select final scale items.



CHECK YOUR PROGRESS

Q.1: a) From which language the word 'scale' originated?

b) Mention the different types of scale

.....

.....

.....

.....

.....

.....

11.5 SCALES TO ANALYSE SOCIAL DISTANCE:

The concept of social distance refers to a continuum described by Robert E. Park as "the grades and degrees of understanding and intimacy which characterize pre-social and social relations generally." We can describe the continuum ranges from close, warm, and intimate contact on the one hand, through indifference, to active dislike, hostility, and rejection on the other hand. While applying the social distance scale two problems must be faced. First, it is necessary to define the remainder of the continuum, that is, to define the measurement as social distance toward something- a social group, a value, or an individual. The second problem,

then, is to measure the social distance which does exist. This technique is quite flexible and can be adapted to variety of needs.

- **The Bogardus Social Distance Scale**

The Bogardus Social Distance Scale is one of the oldest psychological attitude scales. This scale is used for measuring the social distance between nations, races, castes and so on. This scale is named after Emory S, Bogardus, who pioneered in this area. He and his students as well as other sociologists have applied this method to various types of social groups such ethnic minorities, social classes, and occupational types, as well as more abstract social values. A number of suggested relationships may be listed, to which members of an ethnic group may be admitted. The respondent is to indicate as to which racial group is to be admitted by him for each of the specified relationship. The attitude is measured by the closeness of relationship that a respondent is willing to accept or the social distance that he likes to maintain. The Social Distance Scale is an example of a Guttman scale in that it is unidimensional and cumulative. The unidimensional aspect means that the scale items can be used to measure a single theoretical concept and only that concept. For example, in a scale composed of items that measure prejudice, items that measure a different concept would not be included. The items contained in a unidimensional scale can be placed on a continuum. Individual attitude is measured with the social distance indicated by them. The respondents are asked to agree or disagree with various items in it. The items are interconnected in such a way that if a person agrees with the fourth item, he is also likely to agree with the first three items in that regard. One who agrees with the eight items, he is also likely to agree with the first seven items.

In this Bogardus scale, the respondent has to indicate his first feeling. He has to give his reaction to each race as a group and he should not take into account any individual member of a group, best or worst, into account. The social distance can also be calculated mathematically. With the development of the Thurstone and Likert scaling methods, attention has been shifted away from cumulative scaling towards unidimensional scaling which seeks to avoid the influence of the external factors.

- **Sociometric Scale**

The term 'Sociometry' relates to its Latin etymology, 'socius' meaning companion, and 'metrum' meaning measure. Sociometry deals with social relations among groups and individuals. Sociometry is a "method for discovering, describing, and evaluating social status, structure and development through measuring the extent of acceptance or rejection between individuals in groups."

The term is closely associated with the work of the Austrian-born psychiatrist J.L. Moreno, who developed the method as a research and therapeutic technique. Sociometry has come to have several meanings; it is most commonly applied to the quantitative treatment of preferential interpersonal relations, but it is also used to mean the quantitative treatment of all kinds of interpersonal relations. Sociometric techniques have been widely utilized by sociologists, psychologists and psychiatrists in the study of group structure, social status and personal traits. It helps to make more explicit and precise the configuration of group relationships, the characteristics and composition of cliques and other elements in a larger group. A Sociometric measure assesses the attractions within a given group. The basic technique involves asking all group members to identify specific persons within the group they would prefer (or would not prefer) to have as partners in a given activity. Many variations on this technique exist for studying different aspects of social preference. For example, a group's networking structure can be exposed through the Sociometric technique of recording all interactions between group members. Helen Jennings has defined Sociometry as "means of presenting simply and graphically the entire structure of relations existing at a given time among members of a given group." Generally Sociometry is concerned with the social interactions among social groups.

Sociometry is based on the principle of sociometric test. Moreno first applied "sociometric test" in his empirical study regarding the groupings in public school in USA. Sociometric method has been used quite extensively in social research by many writers like Moreno, Jennings, Festinger etc. In

this test, every member of a group may be asked to select a specified number of people from within, who could be made his friend leader and colleague, and so on. Sociometry is fundamentally concerned with the social interactions among social groups. The interaction may be of any type. This method can provide useful data/information about the position of an individual in the group in relation to other groups and sub-groups. Interview and questionnaire are the two main methods for administering sociometric tests. In his test of girls training school, Moreno used this personal interview technique. In the interview each girl was questioned as to her attitude toward those choosing or rejecting her for each criterion. Did she like, dislike, or feel indifferent toward those indication attraction or repulsion toward her? Motivation for attraction and rejection was also investigated in the interview in order to obtain a clearer picture of each girl's position in the social structure. These are easy to administer and are suitable for different situations. Sociometry has been defined as "a method for discovering, describing, and evaluating social status, structure, and development through measuring the extent of acceptance or rejection between individuals in groups." A careful sociometric study may prove very useful in revealing the social interrelations, the pattern of acceptance and rejection and the growth of leadership in a society. This method is appropriate for studying the areas having inter-disciplinary interests. The method can be utilized without a large scale investment of resources and hence, has special relevance in low-developed countries. In sociometric test it is difficult to secure reliable information. This problem is same as the problems occurring in case of attitude scales, case techniques and questionnaires. Sociometric techniques have to be used with discrimination and carefully adapted to each situation.



CHECK YOUR PROGRESS

Q.2: Who used the Social distance scale in Social science research? Discuss about social distance scale.

.....

.....

.....

.....

Q.3: Whose work is associated with the term sociometry? Discuss about sociometric test.

.....

.....

.....

.....

Reliability of Scale:

There are various techniques available in social science research through which the reliability of scale is determined.

- *Retest method*. According to this method, the same scale is applied twice to the same population for the same objectives, and if the two results are similar, the scale is regarded as reliable.

- *Multiple form*. According to this technique, two types of scales are constructed, and they are applied to the same population. In case the two results are more or less similar, the scales may be regarded as reliable.

- *Split-Half Method*. In this method the scale may be divided into two equal parts. Then each part is taken as a whole and measurement is made separately. The correlation between the two scores obtained in this way is estimated and if the degree of correlation is high, the scale may be regarded as reliable.

11.6 IMPORTANCE OF SCALING

In our daily life we rely on different types of standardized tools of measurement. For example, a petrol dispensing machine in a petrol bunk measures the volume of petrol we ask for. A shop keeper uses a weighing machine to measure the required quantity. In the case of measuring

temperature, a thermometer is used which is believed to be a standardized instrument. Similarly height, distance and a whole lot of other things which are part and parcel of our daily life are measured using standardized instruments. Scaling is useful in a number of ways. It improves objectivity. The matter under study can be expressed accurately. Scaling makes the matter concise. A lot of material is expressed with brief and to the point numbers.

Unlike the ones mentioned above, measuring social behaviour (attitudes, perceptions, preferences, or opinion) is difficult for its complexity. It is complex because social behaviour cannot be precisely explained or reduced to certain causal factors. For example, preferences of voters towards candidates in elections are conditioned by several factors like voters' awareness levels, socio-economic background of the voter, and profile of contesting candidates, parties and so on. Hence, it is difficult to single out a factor for preference of a particular candidate by a specific voter.

However, through an intensive and scientific exercise, researcher may be able to explain the attitude of voters in a particular constituency towards contesting candidates. This is possible through a set of statements which is carefully worded and ordered in a scientific manner. Scaling, therefore, helps in assessing the attitudes or opinions of respondents with a higher degree of accuracy. Despite the use of Scaling techniques in psychological measurement and other areas of research; there also have some limitations of the application of scaling techniques.

11.7 DIFFICULTIES IN APPLICATION OF SCALING TECHNIQUES

Application of Scaling technique is not an easy task in social science research. As we know that human nature, behaviour, thinking is not same with each other. For that reason, it is not universally applicable. Human behaviour is heterogeneous and unpredictable. Social phenomena are complex and qualitative in nature. The concepts themselves are not defined properly. Thus quantitative measurement is a baffling affair. Social phenomena cannot be experimented in a controlled way. The phenomena

may be individual, private or esoteric. Therefore, they cannot be properly watched and evaluated. Thus, scale cannot measure all the causative variables involved. We know that human behaviour is flexible and variations are there. Whatever attitude is expressed at a point of time may not be taken as reliable. Even a valid scale may not remain valid during the period of analysis.

In spite of these difficulties, the scaling techniques can be improved to measure the social phenomena. It is important for social science researches.



11.8 LET US SUM UP

In this unit we first discussed the meaning of scale and scaling and also discussed some definitions of scaling.

- Scales are commonly used by people in their day to day activities.
- The word scale comes from the Latin word, 'scala', which means ladder or flight of steps.
- There are four types of scales namely; Nominal scale, Ordinal scale, Interval scale and Ratio scale.
- There are different attitude scales like; point scale, likert scale, Thurstone scale etc.



LET US KNOW

Scale: A scale represents a series of ordered steps at fixed intervals used as a standard of measurement. Scales are used to rank people's judgments of objects, events, or other people from low to high or from poor to good.

Sociometry: The term Sociometry relates to its Latin etymology, 'socius' means companion, and metrum meaning measure. Sociometry deals with social relations among groups and individuals. Sociometry is a "method for discovering, describing, and evaluating social status, structure and development through

measuring the extent of acceptance or rejection between individuals in groups."

Social distance: This scale is made use for measuring the social distance between nations, races, castes and so on. This scale is named after Emory S, Bogardus, who pioneered in this area. He and his students as well as other sociologists have applied this method to various types of social groups such ethnic minorities, social classes, and occupational types, as well as more abstract social values.



11.9 FURTHER READING

1. Young, P. V. 1992. Scientific social surveys and research, New Delhi: Prentice Hall of India.
2. Goode, W. J and Hatt, P.K. 2006. Methods in Social Research. Delhi: Surajeet Publication.
3. Jahoda, Marie, Morton Deutsch, and Stuart W. Cook. 1951. Research Methods in Social Relations. New York: Dryden.
4. Thurstone, L. L, .and E. J. Chave. 1929. The Measurement of Attitudes Chicago: University of Chicago Press.



11.10 ANSWER TO CHECK YOUR PROGRESS

Ans. to Q.No.1: a) Latin word scala

Ans. to Q.No.1: b) Point Scale, Thurstone scale, Likert scale, Guttman scale

Ans. to Q.No.2: a)Emory S, Bogardus

Ans. to Q.No.2: b) Moreno, Jennings, Festinger



11.11 POSIBLE QUESTIONS

A) Short Questions (Answer within 150 words)

Q.1: Define scaling.

Q.2: Write about the attitude scale.

B) Long Questions (Answer within 300-500 words)

Q.1: Discuss the Bogardus and sociometric scale

Q.2: Explain the importance of Scaling techniques? What are the limitations of using scaling method in Social Science Research?

*** ***** ***

UNIT 12: MEASURES OF CENTRAL TENDENCY

UNIT STRUCTURE:

12.1 Learning Objective

12.2 Introduction and Meaning

12.3 Requisites of An Ideal Measure of Central Tendency

12.4 Arithmetic Mean

12.5 Median

12.6 Mode

12.7 Let Sum Up

12.8 Further Reading

12.9 Answer to Check Your Progress

12.10 Possible Questions

12.1 LEARNING OBJECTIVE

After going through this unit, you will be able to

- understand what are the different measures of central tendency
- know what are the different types of mean
- know the different methods of calculating mean
- know what is mode and how to calculate mode
- know what is median and how to calculate median.
- relation between mean, median and mode.

12.2 INTRODUCTION AND MEANING

One of the important objectives of statistical analysis is to determine various numerical measures which describe the inherent characteristics of a frequency distribution. One of such measures is the measures of central tendency which is used to explain data in brief. A measure of central tendency or measure of location is a single value which is used to represent an entire set of data.

12.3 REQUISITES OF AN IDEAL MEASURE OF CENTRAL TENDENCY

According to Prof. Yule, the following are the requisites of a good average or measure of central tendency.

- It should be rigidly defined, i.e, the definition should be clear and unambiguous so that it leads to one and only one interpretation by different persons.
- It should be clearly understandable and easy to calculate, such that it should be computed with sufficient ease and even a non-mathematical person should be able to understand it.
- It should be based on all the observations. In other words, while computing the average, the entire set of data should be used and there should not be any loss of information resulting from not using the available data.
- It should be suitable for further mathematical treatment. In other words, the average should possess some important and interesting mathematical properties so that its use in further statistical theory is enhanced.
- It should be affected as little as possible by fluctuations of sampling. This means that if we take independent random samples of the same size from a given population and compute the average for each of these samples the, for an ideal average, the value so obtained from different samples should not vary much from one another.
- It should not be affected much by extreme observations, i.e., very small or very large observations.

There are five averages or measure of central tendencies – mean, median, mode, geometric mean and harmonic mean. But the syllabus covers only the first three. Hence only mean, median and mode are discussed.

12.4 ARITHMATIC MEAN

Arithmetic of a given set of observations is their sum divided by the number of observations. For example, the arithmetic mean of 5, 10, 12, 18 and 25 is

(5+10+12+18+25)÷5=70÷5=14

In general, if X1,X2,……..Xn are the given observations, then their arithmetic mean, usually denoted by

where "X is the sum of the observations and n is the total number of frequencies.

Calculation of arithmetic mean for Ungrouped data: the arithmetic mean for ungrouped data can be calculated in three ways – direct method, assumed mean method and step deviation method.

- Direct Method: In this method the sum of all the frequencies is divided by the total number of frequencies.

Example 1: The mean of 5, 10, 15 and 20 is

X (bar)=(5+10+15+20)/4

= 50/4

= 12.5

- Assumed mean method:

This method involves the following steps:

a. A number is assumed as mean from the given frequency distribution and denoted as A.

b. Next the deviation from each frequency is calculated, which is denoted by d and the sum of this deviation, denoted by "d is calculated.

c. Next the A.M. is calculated as follows:

x(bar)=A+"d/N

Where A= assumed mean; d= x- A, N= total no. of observations.

c) Step Deviation Method:

In this method the deviation, d= x-A, is divided by a common factor 'c', such that d'= d/c = (x – A)/c is found out. The formula for calculating arithmetic mean by step deviation is:

x(bar)=A+("d'/N) x c

where c = common factor

Where c= common factor

CALCULATION OF ARITHMETIC MEAN FOR GROUPED DATA:

Grouped data may be discrete or continuous. In each case, the methods of calculating mean are different.

- Discrete Series
- Direct Method:

In case of discrete series, frequency against each of the observation is multiplied by the value of the observation. Then the values, so obtained are summed up and divided by the total number of frequencies to get arithmetic mean. Symbolically,

x(bar)="fx/"f = "fx/N

" fx = sum of the product of the value of the variables and their respective frequencies

N = "f = sum of frequencies

- Assumed Mean Method:

In this method the following steps are involved:

a. Take any value from the series of variables X as assumed mean (A).
b. Take the deviation of the values of the variable from the assumed mean and denote it by d i.e. d=X-A
c. Multiply all these values of d by their respective frequencies.

The mean by assumed mean method may be found out by the following formula:

x(bar)=A+"fd/N

Where, A = Assumed mean

" fd = Sum of the product of deviation and frequency

N = "f = total frequency

- Step Deviation Method:

In this method the deviations are divided by a common factor, such that we have d'= d/c = (x – A)/c

One has to follow the following steps:

a. Assume any value of the variable as assumed mean.
b. Subtract the assumed mean from each value of the item and denote it by 'd'.
c. These deviations are to be divided by a common factor to get the step deviations, which is denoted by d'.
d. These deviations (d') are to be divided by their corresponding frequencies (f) such that we get fd' and then find out "fd'.

5. Apply the formula

   x(bar)=A+(“fd’/”f) x c

   Where c=common factor

   N= “f = total frequency

B. Continuous Series:

In case of continuous series, the process of calculating arithmetic mean is same as that in the case of discrete, the only difference being that one has to obtain the mid-points of the class interval. The class intervals may be exclusive, inclusive or of unequal sizes. In each case, the arithmetic mean is calculated in a similar way.

a) Direct Method:

The following steps may be followed in this method-

1. Calculate the mid value and denote it by m.
2. Multiply each of the mid value by their corresponding frequencies to get fm and then calculate “fm.
3. Calculate the arithmetic mean using the following formula –

   x(bar)=”fm/”f

   Where, m= mid value “ fm = sum of the product of mid values and their corresponding frequency “ f= total frequency.

b) Assumed mean method:

The following steps are involved in this method –

1. Obtain the mid values (m) of each class interval.
2. Assume any value from the obtained set of mid values.
3. Calculate the deviations of the mid values from the assumed values and denote it by d.
4. Multiply each deviation (d) with their corresponding frequencies (f) to obtain fd, and also work out “fd.
5. Divide “fd by “f and add the dividend to the assumed mean to get the actual mean.

c) Step deviation method:

In this method deviations are divided by a common factor to simplify the calculations. The formula for calculating average is

x(bar)=A+(“fd’/”f)xc

where d’=d/c=(m-A)/c

A = assumed mean f=frequency

c = common factor N=”f=total frequency

c) Weighted Arithmetic Mean:

Till now the arithmetic mean that we have discussed has given equal importance/weightage to all the items. But it is not always that all items in a distribution are equally important such that equal weightage is given to them. For example, if we want to measure the change of in cost of living of a certain group of people, then simple mean of the prices of the commodities are not equally important. Food items such as wheat and rice are more important than other items. Again, in other items, medicines will be more important than cigarettes. Thus, we will need to assign weights to the items and hence calculate weighted arithmetic mean. Weighted arithmetic mean is the average when different items of a series are given different weights according to their relative importance weights are indicated by *w*.

Let $w_1, w_2, \ldots\ldots, w_n$ be the weights attached too various values $x_1, x_2, \ldots\ldots, x_n$ respectively. Then the weighted arithmetic mean, usually denoted by *xw*(bar) is given by

$xw(bar) = (w_1x_1+w_2x_2+\ldots\ldots..w_nx_n)/w_1+w_2+\ldots\ldots\ldots..+w_n$

$= “wx/”w$

Where, *w*=weight, x_w(bar) = weighted arithmetic mean

ARITHMETIC MEAN OF TWO OR MORE DISTRIBUTION:

If we know the sizes and means of two component series, then we can find the mean of the resultant series obtained on combining the given series.

If n1 and n2 are the sizes and x_1(bar),x_2(bar) are the respective means of two series then mean of the combined series of size n_1+n_2 will be given by–

x(bar) = {n_1x_1(bar) + n_2x_2(bar)/n_1+n_2}/{ n1+n2}

Combined Arithmetic Mean of three series will be

x(bar) = {n_1x_1(bar) + n_2x_2(bar) + n_3x_3(bar) /n_1+n_2}/{ $n_1+n_2+n_3$}

MERITS AND DEMERITS OF MEAN

Merits of Arithmetic mean:

1. It is rigidly defined.
2. It is easy to calculate and easy to understand.
3. It is based on all the observations.
4. It is suitable for further mathematical treatment. It possesses many important mathematical properties because of which it has very wide application in statistical theory.
5. Of all the averages, arithmetic mean is affected least by the fluctuations of sampling.

Demerits of Arithmetic Mean:

1. The Arithmetic Mean is very much affected by extreme values. Two or three vary large values of the variable may unduly affect the value of the arithmetic mean.
2. It cannot be calculated in case of qualitative data.
3. It cannot be determined in open end classes.
4. Sometimes arithmetic mean gives absurd and ridiculous results.
5. It cannot be determined by inspection nor can it be located graphically.

12.5 MEDIAN

Median is the middle value of the series when arranged either in ascending order or in descending order. It is the value in the series which divided the series into two equal parts; one part consists of the values equal to median or smaller than it and the other part having the value equal to the median or larger than it. According to L.R.Connor, “The median is that value of the variable which divided the group into two equal parts, one part comprising all values greater and the other all values lesser than the median.”

CALCULATION OF MEDIAN FOR UNGROUPED DATA

In case of ungrouped data if the total number of items in the series is very less, then it is very easy to calculate median. First of all, values should be arranged in ascending or descending order. Then median can be found out simply by observing the data. Median is denoted by M.

For example, let us consider the values 25,38,32,46,35 arranging them ascending order we have 25,32,35,38,46. Here the median will be 35 because it is the middle most value that divides the series into two equal parts.

Now suppose there are two middle values as in the following series – 25,32,35,37,38,46. Here the two middle values will be 35 and 37. Here, the median will be calculated as follows:

M = (35+37)/2 = 72/2 =36.

On the basis of the above two examples we can summarize the formula for calculating median in the following points:

i. When N of the series is an odd number, median is estimated using the following formula after arranging the data in ascending or descending order.

   M = size of $\{(N+1)/2\}^{th}$ item

ii. When N of the series is an even number, median is estimated using the following formula after arranging the data in ascending or descending order.

   M = {size of $(N/2)^{th}$ item + size of $(N/2+1)^{th}$ item}/2

   Median in case of discrete series

   In case of discrete series we are to calculate cumulative frequencies to determine median. The following steps are involved in the calculation of median in discrete series.

   i. Arrange the data in ascending or descending order.
   ii. Find out the cumulative frequencies.
   iii. Apply the formula and locate the Median at the size of items, in whose cumulative frequency the value of $(N/2+1)^{th}$ item lies.

   Median in case of continuous series

   While calculating median in the continuous series, we should first compute cumulative frequencies. After that median group is located by applying the formula, i.e., size of $(N/2)^{th}$ item. After ascertaining the group of the median the following formula will be applied –

Median = L+{(N/2 – c.f.)/f} x i

Here, L = lower limit of median class.

c.f. = cumulative frequency of the class just before the median class

f = frequency of the median class

I = the class interval of the median class

Steps involved in the calculation of median

i. The data are in ascending or descending order of their class interval.

ii. Cumulative frequencies are computed.

iii. Median class of the series is identified. It corresponds to the cumulative frequency which includes the $(N/2)^{th}$ item.

iv. Then the formula is applied to determine the actual median value.

MERITS AND DEMERITS OF MEDIAN

Merits:

1. It is rigidly defined.
2. Unlike arithmetic mean it is not affected by extreme values.
3. In case of open end classification, it is especially useful because only the position is to be ascertained.
4. Median can sometimes be located by simply inspection and can also be computed graphically.
5. Even if the value of the extremes is not known median can be calculated if the number of items is known.

Demerits:

1. Median indicates the value of the middle term, so, it is not based upon all observations.
2. Median cannot be put further algebraic treatment just as mean.
3. For the calculation of median, we are to arrange the data in ascending or descending order which makes calculation difficult.

4. Median is likely to be affected by the fluctuations of sampling.
5. Median may not be representative of a series in many cases.

12.6 MODE

Another important measure of central tendency of a statistical series is the mode. It is the value which occur largest number of times in a series. According to Coxton and Cowden, "The mode of a distribution is the value at that point around which the items tend to be most heavily concentrated." It is the most typical of a series of values. For example, if a car dealer says Etios is in fashion these days, it means that the sale of Etios is highest and maximum number of people drives Etios. As such Etios may be called the model car because its frequency is highest. Similarly eight students get the following marks in Economics – 38,42,38,38,54,46,42,35. Here the number 38 appears thrice. So the mode is 38.

Let us consider another set of marks – 42,55,68,36,52. Here all the marks appear only once. So mode cannot be determined. In case of discrete frequency distribution mode can be determined simply by observing the values.

Consider the following example

Variable	10	25	35	45	55
Frequency	5	18	25	22	17

Here the highest frequency is 25 and the variable corresponding to 25 is 35. So, mode is 35. It may be pointed out that mode need not be a unique number like mean and median. Sometimes, one distribution may have more than two or more mode.

Determination of mode in case of continuous frequency distribution:

In case of continuous frequency distribution, the class with the highest frequency is identified as the model class. The formula for determining mode is as follows:

$M_0 = L + (f_1+f_2)/(2f_1 – f_0 – f_2) \times I$

Where,

L = Lower limit of model class

f_1 = frequency of the model class

f_2 = frequency of the class preceding the model class

f_0 = frequency of the class succeeding the model class

l = height of the model class

The following things need to be noted about mode:

i. If the frequency of each class is same, then the distribution will not have mode.

ii. In order to determine mode we need not determine mid value or prepare a table.

iii. If mid values are given we have to determine the classes first.

MERITS AND DEMERITS OF MODE:

Merits

i. Mode is easy to calculate as it does not need much mathematical calculation. In individual and discrete series it can be found out through inspection.

ii. It is not affected by extreme values.

iii. To calculate one need not know all the items and frequencies of the distribution. In simple series, it is enough to calculate mode if one knows the item with the highest frequency in the series.

iv. Since mode is the most common item in the series, it is not an isolated example like the median. It cannot be a value which is not found in the series.

v. Mode is the value which occurs most frequently in the series. So, mode is the best representative value of the series.

vi. Mode can be located graphically with the help of histogram.

vii. Mode can be used to describe quantitative as well as qualitative data.

Demerits

i. The value of mode is not based on each and every item of the series.

ii. The value of mode cannot always be determined. It is difficult to find modal class in case of bi – modal and multi modal distribution.

iii. Mode is affected by the fluctuations of sampling.

iv. It is not suitable for further mathematical treatment. From the modal values and the sizes of two or more series, the overall mode of the combined data cannot be determined.

v. Calculation f mode involves cumbersome procedure of grouping the data. Mode is not suitable when relative importance of items is desired.

Relation between Mean, Median and Mode:

i. Median always lies between arithmetic mean and mode.
That is, Arithmetic mean>Median>Mode

ii. In case of symmetrical frequency distribution, mean, median and mode will be identical.

iii. In case of asymmetrical frequency distribution, value of mean, median and mode will be different. Then mode can be determined from arithmetic average and median. In that case, the relationship between mean, median and mode is
Mode = 3 Median – 2Mean.
For example,
If in an asymmetrical distribution, median is 350 and mean is 400, then mode will be
Mode = 3x350 – 2x400
= 1050 – 800 = 250.



CHECK YOUR PROGRESS

Q1: What are the different measure of cultral tendency?

_____

_____

Q2: What are the different methods of calculating mean?

_____

_____

Q3: What is mode?

_____

_____

Q4: What is median?

_____

_____



12.7 LETS SUM UP

- A measure of central tendency or measure of location is a single value which is used to represent an entire set of data.
- Arithmetic of a given set of observations is their sum divided by the number of observations.
- Median is the middle value of the series when arranged either in ascending order or in descending order. It is the value in the series which divided the series into two equal parts; one part consists of the values equal to median or smaller than it and the other part having the value equal to the median or larger than it.
- Mode is the value which occur largest number of times in a series. According to Coxton and Cowden, “The mode of a distribution is the value at that point around which the items tend to be most heavily concentrated.”



12.8 FURTHER READING

1. Medhi, J. 2006. Statistical Methods : An Introductory text. New Age International (p) Ltd.

12.9 ANSWER TO CHECK YOUR PROGRESS

Ans to QNo 1: There are five averages or measure of central tendencies – men, median, mode, geometric mean and harmonic mean.

Ans to QNo 2: (a) Direct method, (b) Assumed mean method (c) Step deviation method

Ans to QNo 3: Another important measure of central tendency of a statistical series is the mode. It is the value which occur largest number of times in a series.

Ans to QNo 4: Median is the middle value of the series when arranged either in ascending order or in descending order. It is the value in the series which divided the series into two equal parts; one part consists of the values equal to median or smaller than it and the other part having the value equal to the median or larger than it.

12.10 POSSIBLE QUESTIONS

A) SHORT QUESTIONS: (Answer within 150 words)

Q.1: What do mean by central tendency?

Q.2: Write the formula for arithmetic mean.

Q.3: Mention two advantages and two disadvantages of mean.

Q.4: Mention two advantages and two disadvantages of mean.

Q.5: Mention two advantages and two disadvantages of mean.

B. LONG QUESTIONS: (Answer within 350-500 words)

Q.1: What are the requisites of a good measure of central tendency?

Q.2: What are the requisites of a god measure of central tendency?

Q.3: What is an arithmetic mean? Write the different methods of calculating it.

Q.4: What is an arithmetic median? Write the different methods of calculating it.

Q.5: What is an arithmetic mode? Write the different methods of calculating it.

*** ***** ***

UNIT 13: MEASURES OF VARIABILITY

UNIT STRUCTURE

13.1 Learning Objectives

13.2 Introduction

13.3 Characteristics of a Good Measure of Dispersion

13.4 Objectives of Measuring Dispersion

13.5 Different Measures of Dispersion

- 13.5.1 Range
- 13.5.2 Quartile Deviation
- 13.5.3 Mean Deviation
- 13.5.4 Standard Deviation

13.6 Lets Sum Up

13.7 Further Reading

13.8 Answer To Check Your Progress

13.9 Possible Questions

13.1 LEARNING OBJECTIVES

After going through this unit, you will be able to know -

- know about dispersion
- know about range
- know about quartile deviation
- know about mean deviation
- know about standard deviation
- recognise the difference between mean deviation and standard deviation
- know the different methods of calculating the various types of dispersion.

13.2 INTRODUCTION

Meaning and Concept

It is not always that measure of central tendency is able to describe a particular distribution completely. We require a further description of the series and therefore the measures of central tendency must be supported

and supplemented by some other measure and one such measure is measure of variability or dispersion. Dispersion or variability is a measure of the extent to which the individual items vary. In other words it gives an idea of compactness and scatteredness of the distribution. If we know the central tendency and dispersion of two or more distribution dispersion is found to be less than that in case of the other than the first distribution is said to be more consistent.

13.3 CHARACTERISTICS OF A GOOD MEASURE OF DISPERSION

A good measure of dispersion should have the following characteristics –

1. It should be rigidly defined such that it is easily understandable.
2. It should be easy to calculate.
3. It should be based on all the observations.
4. It should not be affected by extreme values.
5. It should not be affected by the fluctuations in sampling.
6. It should be amenable to further mathematical treatment.

13.4 OBJECTIVES OF MEASURING DISPERSION

1. Dispersion helps to know the variation of different values of the items from average value of the series.
2. One can test the reliability of the average with dispersion.
3. The degree of variation between two or more series can be found out with dispersion.
4. Measures of variability are very useful for the calculation of skewness, kurtosis, correlation, regression, etc.

13.5 DIFFERENT MEASURES OF DISPERSION

The different measures of dispersion are:

1. Range
2. Average deviation
3. Standard deviation
4. Quartile deviation

13.5.1 RANGE

Range is the difference between the largest (L) and the smallest (S) values in a distribution. Thus

R=L-S

Where, L = largest value of the distribution

M = smallest value of the distribution

Dispersion will be higher with higher values of range.

Co –efficient of range

The co-efficient of range is obtained by dividing the range by the sum of extreme values.

Symbolically,

Co-efficient of range = (L-S)/(L+S)

Steps of calculating Range

1. Find the largest value of the series.
2. Find the smallest value of the series.
3. Apply the formula R=L-S

Calculation of range in individual series

Calculate range on the basis of the following data:

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks	48	52	35	76	81	39	67	45	32	93

Here, Largest value (L) = 93

Smallest value (S) = 32

Range = 93 – 32 =61

Calculation of range in discrete series

Marks	25	30	35	40	45
No. of students	7	3	9	11	13

Here, L = 13

S = 3

Therefore, R = 13 – 3 = 10

Co-efficient of range = (13 – 3) /(13+3)

= 10/26

= 0.26

Calculation of Range in continuous series

Class interval	Frequency
10 – 20	20
20 – 30	40
30 – 40	60
40 – 50	80
50 – 60	100

Here, L = Upper limit of the highest class = 60

S = Lower limit of the lowest class = 10

Therefore, R = 60 – 10 = 50

Merits of range:

1. It is easy to calculate and readily understandable.
2. It takes least time for computation.

Demerits of range:

1. Range considers only the extreme values, but tells nothing about the other values.
2. If in a set of observations, say the heights of children, the extreme values such as the tallest and the shortest remain unchanged and all the other values are changed, the range will remain unaffected.
3. Range is not based on all the observations.
4. Fluctuation in sampling greatly affects range. It varies from sample to sample.
5. Range cannot be used in case of open-end distribution. In open end distribution the lower limit of the lowest class and upper limit of the highest class are not known.

Uses of range:

1. In weather forecasting, where people are mostly interested in knowing the extreme values, range is used.
2. To know the variation in exchange rates, money value, gold prices and share values range is used.
3. Range is also used to study the variation in the quality of manufactured units.

13.5.2 QUARTILE DEVIATION

As range is affected by extreme values, a measure of distribution that is not affected by extreme values is required. Quartile deviation is a measure of distribution which divides the distribution into four equal parts, each containing 25% of the values.

With the help of upper and lower quartiles (i.e. Q3 and Q1, respectively), inter quartile range is calculated. Inter Quartile Range = Q3 – Q1.

Half of inter quartile range is Quartile Deviation. It is also known as Semi – Inter Quartile Range.

Quartile Deviation = Q.D. = (Q3 – Q1)/2

Example: Calculate the quartile deviation for the following data:

Roll No.	1	2	3	4	5	6	7
Marks	28	32	45	16	25	42	36

Solution:

First the marks are arranged in ascending order – 16,25,28,32, 36,42,45.

Now, we calculate Q3 and Q1.

Q1 = Size of $\{(N+1)/4\}^{th}$ item

= Size of $\{(7+1)/4\}^{th}$ item

= Size of 2^{nd} item

Therefore, Q1 = 25

Q3 = Size of 3{(N+1)/4}th item

= Size of $3\{(7+1)/4\}^{th}$ item

= Size of 6^{th} item

Therefore, Q3 = 42

Therefore, Q.D. = (42 – 25)/2

= 17/2

= 8.5

CALCULATION OF QUARTILE DEVIATION FOR DISCRETE SERIES:

EXAMPLE: Find quartile deviation for the following series:

Age	10	11	12	13	14	15	16
No. of participants	8	11	13	14	15	8	7

Solution:

In order to find Q.D. for discrete series we have to calculate cumulative frequency

Age	No. of participants	Cumulative frequency
10	8	8
11	9	17
12	11	28
13	12	40
14	13	53
15	6	59
16	1	60

Q1 = Size of {(N+1)/4}th item

= Size of {(60+1)/4}th item

= Size of 15.25^{th} item

It will lie in the cumulative frequency 17. The value or age corresponding to the cumulative frequency 17 is 11

Therefore, Q1 = 17

Q3 = Size of 3{(N+1)/4}th item

= Size of 3{(60+1)/4}th item

= Size of 45.75th item

It will lie in the cumulative frequency 53. The value or age corresponding to this cumulative frequency is 14.

Therefore, Q3 =

Therefore, Q.D. = (14 – 11)/2

= 3/2

= 1.5

Calculation of Quartile Deviation for continuous series:

Class	10-20	20-30	30-40	40-50	50-60
Frequency	20	36	48	34	28

We first calculate the cumulative frequency:

Class	Frequency	Cumulative frequency
10-20	20	20
20-30	36	56
30-40	48	114
40-50	34	148
50-60	22	170

Q1 = Size of N^{th}/4 item

= Size of 170^{th} /4 item

= Size of 42.5^{th} item

The class containing 42.5^{th} item is 20-30. Now to calculate exact value of Q3 the following formula is used:

Q1 = L+{(N/4 – fc)/f} x i

Where, L = lower limit of the relevant quartile class

fc = value of the c.f. of the class preceding the quartile class

i = interval of the quartile class

f = frequency of the quartile class

Here, L = 20

N/4 = 42.5

fc = 20

i= 10

f=20

Therefore, Q1 = 20 +{(42.5 – 20)/20} x 10

= 20 + {(22.5)/20}x10

= 20 + 11.2

= 31.2

Q3 = size of $3N^{th}/4$ item

= Size of {3x170/4} item

= Size of 510/4 item

= Size of 127.2^{th} item

The class containing 127.2^{th} item is 40-50. Now the following formula is used for Q3.

Q3 = L + {3N/4 – fc}/f x i

Here, L, N/4,fc,f and I denote the same items as in Q1.

Therefore, Q3 = 40 + {127.2 – 114}/34 x 10

= 40 + 113.2/34 x 10

= 40 + 33.02

= 73.02

Therefore, Q.D. = ½(Q3 – Q1)

= ½(73.02 – 31.2)

= 20.91

Merits of quartile deviation:

i. Quartile deviation is based on middle 50% of the observations; hence it is a better measure than dispersion.

ii. It is easy to calculate.

iii. It is easy to understand.

iv. Quartile deviation can be calculated for both open end and close end distributions.

v. It is not unduly affected by extreme values.

Demerits of quartile deviation:

i. It ignores the first 25% and the last 25% of the observations.

ii. Fluctuations in sampling affect quartile deviation. A change in the value of a single item, may, in many cases, affect its value considerably.

iii. It is not capable for further algebraic treatment.

13.5.3 Mean Deviation

Mean deviation of a series is the arithmetic average of the deviations of various items from a measure of central tendency (mean, median or mode).

Theoretically, deviations can be taken from any of the three averages mentioned above but in actual practice mean deviation is calculated either from mean or median. Mean deviation is generally not calculated from Mode because its value is indeterminate and it gives erroneous conclusion. Between mean and median the deviations from median is less than deviations from mean and hence mean deviation from median is better than mean deviation from mean. The value of the mean deviation calculated from the median, is, therefore, always less than the value calculated from mean.

Mean deviation for ungrouped data:

The following steps are involved in the computation of mean deviation for ungrouped data:

i. Compute the mean or median of the series.

ii. Take deviations d=X-A from either mean or median ignoring ± sign.

iii. Obtain total of these deviations " |d | where, |d |= absolute value of the deviation

iv. Divide the above total i.e. in step (iii) by the number of items i.e. M.D. = " |d |/N. If the variable x takes the value $x_1, x_2, \ldots x_n$ then

Mean deviation = $\{|x_1 - A|+|x_2 - A|+\ldots\ldots+|x_n - A|\}/n$

= " $|d|/n$

Where A = mean, median or mode of the x

$d = x - A$

Mean deviation for grouped data in discrete series

The following steps are involved in the computation of mean deviation for grouped data in discrete series:

i. Calculate the specific average from which mean deviation is to be found
ii. Obtain the absolute (positive) deviations (D) of each observation from the specific average.
iii. Absolute deviations (D) are multiplied with the respective frequencies (f) and the products are added up to obtain $f|D|$ of each observation from specific average.
iv. Divide the total ($f|D|$) by the number of items to get mean deviation.

If the variable x takes the values x1,x2,……,xn with respective frequencies f1,f2,f3,…..fn.

Then,

Mean Deviation = $\{f_1|x_1 - A|+f2|x_2 - A|+\ldots.f_n|x_n|\}/N$

Where N = f

Mean deviation for grouped data in continuous series

In case of continuous, the formula for man deviation are same as that for a discrete series for the given continuous frequency distribution, the mid – points of class intervals have to be found out and they are taken as *m.* In this way the continuous series assumes the shape of discrete series. After that all the steps of discrete series are applied.

Symbolically,

Mean deviation from mean or median = $\{ f|m - A|\}/N$

= $\{ f|d|\}/N$

Merits of Mean Deviation:

i. It is rigidly defined.

ii. It is easy to understand and easy to calculate.
iii. It is based on all the observations of the series.
iv. As compared with the standard deviation, it is less affected by extreme values.
v. Mean deviation is based on the deviation from an average. So, it provides a better measure for comparison about formation of different distribution.

Demerits of Mean Deviation:

i. Mean deviation ignores the signs of deviations. As a result this method cannot be used for further algebraic treatment.
ii. Mean deviation cannot be computed for distribution with pen-end classes.
iii. Mean deviation when calculated from mode is not reliable because in many cases mode has no fixed value.
iv. It is difficult to compute when mean, median and mode are in fraction.

13.5.4 STANDARD DEVIATION:

The most commonly used measure of dispersion is standard deviation and was first introduced by Karl Pearson in the year 1983. One weakness of mean deviation is that in its calculation plus and minus signs are ignored. This drawback is removed in the calculation of standard deviation. One of the earliest way of doing away with the algebraic signs is to square the figures and this process is adopted in the calculation of standard deviation. “Standard Deviation is the square root of the arithmetic average of the squares of the deviation measured from the mean. This is generally denoted by sigma of the Greek language.” (Choudhury, et al,2013).

Standard deviation of ungrouped data:

In case of ungrouped data, standard deviation can be calculated by any of the four methods – Actual mean method, Assumed mean method, Direct method, Step Deviation method. The following are the formulas for each method.

i. Actual Mean Method ó= "("[x-x(bar)]^2)/n)

= "("d^2/n)

Where, n= number of items in the series

ó=standard deviation

ii. Assumed mean method

ó = "{("d^2/n)-("d/n)^2}

iii. Direct Method

ó="{"x^2/n – (x bar)^2}

iv. Step Deviation Method

ó= "{("d^2/n) × c}

Standard deviation of grouped data

In case of grouped data, standard deviation can be calculated by any of the two methods – Actual mean method and Assumed mean method. The following are the formulas for each method.

i. Actual Mean Method

In order to calculate standard deviation by this method we should proceed in the following way –

a. Calculate the mean of the distribution

X bar = "fx/"f

If it is a continuous series, then x bar ="fm/"f

b. Then the deviations of the values from the mean shuld be taken

d=x-x bar

In case of continuous series, d = m – x bar should be determined

c. Then multiply the deviations with their corresponding frequencies to get fd values.

d. Calculate fd^2 by multiplying fd values with d values. Sum up them to get fd^2. Of course if we calculate d^2 first and multiply it with f then also we get the same value of fd^2

e. Apply the formula:

Ó= fd^2/n

ii. Assumed mean method:

In order to compute standard deviation by assumed mean method we should proceed in the following way –

a. If it is a discrete series then calculate deviation of the values of the variable from the assumed mean, i.e., d=x-A should be determined.

   If it is a continuous series, calculate mid points f the class first. Then deviation of mid points from the assumed mean, i.e., d=m-A should be determined.

b. Multiple values of d with corresponding frequencies to get fd values and find $\sum fd$.

c. Multiply fd by d to get fd^2 and find $\sum fd^2$.

d. Then calculate standard deviation using the following formula:

   $\acute{o} = \{ \sum fd^2/n - (\sum fd/n)^2\}$

   where $n = \sum f$

VARIANCE: It is the square of standard deviation. It is the measure of variability like standard deviation it is denoted by $\acute{o}^2$. Sometimes it is written as Var (x).

Merits of Standard Deviation:

i. It is rigidly defined.

ii. It depends on all the values of the variable.

iii. Fluctuations in sampling do not influence it much.

iv. Combined standard deviation of two or more groups can be computed.

v. It is capable of further algebraic treatment.

Demerits of Standard Deviation:

i. It is not easy to calculate standard deviation.

ii. It lays greater stress on extreme values and lesser stress n those values which are nearer to them.

Distinction between mean deviation and standard deviation:

i. While calculating mean deviation algebraic signs (i.e. ±) are not taken into account whereas while calculating standard deviation their signs are taken into account.

ii. Mean deviation can be computed from median, mode or mean and its value differs in these cases but the standard deviation is always calculated from the arithmetic mean.

iii. The mathematical properties possessed by standard deviations are far greater than those possessed by mean deviation.



CHECK YOUR PROGRESS

Q1: What is dispension?

_____

_____

Q2: How to calculate range?

_____

_____



13.6 LETS SUM UP

- Dispersion is a measure of the extent to which individual items vary.
- The different measures of dispersion are – range, standard deviation, mean deviation and quartile deviation.
- Range = L – S

  Where, L = largest value of the distribution

  S = smallest value of the distribution
- Quartile Deviation = Q.D. = (Q3 – Q1)/2
- Mean deviation = " |d |/n

  Where, d = x – A

  A = mean, median, mode
- Standard deviation can be calculated by different methods according to the type of data.



13.7 FURTHER READING

1. Gupta S.P. 2010. Statistical Methods. New Delhi: Sultan Chand & Sons.



13.8 ANSWER TO CHECK YOUR PROGRESS

Ans to QNo 1: Dispersion or variability is a measure of the extent to which the individual items vary. In other words it gives an idea of compactness and scatterendness of the distribution.

Ans to QNo 2: Range is the different between the largest (L) and the smallest (S) values in a distribution. Thus

R = L - S

Where, L = largest value of the distribution

M = smallest value of the distribution.



13.9 POSSIBLE QUESTIONS

A) SHORT QUESTIONS (Answer within 150 words)

Q.1: What do you mean by dispersion?

Q.2: What is range? How is it measured?

Q.3: What is mean deviation?

Q.4: What is standard deviation?

Q.5: What is quartile deviation?

Q.6: What are the characteristics of a good measure of dispersion?

Q.7: What are the objectives of dispersion?

Q.8: What is the difference between mean deviation and standard deviation?

B. LONG QUESTIONS (Answer within 350-500 words)

Q.1: What do you mean by dispersion? What are the different measures of dispersion?

Q.2: What is range? Discuss its merits and demerits.

Q.3: What is mean deviation? Discuss its merits and demerits.

Q.4: What is standard deviation? Discuss its merits and demerits.

Q.5: What is quartile deviation? Discuss its merits and demerits.

*** ***** ***

UNIT 14: CORRELATION

UNIT STRUCTURE

14.1 Learning Objectives

14.2 Introduction and Meaning of Correlation

- 14.2.1 Definitions of Correlation
- 14.2.2 Significance of the Study of Correlation
- 14.2.3 Correlation and Causation

14.3 Types of Correlation

- 14.3.1 Positive and Negative Correlation
- 14.3.2 Linear and Non-Linear Correlation

14.4 Methods of Studying Correlation

- 14.4.1 Scatter Diagram Method
- 14.4.2 Karl Pearson's Coefficient of Correlation

14.5 Lets Sum Up

14.6 Further Reading

14.7 Answer to Check Your Progress

14.8 Possible Questions

14.1 LEARNING OBJECTIVES

After going through this unit, you will be able to know,

- What is correlation?
- What are the different types of correlation?
- What is correlation coefficient?

14.2 INTRODUCTION AND MEANING OF CORRELATION

It is not always that in our daily life we confront with problems relating to single variable. We mostly come across situations where two or more variables are in use. Many a times two variables move in such a way that movements in one is accompanied by movements in the other. In other words these variables are interrelated. For instance, there can be some

relation between rainfall and production of wheat; temperature and death of old age persons, price of commodity and its demand or supply, and so on. The degree of relationship between the variables under consideration is measured through the correlation analysis. The measure of correlation, called the correlation coefficient or correlation index summarizes in one figure the direction and degree of correlation (Gupta,S.P.,2010). The correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables.

14.2.1 DEFINITIONS OF CORRELATION

- "Correlation analysis deals with the association between two or more variables."- Simpson & Kafka
- "If two or more quantities vary in sympathy so that movements in one tend to be accompanied by corresponding movements in the other(s) then they are said to be correlated." – L.R. Conner
- "When the relationship is of a quantitative nature, the appropriate statistical tool for discovering and measuring the relationship and expressing it in brief formula is known as correlation." – Croxton & Cowden.
- "Correlation analysis attempts to determine the 'degree of relationship' between variables". – Ya Lun Chau
- "Correlation is an attempt of the covariation between two or more variables". – A.M. Tuttle

14.2.2 Significance of the study of correlation

The significance of the study of correlation may be summarised in the following points:

- Most of the variables show some kind of relationship. For example, there will be some relation between price and supply, income and expenditure, production and distribution and so on.
- With the help of correlation, we can also study as to what way two variables are correlated, i.e., whether they are positively

correlated or negatively correlated or correlation between them is high or low.

- Correlation analysis is helpful in understanding the behaviour of various economic variables.
- It also helps policy makers in taking different policy decisions.
- Correlation analysis also helps the business executive to estimate cost and price.

14.2.3 Correlation and Causation

Correlation analysis helps us in determining the degree of relationship between two or more variables but it tell us nothing about cause and effect relationship. There may be high degree of correlation between two variables, but that does not imply that there is a cause and effect relationship between the two variables. For instance, there may be high correlation between sale of umbrella and number of births in a year. But in reality there is no cause and effect relationship between the two. Thus, in correlation analysis, it is very important that one imposes his or her own rationale logic. The existence of such type of correlation may be due to one of the following reasons:

- The correlation may be due to pure chance, especially in a small sample. As stated above, there may be high correlation between two variables, but there may not be any logic behind their relationship. This is particularly true in case of small samples. Such a correlation may arise either because of pure random sampling variation or because of the bias of the investigator in selecting the sample.
- Both the correlated variables may be influenced by one or more other variables. It is just possible that a high degree of correlation between the variables may be due to some causes affecting each variable or different cause affecting each with the same effect.
- Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other the

effect. There may be high degree of correlation between the variables but it may be difficult to pinpoint as to which is the cause and which is the effect.

14.3 TYPES OF CORRELATION

14.3.1 Positive and Negative Correlation

If the values of the two variables deviate in the same direction, i.e., when the value of one variable increases, the other also increases or when the value of one variable decreases the value of other also decreases, the correlation is said to be positive. On the other hand, when the value of variables moves in the opposite direction, such that when the value of one variable increases the other decreases or when the value of one variable decreases the other increases, the correlation is said to be negative.

Some examples of positive correlation will be –

- Rise of rainfall and sale of umbrellas
- Rise in temperature and sale of ice cream
- Better irrigation facility and production of wheat

Some examples of negative correlation are –

- Increase in income and decrease in the purchase of low priced wheat
- Increase in the price of petrol and decrease in the sale of petrol cars
- Rise in temperature and decrease in the sale of woollen clothes.

14.3.2 Linear and Non – Linear Correlation

The correlation is said to be **linear** between two variables if a unit change in the value of one variable brings about a corresponding change in the value of the other variable over the entire range of values. Let us consider the following example –

X	1	2	3	4	5
Y	3	5	7	8	9

In the above example, for a unit change in the value of X, there is a corresponding change in value of Y by 2 units. The relation between X and Y can be summarized in the following equation –

$Y = 2X + 1$

In general, the above equation can be written as –

$Y = a + bX$

The above equation is the equation for straight line where 'a' is the intercept and 'b' is the slope. Thus, if we plot the values of the two variables in a graph we will get a straight line. This will also be true for the above example.

A correlation is said to be non-linear or curvi – linear if with a unit change in one variable the other variable do not change at a constant rate but at a fluctuating rate. In such a case, when we plot the data on a graph we do not get a straight line but we get a graph.

14.4 METHODS OF STUDYING CORRELATION

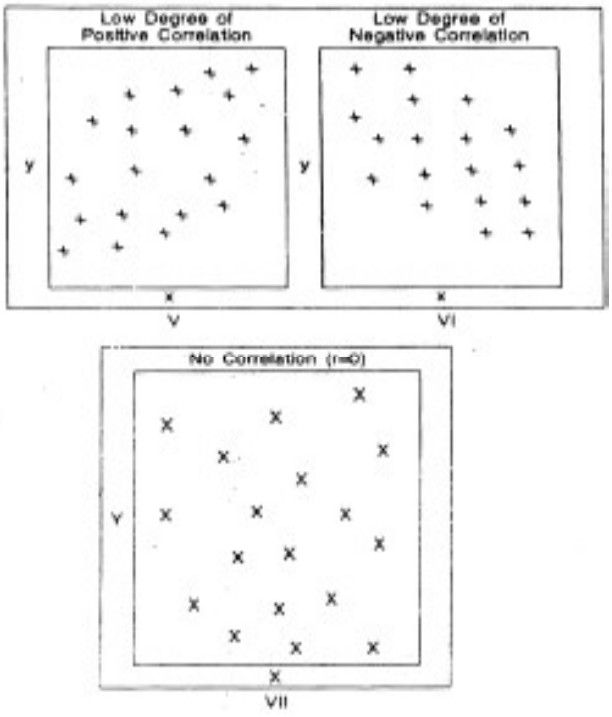
Although there are different methods of studying correlation, we shall confine here to discussing the only two methods of studying only linear correlation. The two methods are –

Scatter diagram method and Karl Pearson's coefficient of correlation.

14.4.1 SCATTER DIAGRAM METHOD

Scatter diagram is the simplest method of studying correlation. In this method, the values of the two variables are plotted in the form of dots in the graph paper and this diagram of dots s obtained is known as scatter diagram. From scatter diagram we can form a fairly good, though rough, idea about the relationship between the two variables:

- If the points are very dense, i.e., very close to each other, a fairly good amount of correlation may be expected between the two variables. On the other hand, if the points are widely scattered, a poor correlation may be expected between them.
- If the points on the scatter diagram reveal any trend (either upward or downward), the variables are said to be correlated and if no trend is revealed, the variables are uncorrelated.
- If there is an upward trend rising from lower left hand corner and going upward to upper right hand corner, the correlation is said to be positive since this reveals that the values of the two variables move in the same direction. If, on the other hand, the points depict a downward trend from upper left hand corner to lower right hand corner, the correlation is said to be negative.
- In particular, if all the points lie on a straight line starting from the left bottom and going up towards the right top, the correlation is perfect and positive, and if all the points lie on a straight line starting from left top and coming down to right bottom, the correlation is perfect and negative.



14.4.2: KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl Pearson suggested a mathematical method for measuring the intensity or the magnitude of linear relationship between two variable series. This method is the most widely used method. It is also known as the product moment correlation coefficient.

Karl Person's correlation coefficient between two variables X and Y, usually denoted by r (X,Y) or r_{xy} or simply *r* is a numerical measure of linear relationship between them and is defined as the ratio of the covariance between X and Y, written as Cov (x,y), to the product of the standard deviations of X and Y. Symbolically,

$r=\{Cov\ (x,y)/\acute{o}_x\acute{o}_y\}$

Properties of coefficient of correlation:

Property I ***Karl Pearson's coefficient of correlation lies between -1 and +1. Symbolically, -1 ≤ r ≤ 1.***

There are two points to be noted in this theorem.

First, it provides a check on our calculations. If in any problem, the obtained value of *r* lies outside the limits ±1, it implies that there is some mistake in our calculation.

Second, *r*=+1 implies perfect positive correlation between the variables and *r*=-1 implies perfect negative correlation between the variables.

Property II ***Correlation coefficient is independent of the change of origin and scale.*** Mathematically, if *X* and *Y* are the given variables and they are transformed to the new variables *U* and *V* by the change of origin and scale viz.,

$u = \{x - A\}/h$ and $v = \{y - B\}/k;\ h>0,k>0$

where *A,B,h* and *k* are constants, *h>0,k>0*; then the correlation coefficient between *x* and *y* is same as the correlation coefficient between *u* and *v* i.e.,

$r(x,y) = r(u,v) \Rightarrow r_{xy} = r_{uv}$

Property III ***Two independent variables are uncorrelated but the converse is not true.***

When two variables x and y are uncorrelated then we have $r_{xy} = 0$. This, however, implies that there is absence of any linear (straight line) relationship between them. They may, however, be related in some other form (other than straight line) e.g., quadratic, logarithmic or trigonometric form.

Assumptions Underlying Karl Pearson's Correlation Coefficient.

- The variables X and Y under study are linearly related. In other words, the scatter diagram of the data will give a straight line curve.
- Each of the variables is being affected by a large number of independent variables so as to produce normal distribution. Variables like height, weight, price, demand, supply, etc. are affected by such forces that a normal distribution is formed.
- There is a cause and effect relationship between the forces affecting the distribution of the items in the two series. If such a relationship is not formed between the variables, i.e., if the variables are independent, there cannot be any correlation.

Interpreting the value of '*r*'

- When r=+1, it means there is perfect positive relationship between the variables.
- When r=-1, it means there is perfect negative relationship between the variables.
- When r=0, it means that there is no relationship between the variables, i.e, the variables are uncorrelated.
- The closer r is to +1 or -1, the closer the relationship between the variables and the closer r is to 0, the less close the relationship.



CHECK YOUR PROGRESS

Q1: What is correlation?

_____

_____

Q2: What are the different types of correlation?

_____

_____



14.5 LETS SUM UP

- The measure of correlation, called the correlation coefficient or correlation index summarizes in one figure the direction and degree of correlation (Gupta,S.P.,2010).
- The correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables.
- There are two types of correlation – a) positive and negative correlation and b) linear and non-linear correlation.
- There are two methods of studying linear relationship between two variables – scatter diagram method and Karl Pearson's coefficient of correlation.
- The sign preceding the value of r describes the nature of relationship between the variables.



14.6 FURTHER READING

1. Gupta S.P. 2010. Statistical Method. New Delhi: Sultan.
2. Medhi. J. 2006. Statistical Method : an introductory text. New Age International (p) Ltd.



14. 7 ANSWER TO CHECK YOUR PROGRESS

Ans to QNo 1: The degree of relationship between the variables under consideration is measured through the correlation analysis. The

measure of correlation, called the correlation coefficient or correlation index summarizes in one figure the direction and degree of correlation.

Ans to QNo 2: (a) Positive and negative correlation

(b) Linear and non-liner correlation



14.8 POSSIBLE QUESTIONS

A) Short answer type (Answer within 150 words)

Q.1: Define correlation.

Q.2: What are different types of correlation?

Q.3: What is linear correlation?

Q.4: What is non-linear correlation?

Q.5: What is positive correlation?

Q.6: What is negative correlation?

Q.7: What are the different methods of studying correlation?

B) Long answer type questions: (Answer within 350-500 words)

Q.1: Discuss the different types of correlation.

Q.2: What is Karl Person's coefficient of correlation? Discuss its properties.

Q.3: What is r? How do you interpret the value of r?

*** ***** ***

Unit 15: REGRESSION

UNIT STRUCTURE

15.1 Learning Objective

15.2 Introduction

15.3 Uses of Regression Analysis

15.4 Objectives of Regression Analysis

15.5 Difference Between Correlation and Regression

15.6 Simmilarity Between Correlation and Regression

15.7 Regression Lines

15.8 Regression Equations

15.9 Lets Us Sum Up

15.10 Further Reading

15.11 Answer to Check Your Progress

15.12 Possible Questions

15.1 LEARNING OBJECTIVE

After going through this unit, you will be able to know

- what is regression
- difference between correlation and regression analysis
- uses of regression
- what are regression lines
- what is regression equation

15.2 INTRODUCTION

The term regression was first used by Sir Francis Galton in 1877. The dictionary meaning of the word regression is the act of returning or going back. Sir Galton introduced the term regression in the paper “Regression towards Mediocrity in Hereditary Stature”, where he studied the relationship between height of father and son. Sir Galton coined the term “regression line” to describe the tendency to regress or going back. Although this term has become popular, but modern writers nowadays prefer

to use the term "estimating line" as regression used for estimating or predicting the value of one variable given the value of another. For example, we know that advertising can have an impact on sales. So producers can use the cost on advertising to predict the amount of sales, or otherwise use the estimated amount of sale to know how much advertising cost needs to be incurred. In other words, regression analysis reveals average relationship between two variables and this makes possible estimation or prediction.

Some definitions of Regression

- "Regression analysis attempts to establish the *'nature of the relationship'* between variables – that is, to study the functional relationship between the variables and thereby provide a mechanism for prediction, or forecasting." – Ya-Lun Chau
- "One of the most frequently used techniques in economics and business research, to find a relation between two or more variables that are related casually, is regression analysis" – Taro Yamane
- The term regression analysis refers to the methods by which estimates are made of the values of a variable from the knowledge of the values of one or more other variables and to the measurement of errors involved in this estimation process". – Morris Hamburg

The above definition reveals that regression analysis is a statistical device with the help of which we are in a position to estimate (or predict) the unknown values of one variable from known values of another variable. The variable which is used to predict the variable of interest is called the independent variable or explanatory variable and the variable we are trying to predict is called the dependent variable or "explained" variable. The independent variable is denoted by X and the dependent variable by Y. The analysis used is called the simple linear regression analysis. It is called simple because there is only one independent variable, and linear because of the assumed linear relationship between the dependent and independent variables. Here the term "linear relationship" indicates that the relationship between dependent and independent variables forms a straight line of the equation Y=a+bx, where a and b are constants.

It may be pointed out that there is only mathematical or functional dependence between the two variables. Mere existence of dependence does not necessarily indicate any cause and effect relationship between the two variables.

15.3 USES OF REGRESSION ANALYSIS:

Regression analysis is widely used in various fields. Its applications are extended to almost all natural, physical and social sciences. Regression analysis is widely used for prediction and forecasting. It is particularly used to understand the relationship between dependent variables and independent variable, and to explore the forms of these relationships. Regression analysis is also widely used in business and in framing economic policies, especially because it has the ability to predict and forecast. It thereby helps in crucial decision making, among others.

15.4 OBJECTIVES OF REGRESSION ANALYSIS

The objectives of regression analysis are as follows:

- Regression analysis provides estimates of values of the dependent variable from values of the independent variable. The device used to accomplish this estimation procedure is the regression line. The regression line describes the average relationship existing between X and Y, i.e., it displays mean values of X for given values of Y. The equation of this line, known as the regression equation, provides estimates of the dependent variable when values of the independent variable are inserted into the equation.
- Another objective of regression analysis is to obtain a measure of the error involved in using the regression line as a basis for estimator on. For this purpose the standard error of estimate is calculated.
- With the help of regression coefficients we can calculate the correlation coefficient.

15.5 DIFFERENCE BETWEEN CORRELATION AND REGRESSION

Correlation and regression are the most widely used economic tools. Both furnish important information, although in most of the situations, it is not clear as to which one should be adopted. The following are some of the points of difference between correlation and regression.

	CORRELATION	REGRESSION
1.	It measures the degree of covariability between X and Y.	It studies the nature of relationship between variables so that we can predict the value of one on the basis of another.
2.	Correlation analysis only provides the information that there is relationship between two variables, asserting the degree of relationship, but tells nothing about dependence.	In regression, on the other hand, one variable is dependent and the other is independent. Thus, it tells us about the mathematical or functional dependence of the two variables.
3.	In correlation analysis r_{xy} is a measure of direction and degree of linear relationship between two variables X and Y. r_{xy} and r_{yx} are symmetric ($r_{xy} = r_{yx}$), i.e., it is immaterial which of X and Y is dependent variable and which is independent variable.	In regression analysis, the regression coefficients *bxy* and *byx* are not symmetric, i.e., *bxy?byx* and hence it definitely makes a difference as to which variable is dependent and which is independent.
4.	There may be correlation between two variables without any logical sense, such as increase in income and weight.	In regression analysis there cannot be any illogical dependence.
5.	Correlation coefficient is independent of change of origin and scale.	Regression coefficient is independent of change of origin but not of scale.

15.6 SIMMILARITY BETWEEN CORRELATION AND REGRESSION

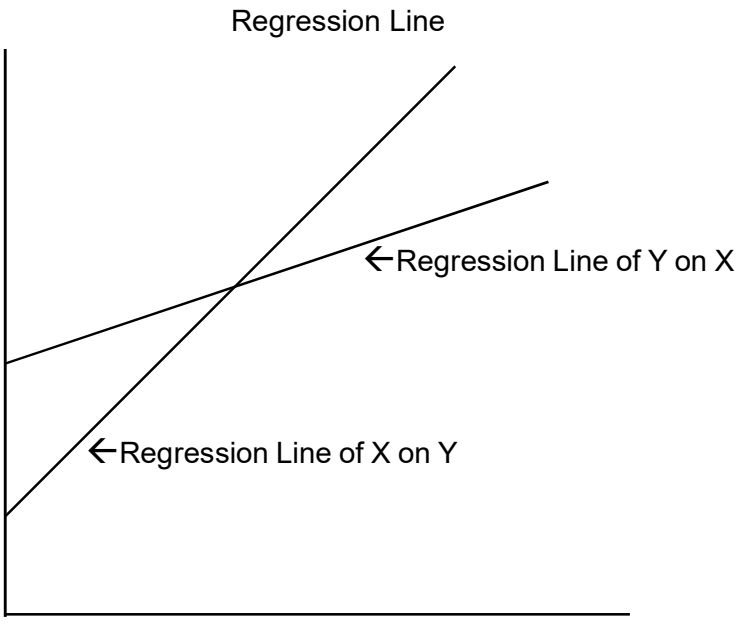
There is also some similarity between correlation and regression analysis.

- The coefficient of correlation (r) takes the same sign as the regression coefficients (b_{xy} and b_{yx}).
- If the value of *b* is significant at a given level of significance *r* is also significant at that level.

15.7 REGRESSION LINES

Let us consider two variables X and Y. We will have two regression lines as the regression of X on Y and the regression of Y on X. The regression line of Y on X gives the most probable values of Y for given values of X and the regression line of X on Y gives the most probable values of X for given values of Y. However, we will have only one regression line, i.e, regression lines will coincide when there is either perfect positive or perfect negative correlation between two variables (r=±1). The further the two regression lines from each other, the lesser will be the degree of correlation between the variables and nearer the two regression lines with each other, the higher will be the degree of correlation between the variables. If the variables are independent, r is zero and the lines of regression are at right angles, i.e., parallel to OX and OY.

It may be pointed out that regression lines cut each other at the point of average of X and Y, i.e., if from the point where both the regression lines cut each other a perpendicular is drawn on the X-axis, we will get the mean value of X and if from that point a horizontal line is drawn on the Y-axis, we will get the mean value of Y.



15.8 REGRESSION EQUATIONS

Regression equations, also known as estimating equations, are algebraic expressions of the regression lines. Since there are two regression lines, there are two regression equations – the regression equation of X on Y is used to describe the variation in the values of X for given changes in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.

Regression equation of Y on X

The regression equation of Y on X is expressed as follows:

$Y = a+bX$

Here, *Y*= dependent variable, i.e., its value depends on X

X=independent variable, i.e., we can take a given value of X and compute the value of Y

a=Y-intercept, because its value is the point at which the regression line crosses the Yaxis or the vertical axis.

b=slope of the line. It represents change in Y variable for a given change in X variable.

'*a*' and '*b*' are called numerical constants because for any given straight line, their values does not change.

Regression equation of X on Y

The regression equation of X on Y is expressed as follows:

$$X = a+bY$$

Here, *X*= dependent variable, i.e., its value depends on X

Y=independent variable, i.e., we can take a given value of X and compute the value of Y

a=X-intercept, because its value is the point at which the regression line crosses the Y-axis or the vertical axis.

b=slope of the line. It represents change in X variable for a given change in Y variable.



CHECK YOUR PROGRESS

Q1: What is the regression equation of Y or X?

Q2: What is the regresion equation of X and Y?



15.9 LETS SUM UP

- Regression analysis reveals the average relationship between two variables and this makes possible estimation or prediction.
- Regression analysis is widely used in various fields. Its applications are extended to almost all natural, physical and social sciences.
- Regression analysis provides estimates of values of the dependent variable from values of the independent variable.
- With the help of regression coefficients we can calculate the correlation coefficient.
- There are similarities as well as differences between correlation and regression.
- The regression equation of Y on X is expressed as follows:
  $Y = a+bX$
- The regression equation of X on Y is expressed as follows:
- $X = a+bY$



15.10 FURTHER READING

1. Gupta S.P. 2010. Statistical Method. New Delhi: Sultan and Sons.



15.11 ANSWER TO CHECK YOUR PROGRESS

Ans to QNo 1: The regression equation of Y on X is expressed as follows:

Here, Y = a + bX

Here, Y = dependent variable, I.e., its value depends on X

X = independent, i.e., we can take a given value of X and compute the value of Y

a = Y - intercept, because its value is the point at which the regression line crosses the Y axis or the vertical axis.

b = slope of the line. It represents change in Y variable for a given change in X variable.

'a' and 'b' are called numerical constants because for any given straight line, their values does not change.

Ans to QNo 2: The regression equation of X on Y is expressed as follows:

X = a + bY

Here, X = dependent variable, i.e., its value depends on X

Y = independent variable, i.e., we can take a given value of X and compute the value of Y

a = X- intercept, because its value is the point at which the regression line crosses the Y-axisor the vertical axis.

b = slope of the line. It represents change in X variable for a given change in Y variable.



15.12 POSSIBLE QUESTIONS

A) Short answer type: (Answer within 150 words)

Q.1: Define regression.

Q.2: Write any two uses of regression analysis.

Q.3: Write any two objectives of regression analysis.

Q.4: Write two similarities of correlation and regression.

Q.6: Write two differences of correlation and regression.

Q.7: What are regression lines?

Q.8: Write the regression equation of X on Y.

Q.9: Write the regression equation of Y on X.

B) Long answer type (Answer within 150 words)

Q.1: What is regression? What are its different uses?

Q.2: What are the objectives of regression analysis?

Q.3: Discuss the differences between correlation and regression.

Q.4: Discuss in detail the regression eqution.

*** ***** ***



REFERENCES (FOR ALL UNITS)

1. Babbie, E. 2001. *The practice of social research* (9th ed.). Belmont, CA: Wadsworth.
2. Bachman, R. & Schutt, R. 2007. *The practice of research in criminology and criminal justice* (3rd ed.). Thousand Oaks, CA: Sage.
3. Corley, K. & Gioia, D. 2011. Building theory about theory building: What constitutes a theoretical contribution? *Academy of Management Review*, *36*, 12-32.
4. Cozby, P. 2009. *Methods in behavioral research*. Boston: McGraw Hill Higher Education.
5. C.R. Kothari 2004 Research Methodology Methods and Techniques, 2nd Edition: New Delhi, New Age International Publisher
6. Dubin, R. (1978). *Theory building* (Rev. ed.). NY: Free Press
7. Guthrie Gerard, *Basic Research Methods: An Entry to Social Science Research*. New Delhi: Sage publication. India. 2010.
8. Gupta S.P. 2010. Statistical Methods. New Delhi: Sultan Chand & Sons.
9. Gupta S.P. 2010. Statistical Method. New Delhi: Sultan.
10. Goode, W. J and Hatt, P.K. 2006. Methods in Social Research. Delhi: Surajeet Publication.
11. Gelso, C. (2006). Applying theories to research: The interplay of theory and research in science. In F.T. Leong & J.T. Austin (Eds.), *The psychology research handbook*. Thousand Oaks, CA: Sage. Retreived from http://www.sage-ereference.com/hdbk_psychrsch/Article_n32
12. Hay, J. & Lee, S. (2009). Theory building through qualitative research: Marshalling opportunities to advance cancer screening efforts. *Health Education & Behavior*, *36*, 145-149. doi: 10.1177/10908109340515
13. Heinen, J. (1985). A primer on psychological theory. *The Journal of Psychology*, *119*, 413-421.
14. Kerlinger, R. (1986). *Foundations of behavioral research*. NY: Holt, Rinehart, & Winston.
15. Jahoda, Marie, Morton Deutsch, and Stuart W. Cook. 1951. Research Methods in Social Relations. New York: Dryden.
16. Kothari C.R. *Research Methods and Techniques.* New Delhi: New Age International Publisher. 2004
17. Lincoln, Y., & Guba, E. (2000). Paradigmatic controversies, contradictions, and emerging confluences. In N. Denzin &Y. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed.) (pp. 163–188). London: Sage.

18. Lynham, S. (2002). The general method of theory-building research in applied disciplines. *Advances in Developing Human Resources*, *4*, 221-241.
19. Locke, E. (2007). The case for inductive theory building. *Journal of Management*, *33*, 867-890. doi: 10.1177/0149206307307636
20. L. Choudhury 2000 An introduction to Statistics, Volume One: Guwahati Kitapghar
21. Medhi, J. 2006. Statistical Methods : An Introductory text. New Age International (p) Ltd.
22. Morrow, R. (1994). *Critical theory and methodology*. Thousand Oaks, CA: Sage.
23. Quinney, R. (1970). *The social reality of crime*. Boston: Little, Brown.
24. Rynes, S. & Gephart, R. (2004). From the editors: Qualitative research and the academy of management journal. *Academy of Management Journal*, *47*, 454-462.
25. Stinchcombe, A. (1968). *Constructing social theories*. Chicago: University of Chicago Press.
26. Sutton, R. & Staw, B. (1995). What theory is not. *Administrative Science Quarterly*, *40*, 371–384.
27. Smith, J., Bekker, H., & Cheater, F. (2011). Theoretical versus pragmatic design in qualitative research. *Nurse Researcher*, *18*, 39-51.
28. Southern, S. & Devlin, J. (2010). Theory development: A bridge between practice and research. *The Family Journal: Counseling and Therapy for Couples and Familes*, *18*, 84-87.doi: 10.1177/1066480709358422
29. Thurstone, L. L, .and E. J. Chave. 1929. The Measurement of Attitudes Chicago: University of Chicago Press.
30. Torraco, R. (2002). Research methods for theory building in applied disciplines: A comparative analysis. *Advances in Developing Human Resources*, *4*, 355-376.
31. Young, P.V. *Scientific Social Surveys and Research*. New Delhi: Prentice Hall, 2009
32. Young, P. V. 1992. Scientific social surveys and research, New Delhi: Prentice Hall of India.
33. Zikmund, W., Babin, B., Carr, J., & Griffin, M. (2008). *Business research methods* (8th ed.). Mason, OH: Cengage.