PREFACE

In a bid to standardise higher education in the country, the University Grants Commission (UGC) has introduced Choice Based Credit System (CBCS) based on five types of courses viz. *core, discipline specific generic elective*, *ability* and *skill enhancement* for graduate students of all programmes at Honours level. This brings in the semester pattern, which finds efficacy in sync with credit system, credit transfer, comprehensive continuous assessments and a graded pattern of evaluation. The objective is to offer learners ample flexibility to choose from a wide gamut of courses, as also to provide them lateral mobility between various educational institutions in the country where they can carry acquired credits. I am happy to note that the University has been accredited by NAAC with grade 'A'.

UGC (Open and Distance Learning Programmes and Online Learning Programmes) Regulations, 2020 have mandated compliance with CBCS for U.G. programmes for all the HEIs in this mode. Welcoming this paradigm shift in higher education, Netaji Subhas Open University (NSOU) has resolved to adopt CBCS from the academic session 2021-22 at the Under Graduate Degree Programme level. The present syllabus, framed in the spirit of syllabi recommended by UGC, lays due stress on all aspects envisaged in the curricular framework of the apex body on higher education. It will be imparted to learners over the *six* semesters of the Programme.

Self Learning Materials (SLMs) are the mainstay of Student Support Services (SSS) of an Open University. From a logistic point of view, NSOU has embarked upon CBCS presently with SLMs in English / Bengali. Eventually, the English version SLMs will be translated into Bengali too, for the benefit of learners. As always, all of our teaching faculties contributed in this process. In addition to this we have also requisitioned the services of best academics in each domain in preparation of the new SLMs. I am sure they will be of commendable academic support. We look forward to proactive feedback from all stakeholders who will participate in the teaching-learning based on these study materials. It has been a very challenging task well executed, and I congratulate all concerned in the preparation of these SLMs.

I wish the venture a grand success.

Professor (Dr.) Subha Sankar Sarkar
Vice-Chancellor

Netaji Subhas Open University

Under Graduate Degree Programme

Choice Based Credit System (CBCS)

Subject : Honours in Geography (HGR)

Course : Statistical Methods in Geography Laboratory and Human Geography Laboratory

Course Code : CC - GR - 05

(Core Course)

First Print : August, 2022

Printed in accordance with the regulations of the
Distance Education Bureau of the University Grants Commission.

Netaji Subhas Open University

Under Graduate Degree Programme

Choice Based Credit System (CBCS)

Subject : Honours in Geography (HGR)

Course : Statistical Methods in Geography Laboratory and Human Geography Laboratory

Course Code : CC - GR - 05

(Core Course)

: Course Writer :

Dr. Sandipan Chakraborty
Associate Professor of Geography
Chandernagar Govt. College

: Course Editor :

Dr. Biraj Kanti Mondal
Assistant Professor of Geography,
NSOU

: Format Editor :

Smt. Tinki Kar Bhattacharya, NSOU

Notification

Kishore Sengupta
Registrar

Netaji Subhas Open University

UG : Geography (HGR)

Course : Statistical Methods in Geography Laboratory and Human Geography Laboratory
Course Code : CC - GR - 05

Module - 1 ❑ Statistical Methods : Theoretical Basis

Module - 2 ❑ Statistical Methods in Geography Laboratory: List of Practical

Module - 3 ❑ Human Geography Laboratory

.

Module - 01
Statistical Methods-Theoretical Basis

Unit-1 ❑ Discrete and continuous data, population and samples, scales of measurements-nominal, ordinal, interval and ratio, Sources of data, collection of data and formation of statistical tables

Structure

1.1 Objectives

- To make the learnees understand about the different types of data.
- To know about the collection of data and understand the statistical table.

1.2 Introduction

The subject Statistics as it seems, is not a new discipline but it is as old as the human society itself. It has been used right from the existence of life on this earth although the sphere of its utility was very much restricted. In the olden days Statistics was regarded as the 'Science Statecraft' and was the by product of the administrative activity of the state. The word Statistics seems to bave been derived from the Latin word 'status' or the Italian word 'statista' or the German word 'statistik' or the French word 'statistique' each of which means a political state. In India, an efficient system of collecting official and administrative statistics existed even 2000 years ago, in particular, during the reign of Chandragupta Maurya (324-300 B.C.). Historical evidences about the prevalence of a very good system of collecting vital statistics and

registration of births and deaths even before 300 B.C. are available in Kautilya's 'Arthashastra'. The records of land, agriculture and wealth statistics were maintained by Todermal, the land and revenue minister in the reign of Akbar (1556-1605 A.D). A detailed account of the administrative and statistical surveys conducted during Akbar's reign is available in the book 'Ain-e-Akbari" written by Abul Fazl (in 1596-97), one of the nine jems of Akbar. Sixteen century saw the application of Statistics for the collection of the data relating to the movements of heavenly bodies–stars and planets– to know about their position and for the prediction of Eclipses. Seventeenth century witnessed the orgin of Vital statistics. Captain John Graunt of London (1620-1674), known as the Father of Vital Statistics, was the first man to make a systematic study of the birth and death statistics.

Modern scientists in the development of the subject of statistics are Englishmen who did pioneering work in the application of Statistics to different disciplines. Francis Galton (1822-1921) pioneered the study of 'RegressionAnalysis' in Biometry; ***Karl Pearson (1857-1936) who founded*** the greatest statistical laboratory in England pioneered the study of 'Correlation Analysis'. His Chi-Square test of Goodness of Fit is the first and most important of the tests of significance in Statistcs; W.S. Gosset with his t-test ushered in an era of exact (small) sample tests. Perhaps most of the work in the statistical theory during the past few decades can be attrubuted to a single person ***Sir Ronald A. Fisher (1890-1936) who applied statistics to a variety of diversified fields such as genetics, biometry, psychology, education and agriculture, etc., and who is rightly termed as the Father of Statistics.*** In addition to enhancing the existing statistical theory he is pioneer in Estimation Theory (Point Estimation and fiducial Inference); Exact (small) Sampling Distributions, Analysis of Variance and Design of Experiments.

Statistics has been defined differently by different authors and each author has assigned new limits to the field which should be included in its scope. We can do no better than give selected definitions of statistics by some authors and then come to the conclusion about the scope of the subject. A.L. Bowley defines, **"Statistics may be called the science of counting".** At another place he difines, **"Statistics may be called the science of averages".** Both these definitions are narrow and throw light only on one aspect of Statistics. According to King, **"The science of statistics is the method of judging collective, natural or social, phenomenon from the results obtained from the analysis or enumeration or collection of estimates".** Many a time counting is not possible and estimates are required to be made. Therefore, ***Boddington defines it*** as "the science of estimates and probabilities". But this definition

also does not cover the entire scope of statistics. The statistical methods are methods for the collection, analysis and interpretation of numerical data and from a basis for the analysis and comparison of the observed phenomena. ***In the words of Croxton & Cowden,*** "Statistics may be difined as the collection, presentation, analysis and interpretation of numerical data". Horace Secrist has given an exhaustive definition of the term Statistics in the plural sense. *According to him:* **"By statistics we mean aggregates of facts affected to a marked extent by a multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standards of accuracy collected in a systematic manner for a predetermined purpose and placed in relation to each other"**. *This definition makes it quite clear that as numerical statement of facts.*

Discrete and Continuous Data

Data can be defined as systematic record of a particular quantity. It is the different values of that quantity represented together in a set. It is a collection of facts and figures to be used for a specific purpose such as a survey or analysis. When arranged in an organized form, can be called information. The source of data (primary data, secondary data) is also an important factor.

Discrete Data :

The term discrete implies distinct or separate. So, Discrete data refers to the type of quantitative data that relies on counts. It contains only finite values, whose subdivision is not possible. It includes only those values that can only be counted in whole numbers or integers and are separate which means the data cannot be broken down into fraction or decimal. Number of students in the school, the number of cars in the parking lot, the number of computers in a computer lab, the number of animals in a zoo, etc. These are data that can take only certain specific value rather than a range of values.

Continuous Data :

Continuous data is described as an unbroken set of observation; that can be measured on a scale. It can take any numeric value, within a finite or infinite range of possible value. Statistically, range refers to the difference between highest and lowest observation. The continuous data can be broken down into fractions and decimal, i.e. it can be meaningfully subdivided into smaller parts according to the measurement precision. These are data that can take values between a certain range with the highest and lowest values. The defference between the highest and lowest value is called the range of data. Age, height or weight of a person, time taken to

complete a task, temperature, time money etc. These are classified as continuous data. Continuous data can be tabulated in what is called a frequency distribution. They can be graphically represented using histograms.

Discrete Data		Continuous Data	
Marks	**No. of Students**	**Height (m)**	**No. of Students**
25	11	1.5 - 1.6	02
35	13	1.4 - 1.5	04
38	08	1.3 - 1.4	05
42	07	1.2 - 1.3	11
50	03	1.1 - 1.2	18

Differences Between Discrete and Continuous Data :

The difference between discrete and continuous data can be drawn clearly on the following grounds :

1. Discrete data is the type of data that has clear spaces between values. Continuous data that falls in a continuous sequence.
2. Discrete data is countable while continuous data is measurable.
3. Discrete data contains distinct or separate values. On the other hand, continuous data includes any value within range.
4. Discrete data is graphically represented by bar graph whereas a histogram is used to represent continuous data graphically.
5. Tabulation of discrete data, done against a single value, is called as an ungrouped frequency distribution. On the contrary, tabulation for continuous data, done against a group of value, called as grouped frequency distribution.
6. Overlapping or mutually exclusive classification, such as 10-20, 20-30, ..., etc. is done for discrete data.
7. In a graph of the discrete function, it shows distinct point which remains unconnected. Unlike, continuous function graph, the points are connected with an unbroken line.

1.3 Population and sample

Whenever we hear the term 'population,' the first thing that strikes our mind is a large group of people. In the same way, in statistics **population** denotes a large

group consisting of elements having at least one common feature. The term is often contrasted with the **sample**, which is nothing but a part of the population that is so selected to represent the entire group.

Population represents the entirety of persons, units, objects and anything that is capable of being conceived, having certain properties. ***On the contrary, the sample is a finite subset of the population,*** that is chosen by a systematic process, to find out the characteristics of the parent set. ***In simple terms, population means the aggregate of all elements under study having one or more common characteristic,*** for example, all people living in India constitutes the population is not confined to people only, but it may also include animals, events, objects, buildings, etc. It can be of any size, and the number of elements or members in a population is know as population size, i.e. if there are hundred million people in India, then the population size (N) is 100 million. The different types of population are :

1. **Finite Population :** When the number of elements of the population is fixed and thus making it possible to enumerate it in totality, the population is said to be finite.
2. **Infinite Population :** When the number of units in a population are uncountable, and so it is impossible to observe all the items of the universe, then the population is considered as infinite.
3. **Existent Population :** The population which comprises of objects that exist in reality is called existent population.
4. **Hypothetical Population :** Hypothetical or imaginary population is the population which exists hypothetically.

Examples :

- The population of workers working in the sugar factory.
- The population of motorcycles produced by a particular company.
- The population of mosquitoes in a town.
- The population of tax payers in India.

By the term sample, we mean a part of population chosen at random for participation in the study. The sample so selected should be such that it represent the population in all its should be free from bias, so as to produce miniature cross-section, as the sample observations are used to make generalizations about the population. ***In other words, the respondents selected out of population constitutes a 'sample', and the process of selecting respondent is known as 'sampling.'*** The units under study ***are called sampling units,*** and the number of units in a sample ***is called sample size.*** While conducting statistical testing, samples are mainly used ***when the sample size is***

too large to include all the members of the population under study.

Difference Between Population and Sample

The difference between population and sample can be drawn clearly on the following grounds :

1. The collection of all elements possessing common characteristics that comprise universe is known as the population. A subgroup of the members of population chosen for participation in the study is called sample.
2. The population consists of each and every element of the entire group. On the other hand, only a handful of items of the population is included in a sample.
3. The characteristic of population based on all units is called parameter while the measure of sample observation is called statistic.
4. When information is collected from all units of population, the process is known as census or complete enumeration. Conversely, the sample survey is conducted to gather information from the sample using sampling method.
5. With population, the focus is to identify the characteristics of the elements whereas in the case of the sample, the focus is made on making the generalisation about the characteristics of the population, from which the sample came from.

1.4 Scales of Measurments

Level of measurement or scale of measure is a classification that describes the nature of information within the values assigned to variables. ***Psychologist Stanley Smith Stevens*** depeloped the best-known classification with four levels or scales of measurement : nominal, ordinal, interval, and ratio. In general the term 'Measurement' is used in narrow sense, but in statistics, the term measuremet is used more broadly and is more appropriately termed scales of measurement. Scales of measurement refer to ways in which variables/numbers are define and categorized. Each scales of measurments has certain properties which in turn determines the appropriateness for use of certain statistical analyses. The four scales of measuremt are nominal, ordinal, interval, and ratio scales are identified so far.

In **nominal** measurement the numerical values just "name" the attribute uniquely. No ordering of the cases is implied. For example, jersey numbers in basketball are measures at the nominal level. A player with number 30 is not more of anything that a player with number 15, and is certainly not twice whatever number 15 is.

In **ordinal** measurement the attributes can be rank-ordered. Here, distances between attributes do not have any meaning. For example, on a survey you might code Educational Attainment as 0=less than high school; 1=some high school.; 2=high school degree; 3=some college; 4=college degree; 5=post college. In this measure, higher number means *more* education. But distance from 0 to 1 same as 3 to 4? of course not. The interval between value is not interpretable in an ordinal measure.

In **interval** measurement the distance between attributes *does* have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. The interval between values is interpretable. Because of this, it makes sense to computer an average of an interval variable, where it doesn't make sense to do so for ordinal scales. But note that in interval measurement ratios don't make any sense - 80 degrees is not twice as hot as 40 degrees (although the attribute value is twice as large).

Finally, in ratio measurement there is always an absolute zero that is meaningful. This means that you can construct a meaningful fraction (or ratio) with a ratio variable. Weight is a ratio variable. In applied social research most "count" variables are ratio, for example, the number of clients in past six months. Why? Because you can have zero clients and because it is meaningful to say that "...we had twice as many clients in the past six months as we did in the previous six months."

It's important to recognize that there is a hierarchy implied in the level of measurement idea. At lower levels of measurement, assumptions tend to be less restrictive and data analyses tend to be less sensitive. At each level up the hierarchy, the current level includes all of the qualities of the one below it and adds something new. In general, it is desirable to have a higher level of measurement (e.g., interval or ratio) rather than a lower one (nominal or ordinal).

Types of Scales

Before we can conduct a statistical analysis, we need to measure our *dependent variable*. Exactly how the measurement is carried out depends on the type of variable involved in the analysis. Different types are measured differently. To measure the time taken to respond to a stimulus, you might use a stop watch. Stop watches are of no use, of course, when it comes to measuring someone's attitude towards a political candidate. A rating scale is more appropriate in this case (with labels like "very favourable," etc.). For a dependent variable such as "favourite color," you can simply note the colour-word (like "red") that the subject offers. Although procedures for measurement differ in many ways, they can be classified using a few fundamental categories. In a given category, all of the procedures share some properties that are

important for you know about. The categories are called "scale types," or just "scales," and are described in this section.

Nominal scales –

When measuring using a nominal scale, one simply names or categorizes. Gender, handedness, favourite color, and religion are examples of variables measured on a nominal scale. The essential point about nominal scales is that they do not imply any ordering among the responses. For example, when classifying people according to their favourite color, there is no sense in which green is placed "ahead of" blue. Responses are merely categorize. Nominal scales embody the lowest level of measurement.

Ordinal scales–

A researcher wishing to measure consumers' satisfaction with their microwave ovens might ask them to specify their feelings as either "very dissatisfied," "somewhat dissatisfied," "somewhat satisfied," or "very satisfied." The item in this scale are ordered, ranging from least to most satisfied. This is what distinguishes ordinal from nominal scales. Unlike nominal scales, ordinal scales allow comparisons of the degree to which two subjects possess the dependent variable. For example, our satisfaction ordering makes it meaningful to assert that one person is more satisfied than another with their microwave ovens. Such an assertion reflects the first person's use of a varbal lable that comes later in the list than the lable chosen by the second person. ***On the other hand, ordinal scales fail to capture important information that will be present in the other scales we examine.*** In particular, the difference between two levels on an ordinal scale cannot be assumed to be the same as the diference between two other levels. In our satisfaction scale, for example, the difference between the responses "very dissatisfied" and "somewhat dissatisfied" is propably not equivalent to the difference between "somewhat dissatisfied" and "somewhat satisfied." Nothing in our measurement procedure allows us to determine whether the two difference reflect the same difference in psychological satisfaction.

Interval scales–

Interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahernheit scale of temperature. The difference between 30 degrees and 40 degrees represents the same temperature difference as the difference between 80 degrees and 90 dgrees. This is because each 10-degree interval has the same physical meaning (in terms of the kinetic energy of molecules). Interval scales are not perfect, however. In particular, they do not have a true zero point even if one of the scaled values happens to carry the name "zero." The Fahrenheit scale

illustrates the issue. Zero degrees Fahrenheit does not represent the complete absence of temperature (the absence of any molecular kinetic energy). In reality, the lable "zero" is applied to its temperature for quite accidental reasons connected to the history of temperature measurement. Since an interval scale has no true zero point, it does not make sense to compute ratios of temperatures. for example, there is no sense in which the ratio of 40 to 20 degrees Fahrenheit is the same as the ratio of 100 to 50 degrees; no interesting physical property is preserved across the ratios.

Ratio scales–

The ratio scale of measurement is the most informative scale. It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured. You can think of a ratio scale as the three earlier scales rolled up in one. Like a nominal scale, it provides a name or category for each object (the numbers serve as lables). Like an ordinal scale, the objects are ordered (in terms of the ordering of the numbers). Like an interval scale, the same difference at two places on the scale has the same meaning. And in addition, the same ratio at two places on the scale also carries the same meaning.

The Fahreheit scale for temperature has an arbitary zero point and is therefore not a ratio scale. However, zero on the kelvine scale is absolute zero. This makes the Kelvin scale a ratio scale. For example, if one temperature is twice as high as another as measured on the Kelvin scale, then it has twice the kinetic energy of the other temperature.

Another exaple of a ratio scale is the amount of money you have in your pocket right now (25 or 55 Rupees,etc.). Money is measured on a ratio scale because, in addition to having the properties of an interval scale, it has a true zero point: if you have zero money, this implies the absence of money. Since money has a true zero point, it makes sense to say that someone with 50 Rupees has twice as much money as someone with 25 Rupees.

1.5 Sources and Types of Data

Statistics is basically a science that involves **data** collection, **data** interpretation and finally **data** validation. **Statistical data** analysis is a procedure of performing various **statistical** operations. When we mean statistical data, we usually refer any type of information, qualitative or quantitative. The Quantitative **data** basically involves descriptive **data**, such as survey **data** and observational **data**. Data may be qualitative or quantitative. Once you know the difference between them, you can know how to use them.

Qualitative Data :

They represent some characteristics or attributes. They depict descriptions that may be observed but cannot be computed or calculated. For example, data on attributes such as intelligence, honesty, wisdom, cleanliness, and creativity collected using the students of your class a sample would be classified as qualitative. They are more exploratory than conclusive in nature.

Quantivative Data :

These can be measured and not simply observed. They can be numerically represented and calculations can be performed on them. For example, data on the number of students playing different sports from your class gives an estimate of how many of the total students play which sport. This information is numerical and can be classified as quantitative.

Classification of Data

The process of arranging data into homogenous groups or classes according to some common characteristics present in the data is called classification. During the process of sorting letters in a post office, the letters are classified according to the cities and further arranged according to streets. So, data can be classified various terms; There are four important bases of classification:

(1) Qualitative Base (2) Quantitative Base (3) Geographical Base (4) Chronological or Temporal Base

(1) Qualitative Base– When the data are classified according to a quality or attribute such as sex, religion, literacy, intelligence, etc.

(2) Quantitative Base– When the data are classified by quantitative characteristics like height weight, age, income, etc.

(3) Geographical Base– When the data are classified by geographical regions or location, like states, provinces, cities, countries etc.

(4) Chronological or Temporal Base– When the data are classified or arranged by their time of occurrence, such as years, months, weeks, days, etc. such as time series data.

Types of classification

(1) One-way Classification– If we classify observed data keeping in view a single characteristic, this type of classification is known as one-way classification. The population of the world may be classified by religion as Muslim, Christian, etc.

(2) Two-way Classification– If we consider two characteristics at a time in order to classify the observed data then we are doing two way classification. The population of the world may be classified by religion and sex.

(3) Multi-way Classification– We may consider more that two characteristics at a time to classify given or observed data. In this way we deal in multi-way classification. The population of the world may be classified by religion, sex and literacy.

1.6 Collection of data and various sources

Data can be collected by diffirent methods and for different purpose. So, data are classified in different manner. The orginal compiler of the data is the primary source. For example, the office of the Registrar General will be the primary source of the decennial population census figures. A secondary source is the one that furnishes the data that were originally compiled by someone else. If the population census figures issued by the office of the Registrar-General are published in the Indian Year Book, this publication will be the secondary source of the population data. The sources of data also are classified according to the character of the data yielded by them. Thus the data which are gathered from the primary sources known as primary data and the one gathered from the secondary source is known as secondary data. When an investigator is making use of figures which he has obtained by field enumeration, he is said to be using primary data and when he is making use of figures which he has obtained from some other source, he is said to be using secondary data. An investigator has to decide whether he will collect fresh (primary) data or he will compile data from the published sources. The former is reliable per se but the latter can be relied upon only by examining the following factors.

(i) source from which they have been obtained;

(ii) their true significance;

(iii) completeness and

(iv) method to collection.

In addition to the above factors, there are other factors to be considered while making choice between the primary or secondary data :

(i) Nature and scope of enquiry.

(ii) Availability of time and money.

(iii) Degree of accuracy requried and

(iv) The status of the investigator i.e., individual, Pvt. Co., Govt. etc.

However, it may may be pointed out that in certain investigations both primary and secondary data may have to be used, one may be supplement to the other.

Methods of Collection o Primary Data

The primary methods of collection of statistical information are the following :

1. Direct Personal Observation,
2. Indirect Personal Observation,
3. Schedules to be filled in by informants
4. Information from Correspondents, and
5. Questionnaires in charge of enumerators

The particular method that is decided to be adopted would depend upon the nature and availability of time, money and other facilities available to the investigation.

Primary Data

They are the data that are ***collected for the first time*** by an investigator for a specific purpose. primary data are 'pure' in the sense that no statistical operations have been performed on them and they are orginal. An example of primary data is the Census of primary data is the Census of India.

Secondary Data

They are the data that are ***sourced from some place*** that has originally collected it. This means that this kind of data has already been collected by some researchers or investigators in the past and is available either in published or unpublished form. This information is impure as statistical operations may have been performed on them already. An example is an information available on the Government of India. Department of Finance's website or in other repositories, books, journals, etc.

Formations of Statistical Table

One of the simplest and most revealing devices for summarising data and presenting them in a meaningful way is to frame a statistical table. It is a systematic arrangement of statistical data in columns and rows in vertical and horizonatal fashion. The number of parts of a table varies from case to case depending upon the given data. A statistical table has at least four major parts and some other minor parts.

(1) The Title

(2) The Box Head (column captions)

(3) The Stub (row captions)

(4) The Body
(5) Prefatory Notes
(6) Foot Notes
(7) Source Notes

The general sketch of table indicating its necessary parts is shown below :

	--- Box Head ---
--- Row Captions ---	---- Column Caption ----
--- Stub Entries ---	--- The Body ---

Foot Notes....

Source Notes...

(1) The Title –

The title is the main heading written in capitals shown at the top of the table. It must explain the contents of the table and throw light on the table, as whole defferent parts of the heading can be separated by commas. There are no full stops in the title.

(2) The Box Head (column captions) –

The vertical heading and subheading of the column are called columns captains. The spaces where these column headings are written is called the box head. Only the first letter of the box head is in capital letters and the remaining words must be written in lowercase.

(3) The Stub (row captions) –

The horizontal headings and sub heading of the row captions and the space where these rows headings are written is called the stub.

(4) The Body –

This is the main part of the table which contains the numerical information classified with respect to row and column captions.

(5) Prefatory Notes –

A statement given below the title and enclosed in brackets usually describes the units of measurement and is called the prefatory notes.

(6) Foot Notes –

These appear immediately below the body of the table providing additional explanation.

(7) Source Notes –

The source notes are given at the end of the table indicating the source the information has been taken from. It includes the information about compiling agency, publication, etc.

General Rules of Tabulation :

* A table should be simple and attractive. There should be no need of further explanation (details).
* Proper and clear headings for columns and rows are necessary.
* Suitable approximation may be adopted and figures may be rounded off.
* The unit of measurement should be well defined.
* If the observations are large in numbers they can be broken into two or three tables.
* Thick lines should be used to separate the data under big classes and thin lines to separate the sub classes of data.
* The process of placing classified data into tabular form is known as tabulation. A table is a symmetric arrabgements of statistical data in rows and cloumns. Rows are horizontal arrangements whereas columns are vertical arrangements. It may be simple, double or complex depending upon the type of classification.

Types of Tabulation :

(1) Simple Tabulation or One-way Tabulation –

When the data are tabulated to one characteristic, it is said to be a simple tabulation or one-way tabulation. Tabulation of data on the population of the

world classified by one characteristic like religion is an example of a simple tabulation.

(2) Double Tabulation or Two-way Tabulation –

When the data are tabulated according to two dharacteristics at a time, it is said to be a bouble tabulation or two-way tabulation. Tabulation of data on the population of the world classified by two characteristics like religion and sex is an example of a double tabulation.

(3) Complex Tabulation –

When the data are tabulated according to many characteristics, it is said to be a complex tabulation. Tabulation of data on the population of the world classified by three or more characteristics like religion, sex and literacy, etc. is an example of a complex tabulation.

The formation of tables is described below –

Simple or one way table

Number of employees in a factory

Age in Years	No.of Employees
<25	32
25 - 35	28
35 - 45	23
45 - 55	19
55 - 65	17
> 65	12

Complex or two way Table

Number of Employees in Factory

Age (yrs.)	EMPLOYEES		Total
	Male	Female	
<25			
20 - 30			
30 - 35			
35 - 40			
40 - 45			
> 45			

1.7 Summary

This is can be summarised that data collection is a continuous process and it enables one to answer relevant questions and evaluate outcomes.

Unit-2 ❑ Theoretical distribution: frequency cumulative frequency, normal. Sampling: need, types and significance and methods of random sampling

Structure

2.1 Objectives

- The learners will learn about the frequency distribution, and significance of sampling.

2.1 Introduction

In the real world, we rarely come across experiments with single outcomes like heads or tails. Mostly a set of events are present and we may carry out the same experiment for n number(s) of time. As a result, we get a collection of outcomes which we can represent in the form of theoretical (or probabillity) distribution. We can further categorize it into **continuous** or **discrete** distribution. By theoretical distribution we mean a frequency distribution, which is obtained in relation to a random variable by some methematical model. The examples of such a distribution are : (i) Binomial distribution, (ii) Poission distribution, (iii) Normal distribution or Expected Frequency distributions of a random variable, which are built up one these distributions, a random exponent is theoretically considered. For these distribution, a random exponent is theoretically assumed to serve as a model, and the probabilities are given by a function of the random variable, called probability function.

All probability distributions can be classified as discrete probability distributions

or as continuous probability distributions, depending on whether they define probabilities associated with discrete variables or continuous variables.

2.3 Discrete vs. Continuous Variables

If a variable can take on any value between two specified values, it is called a **continuous variable;** otherwise, it is called a **discrete variable.** Some examples will clarify the difference between discrete and continuous variables.

* Suppose the fire department mendates that all fire fighters must weigh between 150 and 250 pounds. The weight of a fire fighter would be an example of a continuous variable; since a fire fighter's weight could take on any value between 150 and 250 pounds.
* Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.

Discrete Probability Distributions

If a random variable is a discrete variable, its probability distribution is called as **discrete probability distribution.** Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the random variable X represent the number of Heads that result from this experiment. The random variable X can only take on the values 0, 1, or 2, so it is a discrete random variable. The probability distribution for this statistical experiment appears below.

Table - 2.1

Number of heads	Probability
0	0.25
1	0.50
2	0.25

The above table represent a *discrete* probability distribution because it relates each value of a discrete random variable with its probability distributions.

Continuous Probability Distribution –

If a random variable is a continuous variable, its probability distibution is called a **continuous probability distribution.** A continuous probability distribution differs

from a discrete probability distribution in several ways.

- The probability that a continuous random variable will assume a particular value is zero.
- As a result, a continuous probability distribution cannot be expressed in tabular from.
- Instead, an equation or formula is used to describe a continuous probability distribution.

Most often, the equation used to describe a continuous probability distribution is called a **probability density function.** Sometimes, it is referred to as a **density function**, a **PDF**, or a **pdf.** For a continuous probability distribution, the density function has the following properties :

- Since the continous random variable is defined over a continuous range of values (called the **domain** of the variable), the graph of the density function will also be continuous over that range.
- The area bounded by the curve of the density function and the x-axis is equal to 1, when computed over domain of the variable.
- The probability that a random variable assumes a value between a and b is equal to the area under the density function bounded by a and b.

2.4 Frequency Distribution

Frequency distribution is a table that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency of count of the occurrences of values within a particular group of interval, and in this way, the tables summarizes the distribution of values in the sample. Constructing a frequency distribution table of a survey was taken along a road. In each of 20 homes, people were asked how many cars were registered to their households. The results were recorded as follows:

1, 2, 1, 0, 3, 4, 0, 1, 1, 1, 2, 2, 3, 2, 1, 4, 0, 0.

Steps to be followed for present this date in a frequency distribution table.

1. Divide the results (x) into intervals, and then count the number of results in each interval. In this case, the intervals would be the number of households with no car (0), one car (1), two cars (2) and so forth.
2. Make a table with sparate columns for the interval numbers (the number of cars per household), the talling results, and the frequency of results in each interval. Label these columns Number of cars, Tally and Frequency.

3. Read the list of data from left to right and place a tally mark in the appropriate row. For example, the first result is a 1, so place a tally mark in the row baside where 1 appars in the interval column (Number of cars). The next result is a 2, so place a tally mark in the row beside the 2, and so on. When you reach your fifth tally mark, draw a tally line through the preceding four marks to make your final frequency calculations easier to read.
4. Add up the number of tally marks in each row and record them in the final column entitled Frequency.

Your frequency distribution table for this exercise should look like this :

Table-2.2

Frequency table for the number of cars registered in each household		
Number of cars (x)	Tally	Frequency (f)
0	𝍸	4
1	𝍸 |	6
2	𝍸	5
3	|||	3
4	||	3

Table-2.4

Age (years)	Frequency	Cumulative Frequency
10	3	3
11	18	3 + 18 = 21
12	13	21 +13 = 34
13	12	34 + 12 = 46
14	7	46 + 7 = 53
15	27	53 +27 = 80

Table-2.3

Age (years)	Frequency
10	③
11	⑱
12	⑬
14	⑦
15	㉗

Cumulative Frequency

Cumulative frequency is defined as a running total of frequencies. The frequency of an element in a set refers to how many of that element there are in the set. Cululative frequency can also defined as the sum of all previous frequencies up to the current point. The set of data shows the ages of participants in a certain winter camp. Draw a cumulative frequency table for the data. The cumulative frequency at a certain

point is found by adding the frequency at the present point to the cumulative frequency of the previous point. The cumulative frequency for the first data point is the same as its frequency since there is no cumulative frequency before it. So, a classification of score showing different values of a variate and the corresponding frequency is called frequency distribution. In this regard there are two types of frequency distribution - i) Simple and ii) Grouped frequency distribution.

Simple frequency Distribution

Salary (Rs.)	No of Works
2000	2
5000	4
6500	5
7500	4
8000	3
8500	2
9000	1
10000	1

Grouped Frequency Distribution

Salary (Rs.)	No. of Workers
2000 - 4000	5
4001 - 6000	11
6001 - 8000	15
8001 - 10000	16
10001 - 12000	12

But, when we collect raw data, we should tabulate and arrange the data systematically. When the scores are few and small numbers, we can make a simple frequency distribution and when we have large numbers of score we have to arrange the data in a grouped frequency distributin table by the following items -

i) Class; ii) Class Boundary; iii) Class Interval (Width); iv) Class Frequecy, Total Frequency; v) Mid value or Class Representative; vi) Frequency Density; vii) Frequency Percentage; viii) Smoothed Frequecy.

The most important method of organising and summarising statistical data is by constructing a frequency distribution table. In this method, classification is done by quantitative magnitude. The following steps are to be followed for this purpose.

(A) Raw Data Table - Annual Production of Rice in "00 Kgs.

46	67	23	05	12	36	63	26	48	76	56	31	58
90	32	36	59	54	48	21	58	84	68	65	59	46
53	64	57	65	53	38	58	26	43	45	66	74	16
86	43	36	66	46	58	36	64	58	45	76	74	48
64	58	50	58	95	56	66	44					

(B) Arranged Scores (Data) - Frequency Distribution Table

Class	*Tally*	*Frequency(F)*	*Cumulative Frequency (fc)- Less Than Type*	*Cumulative Frequency(fc)- More Than Type*
01 - 10	/	*01*	*< 10 = 01*	*> 01 = 60 = >01*
11 - 20	//	*02*	*< 20 =01+02=03*	*60 - 01 = 59 = >10*
21 - 30	///	*04*	*< 30 =03+04=07*	*59 - 02 = 57 = >20*
31 - 40	𝍸 //	*07*	*< 40 =07+07=14*	*57 - 04 = 53 = >30*
41 - 50	𝍸 𝍸 //	*12*	*< 50 =14+12=26*	*53 - 07 = 46 = >40*
51 - 60	//// //// ////	*15*	*< 60 =26+15=41*	*46 - 12 = 34 = >50*
61 - 70	𝍸 𝍸 /	*11*	*< 70 =41+11=52*	*34 - 15 = 19 = >60*
71 - 80	////	*04*	*< 80 =52+04=56*	*19 - 11 = 08 = >70*
81 - 90	///	*03*	*< 90 =56+03=59*	*08 - 04 = 04 = >80*
91 - 100	/	*01*	*< 100 =59+01=60*	*04 - 03 = 01 = >90*

In relation to above two tables - (A) and (B) we should consider some concepts regarding frequency distribution.

i) Class or Class interval is classified groups of observations which represent all individual observation within class.

ii) The number of observations simply called frequency in a particular class is known as class frequency. ***The sum of all class frequencies is the total frequency which is the total number of observation (N).***

iii) ***The two ends of a class are called Class Limits (Upper and Lower Class Limit).***

iv) The class boundaries are usually calculated from the class limits by – ***LCB = LCL - 1/2d and UCB = UCL + 1/2D;***

v) The value exactly at the middle of a class interval is known as its mid value. In 01 -10 class, from table, ***mid value = (01 +10)/02 = 5.5***

vi) The width of a class interval is the difference between the class boundaries – ***w = UCB - LCB, in our case it will be 10.5 - 0.5 = 10.***

vii) Frequency Density is the ratio of the class frequency to the width of that class interval.

Frequency Density = Class Frequency / Width
From the given calculated table we have – 5/10 = 0.5

viii) The persentage frequency is the ratio of class frequency to Total frequency expressed as persentage – **% Frequency = (Class frequency / Total Frequency) x 100**

That also Known as – Relative Frequency x 100; Relative frequency is also known as Probability Frequency. From the following table all those expressed above are shown.

Preparation of Frequency Table with its proper explanation.

Class / class Interval (i)	Class Frequency (f)	Class Limit		Class Boundary		Mid value (X)	Width (w)	Frequency Density (w)	% Frequency
		Lower Class (Lc)	Upper Class (Uc)	Lower Boundary (Lb)	Upper Boundary (Ub)				
1 - 10	05	01	10	0.5	10.5	5.5	10	0.5	8.33
11 - 20	11	11	20	10.5	20.5	15.5	10	1.1	18.33
21 - 30	15	21	30	20.5	30.5	25.5	10	1.5	25.00
31 - 40	16	31	40	30.5	40.5	35.5	10	1.6	26.67
41 - 50	13	41	50	40.5	50.5	45.5	10	1.3	21.67
	N=60								100%

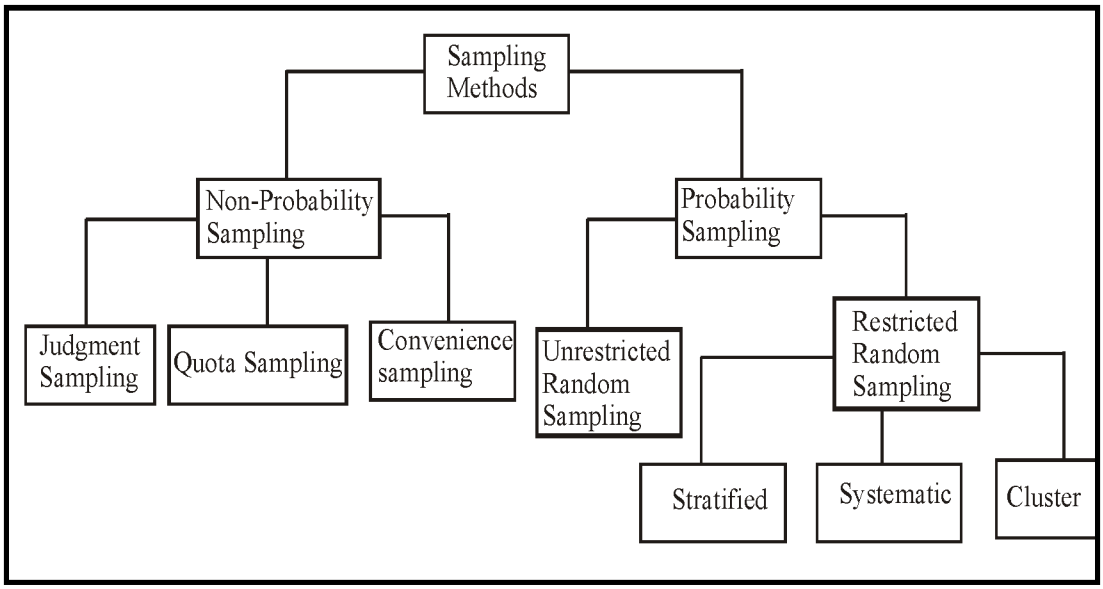
2.5 Sampling : Need, types and significance

The entire vast group of all animate or inanimate individuals, objects, cases or events that posses some from or amount of the specific variable being investigated in a particular experiment, test or survey constitutes the population. It may be infinite and a very large number. It is too laborious, expensive and impracticable also to text for study. For this reason, it is necessary to draw some representative scientifically from those vast group, called sampling.

A sample is difined as a smaller set of data that is chosen and/or selected from a larger population by using a predefined selection method. These elements are known as sample points, sampling units or observation. Creating a sample is an efficient method of conducting research as in most cases, it is impossible or very expensive and

time consuming to research the whole population and hence researching the sample provides insights that can be applied to the whole population.

Sample planning refers to detailed outline of measurements to be taken :



- **At what time** – Decide the time when a survey is to be conducted. For example, taking people views on newspaper outereach before launch of a new newspaper in the area.
- **On Which material** – Decide the material on which the survey is to be conducted. It could be a online poll or paper based checklist.
- **In what manner** – decide the sampling methods which will be used to choose people on whom the survey is to be conducted.
- **By whom** – Decide the person(s) who has to collect the observation.

Sampling plans should be prepared in such a way that the result correctly represent the representative sample of interest and allown all questions to be answered.

Steps For Sampling

Following are the steps involved in sample planning.

- **Identification of parameters** – Identify the attributes/parameters to be measured. Identify the ranges, possible values and required resolution.
- **Choose Sampling Method** – Choose a sampling method with details like how and when samples are to be identified.

- **Select Sample Size**– Select an appropriate sample size to represent the population correctly. Large samples are generally not suitable for proper conclusion.
- **Select storage formats** – Choose a data storage format in which the sampled data is to be kept.
- **Assign Roles and Responsibilities** – Assign role and responsibilities to each person involved in collecting, processing, statistically testing steps.
- **Varify and execute** – Sampling plan should be verifiable. Once verified, pass it to related parties to execute it.

Need For Sampling

When working in the field to collected **samples,** there are different ways to make sure the data is as unbiased and representative of an area as possible. **Sample** collecting is important because it provides the most accurate data without having to spend several years and tons of resources on a single projet. **Sampling** is done in research to be able to produce accurate result. It is impractical and undesirable to study the whole population **and that's why sampling is done.** If the sample is too samll or excessively large, it may lead to incorrect findings. Sampling techniques may be used to find **representative samples to avoid bias.** In practice, the sample size that is selected for a study can have a significant impact on the *quality* of you results/ findings, with sample sizes that are either *too small or excessively large* both potentially leading to incorrect findings. As a result, sample size calculations are sometimes preformed to determine how large your sample size needs to be to avoid such problems.

Sampling Methods

Sampling methods are the ways to choose people from the population to be considered in a sample survey. Samples can be divided based on following criteria.

- **Probability Sampling**– In such samples, each population element has a known probability or chance of being chosen for the sample.
- **Non-probability Sampling** – In such samples, one cannot be assured of having known probability of each population element.

Probability sampling methods –

Probability sampling methods ensures that the sample chosen represent the population correctly and the survey conducted will be statistically valid. Following are the types of probability sampling methods :

- **Simple random sampling.** – This method refers to a method having following properties :

- The population have N objects.
- The sample have n objects.
- All possible samples of n object have equal probability of occurrence.

One example of simple random sampling is lottery method. Assign each population element a unique number and place the numbers in below. Mix the numbers thoroughly. a blind-folded researcher is to select n numbers. Include those population element in the sample whose number has been selected.

- **Stratified sampling** – In this type of sampling method, population is divided into groups called stratabased on certain common characteristic like geography. Then samples are selected from each group a simple random sampling method and then survey is conducted on people of those samples.
- **Cluster sampling** – In this type of sampling method, each population member is assigned to a unique group called cluster. A sample cluster is selected using simple random sampling method and then survey is conducted on people of that sample cluster.
- **Multistage sampling** – In such case, combination of different sampling methods at diffrent stages. For example, at first stage, cluster sampling can be used to choose clusters from population and then simple random sampling can be used to choose elements from each cluster for the final sample.
- **Systematic random sampling** – In this type of sampling method, a list every member of population is created and then first sample element is randomly selected from first k elements. There after, every kth element is selected from the list.

Non-probability sampling methods

Non-probability sampling methods are convenient and cost-savvy. But they do not allow to estimate the extent to which sample statistics are likely to vary from population parameters. Whereas probability sampling methods allow that kind of analysis.

Following are the types of non-probability sampling methods :

- **Voluntary sample** – In such sampling methods, interested people are asked to get involved in a voluntary survey. A good example of voluntary sample in on-line poll of a news show where conducts survey.
- **Convenience sample** – In such sampling methods, surveyor picks people who are easily available to give their inputs. For example, a surveyor chooses a

cinema hall to survey movie viewers. If the cinema hall was selected on the basis that it was easier to reach then it is a convenience sampling method.

Simple Random Sampling—Methods

A simple random sample is defined as one in which each element of the population has an equal and independent chance of being selected. In case of a population with N units, the probability of choosing n sample units, with all possible combinations of N_{cn} samples is given by $1/N_{cn}$ e.g. If we have a population of five elements (A, B, C, D, E) i.e. N 5, and we want a sample of size n = 3, then there are $5_{c3} = 10$ possible samples and the probability of any single unit beign a member of the sample is given by 1/10. Simple random sampling can be done in two different ways i.e. 'with replacement' or 'without replacement.' When the units are selected into a sample sucessively after replacing the selected unit before the next draw, it is a simple random sample with replacement. If the units selected are not replaced before the next draw and drawing of successive units are made only from the remaining units of the population, then it is termed as simple random sample withour replacement. Thus in the former method a unit once selected may be repeated, whereas in the latter a unit once selected is not repeated, due to more statistical efficiency associated with a simple random sample without replacement it is the preferred method. A simple random sample can be drawn through either of the two procedures i.e. through lottery method or through random number tables.

- **Lottery Method** – Under this method units are selected on the basis of random draws. Firstly each member or element of the population is assigned a unique number. In the next step these numbers are written on separate cards which are physically similar in shape, size, colour etc. Then they are placed in a basket and thoroughly mixed. In the last step the slips are taken out randomly without looking at them. The number of slips drawn is equal to the sample size required. Lottery method suffers from few drawbacks. The process of writing N number of slip is cumbersome and shuffling a large number of slips, where population size is very large, is difficult. Also human bias may enter while choosing the slips. Hence the other alternative i.e. random numbers can be used.
- **Random Number Tables Method** – These consist of columns of numbers which have been randomly prepared. Number of random tables are available

e.g. Fisher and Yates Tables, Tippets random number etc. Listed below is a sequence of two digital random numbers from Fisher & Yates table :

61, 44, 65, 22, 01, 67, 76, 23, 57, 58, 54, 11, 33, 86, 07, 26, 75, 76, 64, 22, 29, 35, 74, 49, 86, 58, 69, 52, 27, 34, 91, 25, 34, 67, 76, 73, 27, 16, 53, 18, 19, 69, 32, 52, 38, 64, 81, 79, and 38.

The First step involves : assigning a unique number to each member of the population e.g. if the population comprises of 20 people then all individuals are numbered from 01 to 20. If we are to collect a sample of 5 units then referring to the random number table 5 double digit numbers are chosen. E.g. using the above table the units having the following five numbers will from a sample : 01, 11, 07, 19 and 16. If the sampling is without replacement and a particular random number repeats itself then it will not be taken again and the next number that fits our criteria will be chosen.

Thus a simple random sample can be drawn using either of the two procedures. However in practice, it has been seen that simple random sample involves lots of time and effort and is impractical.

2.6 Summary

- Statisticians attempt to collect samples that are representations of the populations in question.
- The methology used to sample from a larger population depends on the type of analysis being performed.

Unit-3 ❑ Central Tendency-Mean, median, mode, partition values

Structure

3.1 Objective

- The learners will learn about the different forms of central tendency.

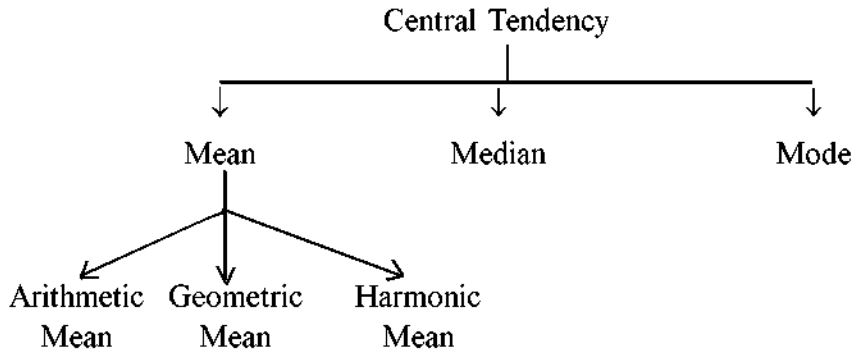
3.2 Introduction

The collected data which we draw from the field are not suitable to draw conclusion about the mass from which it has been taken. Some inferences about the population can be drawn from the frequency distribution of the observed values. Generally, a distribution is categoried by two parameters viz, the location parameter that is central value and the Scale parameter (measures of dispersion). Hence in finding a central value, the data are condensed into a single value around which the largest member of values tend to cluster. Commonly, such a value lies in the centre of the distribution and is termed as central tendency. The objective of an average is to represent a number of variates in a simple and concise manner. So it is a representative figure of the entire data.

So, any arithmetical measure which is intended to represent the centre or central value of a set of observations is known as a measure of central tendency or measure of location. These definition make it clear that the average is a single value in the distribution around which other values congregate, thus gives an 'average' idea of the distribution. "A measure of central tendency is a typical value around which other figures congregate, or which divides their member in half"— Simpson and Kafka (1969)

3.3 Types of Central Tendency

There are various measures of central tendency. The major types are—



The Arithmatic Mean ($\bar{X}$) of the values of a variate x_1, x_2, x_3 .....xn is the sum of the value divided by their number. So,

$$\overline{X} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n} = \left(\frac{\sum x}{n} \right)$$

This A.M. is known as Simple Arithmetic Mean

To find the A.M. of the variate 2, 5, 9, 11, 8, 13, 15 where n = 7

Now, $AM(\overline{X}) = \left[\frac{2+5+9+11+8+13+15}{7} \right]$

A.M. = 9

Another type is weighted Arithmetic mlean or Simply Known as weighted Mean. It the n values of a variate x_1, x_2, x_3, ... x_n are taken f_1, f_2, f_3, ... f_n times, respectively

the weighted Mean $= \frac{f_1x_1 + f_2x_2 + f_3x_3 + ... + f_nx_n}{f_1 + f_2 + f_3 + ... + f_n}$

$$\overline{X}_\omega = \left(\frac{\sum fx}{\sum f} \right)$$

There are various measures of central Tendency. Mean, median, and mode are three kinds of "averages". There are many "averages" in statistics, but, there are three main measuers of **central tendency :** the mode, the median and the mean. Each of these measures describes a **different** indication of the typical or **central** value in the distribution. The mode is the most commonly occurring value in a distribution.

The Mean

The mean is the most common measure of central tendency used by researchers and people in all kinds of professions. It is the measure of central tendency that is also referred to as the average. A researcher can use the mean to describe the data distribution of variables measured as intervals or ratios. A mean is very easy to calculate. One simply has to add all the data values or "scores" and then divide this sum by the total number of scores in the distribution of data. For example, if five families have 0, 2, 2, 3, and 5 children respectively, the mean number of children is (0 + 2 + 2 + 3 + 5)/5 = 12/5 = 2.4. This means that the five households have an average of 2.4 children. The "mean" is the "average" you're used to where you add up all the numbers and then divide by the number of numbers. The "median" is the "middle" value in the list of numbers. To find the median, your numbers have to be listed in numerical order from smallest to largest, so you may have to rewrite your list before you can find the median. The "mode" is the value that occurs most often. If no number in the list is repeated, then there is no mode for the list.

Types of measures of central tendency :

There are five types, namely

1. Arthmetic Mean (A.M.)

2. Median

3. Mode

4. Geometric Mean (G.M)

5. Harmonic Mean (H.M)

WHEN TO USE MEAN, MEDIAN AND MODE

Please use the following summary table to know what the best measure

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Example :

Table :

Variate (x) Income (Rs)	**Frequency (f) (No of workers)**	fx
2	2	4
5	1	5
9	4	36
11	2	22
3	1	13
	10	80

$\therefore$ Weighted mean $\left(\overline{X}_w\right) = \frac{\sum fx}{\sum f} = \frac{80}{10} = 8.00$

Mean from Discrete series – (Short Method)

In this Process $\overline{X}$ is calculated from the following Principles—

$$\overline{X} = A + \frac{\sum fd}{N}$$

A = Assumed mean

N = Total observation

d = Deviation from asumed mean

Example :

Wage (00 Rs.) (X)	**No of Labour (f)**	$d = X - A$	$f \times d$	**Conditions**
49	2	– 13	– 26	N = 25 = (f)
60	8	– 2	– 16	A = 62
62	7	0	0	$\sum fd$ = 47
70	4	+8	+32	$\overline{X} = A + \frac{\sum fd}{N}$
75	3	+13	+39	$= 62 + \frac{47}{25}$
80	1	+18	+18	= 62 + 1·88 = 63·88

So, the average wage of the labour Rs. 63.88.

Mean from continuous series

Mean may be calculated from continuous series by two methods—

(i) Direct Method

Weight (Kg)	No of (f) Students	Mid Point (X)	(fX)	Results
12 – 17	4	14·5	58·0	N = 65
17 – 22	19	19·5	370·5	ΣfX = 16175
22 – 27	21	24·5	514·5	
27 – 32	12	29·5	354·0	$\overline{X} = \frac{\sum fX}{N}$
32 – 37	7	34·5	241·5	= 1617·5/65
37 – 42	2	39·5	79·0	= 24·88

$\therefore$ Arimatic Mean = 24·88

(ii) Shortcut Method

Class	**(X) Mid point**	f	$d = \frac{X - A}{i}$	fd	**Results**
100 – 150	125	20	– 3	– 60	A = Assumed mean
150 – 200	175	30	– 2	– 60	A = 275
200 – 250	225	45	– 1	– 45	i = 50
250 – 300	275(A)	60	0	0	i = Interval
300 – 350	325	40	+ 1	+ 40	
350 – 400	375	25	+ 2	+ 50	
400 – 450	425	10	+ 3	+ 30	

Now according to Principle of shortcut method—

$$\overline{X} = A + \frac{\sum fd}{N} \times i$$

or, $\overline{X} = 275 + \frac{-45}{230} \times 50$

= 275 + (– 9·78(= 275 – 9·78

So, $\overline{X}$ = 265·22

Advantages and Disadvantages of Arithmetic Mean :

Advantages :

(a) It is easy to calculate and simple to understand.

(b) For Counting $\overline{X}$, all the date are utilised.

(c) It is Capable of further mathematical treatment.

(d) It povides a good basis of compare with two or more frequency distribution.

(e) $\overline{X}$ does not necessitate the arrangement of data.

Disadvantages :

(a) It may give cosiderable weight to extreme items.

(b) In some cases arithmetic mean may give misleading impressions.

(c) It can hardly b located by inspection.

Geometric and Harmonic Mean (G.M. and H.M.)

Apart from A. M. Geometric and Harmonic mean have also significant role in the geographical analysis.

The Geometric Mean (GM) of the n Positive values of a variate x_1, x_2, x_3 ..... x_n is the n root of the product of the values, i.e.

$GM = \sqrt[n]{x_1 . x_2 . x_3 ..... x_n}$, It means

$G = (x_1 . x_2 ..... x_n)^{\frac{1}{2}}$. Now taking log on both Sides we find,

$$\log G = \frac{1}{n}\log(x_1 \times x_2 \times x_3 ..... x_n)$$

$$= \frac{1}{n}(\log x_1 + \log x_2 + \log x_3 ..... + \log x_n)$$

$$= \frac{1}{n}\sum \log x \quad .....(i)$$

So G = Antilog $\left[\frac{1}{n}\sum \log x\right]$

Properties of G. M.

1. The Product of n values of a variate is equal to the n-th power of their G. M.
2. The log of G. M. of n observations is equal to the A. M. of log of n observations.

3. The Product of the ratios of each of the n observations to the G. M. is always unity.

Example : Simple series.

(a) Find the G. M. of 111, 171, 191, 212

Now, G.M. $= \sqrt[n]{x_1 \times x_2 \times x_3 \ldots . x_n}$ $\quad n = 4$

$$= \sqrt[4]{111 \times 171 \times 191 \times 212}$$

$$\text{G.M.} = \frac{1}{4}(\log 111 + \log 171 + \log 191 + \log 212)$$

$$= \frac{1}{4}(2{\cdot}0453 + 2{\cdot}2330 + 2{\cdot}2810 + 2{\cdot}3263)$$

$$= \frac{1}{4}(8{\cdot}8856) = 2{\cdot}2214$$

$\therefore$ G.M. = Antilog [2·2214] = 166·5

Geometric Mean = 166·5

(2) Discrete Series : Geometric Mean

x	f	log x	f log x	**Results**
8	3	0·9030	2·7090	N = 48
10	7	1·0000	7·0000	Σf log x = 53·9681
12	10	1·0791	10·7910	GM =
14	12	1·1461	13·7532	
16	9	·2041	10·8369	
18	5	1·2552	6·2760	
20	2	1·3010	2·6020	

So, Geometric mean = A. L (1·1243)

GM = 13·31

(3) **Geometric mean fo Continuous Series**

Class	f	(x) Mid Point	log x	f log x	
10 – 20	10	15	1·1760	11·7600	
20 – 30	12	25	1·3979	16·7748	N = 70
30 – 40	20	35	1·5440	30·8800	$\Sigma f \log x$ =
40 – 50	11	45	1·6532	18·1852	107·6933
50 – 60	10	55	1·7403	17·4030	
60 – 70	7	65	1·8129	12·6903	

$$GM = A.L.\frac{\Sigma(f \log x)}{n}$$

$$= A.L.\frac{107{\cdot}6933}{70}$$

$$= AL\ (1{\cdot}5384)$$

$\boxed{G.M. = 34{\cdot}54}$

Advantages and Disadvantages of G. M.

1. It is not influenced by the extreme items to the same extent as mean.
2. If is rigidly defined and its value is a precise figure.
3. It is also based on all observations and capable of further algebraic treatment.
4. It is useful in Calculating Index number

Disadvantages :

1. It is not Simple to understand, neither easy to calculate.
2. If any value of a Set of abservation is 'O' it cannot be determined.
3. Again, if any value becomes negative, geometric mean become imaginary.

It is used to find average of the rates of changes.

Harmionic mean (H.M.)

The Harmonic mean (H.M.) for n observations x_1, x_2, x_3 ..... x_n is the total number divided by the sum of the reciprocals of the numbers.

i.e., $H = \dfrac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots\ldots\ldots + \frac{1}{x_n}} = \dfrac{n}{\sum \frac{1}{x}}$

Again, $\dfrac{1}{H} = \dfrac{\sum \frac{1}{n}}{n}$ (Reciprocal of H.M. = A.M. of reciprocals of the numbers)

Example :

Calculate H. M. four the following

4, 6, 7, 10, 11

x	$1/x$
20	0·050
30	0·033
40	0·025
50	0·020
60	0·0171
$x = 5$	0·145

$$HM = \frac{N}{\sum \frac{1}{x}} \qquad N = 5$$

$$\sum \frac{1}{x} = 0 \cdot 145$$

$$= \frac{5}{0 \cdot 145} = 34 \cdot 48$$

H.M. = 34·48

For Continuous Series – HM

Class	f	x	f/x
o – 10	2	5	0·400
10 – 20	4	15	0·267
2o – 30	6	25	0·240
3o – 40	5	53	0·143
4o – 50	3	45	0·067

$N = 20, \qquad \Sigma\left(f/x\right) = 1 \cdot 117$

$$HM. = \frac{N}{\Sigma \frac{f}{x}} = \frac{20}{1 \cdot 117}$$

So, **H.M. = 17.905**

Advantages and Disadvantages (H.M.)

1. Like AM and G.M. it is also based on all observations
2. Capable of further algebraic treatment.
3. It is significantly useful while averaging certain types of rates and ratios.

Disadvantages :

1. It is not simple to understand nor can it be calculated with ease.
2. It is usualy a value which may not be a member of the given Set of numbers.
3. It cannot be calculated when there are both negative and positive values.

Median

It is also a significant measure of central tendency. If a set of observations are arranged in order of magnitude (ascending or descending), then the middle most or central value gives the median. So it is called positional average.

Median divides the observations in two equal halves, in such a way that the number of observations Smaller than median is equal to the number greater than it. It is not thus affected by extremely large or small observations. In certain sense, it is the real measure of central tendency.

Computation of Median

For simple or discrete series for n number of observation median can be calculated by $\frac{n+1}{2}$ and $N/2$ for odd and even n respectively after arranging the data in ascending or descending order in order of magnitude.

First we observe the series whether it is odd or even, then apply the Principle accordingly.

(A) Simple series when N is odd.

5, 10, 7, 4, 6, 12, 11.

Ascending order —	4,	5,	6,	7,	10,	11,	12
No. of order —	1	2	3	4	5	6	7

Here $N = 7$

Number of observations is odd

So, $Me = \frac{N+1}{2} = \frac{7+1}{2} = 4$

SO THE VALUE WHICH IS IN 4TH POSITION WILL BE MEDIAN.

The value of 4th Position is 7; So

Me = 7(4th Position value).

When it is even number.

5, 10, 7, 4, 6, 12, 11, 14 — observations

Number of arrangement —	1	2	3	4	5	6	7	8
Vale in ascending order —	4	5	6	7	10	11	12	14
					↑ *Me*			

Here for the even number of values when N = 8, we will use the following Principles

Me = Average value of $n/2$ th item and $\frac{N+1}{2}$ th item.

$$Me = \frac{\frac{N}{2}\text{th} + \frac{N+1}{2}\text{th value}}{2}$$

$$= \frac{\frac{8}{2}+\frac{8}{2}+1}{2} = \frac{4+5}{2} = 4 \cdot 5 \text{th}$$

$$= \frac{7+10}{2} = 8 \cdot 5 \text{ (Mediam)}$$

Me = 8·5

(B) **For discrete Series** (Simple frequency distribution)

x	f	cf
1	7	7
2	12	19
3	17	36
4	19	55
5	21	76
6	24	100
		N = 100

Now, Me =

Value of $= \frac{N+1}{2}$ th item

$= \frac{100+1}{2}$ th item

= 50·5 th items.

It has been observed that the 50·5 is greater than the cumulative frequency 36, but less than the next cf 55 corresponding to x = 4; So all the 19 items (from 37 to 55) have the Same variate 4. And 50·5th item is also one of these 19 item.

(C) **For Continuous Series.**

In this case we are to determine the particular class in which the value of the median curve by using the principles $N/2$ $\left(\text{not by } \frac{N+1}{2}\right)$ as in continuous series $N/2$ divides the area of the curve into two equal halves. After locating median, its magnitude is measures by applying following formula.

$$\left[L_1 + \frac{N/2 - f_c}{fm_c} \times i\right]$$

L_1 = Lower limit of the mediam class

fm = Frequency of the mediam class

f_c = Cumulative frequency just above the mediam class.

N = Total number of observation

Example :

Class	f	f_c	Remarks
0 – 10	5	5_+	Using class limit
10 – 20	7	12_+	
20 – 30	9	21 → fc	
30 – 40	2	33 fm = 12	
40 – 50	10	43	
50 – 60	5	48	
60 – 70	6	54	

Now — N = 54

$$N/2 = 27\text{th value}$$

- Median class (fm) – 30 – 40 and its frequency is 33
- f_c – cumulative frequency (up to) just above the mediam class (30 – 40) = 21
- i = Class interval = 10
- L_1 = Lower Limit = 30

$$Me = 30 + \frac{\frac{54}{2} - 21}{12} \times 10$$

$$= 30 + \frac{6}{12} \times 10 \quad = 30 + 5$$

So, Mediam = 35

Another alternative method : Using class Boundary

Class	Class/Boundary		f	cf	
	Lower	*upper*			
15 – 19	14·5(L)	19·5(U)	5	5	i = 5
20 – 24	19·5	24·5	6	11	N = 59
25 – 29	24·5	29·5	10	21-fc	
30 – 34	29·5	34·5	15-fm	36	$Me = \frac{N}{2}$ th value
35 – 39	34·5	39·5	9	45	So, 59/2 = 29·5th
40 – 44	39·5	44·5	8	53	*value*
45 – 49	44·5	49·5	4	57	fc = 21
50 – 54	49·5	54·5	2	59	fm = 15

Here from the above table we observe that 29·5th value is lying 30–34 class and for this class, class boundary is 29·5 – 34·5 Lower class Boundary 29·5 and upper class boundary is 34·5.

So, According to Principle Median will be—

$$Me = L_1 + \frac{\frac{N}{2} - f_c}{f_m} \times i$$

$$= 29 \cdot 5 + \frac{\frac{59}{2} - 21}{15} \times 5$$

$$= 29 \cdot 5 + \frac{29 \cdot 5 - 21}{15} \times 5$$

$$= 29 \cdot 5 + \frac{8 \cdot 5}{15} \times 5$$

$$= 29 \cdot 5 + 2 \cdot 83 \quad = 32 \cdot 33$$

So, Median = 32·33 (Ans)

$$AL\left[\sum \frac{f \log x}{n}\right]$$

$$= AL\left(\frac{53 \cdot 9681}{48}\right)$$

Advantages and Disadvantages of Median

(i) The median, unlike the mean, is unaffected by the extreme values of the variable.

(ii) It is easy to calculate and simple to understand, particularly in a series of individual observations and a discrete series.

(iii) It is capable of further algebraic treatment. It is used in calculating mean deviation.

(iv) Median can be calculated even if the items at the extreme are not known, but if we know the central items and the total number of items.

(v) It can also be determined graphically.

Disadvantages.

(i) For calculation, it is necessary to arrange the data, whereas other averages do not need any such arrangement.

(ii) It cannot be computed precisely when it lies between two items.

(iii) Process involved to calculate median in case of continuous series is difficult to follow.

(iv) Median is affected more by sampling fluctuations than the mean.

Mode

Mode is the value of the variate which occurs most frequently. It simply represents the most frequent value of a series.

Mode cannot be calculated from a series of individual observations unless it is converted to a discrete series (or Continuous series). In a discrete series the value of the variate having the maximum frequency is the mode.

Calculation – Example :

(A) **For Individual observations**

In this case the Individual observations are to be first converted to discrete series. Then the variate having the maximum will be the mode.

(1) Calculate mode from —

10, 14, 24, 27, 24, 12, 11, 17

Variate	*f*
10	1
11	1
12	1
14	1
17	1
24	2
27	1

Here variate 24 occurs maximum number of items i.e., 2. Hence the model marks are 24 as Mode = 24

MODE MAY BE UNIMODEL AND MULTI MODEL.

In example (i) it is unimodel; but 14, 14, 14, 17, 18, 20, 20, 20, 24

Here 14 occurs 3, and 20 also occurs 3; So although mode here is defined but it is bi-modal type series.

Other methods of Calculating Mode.

(A) **Discrete Series.** (B)

X	f
8	3
12	5
(16)	(16)
20	12
24	8
28	4
32	2

16 has 16 frequencies. So Here mode = 16

For Continuous Series.

By observations or by preparing grouping table and Analysis table, ascertain the model class. Then to find the exact value of mode, applying the following formula:

$$Mo = L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

L_1 = Lower Limit of the model class

f_1 = Frequency of the model class

f_0 = Frequency of the class preceding the modal class

f_2 = Frequency of the class Succeeding the modal class

i = Class interval

Example :

Calculate the mode from the following

Class	f	Results
10 — 20	5	We have here—
20 — 30	8	Modal class — 40 — 50
30 — 40	12-f_0	L_1 = 40 (Lower class limit of the modal class)
40 — 50	16-f_1	f_0 = 12, f_1 = 16, f_2 = 10
50 — 60	10-f_2	i = 10, So, Mode =
60 — 70	8	$L_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$

$$\text{Mode} = 40 + \frac{16-12}{\{(2\times16)-12-10\}}\times10$$

$$= 40 + \frac{4}{10}\times10$$

So, Calculated Mode = 40 + 4 = 44

Alternative Method of Calculating Mode.

We have already Calculated Mean and Median using the respective Principles Karl Pearson, applying the following relationship tried to calculate the mode —

$$\overline{X} - M_o = 3(Me - Md)$$

Or, $M_0 = 3Me - 2\overline{X}$

Or, $M_0 = \left(3Me - 2\overline{X}\right)$

$\bar{X}$ = Mean; Me – Median, Mo = Mode

Example :

Class	f	X	$(f \times X)$	Cf	Results
5 — 10	8	7·5	60	8	
10 — 15	10	12·5	125	18	N = 63
15 — 20	15	17·5	262·5	33	ΣfX = 1227·5
20 — 25	15	22·5	337·5	48	$\overline{X} = \frac{\Sigma fX}{N}$
25 — 30	9	27·5	247·5	57	$= \frac{1227{\cdot}5}{63}$
30 — 35	6	32·5	195	63	$\overline{X} = 19{\cdot}48$

So, $\bar{X}$ = 19·48

$Me = {}^{N}\!/_{2} = {}^{63}\!/_{2}$ = 31·5th value

Mediam class = 15 – 20 (L_1 = 15)

$$Me = L_1 + \frac{N/2 - f_c}{f_m} \times i$$

$$= 15 + \frac{63/2 - 18}{15} \times 5$$

$$= 15 + \frac{31{\cdot}5 - 18}{15} \times 5$$

$$= 15 + \frac{13{\cdot}5}{15} \times 5$$

$= 15 + 4{\cdot}5 \quad = 19{\cdot}5$

Median (*Me*) = 19·5

So, we can calculate Mode (*Mo*) from above value of $\bar{X}$ and Me

$Mo = (3Me - 2\bar{X})$

$= (3 \times 19{\cdot}5) - (2 \times 19{\cdot}48)$

$= 58{\cdot}5 - 38{\cdot}96 \quad = 19{\cdot}54$

So, **Mode (*Mo*) = 19·54**

So

$\bar{X}$ = 19·48

Me = 19·50

Mo = 19·54

Advantage and Disadvantages – Mode

Advantages :

(i) It can often be located by inspection.

(ii) It is not affected by extreme values.

(iii) It is often a really typical value.

(iv) It is simple and precise. It also state an actual item of the series except in a continuous series.

(v) Mode can be determined graphically.

Disadvantages :

(i) It is unsuitable for algebraic treatment.

(ii) When the number of observation is small, the mode may not exist, while the $\bar{X}$ and Me can be calculated.

(iii) The value of Mode is not based on each and every item of series.

3.4 Partition values or fractiles

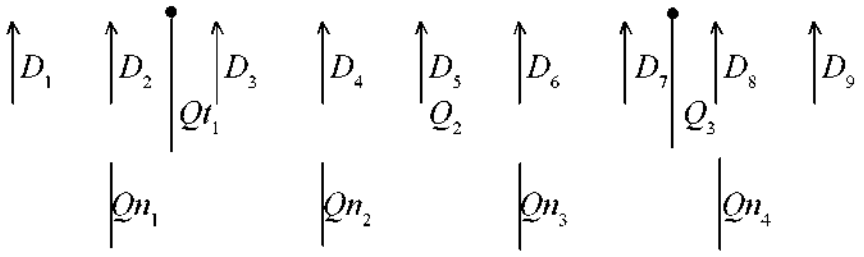
Partition values or alternatively called Fractiles are the magnitudes of those items in an array which divide the number of items thereof into some specified number of equal parts. It is a new approach to the value concept. It links together the asymptotic and the axiomatic approach. Using this approach we prove the existence of a continuous value on different spaces. It is also called quantiles. In statistics partition or fractiles or quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way. We can classify partition values and their relationship in the following way—

Partition values/Quantiles/Fractiles

Quartiles	Quintiles	Deciles	Percentiles
(25%)	(20%)	(10%)	(1%)
4 equal parts	5 equal parts	10 equal parts	100 equal parts
1st, and 3rd Quartiles, 2nd Quartile is median $(Q_1\ Q_2\ Q_3)$	$(Qn_1\ Qn_2\ Qn_3\ Qn_4)$ $Qn_5 = N$	$(D_1, D_2 \ldots\ldots D_9)$ $D_{10} = N$	$P_1, P_2, P_3 \ldots\ldots P_{99}$ $P_{100} = N$

Correlation and Association among partition values.

P_1 ----- P_{10} ----- P_{20} ----- P_{30} ----- P_{40} ----- P_{50} ----- P_{60} ----- P_{70} ---- P_{80} ---- P_{90} ---- P_{99}



$P_{100} = D_{10} = Qt_4 = Qn_5 = 100\%$ = Total observation .

From the Principles of the Median it has been observed that median divides a series in two equal parts. Now in our geographical analysis and also for further study of composition of a series it may require to divide the series into four, five, six ten, thirty five, sixty five, eighty five or hundred parts.

So following the principles of median, as we have seen that Me (median) dividing a series into two equal parts; **So for different objective *Qt* (Quartile)** dividing a series into four equal parts (25%, 50%, 75% and 100%), ***Qn* (Quintites) divides** a series into five equal parts (Qn_1, Qn_2, Qn_3, Qn_4, and Qn_5) (20%, 40%, 60%, 80% and 100%); **Decile (Dn) divides** a series into 10 equal parts (Dn_1 Dn_2 ..... Dn_{10}) for an user purpose; Pn (Percentiles) divides 100 equal parts of a series for specific purpose. Other two parts are also used for the geographical analysis and two other fractile values are imported for dividing a series which are **septiles** which divides a series into seven equal parts and **octiles, divides** a series into eight equal parts. But their use are more specific and also depending on users choice.

Following principles are used for finding the fractile values (Partition)

Quartiles (*Qt*)

For Simple and discrete series (Individual observation)

$$\left.\begin{aligned} Qt_1 &= \frac{N+1}{4}\text{th value} \\ Qt_3 &= \frac{3(N+1)}{4}\text{th value} \end{aligned}\right\} Qt_2 = \text{Median}$$

For continuous series

$Qt_1 = {}^{N}\!/_{4}$ th value and $Qt_3 = {}^{3N}\!/_{4}$ th value

So as per principles

$$\left.\begin{aligned} Qt_1 &= LQ_1 + \frac{{}^{N}\!/_{4} - fc}{fQt_1} \times i \\ Qt_3 &= LQ_3 + \frac{3\,{}^{N}\!/_{4} - fc}{fQt_3} \times i \end{aligned}\right\} Qt_2 = \text{The formula of median as exercised before}$$

***Qt_n* (Quintiles)**

$$Qtn_1 = \frac{N+1}{5}\text{th Value}$$

$$Qtn_4 = \frac{4(N+1)}{5}\text{th Value}$$

* Other values are posible to calculate following the same process.

Dn (Deciles)

$D_1 = \frac{(N+1)}{10}$ th values

$D_4 = \frac{4(N+1)}{10}$ th values

$D_9 = \frac{9(N+1)}{10}$ th values

* Other values are also posible to calculate following the same processes.

***Pn* (Percentiles)**

$Pn_1 = \frac{(N+1)}{100}$ th. Value

$Pn_{15} = \frac{15(N+1)}{100}$ th. Value

$Pn_{67} = \frac{67(N+1)}{100}$ th. Value

$Pn_{99} = \frac{99(N+1)}{100}$ th. Value

* Other values of Percentites are also posible to calculate following the same principles.

Calculation of partition values.

Quartiles, Deciles and Percentites.

(A) Raw Data —

19, 27, 24, 39, 57, 44, 56, 50, 59, 67, 62, 42, 47, 60, 26, 34, 57, 51, 59, 45.

Arranged Series

19, 24, 26, 27, 34, 39, 42, 44, 45, 47, 50, 51, 56, 57, 57, 59, 59, 60, 62, 67.

After arranging the Data Series we got

N = 20 and Range (19 – 67)

Now we calculate for Individual Series.

(i) Qt_1 – (First Quartile) $= \frac{N+1}{4} = \frac{20+1}{4}$ th item

= Size of 5·25th item.

$$= \left\{\text{Size of 5th item} + \frac{1}{4}(\text{Size of 6th itam} - \text{Size of5th item})\right\}$$

$$= 34 + \frac{1}{4}(39-34) = 34 + 1{\cdot}25 = 35{\cdot}25 = Qt_1$$

(ii) Qt_3 = Size of $\frac{3(N+1)}{4}$ th item

= Size of $\frac{3(20+1)}{4}$ th item

= Size of 15·75th item

= Size of 15th item + $\frac{3}{4}$ (Size of 16th item – Size of 15th item)

= 57 + $\frac{3}{4}$ (59 – 57) + 1·50 = 58·50 = Qt_3

D_4 (Fourth Decile)

Size of $\frac{4(N+1)}{10}$ th item = $\frac{4(20+1)}{10}$ th item

= Size of 8·4th item (exactly)

= Size of 8th item + $\frac{4}{10}$ (Size of 9th item – Size of eighth item)

$$D_4 = 44 + \frac{4}{10}(45-44) = 44 + 0{\cdot}4 = 44{\cdot}4$$

P_{60} (Sixty-th Percentile)

Size of $\frac{60(N+1)}{100}$ th item = $\frac{60(20+1)}{100}$ th item

= Size of 12·6th item

= Size of the 12th item + $\frac{6}{10}$ (Size of 13th item – Size of 12th item)

$$= 51 + \frac{6}{10}(56-51) = 51 + 3 = 54{\cdot}00$$

For Discrete Series

Weight (Kg)	frequency	Cf	
40	2	2	
42	6	8	
45	8	16	
50	10	26	N = 98
51	6	32	
54	14	46	
56	12	58	
59	8	66	
60	14	80	
62	12	92	
64	6	98	

Calculation From the Previous Table

Qt_1 = Size of $\frac{N+1}{4}$ th item (N = 98 total frequency)

$= \frac{98+1}{4}$ th item (from the table)

Qt_1 = 24·75th item = 50 Kg (approx)

Qt_3 = Size of $\frac{3(N+1)}{4}$ th item

Qt_3 = 74·25th item = 60 Kg. (approx)

D_4 = Size of $\frac{4(N+1)}{10}$ th item

$= \frac{4(98+1)}{10}$ th item

D_4 = 39·6th item = 54 Kg approx

P_{60} = Size of $\frac{60(N+1)}{100}$ th item

$= \frac{60(98+1)}{100}$ th item

P_{60} = 59·4th item = 59 Kg (approx)

For Continuous Series.

Just like median, the values of quartiles, deciles and percentiles like in various class intervals and the actual values are to be calculated by applying interpolation formulas.

Example —

Calculate Qt_1, Qt_3, D_8 and P_{70} for Quartile, Decile and Percentile.

Class	*f*	ct	Results
40 — 50	3	3	N = 40
50 — 60	5	8	$N/4$ = 10th value
60 — 70	6	16	$3N/4$ = 30th value
70 — 80	9	25	$D_8 = \frac{8N}{10}$ = 32th value
80 — 90	10	35	$P_{70} = \frac{70N}{100}$ = 28th value
90 — 100	4	39	
100 — 110	1	40	

(i) $Qt_1 = L_1 + \frac{N/4 - f_c}{f_{Qt_1}} \times 1$ (60 – 70 = Qt_1 class 10th value lying in this class)

$$= 60 + \frac{40/4 - 8}{8} \times 10$$

$$= 60 + \frac{10-8}{8} \times 10 = 60 + \frac{2}{8} \times 10$$

= 60 + 2·5 = **62·5** = Qt_1

(ii) $Qt_3 = L_3 + \frac{3N/4 - fc}{fQt_3} \times i$ (30th value lying in 80 – 90 class)

$$= 80 + \frac{\frac{3 \times 40}{4} - 25}{10} \times 10$$

$$= 80 + \frac{30 - 25}{10} \times 10$$

Qt_3 = 80 + 5 = **85(Qt_3)**

(iii) D_8 = $\frac{8N}{10}$ th value = 32th value

32th value lying in the 80 – 90 class.

So, $L_{D_1} + \frac{\frac{8N}{10} - fc}{f_{D_8}} \times i$

$$= 80 + \frac{\frac{8 \times 40}{10} - 25}{10} \times 10$$

$$= 80 + \frac{32 - 25}{10} \times 10$$

= 80 + 7 = **87(8th Decile).**

(iv) P_{70} = $\frac{70N}{100}$ = $\frac{70 \times 40}{100}$ = 28th value

28th value is lying in (80 – 90) class

So, $P_{70} = L_{P_{70}} + \frac{\frac{70N}{100} - fc}{f_{P_{70}}} \times i$

$$= 80 + \frac{\frac{70 \times 40}{100} - 25}{10} \times 10$$

$$= 80 + \frac{28 - 25}{10} \times 10$$

$$= 80 + \frac{3}{10} \times 10$$

P_{70} = 80 + 3 = **83 (70th Percentile)**

So, calculated values of

Qt_1 = 62·5

Qt_3 = 85·00

D_8 = 87·00

P_{70} = 83·00

Suitable and ideal measures of Average

According to the previous calculation and discussion it can be ascertained that no one average can be regarded as best or ideal in the true sense of the term. We can state some points in this regard —

(i) A. M. Should be avoided in case of skewed distributions, open end intervals, for averaging speeds and for extreme itemes.

(ii) G. M. is to applied for determining index numbers, for computing average rates of increase or decrease.

(iii) H.M. is useful for finding rates, time etc.

(iv) Median is the best average in open and grouped frequency distribution, in case of Price or Income distribution.

(v) Mode is particularly useful average for discrete series, i.e., number of persons wearing a given size of shoe or number of children per household for a very large frequency, mode is best suited.

3.5 Summary

The central tendency gives a very appropriate conclusion to the data that is being tabulated and to find out the central measure.

Unit-4 ❑ Measures of Dispersion–Mean deviation, Standard Deviation, Co-efficient of Variation

Structure

4.1 Objective

- The learners will come to know about the different measures of dispersion.

4.2 Introduction

The various measures of central tendency represent single direction to represent the entire data. But the central tendency or simply average as we have seen, has its own limitations. There are number of series whose averages may be identical but differ from each other in many ways. So only average cannot explain the data fully as we required. In such cases further statistical analysis of the data is necessary to study these differences. Measures of dispersion help us to explain the characteristics, i.e., the extent to which the items or observations differ from one another and also from central value.

A measure of dispersion is designed to state the extent to which individual observations vary from their average. Here we shall account only the amount of variation (or its degree) and not the direction. The measures of dispersion are also known as 'AVERAGES OF THE SECOND ORDER' because the deviations of the observations from their average are found out, then the average of these deviations is taken to represent the dispersion of a series.

4.3 Types of deviation

Primarily Measures of dispersion are of two types

(A) Absolute measures and (B) Relative Measures.

(A) Absolute measures are of four types —

(i) Range

(ii) Quartile deviation of semi-interquartile range

(iii) Mean deviation (Average Deviation)

(iv) Standard deviation

(B) Among the Relative measures we find the following types.

(i) Co-efficient of quartile deviation.

(ii) Co-efficient of Mean Deviation.

(iii) Co-efficient of variation. (cv)

(iv) Co-efficient of Range.

(A) Range and Co-efficient fo Range

Range is the difference between the two extreme items, i.e., it is the difference between the maximum value and minimum value in a series. If a series is represented by

15, 21, 5, 30, 40, 51, 60, 25, 35, 55

After arrangement—we get

5, 15, 21, 25, 30, 35, 40, 51, 55, 60

Here L_1 = Lower value = 5

L_2 = Highest value = 60

So, Range = $L_2 - L_1 = 60 - 5 = 55$

Range is very simple and easy to understand but it is mostly affected by extreme values and does not depend on all the observations, but only on the extreme values.

Co-efficient of range can be calculated by the following formula—

$$CR = \frac{L_2 - L_1}{L_2 + L_1} \times 100$$

or, $\left[\frac{R}{L_2 + L_1} \times 100\right]$ $\quad R = L_2 - L_1$

So from the Previous example we get—

$$CR = \frac{55}{60+5} \times 100 \quad = \frac{55}{65} \times 100$$

CR = 84·62% It's use is very limited.

Quartile Deviation

Q. D. is a significent absolute measure of dispersion. As we know that Quartile divides four equal parts in any Series. The Quartile deviation is half of the difference between the upper and lower Quartile. If Q_1, Q_2, Q_3 are three Quartiles Q_1 and Q_3 are called lower and upper Quartiles respectively. Q_2 is the median and divide the series into two equal parts. By **Inter-quartile range**, we understand the difference between two quartiles (i.e., $Q_3 - Q_1$), and half of this means—

Semi Inter-quartile range.

Since 50% of the observations lie between two quartiles, as Such Interquartile range gives a fair measure of variability. Quartile Deviations (*Q.D.*) is an absolute measure of dispersion if it is divideded by average value of two quartiles, we will find **Coefficient of Quartile Deviation.** Symbolically, Co-efficient of quartile deviations =

$$\frac{\frac{1}{2}(Q_3 - Q_1)}{\frac{1}{2}(Q_3 + Q_1)} = \frac{(Q_3 - Q_1)}{(Q_3 + Q_1)}$$

Example :

For individual observation : Simple series

Find the quartile deviation and Co-efficient of quartile deviation of the following series

X = 11, 12, 14, 17, 19, 21, 27, 28, 30, 32, 33

Here n = 11; So Q_1 and Q_1 will be

$$Q_1 = \text{Size of } \frac{n+1}{4} \text{th item} = \frac{11+1}{4} = 3\text{rd item}$$

Q_1 = 14 (First Quartile)

$$Q_3 = \text{Size of } \frac{3(n+1)}{4} \text{th item} = \text{Size of the 9th item} = \frac{3(11+1)}{4} = 9\text{th item}$$

So Q_3 = 30 (3rd Quartile)

$\therefore$ Quartile Deviation $= \frac{30-14}{2} = \frac{16}{2} = 8$

QD = 8

Again Coefficient of Quartile Deviation

$= \frac{30-14}{30+14} = \frac{16}{44} = 0 \cdot 363$

$\therefore$ **Coefficient of QD = 0·363**

and Semi Interquartile range = $(Q_3 - Q_1)$

$(Q_3 - Q_1)$ = (30 – 14) = 16

For Discrete Series — Example :

Wages (x) — 12 14 17 21 27 30 36

No of workers (f)— 4 6 8 7 12 10 4

X	f	Cf	**Results**
12	4	4	
14	6	10	N = 57, Here—
17	8	18	$Q_1 = \frac{N+1}{4}$ th item
21	7	25	$Q_3 = \frac{3(N+1)}{4}$ th item
27	12	37	cf = cumulative
30	10	47	frequency
36	4	51	

So, according to principle

$Q_1 = \frac{N+1}{4}$ th item $= \frac{51+1}{4} = \frac{52}{4} =$

13th item of the Series = **17**

Q_1 = 17

Q_3 = Size of $\frac{3(N+1)}{4} = \frac{3 \times 52}{4}$ th item

= Size 39th item = 30

Here $QD = \frac{30-17}{2} = \frac{13}{2} = 6 \cdot 5$

(C) **For continuous series :**

Production *Kg*(*X*)	**Number of plots (*f*)**	***Cf***		**Results**
5 – 10	18	18		N = 160
10 – 15	30	48	$\rightarrow Q_1$	$Q_1 = \frac{N}{4} = \frac{160}{4} = 40$
15 – 20	46	94		
20 – 25	28	122	$\rightarrow Q_1$	$Q_3 = \frac{3N}{4} = \frac{3 \times 160}{4} = 120$ th
25 – 30	20	142		
30 – 35	12	154		
35 – 40	6	160		

So, according to principle $N = 160$, $L_{Q_1} = 10$, $f_c = 18$, $f_{Q_1} = 30$, $i = 5$ } for explation see calculation for median.

$$Q_1 = LQ_1 + \frac{N/4 - f_c}{f_{Q_1}} \times i$$

$$= 10 + \frac{\frac{160}{4} - 18}{30} \times 5$$

$$= 10 + \frac{40-18}{30} \times 5$$

$$= 10 + \frac{22}{30} \times 5 \quad = 10 + 3 \cdot 67 \quad = \boxed{\mathbf{13 \cdot 67 = Q_1}}$$

$$Q_3 = L_{Q_3} + \frac{\frac{3N}{4} - f_c}{f_{Q_3}} \times i$$

$$= 20 + \frac{\frac{3 \times 160}{4} - 94}{28} \times 5$$

$$= 20 + \frac{120-94}{28} \times 5$$

$$= 20 + \frac{26}{28} \times 5 \quad = 20 + 4 \cdot 64 \quad = 24 \cdot 64$$

$N = 160$, $L_{Q_3} = 20$, $f_C = 94$, $C = 5$, $f_{Q_3} = 28$

$\boxed{Q_3 = 24{\cdot}64}$

$$QD = \frac{Q_3 - Q_1}{2}$$

$$= \frac{24{\cdot}64 - 13{\cdot}67}{2} = \frac{10{\cdot}97}{2}$$

$$= 5{\cdot}485$$

$\therefore$ QD = 5·49 (approx)

Coefficient of Quartile Deviation

$$CQD = \left[\frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100\right]$$

$$CQD = \frac{24{\cdot}64 - 13{\cdot}67}{24{\cdot}64 + 13{\cdot}67} \times 100 = \frac{10{\cdot}97}{38{\cdot}31} \times 100 = 28{\cdot}63\% \quad CQD = 28{\cdot}63\%$$

Advantages : It is Superior to range as measures of dispersion. In case of open-end distributions, it can be computed. It is not affected by the presence of extreme values.

Disadvantages : QD is neither based on all the observations nor is it capable of further algebraic treatment. Its value is much affected by sampling fluctuations.

Mean (Average Deviation) Deviation

Among the methods of absolute measures of dispersion range and QD are calculated based on only two points of a series — extreme values in case of range and quartiles for quartile deviation and are not based on all the observations. Mean deviation and Standard deviation, however, are computed by taking into account all the observations of the series.

Definition

Mean deviation (MD) of a Series is the arithmetic average of the deviations of various items from the median or mean of that series. Median is preferred since the sum of the deviations from the median is less than that from the mean. MD is also known as First Moment of dispersion.

In any series if X_1, X_2, X_3 .....X_n are values and their A.M., Me and Mo are $\bar{X}$, Me and Mo, Mean deviation will be

(i) $MD_{\bar{X}} = \frac{\sum|X - \bar{X}|}{N}$ ..... (i)

(ii) $MD_{Me} = \frac{\sum|X - Me|}{N}$ ..... (ii)

(iii) $MD_{MO} = \frac{\sum|X - M_O|}{N}$ ..... (iii)

Example :

(A) For Simple Series

Calculate Mean Deviation.

55, 34, 56, 40, 75, 81, 58, 65, 44, 72

Arranged Series

34, 40, 44, 55, 56, 58, 65, 72, 75, 81

X	$\lvert X - \bar{X}\rvert$ = $\lvert X - 58\rvert$	$\lvert X - Me\rvert$ = $\lvert X - 57\rvert$	$\lvert X - Mo\rvert$ = $\lvert X - 55\rvert$	Results
34	24	23	21	$\bar{X} = \frac{\sum X}{N} = \frac{580}{10} = 58$
40	18	27	15	
44	14	13	11	$Me = \frac{N/2 + \frac{N}{2} + 1}{2}$
55	3	2	0	
56	2	1	1	$= \frac{\frac{10}{2} + \frac{10}{2} + 1}{2}$
58	0	1	3	
65	7	8	10	$= \frac{5\text{th} + 6\text{th}}{2}$
72	4	15	17	
75	7	18	20	$= \frac{56+58}{2} = \frac{114}{2}$
81	23	24	26	**Me = 57·00**
$\sum X = 580$ N = 10	$\sum\lvert X - \bar{X}\rvert$ = 122	$\sum\lvert X - Me\rvert$ = 122	$\sum\lvert X - Mo\rvert$ = 124	N = 10

So,

(1) $M.D_{\bar{X}}(\text{Mean}) = \frac{\sum|X - \bar{X}|}{N} = \frac{122}{10} = 12{\cdot}2$

(2) $M.D_{Me}(\text{Median}) = \frac{\sum|X - Me|}{N} = \frac{122}{10} = 12{\cdot}2$

(3) $M.D_{MO}(\text{Mode}) = \frac{\sum|X - Mo|}{N} = \frac{124}{10} = 12{\cdot}4$

$Mo = 3Me - 2\bar{X}$

$= 3 \times 57 - 2 \times 58$

$= 171 - 116 = 55$

(2) **Continuous Series**

Class	f	**Mid point** (X)	fX	$\lvert X - \bar{X}\rvert$ $\bar{X} = 31{\cdot}27$	$f\lvert X - \bar{X}\rvert$
1 – 10	8	5·5	44·00	25·77	206·16
11 – 20	12	15·5	186·00	15·77	189·24
21 – 30	17	25·5	433·50	5·77	98·09
31 – 40	14	35·5	497·00	4·23	59·22
41 – 50	9	45·5	409·50	14·23	128·07
51 – 60	7	55·5	388·50	24·23	169·61
61 – 70	4	65·5	262·00	34·23	136·92
	N = 71		2220.5		987.31

$\bar{X} = \frac{\sum fX}{N} = \frac{2220{\cdot}5}{71} = 31{\cdot}27$

$M.D_{\bar{X}} = \frac{\sum f|X - \bar{X}|}{N}$ (From the Table)

$= \frac{987{\cdot}31}{71} = 13{\cdot}905$

$MD_{\bar{X}} = 13{\cdot}91$ (approx)

Coefficient of Mean Deviation

Following the example Table (A) we have calculated $\bar{X}$, *Me* and *Mo* where $\bar{X}$ = 58, *Me* = 57 and *Mo* = 55 and $MD_{\bar{X}}$ = 12·2, MD_{Me} = 12·2 and MD_{MO} = 12·4 Using the above result we will now calculate the coefficient of MD about Mean, Mediam and Mode.

(i) $CMD_{\bar{X}} = \frac{MD_{\bar{X}}}{\bar{X}} \times 100$

$= \frac{12{\cdot}2}{58} \times 100$ **= 21·0344%**

(ii) $CMD_{Me} = \frac{MD_{Me}}{Me} \times 100$

$= \frac{12{\cdot}2}{57} \times 100$ **= 21·4035%**

(iii) $CMD_{Mo} = \frac{MD_{Mo}}{Mo} \times 100$

$= \frac{12{\cdot}4}{55} \times 100$ **= 22·545%**

M. D. – Advantages and Disadvantages.

(1) It is based on all the observations. Any change in any item would change the value of mean deviation.

(2) It is readily understood, simple to calculate. It is the average of the deviation from a measure of central tendency.

(3) M.D. is less affected by the extreme items than the Standard Deviation.

But M.D. ignores the algebraic signs of the deviations, and as such it is not capable of further algebraic treatement. Moreover, it is not accurate measure, particularly when it is calculated from mode. It is not also popular as standard Deviation.

Standard Deviation (S.D.)

As measures of dispersion Standard deviation is very much significant and in geographical analysis it has also wide range of uses. In calculating mean deviation, we ignored the algebraic signs, which is mathematically unjustified and illogical also. This drawback is removed in calculating standard deviation, usually denoted by 'δ' (Small Sigma).

Standard Deviation (S.D.) is the Square root of the arithmetic mean of the Squares of all the deviations from the mean. Briefly it is the root-mean-squeare deviation from the mean.

In X variable with N number of values (like X_1, X_2, X_3.....X_n) with arithmatic mean x, the S.D. will be

$$\sigma = \sqrt{\frac{\Sigma\left(X - \overline{X}\right)^2}{N}}$$

Computation of standard deviation may be done in two ways—

(a) By taking deviations from actual mean.

(b) By taking deviations from assumed mean.

For – (a) — Steps are—

(1) Find the actual mean

(2) Find the deviation from the mean

(3) Make Squares of the deviation and add up.

(4) Divide the addition by total number of items and find Square root.

For – (b) — Steps are

(1) Find the deviations of the items from an assumed mean and denote it by d. Find also Σd,

(2) Square the deviations find $\mathring{a}d^2$.

(3) Apply the following rule to find the S.D.

$$SD(\sigma) = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

Example : For (a)

X	$X - \bar{X}$	$(X - \bar{X})^2$
5	– 7	49
7	– 5	25
9	– 3	9
11	– 1	1
13	+ 1	1
15	+ 3	9
16	+ 4	16
20	+ 8	64
96		174

$$\bar{X} = \frac{\Sigma X}{N} = \frac{96}{8} = 12$$

$$\partial = \sqrt{\frac{\Sigma\left(X - \bar{X}\right)^2}{N}}$$

$$= \sqrt{\frac{174}{8}}$$

$$= \sqrt{21{\cdot}75}$$

$$\boxed{\partial = 4{\cdot}66}$$

$N = 8$

$\Sigma X = 96 \quad \therefore \quad \partial = 4{\cdot}66$

Example : For (b)

Applying the same example for (b)

X	A	$(X - A) = d$	d^2
5		– 8	64
7		– 6	36
9		– 4	16
11	Assumed	– 2	4
13 →	(13)	0	0
15	Mean	+ 2	4
16		+ 3	9
20		+ 7	49
$N = 8$		$\Sigma d = -8$	$\Sigma d^2 = 182$

So, Standard Deviation for this case

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

$$= \sqrt{\frac{182}{8} - \left(\frac{-8}{8}\right)^2}$$

$$= \sqrt{22{\cdot}75 - 1}$$

$$= \sqrt{21{\cdot}75}$$

$$\sigma = 4{\cdot}66$$

Now for continous series the standard deviation will be calculated by the following rule.

Class	***f***	**(*X*) Mid point**	***fX***	$\bar{X}$ **= 71·40** $(X - \bar{X})$	$(X - \bar{X})^2$	$f(X - \bar{X})^2$
40 – 50	104	45	4680	– 26·4	696·96	72483·84
50 – 60	117	55	6435	– 16·4	268·96	31468·32
60 – 70	223	65	4495	– 6·4	40·96	934·08
70 – 80	132	75	9900	3·6	12·96	1710·72
80 – 90	224	85	19040	13·6	184·96	443·04
90 – 100	109	95	0355	23·6	556·96	60708·64
	N = 909	5	ΣfX = 64905			26936·64

$$\bar{X} = \frac{\Sigma fX}{N} = \frac{64905}{909} = \boxed{71{\cdot}40 = \bar{X}}$$

$$\sigma = \sqrt{\frac{\Sigma f\left(X - \bar{X}\right)^2}{N}}$$

$$= \sqrt{\frac{26936{\cdot}64}{909}}$$

$$\sigma = \sqrt{238{\cdot}64}$$

$$\boxed{\sigma = 15{\cdot}45}$$

Example — Alternative Method.

We can calculate S.D. by Shortcut method also. Here is an example of this method.

Class	*f*	**Mid point** (*X*)	***d* = deviation** $d = \frac{X - A}{i}$	*fd*	fd^2	**Results**
10 – 19	3	14·5	– 3	– 9	27	N = 500
20 – 29	61	24·5	– 2	– 122	224	
30 – 39	223	34·5	– 1	– 223	223	
40 – 49	137	44·5	0	0	0	A = 44·5
50 – 59	53	54·5	+ 1	+ 53	53	i = 10
60 – 69	19	64·5	+ 2	+ 38	76	$\sum fd$ = –251
70 – 79	4	74·5	+ 3	+ 12	36	$\sum fd^2$ = 659
	N = 500			– 25	659	

So, According to above results —

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$$

$$\sigma = \sqrt{\frac{659}{500} - \left(\frac{-251}{500}\right)^2} \times i$$

$$\sigma = \sqrt{\frac{659}{500} - \frac{63001}{250000}} \times 10$$

$$\sigma = \sqrt{1{\cdot}318 - 0{\cdot}252} \times 10$$

$$\sigma = \sqrt{1{\cdot}066} \times 10 \qquad = 10 \times 1{\cdot}03$$

$$\sigma = \mathbf{10{\cdot}03}$$

4.4 Coefficient of Variation – (C.V.)

C. V. is the ratio of the Standard Deviation to the Mean expressed as percentage. This relative measure was first proposed by Prof. Karl Pearson. According to him, C.V. is the percentage variation in the mean, while S.D. is the total variation in the mean.

Symbolically, C.V. Can be expressed as—

$\text{C.V.} = \frac{\sigma}{X} \times 100$, i.e. the Co-efficient of Standard deviation multiplied by 100. The C.V. is also know as co-efficient of variability.

Example — C.V.

Class	f	X	$d = (X - A)$ $A = 25$	fd	fd^2
1 – 10	10	5	– 20	– 200	400
11 – 20	3	15	– 10	– 30	300
21 – 30	2	25(A)	0	0	0
31 – 40	1	35	+ 10	10	100
41 – 50	4	45	+ 20	80	1600
	N = 20			Σfd = – 140	Σfd^2 = 600

Now — $\sigma = \sqrt{\frac{\Sigma fd^2}{4} - \left(\frac{\Sigma fd}{N}\right)^2}$

$$\sigma = \sqrt{\frac{6000}{20} - \left(\frac{-140}{20}\right)^2}$$

$= \sqrt{300-(7)^2} \quad = \sqrt{251} \quad = 15{\cdot}84$ **σ = 15·84**

Here $\bar{X}$ will be calculated by the following rules

$$\overline{X} = A + \frac{\Sigma fd}{N}$$

$$= 25 + \frac{-140}{20}$$

= 25 + (–7) = 25 – 7 **$\bar{X}$ = 18**

So, $CV. = \frac{\sigma}{\overline{X}} \times 100 = \frac{15{\cdot}84}{18} \times 100$

$C.V.$ = 88%

With the help of C.V. we can compare the status between two series of data.

Example :

Gr A	58	59	60	65	66	52	75	31	46	48
Gr B	56	87	89	46	93	65	44	54	78	68

In order to find out the more consistent between the two groups, we are to compare their averages and then the C.V. for comparison.

For Gr A —

$$\overline{X} = \frac{\Sigma X}{N} = \frac{560}{10} = 56$$

$$\sigma = \sqrt{\frac{\Sigma d^2}{N}} = \sqrt{\frac{1376}{10}} = \sqrt{137{\cdot}6}$$

σ = 11·73 (From the table stated below)

$$CV. = \frac{11{\cdot}73}{56} \times 100 = 20{\cdot}94\%$$

For Group – B

$$\overline{X} = \frac{\Sigma X}{N} = \frac{680}{10} = 68$$

$$\sigma = \sqrt{\frac{\Sigma d^2}{N}} = \sqrt{\frac{2936}{10}} = \sqrt{293{\cdot}6} = 17{\cdot}14$$

$\sigma = 17{\cdot}14$

$$C.V. = \frac{17{\cdot}14}{68} \times 100 = 25{\cdot}21\%$$

Calculation Table :

X	d	d^2	X	d	d^2	**Results**
58	2	4	56	– 12	144	**Gr – A**
59	3	9	87	19	361	$\Sigma X = 560$
60	4	16	89	21	441	$N = 10$
65	9	81	46	– 22	484	$\Sigma d^2 = 1376$
66	10	100	93	25	625	
52	– 4	16	65	– 3	9	**Gr – B**
75	19	361	44	– 24	576	$N = 10$
31	– 25	625	54	– 14	196	$\Sigma X = 680$
46	– 10	100	78	10	100	$\Sigma d^2 = 2936$
48	– 8	64	68	0	0	
560		1376	680		2936	

Explanation :

Average ($\bar{X}$) of Group B are higher than that of Group A, So, status of group B is better than Group A, again, since C.V. of Group A is less than the Group B, so status of Group A is more consistent.

4.5 Advantages and Disadvantages of S.D.

(i) S.D. is based on all observations and is rigidly defined.

(ii) It is useful for algebraic treatment and possesses many mathematical properties like probability.

(iii) It is less affected by fluctuation of sampling than most other measures of dispersion.

(iv) For Comparing variability of two or more series, C.V. is considered as most appropriate and this is based on S.D. and $\bar{X}$–.

Disadvantages :

(i) It is not easy to understand and not so simple to calculate.

(ii) It gives more weight to the extremes and less to the items nearest to X–, since the squares of the deviations of higher sizes would be proportionately greater than that which are comparatively small.

Still the S.D. is the best measure of dispersan in comparison to other and should be used wherever possible.

4.6 Summary

These techniques are very much essential and useful in computation of central tendency of data.

Unit-5 ❑ Association- and Correlation: Rank correlation Product moment correlation

Structure

5.1 Objectives

- The learners will know about the correlation and the association among values.

5.1 Introduction

We, the Geographers always work with more than single variables. We have already worked with single variable by measures of central tendency, measures of dispersion etc., for calculation and analysis. But in practice, we also face a large number of problems involving the use of two or more variables. But in Geographical analysis is it necessary to deal with more than two variables. If two sets of variables vary in such a way that changes of one set are related by changes in the other, than these sets are said to be correlated. In general with the increase of height of a person the weight also increases. But in real world it may or may not, So how much they are correlated or interdependent to each other, should be observed.

5.3 Correlation and Association

So, Correlation and Association are meant for the relationship between two variables where with the changes in the values of one variables, the values of other variable also change. Correlation is also called co-variation.

According to nature and types correlation may be— (i) Positive, (ii) Negative and (iii) Zero correlation or No Correlation. A correlation is said to be positive, when high values of one variable are accompanied by the high values of the other, and that low

values of one are accompanie by low values of the other. But in negative correlation high values of one variable are accompanied by the low values of the other. If the values of two variables change in opposite directions, then it is negative correlation. In zero or no correlation the paired observations are only scattered. No association is found between paired obervations.

Or–I–Example

Positive Correlation

x	y
4	10
6	14
8	18
10	22
12	26

Negative Correlation

x	y
10	20
15	18
20	16
25	14
30	12

Moreover, when only two variables are studied it is a Simple Correlation. When there are three or more variables for composison it is multiple or partial correlation. In multiple correlation, three or more variables are studied. Correlation may be of Linear or nonlinear or Curvilinear.

Correlation will be linear, it the variations in the values of two variables are in a constant ratio. Conversely, it the variation of values of the variables do not bear a constant ratio, we will find nonlinear or curvilinear correlation.

5.4 Methods of Correlation

A number of methods are used for correlating variables —

(i) Scatter diagram method

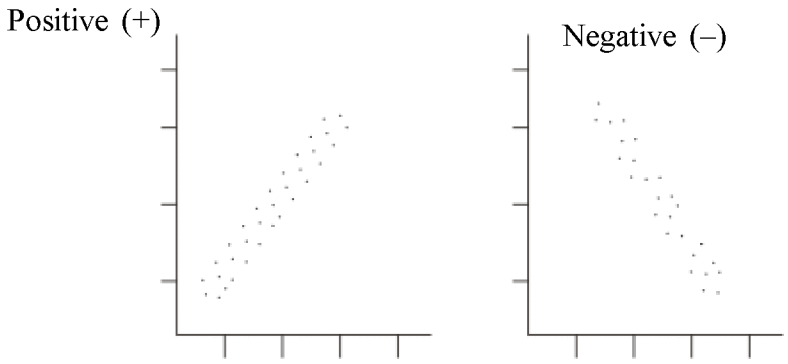
(ii) Karl Pearson's coefficient of correlation or product moment correlation.

(iii) Rank correlation by C.E. Spearman

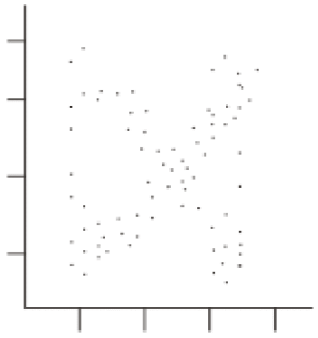
Scatter diagram method is a special type of dot chart. Here given variables (x, y) are plotted in a graph in the form of dots. From this plotting we can get some trend of the distribution of dots on the coordinate, either upward or downward.

5.5 Correlation Coefficient

A Correlation Co-efficient is a statistical measure of the degree to which changes



to the value of one variable predict change to the value of another. It may be positively or negatively correlated and are expressed as values between (+1) and (–1). Lesser degrees of correlation are expressed as non-zero decimals.



Correlation Coefficient	Degree of Correlation
± 1·00	Perfect positive or negative correlation
± 0·80 – 1·00	Highly positive or negative
± 0·60 – 0·80	Moderately High Positive or negative correlation
± 0·40 – 0·60	Moderate correlation
± 0·20 – 0·40	Low correlation
± < 0·20	Negligible correlation

Correlation of – 0·97 is a strong negative while a correlation of 0·10 would be a weak positive correlation.

Karl Pearson's coefficient of correlation.

This Coefficient of Correlation, popularly known as Pearsonian Coefficient of Correlation, can measure the extent of relationship between two sets of data. This can be done by the following way :

(1) Deviations taken from the actual A.M. By this method Pearsonian Coefficient of Correlation is found by the following principles

$$rxy = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} \Big/ \text{Where } \begin{matrix} x = X - \overline{X} \\ y = Y - \overline{Y} \end{matrix}$$

In both cases deviation from Actual Arithmetic mean has been taken.

Example :

x –	1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11
y –	4,	7,	10,	13,	16,	19,	22,	25,	28,	31,	34

X	Y	$x = X - \overline{X}$	$y = Y - \overline{Y}$	x^2	y^2	xy
1	4	–5	–15	25	225	75
2	7	–4	–12	16	144	48
3	10	–3	–9	9	81	27
4	13	–2	–6	4	36	12
5	16	–1	–3	1	9	3
6	19	– 0 –	– 0 –	0	0	0
7	22	1	3	1	9	3
8	25	2	6	4	36	12
9	28	3	9	9	81	27
10	31	4	12	16	144	48
11	34	5	15	25	225	75

$\Sigma X = 66$
$\Sigma Y = 209$
$\Sigma xy = 330$
$\Sigma x^2 = 110$
$\Sigma y^2 = 990$

For previous (above) tabl

$$\overline{X} = \frac{\Sigma X}{n} = \frac{66}{11} = 6$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{209}{11} = 19$$

Now, $$y = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \times \Sigma y^2}} = \frac{330}{\sqrt{110 \times 990}} = \frac{330}{\sqrt{108900}}$$

$$= \frac{330}{330} = 1$$

So, The Correlation Coefficient of the variables X and Y is perfectly positive.

The Correlation Coefficient may also be written as

$$r_{xy} = \frac{\frac{1}{n}\Sigma x'y'}{\sigma x \sigma y} = \left[\frac{\Sigma\left(x-\bar{x}\right)\left(y-\bar{y}\right)}{n \times \sigma x \times \sigma y}\right]$$

Where n = number of pairs of observations.

$$\sigma x = \sqrt{\frac{\Sigma\left(x-\bar{x}\right)^2}{n}}$$

$$\sigma y = \sqrt{\frac{\Sigma\left(y-\bar{y}\right)^2}{n}}$$

The numerator $\frac{1}{n}\Sigma x'y' = \frac{1}{n}\sum\left(x-\bar{x}\right)\left(y-\bar{y}\right)$ is called covariance between the two variables x and y and is written as cov (x, y).

So we may also write $y = \frac{\text{cov}\left(x, y\right)}{\sqrt{v(x)} \times \sqrt{v(y)}}$

Since $\sigma x = \sqrt{v(x)}$ and $\sigma y = \sqrt{v(y)}$

where, v = variance

So, after extension we may write correlation coefficient as—

$$r_{xy} = \frac{\frac{1}{n}\Sigma xy - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\Sigma x^2 + \bar{x}^2} \times \sqrt{\frac{1}{n}\Sigma y^2 + \bar{y}^2}}$$

$$= \frac{\frac{1}{n}\sum xy - \frac{\sum x}{n} \times \frac{\sum y}{n}}{\sqrt{\frac{1}{n}\sum x^2 - \left(\frac{\sum x}{n}\right)^2} \times \sqrt{\frac{1}{n}\sum y^2 - \left(\frac{\sum y}{n}\right)^2}}$$

Since $\bar{x} = \frac{\sum x}{n}$ and $\bar{y} = \frac{\sum y}{n}$

So, it can be written as —

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}} \quad \text{.....(A)}$$

For computational purpose the expression (A) is generally used.

Coefficient of Correlation directly from ungrouped data.

X	Y	X^2	Y^2	XY	
48	88	2304	7744	4224	$\Sigma X = 382$
32	80	1024	6400	2560	$\Sigma Y = 776$
36	78	1296	6084	2808	
34	74	1156	5476	256	$\Sigma X^2 = 14876$
39	74	1521	5476	2886	$\Sigma Y^2 = 60444$
37	75	1369	5625	2775	
41	78	1681	6084	3198	$\Sigma XY = 29832$
45	83	2025	6889	3735	
40	75	1600	5625	3000	N = 10
30	71	900	5041	2130	

Here Pearsonian Correlation Coefficient

$$r_{xy} = \frac{10 \times 29832 - 382 \times 776}{\sqrt{10 \times 14876 - (382)^2} \times \sqrt{10 \times 60444 - (776)^2}}$$

$$= \frac{298320 - 296432}{\sqrt{148760 - 145924} \times \sqrt{604440 - 602176}}$$

$$= \frac{1888}{\sqrt{2832} \times \sqrt{2264}} = \frac{1888}{53{\cdot}3 \times 47{\cdot}6}$$

$$= \frac{1888}{2537{\cdot}08} = 0{\cdot}74$$

So, r_{xy} = 0·74 (High positive correlation)

Following properties are to be mentioned with respect to r_{xy} i.e., Pearsonian Correlation Coefficient.

(A) Correlation Coefficient is a pure number, i.e. it is independent to the unit of measurement if variable.

(B) The Correlation Coefficient does not depend on origin of reference or scale of measurement.

(C) The Correlation Coefficient lies between (–1) to (+1) and it has also been proved mathematically.

Rank Correlation : C.E.Spearman

Geographers usually deal with the quantitative data but in socio-cultural field there are some attributes which cannot be measured by quantity. In such cases individuals in the group can be arranged in order and hence obtaining for each individual a number indicating the rank in the group. This method was developed by C.E. Spearman, a British Psychologist, in 1904.

Steps for calculating the Coefficient of Correlation by Rank Method. It is assigned by (ρ)

(1) Two or more attributes or variables may be involved. Rank may be assigned either in ascending or in descending order.

(2) Assign ranks to various items of the series.

(3) Find differences of the ranks (d)

(4) Find out the Square of the differences (d^2)

(5) If two items have equal value, in that case two items will be given average rank of the ranks. Further more, the immediate rank will not be given for next value.

(6) Using the following rule we can find r_{xy} for getting the rank correlation coefficient.

$$\rho = 1 - \frac{6(\sum d)^2}{n^3 - n}$$

Where n = number of pairs of observations.

Significance of Spearman's Coefficient

The value of this Co-efficient ranges between +1 and –1. It r_{xy} = +1, there is

complete agreement in the order of ranks and the ranks are in the same direction. Again if $r_{xy} = (-1)$, there is complete agrement in the order of ranks and they are in opposite directions.

Example

Block ID	No of crimes (male)	No of crimes female	R_1	R_2	D	D^2
A	2	1	2	1	1	1
B	5	3	5	3	2	4
C	3	5	3	5	2	4
D	7	8	7	8	1	1
E	9	7	9	7	2	4
F	1	2	1	2	1	1
G	4	6	4	6	2	4
H	8	4	8	4	4	16
I	6	9	6	9	3	9
J	10	10	10	10	0	0

$\sum D^2 = 44$, $n = 10$

According to rule — $\rho = 1 - \frac{6\sum(d)^2}{n(n^2-1)}$

$$\rho = 1 - \frac{6 \times 44}{10 \times (10^2 - 1)}$$

$$= 1 - \frac{4}{15} \quad = 1 - 0{\cdot}266 \quad = 0{\cdot}73$$

(B) Where in a variable two items are of same value.

in '000 Rs (1) Salary of Male	in '000 Rs (2) Salary of Female	R_1	R_2	d	d^2	Results
15	40	2	6	– 4	16	
20	30	3·5	4	– 0·5	0·25	

28	50	5	7	– 2	4	
12	30	1	4	–3	9	$\sum d^2 = 81{\cdot}50$
40	20	6	2	4	16	
60	10	7	1	6	36	
20	30	3·5	4	– 0·5	0·25	
80	60	8	8	0	0	

For equal ranks some adjustment in the above formula is required i.e., to add—

$\frac{1}{2}\left(m^3 - m\right)$ with $\sum d^2$ where

m = number of items whose ranks are common.

Here $r_{xy} = 1 - \frac{6\left\{\sum d^2 + \frac{1}{12}\left(m^3 - m\right) + \frac{1}{12}\left(m^3 - m\right)\right\}}{n^3 - n}$

The item 20 is repeated 2 times in X – series i.e., $m = 2$ in X-series and again $m = 3$ in Y – series

$$\therefore \quad r_{xy} = 1 - \frac{6\left\{81{\cdot}5 + \frac{1}{2}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)\right\}}{8^3 - 8}$$

$$= 1 - \frac{6\{81{\cdot}5 + 0{\cdot}5 + 2\}}{504}$$

$$= 1 - \frac{504}{504} \quad = 1 - \frac{504}{504} \quad = 1 - 1 \quad = 0$$

= No Correlation.

5.6 Summary

- The time series analysis for any geographical data to be studied is very significant to show the timely variation and changes of that data.
- It plays a vital role in analysing the trend of data to be dealt with.

Unit-6 ❑ Linear Regression

Structure

6.1 Objectives

By correlation the students can establish the relation and association between two variables. But in geographical analysis we are also interested to predict the value of one variable for the given value of the other.

6.2 Introduction

The term 'regression' means, going back or study, as established by F. Galton. He studied regarding the height of fathers and their sons revealed an interesting relationship. The deviations of mean heights of the sons from the mean height of the race were less than the deviations in the mean height of the fathers from mean height of the race. When the fathers were above or below the mean, the sons tended to go back or regress towards the mean.

6.3 Regression

Galton represented the average relationship between these variables graphically and called the line as the line of regression. Regression lines give idea on the correlation of two series. The regression amalysis helps in following ways —

(i) To estimate or predict the values of dependent variables from values of independent variables.

(ii) To abtain the measure of error involved in using the regression line as a basis of estimation.

(iii) To obtain a measure of association or Correlation that exists between the two variables.

So, there is a close relationship between Correlation and regression. Correlation Coefficient is a **measure of degree of relationship between X and Y,** where as, the regression analysis reveals the study of **nature of relationship** between the variables.

Regression Equations :

To find out the regression equations for two variables we are to consider two regression lines — (i) x on y and (ii) y on x. Regression equations are algebraic expression of regression lines. For two regression lines there will be two regression equations.

Regression equation of X on Y

To develop this, we will follow

$$x = a + by$$

x and y are two variables and a, b are two constants. Now to determine the constants a and b we are to solve the following normal equations

$\Sigma x = na + b\Sigma y$.....(i)

$\Sigma xy = a\Sigma y + b\Sigma y^2$.....(ii)

n = number of observed pair of values.

(ii) **Regression equation of y and x**

For this we will follow

$y = a + bx$, where the value of a and b are to be attained by solving.

$\Sigma y = na + b\Sigma x$ .....(i)

$\Sigma xy = a\Sigma x + b\Sigma x^2$.....(ii)

Example : From the following two variables we will try to obtain the two regression equations.

x — 6, 2, 10, 4, 8

y — 9, 11, 5, 8, 7

Computation of Regression Equations.

x	y	xy	x^2	y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49

For equation of x on y — $x = a + by$ and the subsequent normal equations are

$\Sigma x = na + b\Sigma y, \qquad \Sigma xy = a\Sigma y + b\Sigma y^2$

now substituting the values on respective eg. equations, we got—

$30 = 5a + 40b$ .....(i)

$214 = 40a + 340b$ .....(ii)

For eliminating a we are to multiply (i) by 8 and subtracting from (ii) we get

$$-26 = 20b$$

$$\text{or, } \boxed{b = 1{\cdot}3}$$

Now putting this value of b in (i) we get —

$$30 = 5a + 40(-1{\cdot}3)$$

$$\text{or, } \boxed{a = 16{\cdot}4}$$

$\therefore$ The regression lime of x on y is

$$\boxed{x = 16{\cdot}4 - 1{\cdot}3y}$$

Now for the regression equation y on x we get

$y = a + bx$ with normal equations —

$\Sigma y = na + b\Sigma x$ .....(iii)

$\Sigma xy = a\Sigma x + b\Sigma x^2$ .....(iv)

Putting the values we get —

$40 = 5a + 30b$ .....(v)

$214 = 30a + 220b$ .....(vi)

Following the same process as previous eg.

We get multiplying (v) by 6 and Sub traching it from (vi) we have

$-26 = 40b$ or $b = -0{\cdot}65$

Putting the value of b in e.g. (v) we have

$40 = 5a + 30(-0{\cdot}65)$

or $a = 11{\cdot}9$

Hence the regression line of y on $x = y = 11{\cdot}9 - 0{\cdot}65x$

6.4 Properties of Linear Regression equations

(i) the linear regression equation of y on x is $y - \bar{y} = byx\ (x - \bar{x})$ and that of x on y is $x - \bar{x} = bxy(y - \bar{y})$ where byx and bxy are known as regression coefficients of y on x and x on y respectively.

So, we can express —

$$byx = \frac{\text{cov}(x, y)}{\sigma_x^2} = \left\{ y \frac{\sigma_y}{\sigma_x} \right\} \text{and}$$

$$byx = \frac{\text{cov}(x, y)}{\sigma_y^2} = \left\{ y \frac{\sigma_x}{\sigma_y} \right\}$$

(ii) The product of two regression coefficients is equal to the square of the correlation coefficient, i.e.,

$$(byx \times bxy) = y \frac{\sigma_y}{\sigma_x} \times y \frac{\sigma_x}{\sigma_y} = y^2$$

(iii) Regression Coefficient and Correlation Coefficient, i.e., *byx*, *bxy* and *y* have the same sign, i.e., if both the regression Co-efficients have a negative sign, *y* will also be negative, and again if the regression co-efficients are both positive, then *y* will be positive. If the *y* is '0', then *byx* and *bxy* will be zero.

(iv) Two regression lines always intersect at $(\bar{x}, \bar{y})$. The slope of regression line of *y* on *x* is *byx* and that of *x* on *y* is $\frac{1}{bxy}$.

(v) Two regression equations may be written as $\frac{y - \bar{y}}{\sigma_y} = y \frac{x - \bar{x}}{\sigma x}$ and $\frac{x - \bar{x}}{\sigma_x} = y \frac{y - \bar{y}}{\sigma_y}$ which are different.

But if $y = \pm 1$, the two equations become identical. Again if $y = 0$ then we find $y = \bar{y}$ and $x = \bar{x}$. In that case *y* or *x* cannot be estimated from linear regression equations.

6.5 Summary

Thus can be concluded that in linear regression the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data linear regression is widely used in biological, behavioral and social sciences to describe possible relationships between variables.

Unit-7 ❑ Time Series Analysis

Structure

7.1 Objectives

- The learners will learn about the time series analysis and the various characteristic.

7.2 Introduction

Analysis of time series data is a significant approach for geographical research. Spatio-temporal data play decisive role for geographical analysis. A time series is defined as data arranged chronologically. Such a series of observations disclosed the changes or variations in the value of the variable due to changes in time. The time series data are playing increasingly significant role in all kinds of economic, social, cultural and other geographical activities.

7.3 Characteristics of time series

A time series data may best be studied by plotting them on a graph paper. After plotting one can notice the following trends of the data —

(i) Some movements exhibiting persistent growth or decline.

(ii) Some movements are regular and periodic in nature with period not more than one year.

(iii) Some fairly regular and periodic with period of duration of more than a year, and Finally, some irregular, mild or violent movements.

Thus, a time series, in general is the result of four types of movements—

(i) **A basic or secular trend** which is a smooth, regular and long term upward or downward movement in the data. It reveals the general tendency of the data.

(ii) **Seasonal variation** is a short term periodic movement whose period is not longer than a year. It is uniform and regular in nature. This short term movement is mainly due to the climatic changes or to holidays or to social customs, trading and other habits of the people.

(iii) Cyclic variations are oscillatory movements with a period more than a year. Such movements do not ordinarily exhibit regular periodicity. Most of the geographic activities have four distinct phases — (1) Prosperity, (2) Decline, (3) Depression and (4) Recovery. These four phases are generated by factors other than changes in climate, social customs and those which create scasonal variations.

(iv) Irregular variations are of two types — Catastrophic and Accidental.

Catastroophic variations are due to specific events such as fires, earthquakes, floods etc. The accidental variations are due to multiplicity of causes of unknown origin.

For time series analysis trends are commonly measured by the following methods—

(i) Graphic Methods

(ii) Semi – Average Method

(iii) Moving Average Method

(iv) Method of Least Square

Among all those methods moving average method is the simplest of smoothing out fluctuations and abtaining the trend values with fair degree of accuracy. Moving averages are number of arithmetic averages calculated from the time series data, each based on a fixed number of cusequtive observations.

The objective of the moving average method is to smooth out Cyclical, Seasonal and Irregular variations of the time series data in order to isolate the trend. Moving average, in general, cannot eliminate irregular fluctuations but it only reduces them.

7.4 Merits and Demerits

(1) It is flexible and not subjective and simple to understand.

(2) It is appropriate only when the trend is linear. In non-linear cases it may over-estimate the trend value.

(3) Cyclical fluctuation may be eradicated completely if the cycles are regular and the period of moving average be equal to or an integral multiple of the period of fluctuation.

(4) Trend value cannot be determined for some periods at the beginning and at the end.

(5) This method is very sensitive to a few very high and low values which the series may contain.

Example :

Calculate the four yearly moving averages from the following observation.

Year		Production (000 tons)	
2001	—	506	
2002	—	620	
2003	—	1036	
2004	—	673	
2005	—	588	
2006	—	696	
2007	—	1116	The processes of calculation are somewhat odd number of years and even number of years.
2008	—	738	
2009	—	663	
2010	—	773	
2011	—	1189	
2012	—	818	
2013	—	745	
2014	—	845	
2015	—	1276	

(1) Year	Y	(2) Year moving Total	(3) 4 Year Moving Average = (Col 2/4)	(4) 2 item moving total	(5) Col 4/2 4 year centred moving average
2001	506				
2002	620	2835	708·75		
				(2003) 1438·00	719·00
2003	1036	2917	729·25		
				(2004) 1477·5	738·75
2004	673	2993	748·25		
				(2005) 1516·5	758·25
2005	588	3073	768·25		
				(2006) 1552·75	77·38
2006	696	3138	784·50		
				(2007) 1587·75	793·87
2007	1116	3213	803·25		
				(2008) 1625·75	812·87
2008	738	3290	822·50		
				(2009) 1663·25	831·67
2009	663	3363	840·75		
				(2010) 1701·50	850·75
2010	773	3443	860·75		
				(2011) 1742·00	871·00
2011	1189	3525	881·25		
				(2012) 1780·50	890·25
2012	818	3597	899·25		
				(2013) 1820·25	910·12
2013	745	3684	921·00		
2014	845				
2015	1276				

Example : 2 — Compute fiveyearly moving averages from the following.

Year –	2004	2005	2006	2007	2008	2009	2010	2011	2012
Yearly – Production (000 ton)	6·4	4·3	4·3	3·4	4·4	5·4	3·4	2·4	1·4

Calculation

Year	Production	5 Year Moving Total	5 Year Moving Average
2004	6·4	—	—
2005	4·3	—	—
2006	4·3	22·8	4·56
2007	3·4	21·8	4·36
2008	4·4	20·9	4·18
2009	5·4	19·0	3·80
2010	3·4	—	—
2011	2·4	—	—
2012	1·4	—	—

First moving total – 6·4 + 4·3 + 4·3 + 3·4 + 4·4 = 22·8

First moving Average – 22·8/5 = 4·56

All are calculated in this processes.

7.5 Summary

- It can be said that, a time series is a sequence taken at successive equally spaced points im time.
- It is a statistical method of analyzing data from repeated observations on a single unit or individual at regular intervals.

Module-02

Statistical Methods in Geography
Laboratory : List of Practical

Unit 1 ❑ Construction of Data Matrix with each row representing an aerial unit (districts/ Blocks/Mouzas/Towns) and columns representing relevant attributes

Structure

1.1 Objective

- The learners will come to know about data matrix.

1.2 Introduction

Data matrix is a rectangular array of data variables, which may be numerical, classificatory, or alphanumeric. The data matrix forms the input structure upon which statistical procedures for regression analysis, analysis of variance, multivariate analysis, cluster analysis, or survey analysis will operate. Each attributes accounts for a column in the **geographical matrix.**

1.3 Data Matrix

Geographical data can be arranged in two distinct tabular forms : a spatial structure **data matrix** and a spatial interaction data matrix. These two forms are also referred to as a **geographical data matrix** and a spatial behaviour **data matrix** respectively. In all cases of interpretation, data needs to be presented as a rectangular data matrix. Each column in a data matrix contains a variable (indicator, measurement, questions in a survey) and each row an observation (case).

Each cell contains a single value for a particular variable and observation, e.g. the GDP per capita for Albania. If the value is not available, the cell content will contain show somehow that the value is missing (missing value indicator, all statistically oriented software will automatically skip that kind of value in computations.

Here's schematic representation of a Data Matrix.

Country	**Continent**	**Cont Num**	**GDP per capita**	**Variable$_j$**	**Variable$_k$**	
Afganistan	*Asia*	*1*	*value*	*value*	*value*	*value*
Albenia	*Europe*	*3*	*value*	*value*	*value*	*value*
...	...	...	...	...	...	...
country$_n$	...	*value*	*value*	*value*	*value*	*value*

Each column has a name (Variable Name) used to refer to it. All values of a particular variable have to be of the same type, i.e. numeric or strin. In the example above obviously the country and continent names are strings, and GDP per capita contains numerical information. The rectangular data matrix is mandatory for statistical analysis ; if data is presented in a different way it has to be restructured first to produce a rectangular data matrix.

Data Matrix — Columns Representing religious categories and Rows representing Spatial Units.

India : Percentage of Different Religious Communities in Total Population, 2001

	Hindus	**Muslims**	**Sikhs**	**Christians**	**Jains**	**Buddhists**	**Others**
India	80.46	13.43	1.87	2.34	0.41	0.77	0.65
Jammu & Kashmir	29.63	66.97	2.04	0.20	0.02	1.12	0.00
Himachal Pradesh	95.43	1.97	1.19	0.13	0.02	1.25	0.01
Punjab	36.94	1.57	59.19	1.20	0.16	0.17	0.04
Chandigarh	28.61	3.95	16.12	0.85	0.29	0.15	0.03
Uttaranchal	84.96	11.92	2.50	0.32	0.11	0.15	0.01
Haryana	88.23	5.78	5.54	0.13	0.27	0.03	0.01
Delhi	82.00	11.72	2.50	0.32	0.11	0.15	0.01
Rajasthan	88.75	8.47	1.45	0.13	1.15	0.02	0.01
Uttar Pradesh	80.61	18.50	0.41	0.13	0.12	0.18	0.01
Bihar	83.23	16.53	0.03	0.06	0.02	13.03	30.73
Sikkim	60.93	1.42	0.22	6.68	0.03	28.11	2.39
Arunachal Pradesh	34.60	1.88	0.17	18.72	0.02	13.03	30.73
Nagaland	7.70	1.76	0.06	89.97	0.11	0.07	0.31
Manipur	46.01	8.81	0.08	34.04	0.07	0.09	10.86
Mizoram	3.55	1.14	0.04	86.97	0.02	7.93	0.27
Tripura	85.62	7.95	0.04	3.20	0.01	3.09	0.04
Meghalaya	13.27	4.28	0.13	70.25	0.03	0.20	11.53

	Hindus	Muslims	Sikhs	Christians	Jains	Buddhists	Others
Assam	64.89	30.92	0.08	3.70	0.09	0.19	0.09
West Bengal	72.47	25.25	0.08	0.64	0.07	0.30	1.12
Jharkhand	68.57	13.85	0.31	4.06	0.06	0.02	13.04
Orissa	94.35	2.07	0.05	2.44	0.02	0.03	0.98
Chhattisgarh	94.70	1.97	0.33	1.92	0.27	0.31	0.46
Madhya Pradesh	91.15	6.37	0.25	0.28	0.90	0.35	0.68
Gujarat	89.09	9.06	0.09	0.56	1.04	0.04	0.06
Damman & Diu	89.69	7.76	0.09	2.13	0.17	0.08	0.07
Dadra & Nagar Haveli	93.52	2.96	0.06	2.75	0.39	0.21	0.04
Maharashtra	80.37	10.60	0.22	1.09	1.34	6.03	0.24
Andhra Pradesh	89.01	9.17	0.04	1.55	0.05	0.04	0.01
Karnataka	83.86	12.23	0.03	1.91	0.78	0.74	0.22
Goa	65.78	6.84	0.07	26.68	0.06	0.05	0.03
Lakshadweep	3.66	95.47	0.01	0.84	0.00	0.00	0.00
Kerala	56.16	24.70	0.01	19.02	0.01	0.01	0.01
Tamil Nadu	88.11	5.56	0.02	6.07	0.13	0.01	0.01
Pondicherry	86.77	6.09	0.01	6.95	0.10	0.01	0.02
Andaman & Nicobar	69.24	8.22	0.45	21.67	0.01	0.12	0.07

Source of Data : Registrar General and Census Commissioner, India (2004) The First Report on Religious Data, Census of India, New Delhi.

Data Matrix of Major Religious Composition of India for Computation and Representation of frequency Distribution

Sl. No.	Hindu	Muslim	Others
01	80	13	07
02	29	67	04
03	95	02	03
04	37	02	61
05	79	04	17
06	85	12	03
07	88	06	06
08	82	12	06
09	89	08	03
10	81	18	01
11	82	17	01
12	61	01	38
13	35	02	63
14	08	02	90

15	46	09	65
16	04	01	95
17	86	08	06
18	13	04	83
19	65	31	04
20	72	25	03
21	69	14	17
22	94	02	04
23	95	02	03
24	91	06	03
25	89	09	02
26	90	08	02
27	94	03	03
28	80	11	09
29	89	09	02
30	84	12	04
31	66	07	27
32	04	95	01
33	56	25	19
34	88	06	06
35	87	06	07
36	69	08	23

Following the data matrix stated in the previous table religious composition of India has been shown. Three major religious categories are stated

Total No. of spatial unit = 36

So N = 36

Range – 04 – 95

No of Classes = 1 + 3.3 Log N

= 1 + 3.3 Log (36) = 6.1357

= clasess

$$\frac{95-04}{36}=15(\text{approx})$$

So, width of the class (w/i) = 15

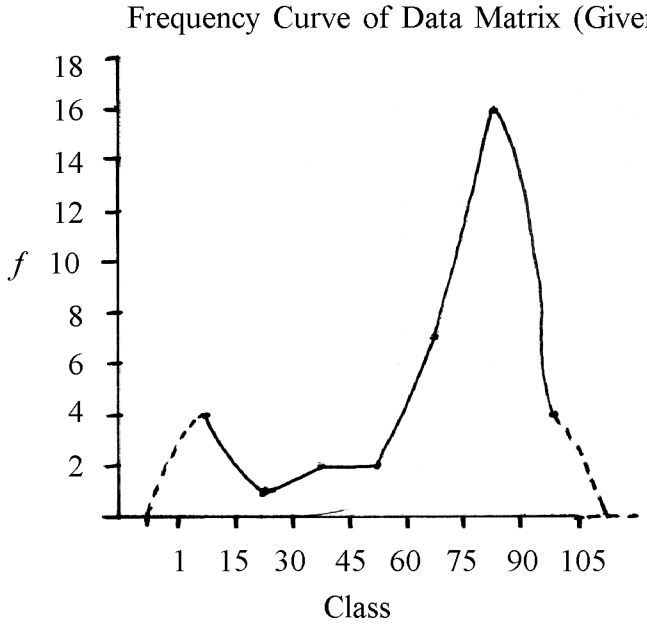
Class	tally	f	x	Relative frequency	Smoothed frequency	Cumulative frequency
1–15	IIII	04	7.5	0.11	1.666	04
16–30	I	01	22.5	0.03	2.333	04
31–45	II	02	37.5	0.06	1.666	07
46–60	II	02	52.5	0.06	3.666	09
61–75	IIII II	07	67.5	0.19	8.333	16
76–90	IIII IIII IIII I	16	82.5	0.44	9.000	32
91–105	IIII	04	97.5	0.11	6.666	36
		N=36		1.002		

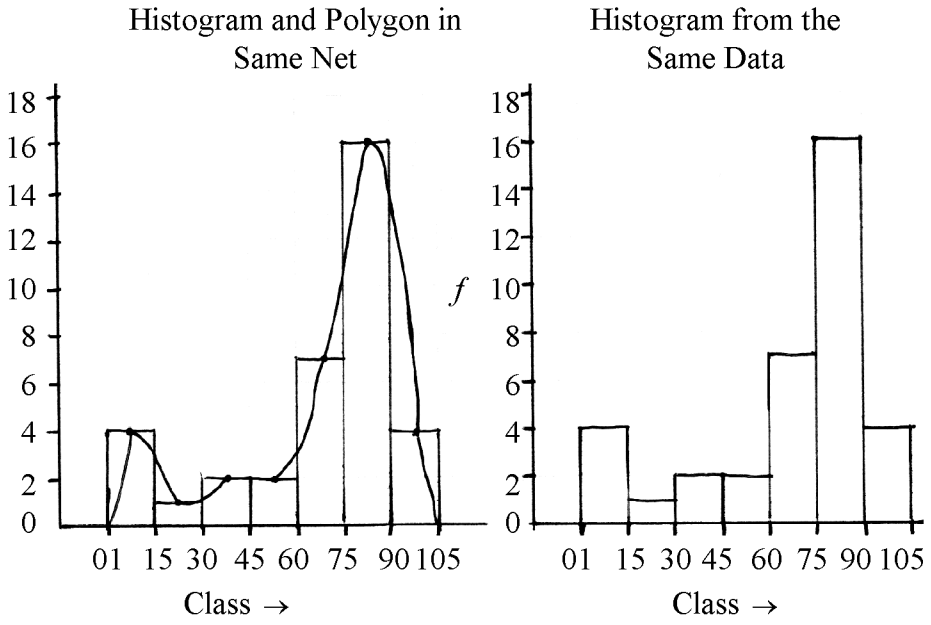
Cumulative of frequency for cumalative Frequency curve, agive (More than amd less than type).

Class	f	Less than LCL	f	UCL	More than UCL(f)	Σfx	(X)
1–15	04	01	00	01	36	30	7.5
16–30	01	16	04	15	32	22.5	22.5
31–45	02	31	05	30	31	75	37.5
46–60	02	46	07	45	29	105	52.5
61–75	07	61	09	60	27	105	67.5
76–90	16	76	16	75	20	47.25	67.5
91–105	04	91	32	90	04	1320	82.5
				105	00	330	97.5
	36	105	36		=36		

LCL = Lower Class Limit

UCL = Upper Class Limit





1.4 Summary

Data Matrix is an important tool in statistical analysis. It is a two-dimensional code consisting of black and white cells or dots arranged in either a square or rectangular pattern also known as matrix.

Unit 2 ❑ Frequency Table – Computation and Interpretation

Structure

2.1 Objective

The learners will learn about the computation of frequency table.

2.2 Introduction

Collected and classified data are presented in a form of frequency distribution. ***Frequency distribution is simply a table in which the data are grouped into classes on the basis of common characteristics and the number of cases which fall in each class are recorded.*** It shows the frequency of occurrence of different values of a single variable. ***A frequency distribution is constructed to satisfy three objectives :***

(i) to facilitate the analysis of data,

(ii) to estimate frequencies of the unknown population distribution from the distribution of sample data, and (iii) to facilitate the computation of various statistical measures.

2.3 Types of frequency distribution

Frequency distribution can be of two types :

1. Univariate Frequency Distribution
2. Bi-variate Frequency Distribution.

Univariate distribution incorporates different values of one variable only whereas the ***Bivariate frequency distribution incorporates*** the value of two variables. ***The Univariate frequency distribution is further classified into three categories :***

(i) Series of individual observations,

(ii)Discrete frequency distribution, and

(iii) Continuous frequency distribution

Series of individual observations, is a simple listing of items of each observation. If marks of 14 students in statistics of a classare given individually, it will form a series of individual observations.

Marks obtained in Statistics in a University

Roll Nos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Marks :	60	71	80	41	81	41	85	35	98	52	50	91	30	88

Marks in
Ascending Order30,35, 41, 41, 50, 52, 60, 71, 80, 81, 85, 88, 91, 98
Marks in
Descending Order 98, 91, 88, 85, 81, 80, 71, 60, 52, 50, 41, 41, 35, 30

Discrete Frequency Distribution :

In a discrete series, the data are presented in such a way that exact measurements of units are indicated. In a discrete frequency distribution, we count the number of times each value of the variable in data given to you. This is facilitated through the technique of tally bars. With the help of this technique we will represent the discrete frequency distribution in the following manner.

We are given marks of 42 students :

55 51 57 40 26 43 46 41 46 48 33 40 26 40 40 41
43 53 45 53 33 50 40 33 40 26 53 59 33 39 55 48
15 26 43 59 51 39 15 45 26 15

We can construct a discrete frequency distribution from the above given marks.

Marks of 42 students

Marks	Tally mark	Frequency	Mark	Tally	Frequency
15	III	3	46	//	02
26	𝍸	5	48	//	02
33	IIII	4	50	/	01
39	II	2	51	//	02
40	𝍸	5	53	///	03
41	II	2	55	///	03
43	III	3	57	/	01
45	II	2	59	//	02
					Total 42

Continuous Frequency Distribution :

If the identity of the units about a particular information collected, is neither relevant nor is the order in which the observations occur, then the first step of condensation is to classify the data into different classes by dividing the entire group of values of the variable into a suitable number of groups and then recording the number of observations in each group. Thus, we divide the total range of values of the variable (marks of 42 students) i.e. 59–15 = 44 into groups of 10 each, then we shall get (42/10) 5 groups and the distribution of marks is displayed by the following frequency distribution.

Marks of 42 Students

Marks (×)	Tally Bars	Number of Students (f)
15-25	///	3
25-35	𝍸 ////	9
35-45	𝍸 𝍸 //	12
45-55	𝍸 𝍸 //	12
55-65	𝍸 /	6
	Total	42

The various groups into which the values of a variable are classified are known classes, the length of the class interval (10) is called the width of the class. Two values, specifying the class are called the class limits. The presentation of the data into continuous classes with the corresponding frequencies is known as continuous frequency distribution. There are two methods of classifying the data according to class intervals :

(i) exclusive method and

(ii) inclusive method

In an exclusive method, the class intervals are fixed in such a manner that upper limit of one class becomes the lower limit of the following class. Moreover, an item equal to the upper limit of a class would be excluded from that class and included in the next class. The following data are classified on this basis.

Income (Rs.)	**No. of Persons**
200–250	50
250–300	100
300–350	70
350–400	130
400–450	50
450–500	100
Total	500

It is clear from the example that the exclusive method ensures continuity of the data in as much as the upper limit of one class is the lower limit of the next class. Therefore, 50 persons have their incomes between 200 to 249.99 and a person whose income is 250 shall be included in the next class of 250-300. According to the inclusive method, an item equal to upper limit of a class is included in that class itself. The following table demonstrates this method.

Income (Rs.)	**No. of Persons**
200–249	50
250–299	100
300–349	70
350–399	130
400–449	50
450–499	100
	Total 500

Hence in the class 200 – 249, we include persons whose income is between Rs. 200 and Rs. 249.

2.4 Principles for Constructing Frequency Distributions

In spite of the great importance of classification in statistical analysis, no hard and fast rules are laid down for it. A statistician uses his discreation for classifying a frequency distribution and sound experience, wisdom, skill an aptness for an appropriate classification of the data. However, the following guidelines must be consideried to construct a frequency distribution :

1. Type of classes : The classes should be clearly defined and should not lead to any ambiguity. They should be exhaustive and mutually exclusive so that any value of variable corresponds to only class.

2. Number of classes : The choice about the number of classes in which a given frequency distribution should he divided depends upon the following things;

(i) The total frequency which means the total number of observation in the distribution.

(ii) The nature of the data which means the size or magnitude of the values of the variable.

(iii) The desired accuracy.

(iv) The convenience regarding computation of the various descriptive measures of the frequency distribution such as means, variance etc.

The number of classes should not be too small or too large if the clases are few, the classification becomes very broad and rough which might obscure some important

features and characteristics of the data. The accuracy of the results decreases as the number of classes becomes smaller. On the other hand, too many classes will result in a few frequencies in each class. Hence a balance shold be maintained between the loss of information in the first case and irregularity of frequency distribution in the second case, to arrive at a suitable number of classes. Normally, the number of classes should not be less than 5 and more than 20. Prof. Sturge has given a formula : $k = 1 + 3.322 \log n$; where k refers to the number of classes and n refers to total frequencies or number of observations. The value of k is rounded to the next higher integer,

if $n = 100$ $K = 1 + 3{\cdot}322 \log 100 = 1 + 6{\cdot}644 = 8$

Further, the number or class intervals should be such that they give uniform an unimodal distribution which means that the frequencies in the given classes increase and decrease steadily and there are no sudden jumps. The number of classes should be an integer preferably 5 or multiplies of 5, 10, 15, 20, 25 etc. which are convenient for numerical computations.

3. Size of Class intervals :

Because the size of the class interval is inversely proportional to the number of classes in a given distribution, the choice about the size of the class interval will depend upon the sound subjective judgement of the statistician. An approximate value of the magnitude of the class interval say i can be calculated ***with the help of Sturge's Rule,*** where i stands for class magnitude or interval, Range refers to the difference between the largest and smallest value of the distribution, and n refers to total number of observations. If we are given the following information ; n = 400, largest item = 1300 and smallest item = 340, then the size of class intervals should be taken as 5 or multiplies of 5, 10, 15, or 20 for easy computations of various statistical measures of the frequency distribution, class intervals should be so fixed that each class has a convenient mid-point around which all the observations in that class cluster. It means that the entire frequency of the class is concentrated at the mid value of the class. It is always desirable to take the class intervals of equal or uniform magnitude throughout the frequency distribution.

4. Class Boundaries :

If in a grouped frequency distribution there are gaps between the upper limit of any class and lower limit of the succeeding class (as in case of inclusive type of classification), there is a need to convert the data into a continuous distribution by applying a correction factor for continuity for determining new classes of exclusive

type. The lower and upper class limits of new exclusive types classes are called class boundaries.

Marks	Class Boundaries
20-24	(20 – 0.5, 24 + 0.5) i.e., 19.5 – 24.5
25-29	(25 – 0.5, 29 + 0.5) i.e., 24.5 – 29.5
30-34	(30 – 0.5, 34 + 0.5) i.e., 29.5 – 34.5
35-39	(35 – 0.5, 39 + 0.5) i.e., 34.5 – 39.5
40-44	(40 – 0.5, 44 + 0.5) i.e., 39.5 – 44.5

5. Mid-value or Class Mark :

The mid value or class mark is the value of a variable which is exactly at the middle of the class. The mid-value of any class is obtained by dividing the sum of the upper and lower class limits by 2.

Mid value of a class = 1/2 [Lower class limit + Upper class limit]

The class limits should be selected in such a manner that the observations in any class are evently distributed throughout the class interval so that the actual average of the observations in any class is very close to the mid-value of the class.

6. Open End Classes :

The classification is termed as open end classification if the lower limit of the first class or the upper limit of the last class or both are not specified and such classes in which one of the limits is missing are called open end classes. For example, the classes like the marks less than 20 or age above 60 years. As far as possible open end classes should be avoided because in such classes the mid-value cannot be accurately obtained. But if the open end classes are inevitable then it is customary to estimate the class mark or mid-value for the first class with reference to the succeeding class. In other words, we assume that the magnitude of the first class is same as that of the second class.

(A) Raw Data matrix - Production ('000 kgs) of pulses against selected Villages of a Block)

46	67	23	05	12	36	63	26	48	76	56	31	58
90	32	36	59	54	48	21	58	84	68	65	59	46
53	64	57	65	53	38	58	26	43	45	66	74	16
86	43	36	66	46	58	36	64	58	45	76	74	48
64	58	50	58	95	56	66	44					

(B) Arranged Scores (Data) – Frequency Distribution Table

Class	*Tally*	*Frequency (F)*	*Cumulative Frequency (fc)— (Less than type)*	*Cumulative Frequency (fc)— (More than type)*
01–10	/	01	< 10 = 01	60 = > 01
11-20	//	02	< 20 = 01÷02=03	60 – 01 = 59 = > 10
21-30	////	04	< 30 = 03÷04=07	59 – 02 = 58 = > 20
31-40	𝍸 //	07	< 40 = 07÷07=14	57 – 04 = 53 = >30
41-50	𝍸 𝍸 //	12	< 50 = 14÷12=26	53 – 07 = 46 = >40
51-60	𝍸 𝍸 𝍸	15	< 60 = 26÷15=41	46 – 12 = 34 = >50
61-70	𝍸 𝍸 /	11	< 70 = 41÷11=52	34 – 15 = 19 = >60
71-80	////	04	< 80 = 52÷04=56	19 – 11 = 08 = >70
81-90	///	03	< 90 = 56÷03=59	08 – 04 = 04 = >80
91-100	/	01	< 100 = 59÷01=60	04 – 03 = 01 = >90

Preparation of Frequency Table with its proper explanation.

Class/class Interval (i)	Class Frequency (f)	Class Limit		Class Boundary		Mid Value (X)	Width (W)	Frequency density (f)	% Frequency
		Lower Class (Lc)	Upper Class (Uc)	Lower Class Boundary (LB)	Upper Class Boundary (UB)				
1-10	05	01	10	0.5	10.5	5.5	10	0.5	8.33
11-20	11	11	20	10.5	20.5	15.5	10	1.1	18.33
21-30	15	21	30	20.5	30.5	25.5	10	1.5	25.00
31-40	16	31	40	30.5	40.5	35.5	10	1.6	26.67
41-50	13	41	50	40.5	50.5	45.5	10	1.3	21.67
	N=60								100%

02 – Following data matrix representing the percentages of population who are worker i.e. the selected mouzas of Ghatal Block, West Midnapur District.

55	60	38	30	26	27	28	72	85	25	12	63
61	41	31	48	71	30	32	73	60	52	58	45
36	22	19	28	38	48	58	68	79	78	39	44
42	32	42	52	36	73	14	30	33	34	42	75
19	29	39	49	59	69	79	80	30	40	20	29

(i) Maximum value = 85, Minimum Value = 12
Range = 85 – 12 = 73

(ii) Number of classes = 1 + 3.3 Log N (N = 60)
$1 + 3.3 \times \log 60 = 1 + 3.3 \times 1.7781$
$\therefore$ K = 1 + 5.92 = 6.92 = 7 (approx)
So here we can consider 7 or 8 classes.
We will take 8 classes here.

(iii) For this case class interval will be

$$\text{C.I.} = \frac{\text{Range}}{\text{K}} = \frac{73}{8} = 9.125 = 10$$

Class Interval = 10

It is hereby noted that we can take 7 classes here in place of 8 and we can also consider such flexibility in relation to number of classes and class interval for our purpose of analysis.

Based on the data matrix stated in the previous page, we now construct frequency table and compute the same through following table.

A.

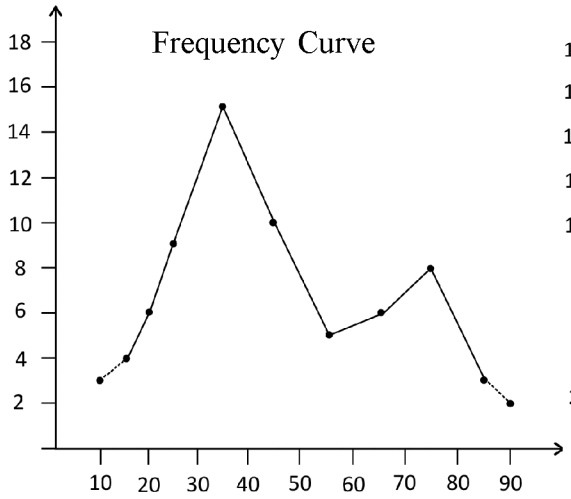
Class	Tally	f	
10-19	////	04	
20-29	////	09	N = 60
30-39	/|| /|| /||	15	This table has been prepared
40-49	/|| /||	10	by inclusive method
50-59	/||	05	
60-69	/	06	Class Interval = 10
70-79	///	08	
80-89	///	03	

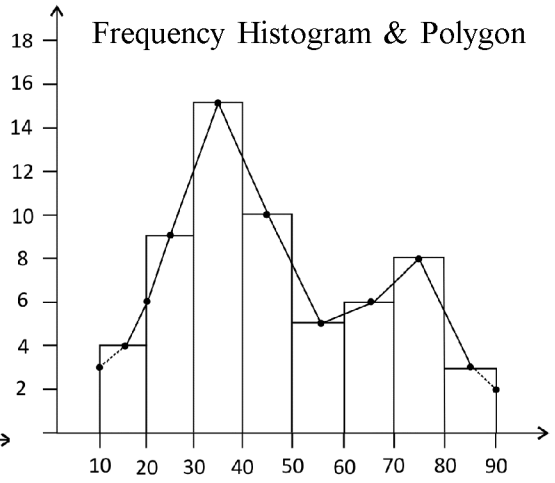
B. Frequency Table by Exclusive Method

Class	Tally	f	
10-20	////	04	N = 60
20-30	/|| ////	09	This table has been prepared
30-40	/|| /|| /||	15	by Exclusive method.
40-50	/|| /||	10	
50-60	/||	05	Here 20 exclusted from the
60-70	/	06	first class and included in the
70-80	///	08	second class
80-90	///	03	

C. Computation of Complete Frequency Table

Class	f	Midpoint (X)	Coumulative Frequency Cf	Frequency Density fd	
10-20	4	15	4	4/10 = 0.4	
20-30	9	25	4+9 = 13	9/10 = 0.9	$(X) = \frac{L+U}{2}$





30-40	15	35	13+15 = 28	15/10 = 0.15	$= \frac{10+20}{2}$
40-50	10	45	28+10 = 38	10/10 = 1.00	= 15
50-60	5	55	38+5 = 43	5/10 = 0.5	Calculation of
60-70	6	65	43+6 = 49	6/10 = 0.6	Midpoint
70-80	8	75	49+8 = 57	8/10 = 0.8	L = Lower Limit
80-90	3	85	57+3 = 60	3/10 = 0.3	U = Upper Limit

D. Cumulative Frequency

Class	Less than	f	More than	f	
10-20	20	4	10	60	> 10 = 60
20-30	30	13	20	57	> 20 = 60 – 4 = 56
30-40	40	28	30	47	> 30 = 56 – 9 = 47
40-50	50	38	40	32	> 40 = 47 – 15 = 32

50-60	60	43	50	22	> 50 = 32 – 10 = 22
60-70	70	49	60	17	> 60 = 22 – 5 = 17
70-80	80	57	70	11	> 70 = 17 – 6 = 11
80-90	90	60	80	3	> 80 = 11 – 8 = 3
			90	0	> 90 = 3 – 3 = 0

Smoothed Frequency Distribution

Class	f	Smoothed
10-20	$4 - f_1$	4.33
20-30	$9 - f_2$	9.33
30-40	$15 - f_3$	11.33
40-50	$10 - f_4$	10
50-60	$5 - f_5$	7
60-70	$6 - f_6$	6.33
70-80	$8 - f_7$	5.67
80-90	$3 - f_8$	3.67

S_f = Smoothed frequency

For class – 10 – 20 =

$f_0 + f_1 + f_2 / 3 = 4.33$

For class – 20 – 30 =

$f_1 + f_2 + f_3 / 3 = 9.33$

For class – 30 – 40 =

$f_2 + f_3 + f_4 / 3 = 11.33$

For class – 80 – 90 =

$f_7 + f_8 + f_0 / 3 = 3.67$

Class	f	% f	Probability f
10-20	4	6.67	0.07
20-30	9	15	0.15
30-40	15	25	0.25
40-50	10	16.67	0.17
50-60	5	8.33	0.08
60-70	6	10	0.10
70-80	8	13.33	0.13
80-90	3	5	0.05
	N = 60	100%	1.00

N = 60 = 100%

% frequency =

$$\frac{\text{Class frequency}}{\text{N}} \times 100$$

Probability f =

$$\frac{\text{Class frequency}}{\text{N}}$$

For % f – N = 100%

For probability f – N = 1

Presentation of Different types of Frequency Curve, Polygon, Histogram Frequency Curve

A. Following Table

No. of Days absent	**12-15**	**15-18**	**18-21**	**21-24**	**24-27**	**27-30**	**30-33**	**33-36**
No. of Labours	**8**	**23**	**17**	**18**	**8**	**4**	**01**	**01**

Class	Mid point	*f*	f_{sm} — N = 80 / *f*	%*f*
12-15	13.5	8	0 + 8 + 23/3 = 10.3	10
15-18	16.5	23	8 + 23 + 17/3 = 16	28.75
18-21	19.5	17	23 + 17 + 18/3 = 19.33	21.25
21-24	22.5	18	17 + 18 + 8/3 = 14.33	22.5
24-27	25.5	8	18 + 8 + 4/3 = 30	10
27-30	28.5	4	8 + 4 + 1/3 = 4.33	5
30-33	31.5	01	4 + 1 + 1/3 = 2	1.25
33-36	34.5	01	1 + 1 + 0/3 = 0.67	1.25
				100%

Following the table (A) Frequency curve and Frequency Polygon will be drawn. Smoothed Frequency curve will also be drawn from the same table.

X = Mid point, f = frequency of the class fsm = Smoothed frequency

(i) Diagram for Frequency curve

(ii) Diagram for Smoothed frequency curve

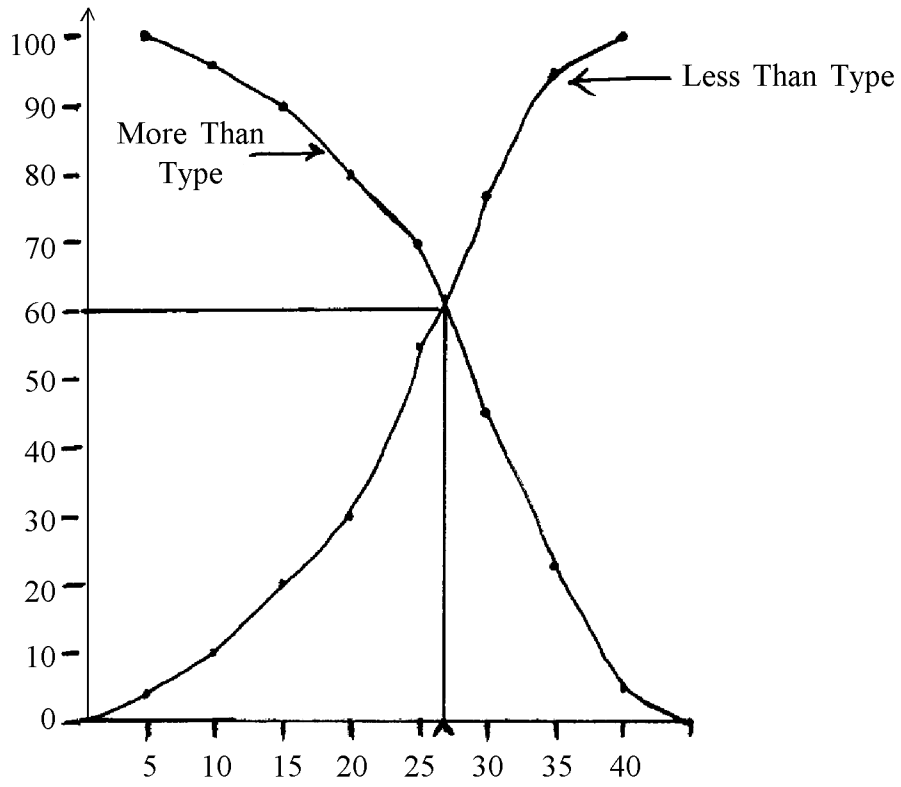
(iii) Diagram for Frequency Polygon

(iv) % Frequency Curve.

(iii) For more than and Less than type agive in one net.

Less than Type	Cf	More than Type	cf
5	4	0	100
10	10	5	96
15	20	10	90
20	30	15	80
25	55	20	70

30	77	30	23
35	95	35	5
40	100	40	0



OGIVE- Cumulative Frequency Curve

2.5 Summary

The frequency table gives the description of the various attributes of the frequency distribution which facilitate the analysis of data.

Unit 3 ❑ Measures of Central Tendency

Structure

3.1 Objectives

- The learners will know about the measures of central tendency and their use.

3.2 Introduction

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. The mean (often called the average) is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode. **The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.**

3.3 Mean

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data. The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have in values in a data set and they have values x_1, x_2, ..., x_n, the sample mean, usually denoted by $\bar{X}$ (pronounced x bar), is :

$$\bar{X} = \frac{(x_1 + x_2 + ... + x_n)}{n}$$

This formula is usually written in a slightly different manner using the Greek capital letter, $\sum$, pronounced "sigma". which means "sum of....".

$$\bar{x} = \frac{\sum x}{n}$$

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

When not to use the mean

The mean has one main disadvantage that it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below :

Staff	01	02	03	04	05	06	07	08	09	10
Salary ('000)	15	18	16	14	15	15	12	17	90	95

The mean salary for these ten staff is Rs. 30,000 (approx). However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers hlave salaries in the 12,000 to 18,000 range. The mean is being skewed by the two large salaries. Therefore, in this situation, we would like to have a better measure of central tendency in this situation.

Another time ***when we usually prefer the median over the mean (or mode) is when our data is skewed (i.e., the frequency distribution for our data is skewed).*** If we consider the normal distribution as this is the most frequently assessed in statistics— when the data is perfectly normal, the mean, median and mode are identical. However, as the data becomes skewed the mean loses its ability to provide the best central location for the data because the skewed data is dragging it away from the typical value. However, the median best retains this position and is not as strongly influenced by the skewed values.

3.4 Median

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers an skewed data. In order to calculate the median, suppose we have the data below :

65 55 89 56 35 14 56 55 87 45 92

We first need to rearrange that data into oder of magnitude (Ascending order) :

14 35 45 55 55 **56** 56 65 87 89 92

Our median mark is the middle mark— in this case, 56 (highlighted in bold). It is the middle mark because there are 5 scores before and after it. But what happens when you have an even number of scores ? So, for this we can take the example below :

65 55 89 56 35 14 56 55 87 45

We again rearrange that data into order of magnitude (Ascending order) :

14 35 45 55 55 56 56 65 87 89

Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

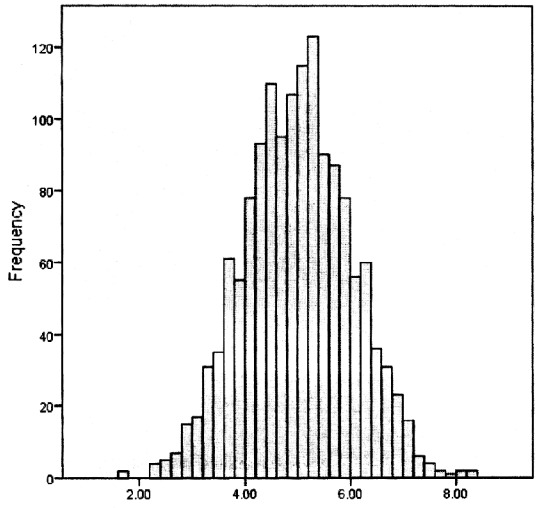
3.5 Mode

The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram. You can, therefore, sometimes consider the mode as being the most popular option. An example of a mode is presented below. Normally, the mode is used for categorical data where we wish to know which is the most common category, as illustrated by data matrix presented below.

We are now stuck as to which mode best describes the central tendency of the data. This is particularly problematic when we have continuous data because we are more likely not to have any one value that is more frequent than the other. For example, consider measuring 30 peoples' weight (to the nearest 0.1 kg). How likely is it that we will find two or more people with exactly the same weight (e.g., 67.4 kg)? The answer, is probably very unlikely– many people might be close, but with such a small smaple (30 people) and a large range of possible weights, you are unlikely to find two people with exactly the same weight ; that is, to the nearest 0.1kg. This is why the mode is very rarely used with continuous data. Another problem with the mode is that it will not provide us with a very good measure of central tendency when the most common mark is far away from the rest of the data in the data set, as depicted in the presentation from data matrix.

Skewed Distributions and the Mean and Median

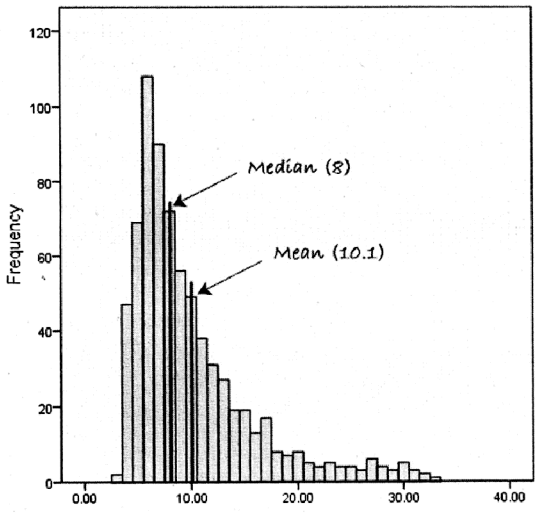
We often test whether our data is normally distributed because this is a common assumption underlying many statistical tests. An example of a normally distributed set of data is presented below. When you have a normally distributed sample you can legitimately use both the mean or the median as your measure of central tendency. In fact, in any symmetrical distribution the mean, median and mode are equal. However, in this situation, the mean is widely preferred as the best measure of central tendency because it is the measure that includes all the value in the data set for its calculation,



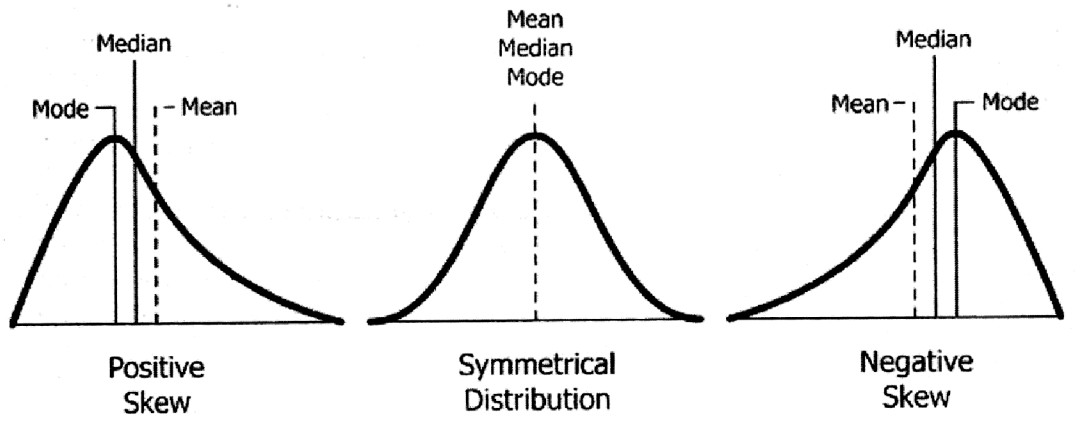
Example of Normal Distribution

and any change in any of the scores will affect the value of the mean. This is not the case with the median or mode.

We find that the mean is being dragged in the direct of the skew. In these situations, the median is generally considered to be the best representative of the central location of the data. The more skewed the distribution, the greater the difference between the median and mean, and the greater emphasis should be placed on using the median as opposed to the mean. A classic example of the above right-skewed distribution is income (salary), where higher-earners provide a false representation of the typical income if expressed as a mean and not a median.



Example of Skewed Distribution



However, when our data is skewed, for example, as with the right-skewed data set below :

Presentation of different types of Frequency Distribution and the relative position of Mean, Median and Mode.

Table – 3.1

Calculation and Presentation of Measures of Central Tendency from the data matrix.

(A) Raw Data matrix-production ('000 kgs) of Pulses against selected villages of a Block)

46	67	23	05	12	36	63	26	48	76	56	31	58
90	32	36	59	54	48	21	58	84	68	65	59	46
53	64	57	65	53	38	58	26	43	45	66	74	16
86	43	36	66	46	58	36	64	58	45	76	74	48
64	58	50	58	95	56	66	44					

(B) Arranged Scores (Data)- Frequency Distribution Table (Table –3.2)

Class	*Tally*	*Frequency (F)*	*Cumulative Frequency (fc)— (Less than type)*	*Cumulative Frequency (fc)— (More than type)*
01–10	/	01	< 10 = 01	60 = > 01
11-20	//	02	< 20 = 01+02=03	60 – 01 = 59 = > 10
21-30	////	04	< 30 = 03+04=07	59 – 02 = 58 = > 20

31-40	𝍸 //	07	< 40 = 07+07=14	57 – 04 = 53 = >30
41-50	𝍸 𝍸 //	12	< 50 = 14+12=26	53 – 07 = 46 = >40

Class	*Tally*	*Frequency (F)*	*Cumulative Frequency (fc)— (Less than type)*	*Cumulative Frequency (fc)— (More than type)*
51-60	𝍸 𝍸 𝍸	15	< 60 = 26+15=41	46 – 12 = 34 = >50
61-70	𝍸 𝍸 /	11	< 70 = 41+11=52	34 – 15 = 19 = >60
71-80	////	04	< 80 = 52+04=56	19 – 11 = 08 = >70
81-90	///	03	< 90 = 56+03=59	08 – 04 = 04 = >80
91-100	/	01	< 100 = 59+01=60	04 – 03 = 01 = >90

(Table –3.3)

Class	**frequency**	**Cf**	**X(Mid Point)**	**fx**	**Result**
10-20	2	2	15	30	
20-30	4	6	25	100	N = 30 = Σf
30-40	8	14	35	280	Σfx = 1210
40-50	10	24	45	450	
50-60	4	28	55	220	
60-70	2	30	65	130	

(i) Mean ($\bar{X}$) = Σ fx/N = 1210/30 = 40.33

(ii) Calculation and Presentation of Median (Me)

From the given table (A)

$$Me = \frac{N}{2} = \frac{30}{2} = 15 \text{ th value}$$

So, From Table 3.4

(Table –3.4)

Class	f	Cf Less than	Cf More than	Results
10-20	2	0	30	$\frac{N}{2} = \frac{30}{2} = 15$
		2	28	

20-30	4	6	24	$L_1 = 40$, i = 10
30-40	8	14	16	$f_m = 10$
40-50	10(me)	24	6	$f_c = 14$
50-60	4	28	2	M_c = Model class = 40 – 50
				M_f = Frequency of model
60-70	2	30	0	Class = 10

$$\text{Me} = L_1 + \frac{\frac{N}{2} - f_c}{fm} \times i \rightarrow 40 + \frac{15-14}{10} \times 10 = 40 + \frac{1}{10} \times 10 = 40 + 1 = 41 \text{ (Median)}$$

$$\text{Mo (Mode)} = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

Δ_1 = Difference between Model class and Previous class

Δ_2 = Difference between Model class and Following class.

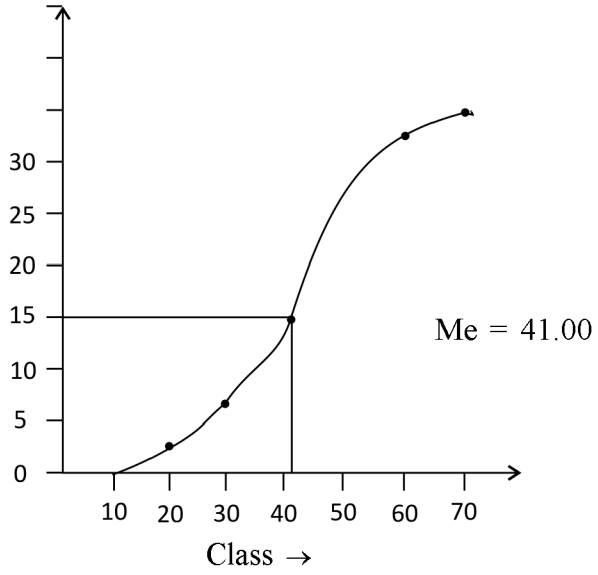
$$\text{So, Mo} = 40 + \frac{(10-8)}{(10-8)+(10-4)} \times i$$

$$= 40 + \frac{2}{2+6} \times 10 = 40 + \frac{2}{8} \times 10 = 40 + 2.5 = 42.5$$

So, $\bar{x} = 40.33$
Me = 41.00
Mo = 42.50

All will be represented Graphically and can be frquency curve.

Graphical representation of Median (Me) (Table 3.4)



Graphical representation of Mode (Mo) (Table 3.4)

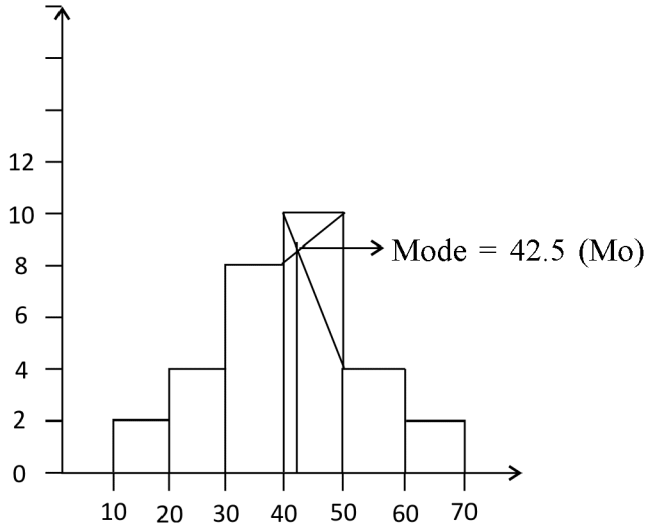
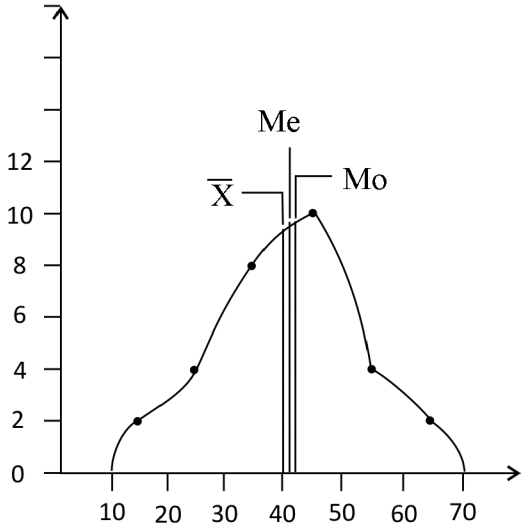


Table 3.4

Graphical representation of X, Me and Mo on Normal Frequency Curve



3.6 Summary

In conclusion it can be said that the existence of outliers in a distribution, the mean can still be an appropriate measure of central tedency. Several common regression techniques can help reduce the influence of outliers on the mean value.

Unit 4 ❑ Measures Of Dispersion

Structure

4.1 Objective

- The learners will learn about different measures of dispersion and their characteristics.

4.2 Introduction

Suppose you are given a data series. Someone asks you to tell some interesting facts about thid data series. How can you do so ? You can say you can find the mean, the median or the mode of this data series and tell about its distribution. But is it the only thing you can do ? Are the central tendencies the only way by which we can get to know about the concentration of the observation ? The answer is no, so, we have to go through Measures of dispersion. As the name suggests, the measure of dispersion shows the scattering of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data. The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations.

Suppose you have four datasets of the same size and the mean is also same. In all the cases the sum of the observations will be the same. Here, the mesaure of central tendency is not giving, a clear and complete idea about the distribution for the four given sets.

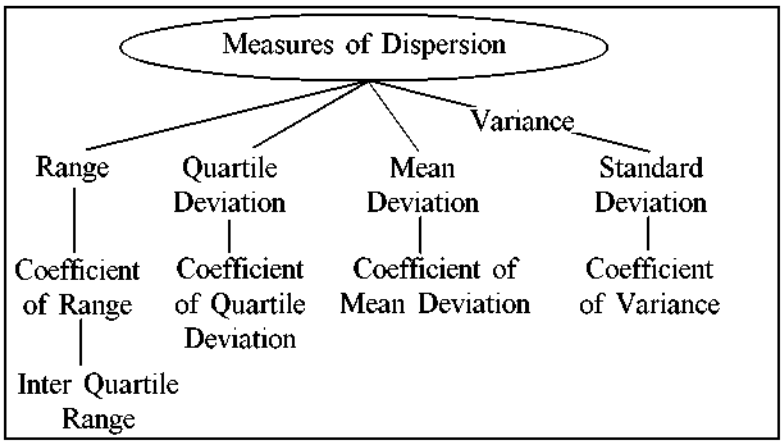
Can we get an idea about the distribution if we get to know abut the dispersion of the observations from one another within and between the datasets ? The main idea about the measure of dispersion is to get to know how the data are spread. It shows how much the data vary from their average value.

4.2 Characteristics of Measures of Dispersion

- A measure of dispersion should be rigidly defined.
- It must be easy to calculate and understand
- Not affected much by the fluctuations of observations
- Based on all observations

4.3 Classification of Measures of Dispersion

Measures of dispersion can be classified in the following manner—



The measure of dispersion is categorized as :

(i) An absolute measure of dispersion :

The measures which express the scattering of observation ***in terms of distances i.e., range, quartile deviation.***

The measure which expresses the variations ***in terms of the average of deviations of observations like mean deviation and standard deviation.***

(ii) A relative measure of dispersion :

We use a relative measure of dispersion for comparing distribution of two or more data set and for unit free comparison. ***They are the coefficient of range, the coefficient of mean deviation, the coefficient of quartile deviation, the coefficient of variation, and the coefficient of standard deviation.***

Range

A range is the most common and easily understandable measures of dispersion.

It is the difference between two extreme observations of the data set. If X_{max} and X_{min} are the two extreme observations then

$$\text{Range} = X_{max} - X_{min}$$

Merits of Range

- It is the simplest of the measure of dispersion
- Essay to calculate
- Easy to understand
- Independent of change of origin

Demerits of Range

- It is based on two extreme observations. Hence, get affected by fluctuations
- A range is not a reliable measure of dispersion
- Dependent on change of scale

Quartile Deviation

The quartiles divide a data set into quarters. The first quartile, (Q_1) is the middle number between the smallest number and the median of the data. The second quartile, (Q_2) is the median of the data set. The third quartile, (Q_3) is the middle number between the median and the largest number. Quartile deviation or semi-intger-quartile deviation is

$$Q = {}^1/_2 \times (Q_3 - Q_1)$$

Merits of Quartile Deviation

- All the drawbacks of Range are overcome by quartile deviation
- It uses half of the data
- Independent of change of origin
- The best measure of dispersion for open-end classification

Demerits of Quartile Deviation

- It ignores 50% of the data
- Dependent on change of scale
- Not a reliable measure of dispersion

Mean Deviation

Mean deviation is the arithmetic mean of the absolute deviations of the observations from a measure of central tendency. If x_1, x_2, ..., x_n are the set of observation, then the mean deviation of x about the average A (mean, median, or mode) is

Mean deviation from average A = $1/n\ [\Sigma_i|x_i - A|]$

For a grouped frequency, it is calculated as :

Mean deviation from average A = $1/N\ [\Sigma_i\ f_i\ |x_i - A|]$, $N = \Sigma f_i$

Here, x_i and f_i are respectively the mid value and the frequency of the i^{th} class interval.

Merits of Mean Deviation

- Based on all observations
- It provides a minimum value when the deviations are taken from the median
- Independent of change of origin

Demerits of Mean Deviation

- Not easily understanadable
- Its calculation is not easy and time-consuming
- Dependent on the change of scale
- Ignorance of negative sign creates artificiality and becomes useless for further mathematical treatment

Standard Deviation

A standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by a Greek letter sigma. σ, It is also referred to as root mean square deviation. The standard deviation is given as

$$\sigma = [(\Sigma_i (y_i - \bar{y})/n]^{1/2} = [(\Sigma_i\ y_i^2/n) - \bar{y}^2]^{1/2}$$

For a grouped frequency distribution, it is

$$\sigma = [(\Sigma_i\ f_i (y_i - \bar{y})/N]^{1/2} = [(\Sigma_i f_i\ y_i^2/n) - \bar{y}^2]^{1/2}$$

The square of the standard deviation is the variance. It is also a measure of dispersion.

$$\sigma^2 = [(\Sigma_i (y_i - \bar{y})/n]^{1/2} = [(\Sigma_i\ y_i^2/n) - \bar{y}^2]$$

For a grouped frequency distribution, it is

$$\sigma^2 = [(\Sigma_i f_i (y_i - \bar{y})/N]^{1/2} = [(\Sigma_i f_i\ x_i^2/n) - \bar{y}^2]$$

If instead of a mean, we choose any other arbitrary number, say A, the standard deviation becomes the root mean deviation.

Merits of Standard Deviation

- Squaring the deviations overcomes the drawback of ignoring signs in mean deviations

- Suitable for further mathematical treatment
- Least affected by the fluctuation of the observations
- The standard deviation is zero if all the observations are constant
- Independent of change of origin

Demerits of Standard Deviation

- Not easy to calculate
- Difficult to understand for a layman
- Dependent on the change of scale

4.4 Coefficient of Dispersion

Whenever we want to compare the variability of the two series which differ widely in their averages.

Also, when the unit of measurement is different. We need to calculate the coefficients of dispersion along with the measure of dispersion. The coefficients of dispersion (C.D.) based on different measures of dispersion are

Based on Range $= (X_{max} - X_{min}) / (X_{max} + X_{min})$.

C.D. based on quartile deviation $= (Q_3 - Q_1) / (Q_3 + Q_1)$.

Based on mean deviation = Mean deviation / average from which it is calculated.

For Standard deviation = S.D. / Mean

Measures of Dispersion and Their Presentation Table-T

Class	f	Cf	Q	Results
5-10	18	18		
10-15	30	48	Q_1	$N = 160$
15-20	46	94		Q_1 class $-10-15$
20-25	28	122	Q_3	Q_3 class $-20-25$
25-30	20	142		$Q_1 = \frac{N}{4} = \frac{160}{4} = 40$th value $(10-15)$
30-35	12	154		$Q_3 = \frac{3N}{4} = \frac{3\times160}{4} = 120$th value $(20-25)$
35-40	06	160		

(i) $Q_1 = L_{Q_1} + \frac{N/4 - f_c}{f_{Q_1}} \times i = 10 + \frac{\frac{160}{4} - 18}{30} \times 5$

$$= 10 + \frac{40-18}{30} \times 15 = 10 + \frac{22}{30} \times 5$$

$$Q_1 = 10 + 3.67 = \boxed{13.67}$$

(ii) $$Q_3 = L_{Q_3} + \frac{\frac{3N}{4} - f_c}{fQ_3} \times i = 20 + \frac{\frac{3\times160}{4} - 94}{28} \times 5$$

$$= 20 + \frac{120-94}{28} \times 5 = 20 + \frac{26}{28} \times i$$

$$Q_3 = 20 + 4.64 = \boxed{24.64}$$

(iii) $$Q_D = \frac{Q_3 - Q_1}{2} = \frac{24.64 - 13.67}{2}$$

$$Q_D = \frac{10.97}{2} = 5.485 = 5.49$$

Table – B

From the following table calculate and represent Q_1, Q_3, D_6 and P_{80}.

Class	*f*	*Cf*	*Results*
40-50	4	4	N = 95
50-60	12	16	Q_1 = N/4 = 95/4 = 23.75th
60-70	18	34	Q_3 = 3N/4 = $\frac{3\times95}{4} = \frac{285}{4} = 71.25$ th
70-80	25	59	
80-90	20	79	$D_6 = \frac{6N}{10} = \frac{6\times95}{10} = \frac{570}{10} = 57$ th
90-100	10	89	
100-110	6	95	$P_{80} = \frac{80N}{100} = \frac{80\times95}{100} = 76$ th
Q_1 class = 60-70, Q_3 class = 80-90, D_6 class = 70-80, P_{80} class = 80-90			

(i) $Q_1 = L_1 + \frac{N/4 - f_c}{fQ_1} \times i$

$$Q_1 = 60 + \frac{94/4 - 16}{18} \times 10 = 60 + \frac{23.75 - 16}{18} \times 10 = 60 + \frac{7.75}{18} \times 10 = 60 + 4.31 = \boxed{64.31} = Q_1$$

$$Q_3 = L_1 + \frac{\frac{3N}{4} - 59}{20} \times 10$$

$$= 80 + \frac{\frac{3 \times 95}{4} - 59}{20} \times 10 = 80 + \frac{71.25 - 59}{20} \times 10$$

$$= 80 + \frac{12.25}{2} = 80 + 6.13 = 86.13 \quad Q_3 = \boxed{86.13} = Q_3$$

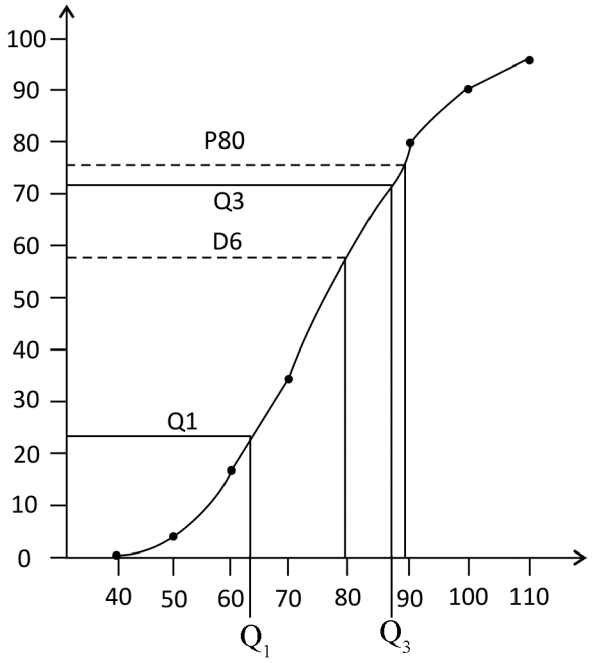
(ii) $$D_6 = L_1 + \frac{\frac{6N}{10} - f_c}{f_{D6}} \times i = 70 + \frac{\frac{6 \times 95}{10} - 34}{25} \times 10 = 70 + \frac{57 - 34}{25} \times 10$$

$$= 70 + \frac{23}{25} \times 10 = 70 + \frac{230}{25} = 70 + 9.20$$

$$\boxed{D_6 = 70 + 9.20 = 79.20 (D_6)}$$

(iii) $$P_{80} = L_1 + \frac{\frac{80N}{100} - f_c}{f_{p80}} \times i = 80 + \frac{\frac{80 \times 95}{100} - 59}{20} \times 10 = 80 + \frac{76 - 59}{20} \times 10 = 80 + \frac{17}{2} = 80 + 8.50 = 88.50$$

$$\boxed{P80 = 88.50}$$



For representation See Diagram.

4.4 Summary

Each of the measures of central tendency describes a different indication of the typical or central value in the distribution. It also condenses the data set down to one representative value, which is useful when one is working with large amounts of data.

Unit 6 ❑ Plotting of Scatter Diagram and Regression Line based on Sample Data

Structure

6.1 Objective

- The learners will learn about the plotting of scatter diagram and regression line.

6.2 Introduction

Scatter diagram is a graphical representation of a suitable pair of data in form of dots. Each pair is denoted by x and y variables the x is noted as independent and y is noted by dependent variable, though always it is not possible to determine which one is x and which one y.

6.3 Scatter Diagram

The scatter diagram serves as a useful tool in the study of relation and also for assessing how marked the relationship is. We will try to represent the following set of data for scatter diagram.

X	56, 73, 65, 80, 35, 62, 36, 40, 92, 45
Y	75, 80, 56, 82, 45, 65, 30, 25, 90, 50

At a glance the scatter diagram reveals that there exists a tendency for small values of x to be associated with small values of y, and so also for large values of x and y.

6.4 Regression

From correlation and association it is possible to get the degree of relationship i.e., distance between two variables.

Regression analysis is used to model and analyse numerical data consisting of values of an independent variable X (the variable that we fix or choose deliberately) and dependent variable Y.

The main purpose of finding a relationship is that the knowledge of the relationship may enable events to be predicted and perhaps controlled.

6.5 Correlation Coefficient

To measure the strength of the linear relationship between X and Y the sample correlation coefficient r is used.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\, S_{xy}}},$$

$$S_{xy} = n\Sigma xy - \Sigma x \Sigma y,$$

$$S_{xx} = n\Sigma x^2 - (\Sigma x)^2,\ S_{yy} = n\Sigma y^2 - (\Sigma y)^2$$

where x and y observed values of variables X and Y respectively.

Important notes

(1) If the calculated r value is positive then the slope will rise from left to right on the graph. If the calculated values of r is negative the slope will fall from left to right.

(2) The r value will always lie between— 1 and + 1. If you have an r value outside of this range you have made an error in the calculations.

(3) Remember that a correlation does not necessarily demonstrate a causal relationship. A significant correlation only shows that two factors vary in a related way (positively or negatively).

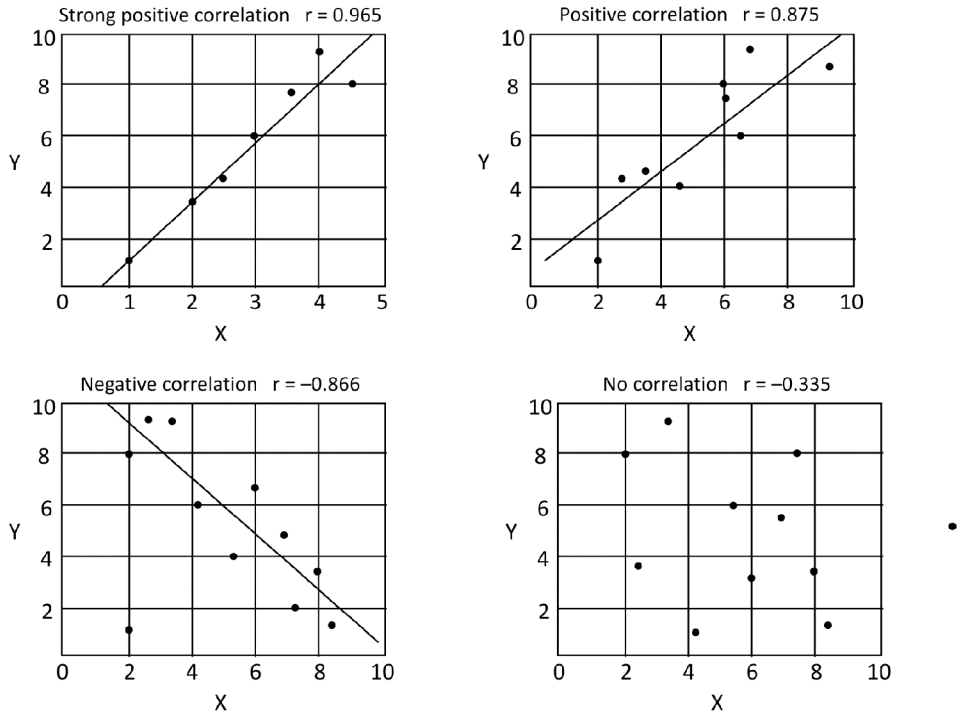
(4) The formula above can be rewritten as

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y},\quad \sigma_x = \sqrt{\frac{1}{n}\Sigma x^2 - \bar{x}^2},\quad \sigma_y = \sqrt{\frac{1}{n}\Sigma y^2 - \bar{y}^2}$$

$$\sigma_{xy} = \frac{1}{n}\Sigma xy - \overline{xy},\quad \bar{x} = \frac{1}{n}\Sigma x,\quad \bar{y} = \frac{1}{n}\Sigma y$$

Scatter Diagrams

Scatter diagrams are used to graphically represent and compare two sets of data. The INDEPENDENT VARIABLE is usually plotted on the X axis. The **dependent variable** is plotted on the Y axis. By looking at a scatter diagram, we can see whether there is any connection (correlation) between the two sets of data. A scatter plot is a useful summary of a set of bivariate data (two variables), usually drawn before working out a linear correlation coefficient or fitting a regression line. It gives a good visual picture of the relationship between the two variables, and aids the interpretation of the correlation coefficient or regression model.



From plots one can see that if the more the points tend to cluster around a straight line and the higher the correlation (the stronger the linear relationship between the two variables). If there exists a random scatter of points, there is no relationship between the two variables (very low or zero correlation).

Very low or zero correlation could result from a non-linear relationship between the variables. If the relationship is in fact non-linear (points clustering around a curve, not a straight line), the correlation coefficient will not be a good measure of the strength. A scatter plot will also show up a **non-linear relationship** between the two variables and whether or not there exist any outliers in the data.

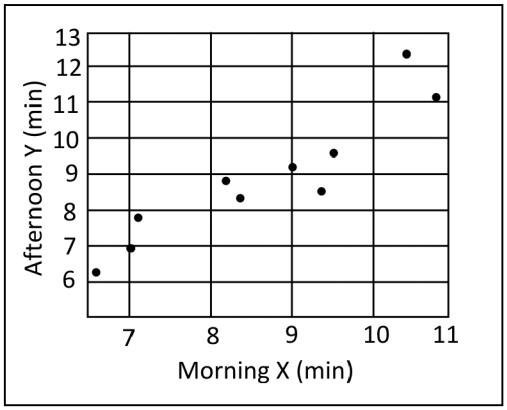
Example 1

Determine on the basis of the following data whether there is a relationship between the time, in minutes, it takes a person to complete a task in the morning X and in the late afternoon Y.

Morning (x) (min)	8.2	9.6	7.0	9.4	10.9	7.1	9.0	6.6	8.4	10.5
Afternoon (y) (min)	8.7	9.6	6.9	8.5	11.3	7.6	9.2	6.3	8.4	12.33

Solution :

The data set consists of n = 10 observations.



Step 1.

To construct the scatter diagram for the given data set to see any correlation between two sets of data.

From the scatter diagram we can conclude that it is likely that there is a linear relationship between two variables.

Step 2.

Set out a table as follows and calculate all required values

Σx, Σy, Σx^2, Σy^2, Σxy.

Morning (x) (min)	Afternoon (y) (min)	x^2	y^2	xy
8.2	8.7	67.24	75.69	71.34
9.6	9.6	92.16	92.16	92.16
7.0	6.9	49.00	47.61	48.30
9.4	8.5	88.36	72.25	79.90
10.9	11.3	118.81	127.69	123.17
7.1	7.6	50.41	57.76	53.96
9	9.2	81.00	84.64	82.80
6.6	6.3	43.56	39.69	41.58
8.4	8.4	70.56	70.56	70.56
10.5	12.33	110.25	151.29	129.465
$\Sigma x = 86.7$	$\Sigma y = 88.8$	$\Sigma x^2 = 771.35$	$\Sigma y^2 = 819.34$	$\Sigma xy = 792.92$

Step 3.

Calculate

$$S_{xy} = n\sum xy - \sum x \sum y = 10 \times 792.92 - 86.7 \times 88.8 = 230.24$$

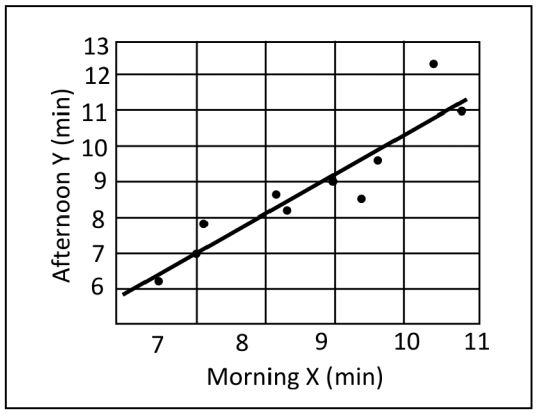
$$S_{xx} = n\sum x^2 - (\sum x)^2 = 10 \times 771.35 - (86.7)^2 = 196.61$$

$$S_{yy} = n\sum y^2 - (\sum y)^2 = 10 \times 819.34 - 88.8^2 = 307.96$$

Step 4.

Finally we obtain correlation coefficient r

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}} = \frac{230.24}{\sqrt{196.61 \times 307.96}} = 0.9357$$



The correlation coefficient is closed to 1 therefore the linear relationship exists between the two variables.

It would be tempting to try to fit a line to the data we have just analysed – producing an equation that shows the relationship. The method for this is called **linear regression.** By using linear regression method the line of best fit is

Regression equation : y = 1.171x – 1.273

This line is shown on the above graph.

6.5 Linear regression analysis : fitting a regression line to the data

When a scatter plot indicates that there is a strong linear relationship between two variables (confirmed by high correlation coefficient), we can fit a straight line to this data which may be used to predict a value of the dependent variable, given the value of the independent variable.

Recall that the equation of a **regression line** (straight line) is

$$y = a + bx$$

$$b = \frac{S_{xy}}{S_{xx}} \quad a = \bar{y} - b\bar{x} = \frac{\sum_i y_i - b\sum_i x_i}{n}$$

So, let us consider the following data

Example 2

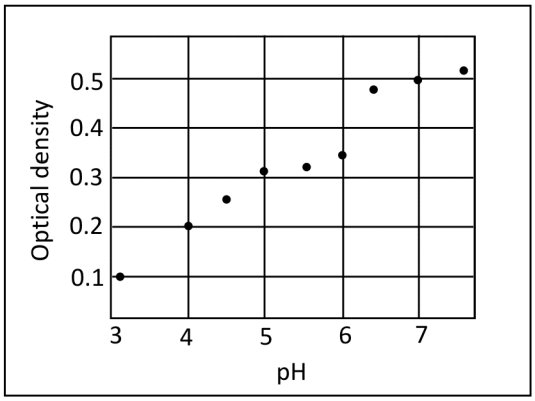
Suppose that we had the following results from an experiment in which we measured the growth of a cell culture (as optical density) at different pH levels.

pH	3	4	4.5	5	5.5	6	6.5	7	7.5
Optical density	0.1	0.2	0.25	0.32	0.33	0.35	0.47	0.49	0.53

Find the equation to fit these data.

Solution :

We can follow the same procedures for correlation, as before



The data set consists of n = 9 observations.

Step 1 : To construct the scatter diagram for the given data set to see any correlation between two sets of data.

These results suggest a linear relationship.

Step 2.

Set out a table as follows and calculate all required values

Σx, Σy, Σx^2, Σy^2, Σxy.

pH (x)	Optical density (y)	x^2	y^2	xy
3	0.1	9	0.01	0.3
4	0.2	16	0.04	0.8
4.5	0.25	20.25	0.0625	1.125
5	0.32	25	0.1024	1.6
5.5	0.33	30.25	0.1089	1.815
6	0.35	36	0.1225	2.1
6.5	0.47	42.25	0.2209	3.055
7	0.49	49	0.240	3.43
7.5	0.53	56.25	0.281	3.975
x = 49	y = 3.04	x^2 = 284	y^2 = 1.1882	xy = 18.2
x = 5.444	y = 0.3378			

Step 3.

Calculate

$$S_{xy} = n\sum xy - \sum x \sum y = 9 \times 18.2 - 49 \times 3.04 = 163.8 - 148.96 = 14.84.$$

$$S_{xx} = n\sum x^2 - (\sum x)^2 = 2556 - 2401 = 155.$$

$$S_{yy} = n\sum y^2 - (\sum y)^2 = 10.696 - 9.242 = 1.454$$

Step 4.

Finally we obtain correlation coefficient r

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}} = \frac{14.84}{\sqrt{155 \times 1.454}} = 0.989$$

The correlation coefficient is closed to 1 therefore it is likely that the linear relationship exists between the two variables. To verify the correlation r we can run a hypothesis test.

6.7 Regression Analysis

Regression analysis is a powerful statistical method that allows you to examine the relationship between two or more variables of interest.

While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable.

Regression analysis provides detailed insight that can be applied to further improve products and services.

What is regression analysis and what does it mean to perform a regression ?

Regression analysis is a reliable method of identifying which variables have impact on a topic of interest. The process of performing a regression allow you to confidently determine which factors matter most, which factors can be ignored, and how these factors influence each other.

In order to understand regression analysis fully, it's essential to comprehend the following terms :

Dependent Variable : This is the main factor that you're trying to understand or predict.

Independent Variables : These are the factors that you hypothesize have an impact on you dependent variable.

The methods of regression analysis are explained in the following paragraphs.

6.8 Scatter Diagram Method

Definition : The Scatter Diagram Method is the simplest method to study the correlation between two variables wherein the values for each pair of a variable is plotted on a graph in the form of dots thereby obtaining as many points as the numbe of observations. Then by looking at the scatter of several points, the degree of correlation is ascertained.

The degree to which the variables are related to each other depends on the manner in which the points are scattered over the chart. The more the points plotted are scattered over the chart, the lesser is the degree of correlation between the variables.The more the points plotted are closer to the line, the higher is the degree of correlation. The degree of correlation is denoted by "r".

The following types of scatter diagrams tell about the degree of correlation between variable X and variable Y.

1. Perfect Positive Correlation (r = + 1) : The correlation is said to be perfectly positive when all the points lie on the straight line rising from the lower left-hand corner to the upper right-hand corner.

2. Perfect Negative Correlation (r = –1) : When all the points lie on a straight line falling from the upper left-hand corner to the lower right-hand corner, the variables are said to be negatively correlated.

3. High Degree of +Ve Correlation (r = + High) : The degree of correlation is high when the points plotted fall under the narrow band and is said to be positive when these show the rising tendency from the lower left-hand corner to the upper right-hand corner.

4. High Degree of –Ve Correlation (r = –High) : The degree of negative correlation is high when the point plotted fall in the narrow band and show the declaining tendency from the upper left-hand corner to the lower righ-hand corner.

5. Low Degree of +Ve Correlation (r = +Low) : The correlation between the

variables is said to be low but positive when the points are highly scattered over the graph and show a rising tendency from the lower left-hand corner to the upper right-hand corner.

6. Low Degree of –Ve Correlation (r = + Low) : The degree of correlation is low and negative when the points are scattered over the graph and the show the falling tendency from the upper left-hand corner to the lower right-hand corner.

7. No Correlation (r = 0) : The variable is said to be unrelated when the points are haphazardly scattered over the graph and do not show any specific pattern. Here the correlation is absent and hence r = 0.

Thus, the scatter diagram method is the simplest device to study the degree of relationship between the variables by plotting the dots for each pair of variable values given. The chart on which the dots are plotted is also called as a **Dotogram.**

Scatter Diagram

The Scatter Diagram is known by many names, such as scatter plot, scatter graph, and correlation chart. This diagram is drawn with two variables, usually the first variables is independent and the second variable is dependent on the first variable. The scatter diagram is used to find the correlation between these two variables. This diagram helps you determine how closely the two variables are related. After determining the correlation between the variables, you can then predict the behaviour of the dependent variable based on the measure of the independent variable. This chart is very useful when one variable is easy to measure and the other is not.

Example from daily incident

You are analyzing the pattern of accidents on a highway. You select the two variables : motor speed and number of accidents, and draw the diagram.

Once the diagram is completed, you notice that as the speed of vehicle increases, the number of accidents also goes up. This shows that there is a relationship between the speed of vehicles and accidents happening on the highway.

According to the PMBOK* Guide 6th edition, "A scatter diagram is a graph that shows the relationship between two variables. Scatter diagrams can demonstrate a relationship between any element of a process, environment, or activity on one axis and a quality defect on the other axis."

Since this diagram shows you the correlation between the variables, it is also known as a correlation chart.

Usually the independent variable is plotted along the horizontal axis (abcissa) (x-axis) and the dependent variable is plotted on the vertical axis (y-axis) ordinate. The independent variable is also known as the control parameter because it influences the behaviour of the dependent variable.

It is not necessary for one parameter to be a controlling parameter. You can draw the scatter diagram with both variables independent to each other. In this case you can draw any variable on any axis.

Type of Scatter Diagram

The scatter diagram can be categorized into several types ; however, we will discuss the two types that will cover most scatter diagrams used in geographical analysis. The first type is based on the type of correlation, and the second type is based on the slope of trend.

According to the type of correlation, scatter diagrams can be divided into following categories :

Scatter Diagram with No Correlation

Scatter Diagram with Moderate Correlation

Scatter Diagram with Strong Correlation

Scatter Diagram with No Correlation

This type of diagram is also known as "Scatter Diagram with Zero Degree of Correlation".

In this type of scatter diagram, data points are spread so randomly that you cannot draw any line through them.

In this case you can say that there is no relation between these two variables.

Scatter Diagram with Moderate Correlation

This type of diagram is also known as "Scatter Diagram with Low Degree of Correlation".

Here, the data points are little closer together and you can feel that some kind of relation exists between these two variables.

Scatter Diagram with Strong Correlation

This type of diagram is also known as "Scatter Diagram with High Degree of Correlation". In this diagram, data points are grouped very close to each other such that you can draw a line by following their pattern. In this case you will say that the

variables are closely related to each other. As discussed earlier, you can also divide the scatter diagram according to the slope, or trend, of the data points :

Scatter Diagram with Strong Positive Correlation

Scatter Diagram with Weak Positive Correlation

Scatter Diagram with Strong Negative Correlation

Scatter Diagram with Weak Negative Correlation

Scatter Diagram with Weakest (or no) Correlation

Strong positive correlation means there is a clearly visible upward trend from left to right ; a strong negative correlation means there is a clearly visible downward trend from left to right. A weak correlation means the trend, up of down, is less clear. A flat line from left to right is the weakest correlation, as it is neither positive nor negative and indicates the independent variable does not affect the dependent variable.

Scatter Diagram with Strong Positive Correlation

This type of diagram is also known as Scatter Diagram with Positive Slant. In positive slant, the correlation will be positive, i.e. as the value of x increases, the value of y will also increase. You can say that the slope of straight line drawn along the data points will go up. The pattern will resemble the straight line. For example, if the temperature goes up, cold drink sales will also go up.

Scatter Diagram with Weak Positive Correlation

Here as the value of x increases the value of y will also tend to increase, but the pattern will not closely resembles a straight line.

Scatter Diagram with Strong Negative Correlation

This type of diagram is also known as Scatter Diagram with Negative Slant. In negative slant, the correlation will be negative, i.e. as the value of x increases, the value of y will decrease. The slope of a straight line drawn along the data points will go down.

For example, if the temperature goes up, sales of winter coats goes down.

Scatter Diagram with Weak Negative Correlation

Here as the value of x increases the value of y will tend to decrease, but the pattern will not be as well defined.

Scatter Diagram with No Correlation

In this type of chart, you are not able to see any kind of relationship between the two variables. It might just be a series of points with no visible trend, or it might simply be a straight, flat row of points. In either case, the independent variable has no effect on the second variable (it is not dependent).

Limitations of a Scatter Diagram

The following are a few limitations of a scatter diagram :

- Scatter diagrams are unable to give you the exact extent of correlation.
- Scatter diagram does not show you the quantitative measure of the relationship between the variable. It only shows the quantitative expression of the quantitative change.
- This chart does not show you the relationship for more than two variables.

Benefits of a Scatter Diagram

The following are a few advantages of a scatter diagram :

- It shows the relationship between two variables.
- It is the best method to show you a non-linear pattern.
- The range of data flow, i.e. maximum and minimum value, can be easily determined.
- Observation and reading is straight forward.
- Plotting the diagram is relatively simple.

Extraction of Regression Equation for drawing Best Fit Line

Example : Extract Regression equation and find necessary calculation for drawing best fit line from the following data

Weight(x)	Length (y)	x^2	y^2	xy
1	2	1	4	2
4	5	16	25	20
3	8	9	64	24
4	12	16	144	48
8	14	64	196	112
9	19	81	361	152
8	22	64	484	176
x = 37	y = 82	x^2 = 251	y^2 = 1278	xy = 553

Here we will extract estimated y (Yc) = a + bx, with the help of the following normal equation—

$\Sigma y = na + b\ \Sigma x$ ...(i) and **$n = 7, \Sigma x = 37, \Sigma y = 82, \Sigma xy = 553, \Sigma x^2 = 251$**

$\Sigma xy = a\Sigma x + b\ \Sigma x^2$ ...(ii)

from the given two variables x and y we can get all extractions excepting a and b, which are two resultant expressions,

The regression equation for y on x is : y = a + bx, where b is the slope and a is the intercept (the point where the line crosses the y axis)

Substituting coefficients in equation no (i), we get,

82 = 71 + 37b .....(i) and

553 = 37a + 251b.....(ii)

For the extraction of 'a' and 'b', we will associate equation no (i) and (ii) in the following manner—

Multiplying coefficient of eq (i) with (ii) and (ii) with (i)

3034 = 259a + 1369b .....(i)

Subtracting the result

3871 = 259a + 1757b .....(ii)

–837 = –388b –388b = –837

b = 2.16

Now for the extraction of 'a' we will substitute the value of 'b' in eq (i)

82 = {7a + (37 × 2.16)}

82 = 7a + 79.92

82 = 7a + 79.92

or, 7a + 79.92 = 82

or, 7a + 82 – 79.92

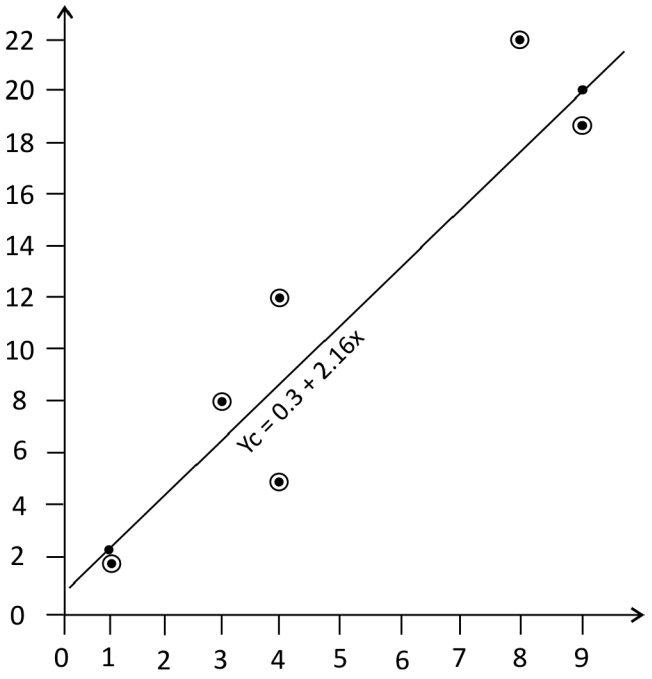
$$\text{or, } a = \left(\frac{82 - 79.92}{7}\right)$$

or, a = 0.30 ; b = 2.16

X	Y	bX	Y' = 0.30 + bX	Remarks
1	2	2.16	0.30 + 2.16 = 2.46	
4	5	8.64	0.30 + 8.64 = 8.94	a = 0.30
3	8	6.48	0.30 + 6.48 = 6.78	b = 2.16
4	12	8.64	0.30 + 8.64 = 8.94	

8	14	17.28	0.30 + 17.28 = 17.58
9	19	19.44	0.30 + 19.44 = 19.74
8	22	17.28	0.30 + 17.28 = 17.58

Extraction of Best Fit Line 'Y' and 'X' (Estimated 'y')



6.9 Summary and Conclusion

Regression analysis is one of the most important tool of enjoying bivariate and multivariate data. It visualizes the relationship between two variables. It provides the data to confirm a hypothesis that two variables are related.

Unit 7 ❑ Drawing of Time Series graphs and Trend Line by Moving Average Method

Structure

7.1 Objectives

- To make the learners understand the drawing techniques of trend lines and time series.

7.2 Introduction

A time series is a set of observation is taken at specific times, usually at equal intervals and refers to the chronologically ordered values of a variable. A time series data may best be studied by plotting them on a graph paper.

7.3 Representation

Representing of a time series data can be a significant tool for analysing geographical data. Among all those methods moving average method is the simplest of smoothing out fluctuations and obtaining the trand volumes with fair degree of accuracy. The objective of the moving average method is to smooth out cyclical, seasonal and irregular varifications of the time series data in order to isolate the trend.

Representation by Moving Average Method
Example-1 : Three Years Moving Average (Odd Year)

Year	1995	1996	1997	1998	1999	2000	2001	2002
Production (Tonnes)	50	55	62	70	78	80	87	95

Year	*Production Tonnes*	*3 Years Moving Total*	*3 Years Moving Average*
1995	50		
1996	55	167	55.67
1997	62	187	62.33
1998	70	210	70.00
1999	78	228	76.00
2000	80	245	81.67
2001	87	262	87.33
2002	95		

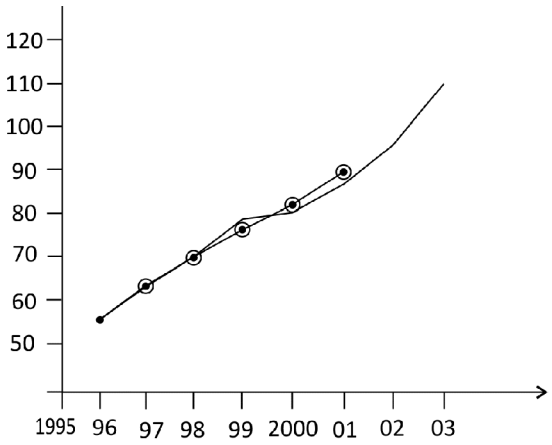
* Graphical representation of the above data (see graph).

Example - 2 : 4 - years Moving Average

Year **01**	*Production Tonnes* **02**	*4 Years Moving Total* **03**	*4 Years Moving Average* **04 = (03/4)**	*Centered Average* **05**
1995	50			
1996	55			
		237	59.25	
1997	62			62.75
		265	66.25	
1998	70			69.38
		290	72.50	
1999	78			75.63
		315	78.75	
2000	80			81.88
		340	85.00	
2001	87			89.00
		372	93.00	
2002	95			
2003	110			

Centered Average = $\left(\frac{\text{Two consecutive item total}}{2}\right)$

Col No. 5 = $\left(\frac{59.25+66.25}{2}\right) = 62.75$



7.3 Summary

The time series allows the learners to understand the fluctuations of data through the graphs.

6.10. Model Questions

Module-02 : Statistical Methods in Geographical Lab oratory

List of Practical

UNIT–1

1. What is data matrix ? What are the major attributes of the data matrix ? Cosntruct a data matrix presenting the following data in Rowas and Columns from the following information collected from 12 mouzas of a block in West Medinipore District.

(i) Total Population = (Male + Female) — (1) 123 + 115, (2) 256 + 214, (3) 245 + 204 (4) 335 + 290, (5) 115 + 92, (6) 415 + 315, (8) 295 + 215, (9) 340 + 310, (10) 180 + 152, (11) 145 + 112, (12) 316 + 210, (13) 425 + 380, (14) 385 + 310, (15) 220 + 180

(ii) % of SC – 15, 25, 18, 30, 12, 20, 18, 35, 38, 42, 24,14, 26, 18, 36

(iii) % of Hindu – 89, 85, 95, 82, 66, 84, 90, 92, 88, 78, 68, 80, 93, 84, 85

(iv) % of Worker – 55, 58, 45, 40, 35, 55, 50, 60, 68, 42, 48, 52, 53, 58, 52

UNIT–II

2. The following is an arrangement of ***Raw data matrix of % of tertiary worker of 50 Municipal Towns***. Based on this table arrange the data properly and tabulate to – (i) Prepare a frequency table ; (ii) Compute different types of frequency ; (iii) Cumulative frequency (Less than and More than type). Represents the computation by (a) frequency Curve ; (b) Histogram ; (c) Polygon ; (d) Ogive.

52	42	34	53	56	59	45	46	49	58
62	68	85	82	56	48	35	39	61	69
65	78	86	88	59	42	47	55	69	78
36	38	44	80	86	87	65	68	62	56
42	48	55	66	77	88	75	54	61	60

3. Compute the relative frequency, Cumulative % frequency (more than and Less than type) from the above data matrix prepared.

UNIT–III, IV, V

4. With the help of following table calculate arithmetic Mean and Geometric Mean and interpret the same to compare the same.

Production of Tea ('00 kgs) of Two Tea gardens.

Sl No.	T.G.-I	T.G.-II
01	5	11
02	4	22
03	9	15
04	25	26
05	12	35
06	15	36
07	16	34
08	18	23
09	22	28
10	26	29
11	23	16
12	30	14
13	32	28
14	36	35
15	45	46
16	32	45
17	48	48
18	52	56
19	50	54
20	51	58
21	56	62
22	58	56
23	59	68
24	48	52
25	42	69
26	38	54
27	36	47
28	22	54

29	26	34
30	29	39
31	35	38
32	34	56
33	41	66
34	9	64
35	8	12
36	23	22

5. Calculate Median and Mode from the following table. Show median and mode graphically and proof you result mathematically.

Class	Frequency
31-33	6
34-36	4
37-39	7
40-42	10
43-45	13
46-48	16
49-51	14
52-54	12
55-57	11
58-60	5

6. What is the relationship among A.M. G.M., and H.M. describe the uses of geometric mean.

Find the Arithmetic Mean (A.M.) of the series –

2, 6, 7, 4, 8, 9, 5, 6, 1, 15, 12, 18

7. Find the Geometric Mean of

111, 112, 123, 167, 156, 198, 216 having weighted by 3, 2, 4, 5, 7, 8, 9

8. Find the Median of the following series —

4, 6, 7, 12, 14, 9, 17, 19, 16, 8, 21, 26

9. Compute the Median from the following table and represent it graphically to justify your result.

Class	frequency	Class	Frequency
110-120	4	160-170	46
120-130	9	170-180	38
130-140	15	180-190	26

140-150	20	190-200	14
150-160	32	200-210	7

10. From the following table compute – **(i) Mean, Median, Mode ; (ii) First and Third quartiles ; (iii) 4th, 6th and 8th Deciles ; (iv) 19th, 39th, 65th, 85th Percentiles and (v) Represent Them graphically to justify your calculation.**

Production ('00kgs)	Frequcney	Production ('00	Frequency
20-24	3	48-52	9
24-28	6	52-56	6
28-32	10	56-60	5
32-36	16	60-64	2
36-40	28		
40-44	18		
44-48	12		

11. (a) Following table in question 4, draw frequency curve, smooth frequency curve, cumulative frequency curve.

(b) Following Table in question number 10, draw Histogram, Polygon, Ogive (More than and Less than type). Show inter-quartile range on the graph.

(c) Compute Quartile Deviation from the same table.

12. Calculate Quartile deviation from the following Series–
10, 5, 8, 12, 15, 20, 19, 18, 26, 28, 25, 8

13. Calculate Mean Deviation from the following Series–
55, 34, 56, 42, 78, 82, 60, 65, 48, 75

14. Compute Standard Deviation from the following series –
5, 8, 10, 14, 15, 16, 20, 24, 26, 30

15. Compute ***Quartile Deviation, Mean Deviation and Standard Deviation*** from the following table.

Class	Frequency
130-140	12
140-150	19
150-160	30
160-170	40
170-180	32
180-190	18
190-200	14

16. From the data given below find which series is more consistent.

Class	Series A	Series B
20-30	6	20
30-40	12	15
40-50	16	38
50-60	26	42
60-70	35	19
70-80	26	8
80-90	14	6

UNIT–VI & VII

17. Estimate the Correlation coefficient and Regression coefficient from the following series

X – 27	28	21	26	30	22	30	31	22	25	30
Y – 38	32	35	36	29	29	40	46	36	38	38

18. From the following data series obtain the regression equation of Y on X and X on Y

X –	91	97	108	122	69	128	69	75	121	69
Y –	71	75	69	99	65	90	75	70	82	50

19. Compute the correlation coefficient from the following data and interpret the result.

X –	80	86	90	85	88	90	96	93	90	99
Y –	145	135	136	118	134	129	116	94	110	93

20. Find the Spearmen's Rank correlation coefficient from the following data.

X –	10	25	9	15	20	22	28	30	36	34
Y –	15	36	18	23	26	42	36	18	46	15

Interpret the Result.

21. Compute the ***Three years Moving Average*** from the following data.

Year	Production (Tons)
1995	50
1996	55
1997	62
1998	70
1999	76
2000	82

2001	88
2002	96

22. Compute 4 – years Moving average from the following data.

Year	Production (Tons)
1995	56
1996	59
1997	66
1998	75
1999	79
2000	85
2001	88
2002	98
2003	112

23. Represent the Trend of production from the results obtained from the calculation of Table from question no. 21 and 22.

24. Using least square method draw trend line from the following information.

Year	Production
1992	104
1993	102
1994	106
1995	115
1996	114
1997	120
1998	126
1999	132
2000	138

(a) Estimate the production of 2003 and 2006 respectively

(b) Represent the data and results graphically to justify your mathematical results.

Module-03

Human Geography/Laboratory

Unit-1 ❑ Spatial variation in continent or country-level religious composition by divided proportional circles.

Structure

1.1 Objective

- The learners will learn about the techniques of computing and tabulation.

1.2 Introduction

Religious composition in any country or state reflects signficant cultural trait in any state or region. **Religion** in a cultural system of designated behaviors and practices, morals, world views, texts, sanctified places, prophecies, ethics, or organizations, that relates humanity to supernatural, transcendental, or spiritual elements. However, there is no scholarly consensus over what precisely constitutes a religion. Among people who do identify with a religion, however, there has been little, if any change on many measures of religious belief. People who are affiliated with a religious tradition are as likely now as in the recent past to say religion is very important in their lives and to believe in heaven. They also are as likely to believe in God, although the share of religiously affiliated adults who believe in God with absolute certainty has declined somewhat.

1.3 Tabulation

West Bengal is home to people belonging to a number of different religions. In fact, people of almost all religions practiced in India live in West Bengal. However, Hindus and Muslims from the major chunk of the state's population. Hundus make up about 72.5% of the total population in West Bengal, while Muslims comprise about 25% of the population. The other minority communities in the

state include Christians, Buddhists, Sikhs and Jains, which together comprise less than 1% of the entire population. About 2% of West Bengal population is made up of tribal people. All these people live here with immense harmony and peace. Eid, Durga Pooja and Christmas are celebrated with same zeal and fevor. The variety in the religious beliefs and traditions of people in West Bengal make it an interesting culture-conglomerate.

T–1.1

Broad Religious Composition (2011) of West Bengal (District Wise)

District	Hindu	Muslim	Others
Uttar Dinajpur	51.72	47.36	0.92
South 24-Parganas	65.86	33.24	0.90
Purulya	83.42	7.12	9.46
North 24-Parganas	75.23	24.22	0.55
Nadia	73.75	25.41	0.84
Murshidabad	35.93	63.67	0.40
Medinipur	85.58	11.33	3.09
Malda	49.28	49.72	1.00
Kolkata	77.68	20.27	2.05
Koch Bihar	75.50	24.24	0.26
Jalpaiguri	83.30	10.85	5.85
Hugli	83.63	15.14	1.23
Haora	74.98	24.44	0.58
Darjeeling	76.92	5.31	17.77
Dakshin Dinajpur	74.01	24.02	1.97
Birbhum	64.49	35.08	
Barddhaman	78.89	19.78	1.33
Bankura	84.35	7.51	8.14
West Bengal	**72.47**	**25.25**	**2.28**

Source- Censusindia.com

T–1.2

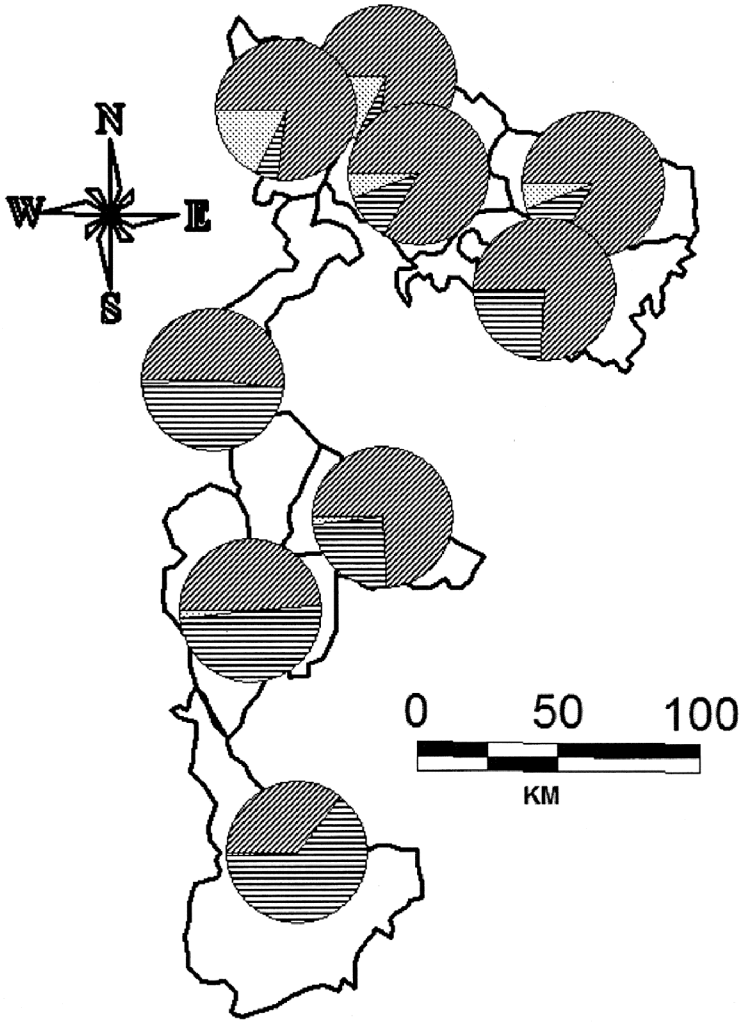
Broad Occupational categories of West Bengal (North Bengal Districts)

Sl. No.	Dist(s)	Cultivators	Agricultural labours	Household Industrial workers	Other workers
01	Darjeeling	51984	32087	10517	438267
02	Jalpaiguri	181104	212657	18115	741889
03	Koch Bihar	312014	272435	26147	275761
04	Uttar Dinajpur	226346	294195	21947	281854
05	Dakhin	170682	174690	20185	160613
	Dinajpur	170682	174690	20485	160613
06	Malda	219241	322452	100383	410919
07	Murshidabad	346103	627126	315687	733221

Source - WB District Census, 2011

Map showing **occupational structure** of northern districts of West Bengal with the help of the above data.

PIE-DIAGRAM SHOWING RELIGIOUS COMPOSITION OF WEST BENGAL (PART)



Pie-Diagram Showing Religious Composition of West Bengal (North)

1.4 Summary

From the analysis it can be said that the variation in the data can be shown by the proportional circles and is on of the best methods of analyzing date.

Unit-2 ❑ Decadal Growth Rate of Population

Structure

2.1 Objective

● The learners will be able to learn about the decadal growth rate of population and its measure.

2.2 Introduction

The growth rate of population in any region is influenced by various factors, which includes soil conditions, development of agriculture, growth of industries, growth of urbanization & development of transport facilities. The concept of population growth or change refers to the growth of the human population in a particular area during a specific period of time. The growth may be positive of negative. It mainly depends upon three factors - migration, birth and death. Such a growth or change can be measured both in terms of absolute numbers and in terms of percentage. It gives us a general idea of the development of the region and socio-economic characteristics of the region. Therefore, the study of population growth is most important and needful aspect in population geography.

In 21th century, many developed and developing countries of the world are facing an ever increasing pressure of polulation upon land, water, mineral and energy resources. under these critical conditions planning of growth of human resource is the only solution. For that reason, the study of growth of population or spatio-temporal changes in population is most useful.

2.3 The decadal growth rate

The decadal growth rate is a vital part of the census operations. It gives an overview of the percentage of total population growth in a particular decade. Growth

rate is basically increase of people in a country, state or a city. There are records that keep track of the increase and decrease in population and it is called "decadal" as a decade consists of a period of 10 years. Thus, the decadal growth rate gives an overview of the total population growth in a particular decade. High rates of population growth contribute worsening economic conditions, political instability and eventually may lead to the collapse of social and economic systems.

To calculate **growth rate,** start by subtracting the past value from the current value. Then, divide that number by the past value. Finally, multiply your answer by 100 to express it as a **percentage.** For this purpose, we can take example; suppose for any country we have population value of two census year, and we can calculate the decadal growth rate of population in the following method.

population of A country in the year of – 2001 – P_{01}
and for the same country and for the year – 2011 – P_{02}
So, with help of this two population values, we have –

Decadal Growth Rate = $\left[\left(P_{02} - P_{01}\right) / P_{01} \times 100\right]$

Taking into consideration of the population value and with the help of this principle, *we can get the Decadal growth rate of population.*

T- 2.1 Table – Population of West Bengal

Sl. No.	District	Area Sq.Km.	Population 2001			Population 2011			Decennial Growth Rate (%)		Populatiom Density Per Sq.Km.	
			P	M	F	P	M	F	1991-2001	2001-2011	2001	2011
1	2	3	4	5	6	7	8	9	10	11	12	13
	West Bengal	188,752	80176197	41465985	38710212	91437736	4692738	44420347	17.77	13.93	903	1029
1	Darjiling	3,149	1609172	830644	778528	1842034	934796	907238	23.79	14.47	511	585
2	Jalpaiguri	6,227	3401173	1751145	1650028	3869675	1980068	1889607	21.45	13.77	546	621
3	Koch Bihar	3,387	2479155	1272094	1207061	2822780	1453590	1369190	14.19	13.86	732	833
4	Uttar Dinajpur	3,140	2441794	1259737	1182057	3000849	1550219	1450630	28.72	22.90	778	956
5	Dakshin Dinajpur	2,219	1503178	770335	732843	1670931	855104	815827	22.15	11.16	677	753
6	Maldah	3,733	3290468	1689406	1601062	3997970	2061593	1936377	24.78	21.50	881	1071
7	Murshidabad	5,324	5866569	3005000	2861569	7102430	3629595	3472835	23.76	21.07	1102	1334
8	Birbhum	4,545	3015422	1546633	1468789	3502387	1791017	1711370	17.99	16.15	663	771
9	Barddhaman	7,024	6895514	3588376	3307138	7723663	3975356	3748307	13.96	12.01	982	1100

10	Nadia	3,927	4604827	2366853	2237974	5168488	2655056	2513432	19.54	12.24	1173	1316
11	North Tweenty Four Parganas	4,094	8934286	4638756	4295530	10082852	5172138	4910714	22.69	12.86	2182	2463
12	Hugli	3,149	5041976	2589625	2452351	5520389	2819100	2701289	15.77	9.49	1601	1753
13	Bankura	6,882	3192695	1636002	1556693	3596292	1840504	1755788	13.82	12.64	464	523
14	Purulya	6,259	2536516	1298078	1238438	2927965	1497656	1430309	14.02	15.43	405	468
15	Haora	1,467	4273099	2241898	2301201	4841638	2502453	2339185	14.57	13.31	2913	3300
16	Kolkata	185	4572876	2500040	2072836	4486679	2362662	2124017	3.93	-1.88	24718	24252
17	South Tweenty Four Parganas	9,960	6906689	3564993	3341696	8153176	4182758	3970418	20.85	18.05	693	819
18	Paschim Medinipur	9,345	5193411	2648048	2545363	5943300	3032630	2910670	15.76	14.44	556	636
19	Purba Medinipur	4,736	4417377	2268322	2149055	5094238	2631094	2463144	14.87	15.32	933	1076

T– 2.2 : Decadal Variation of Population by District (% Change - Positive or Negative)

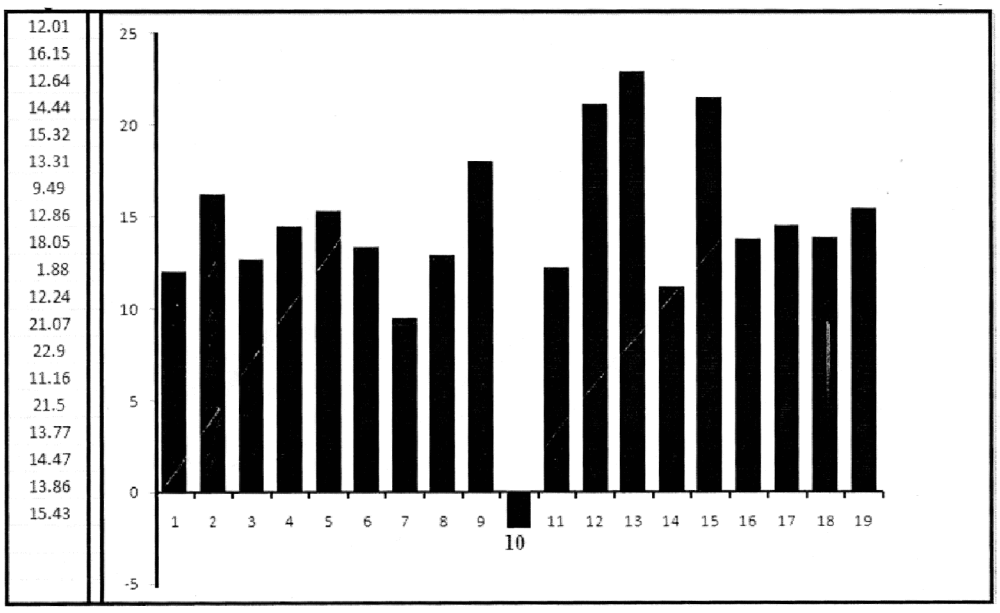
District		2001-2011
(1)		(7)
01	Burdwan	12.01
02	Birbhum	16.15
03	Bankura	12.64
04	Pachim Midnapor	14.44
05	Purba Midnapore	15.32
06	Howrah	13.31
07	Hooghly	9.49
08	24-Parganas (N)	12.86
09	24-Parganas (S)	18.05
10	Kolkata	–1.88
11	Nadia	12.24
12	Murshidabad	21.07
13	Uttar Dinajpur	22.90

14	Dakshin Dinajpur	11.16
15	Malda	21.50
16	Jalpaiguri	13.77
17	Jarjeeling	14.47
18	Coochbehar	13.86
19	Purulia	15.43

Source : Directorate of Census Operation, West Bengal.

DECADAL VARIATION OF POPULATION (2001 – 2011)

Growth rate



DECADAL VARIATION OF POPULATION – WEST BENGAL – 2001 TO 2011

– Horizantal axis – number of Districts 01-19 Districts

– Vertical axis – Growth Rate

– 10 – Kolkata Observed Negative growth rate.

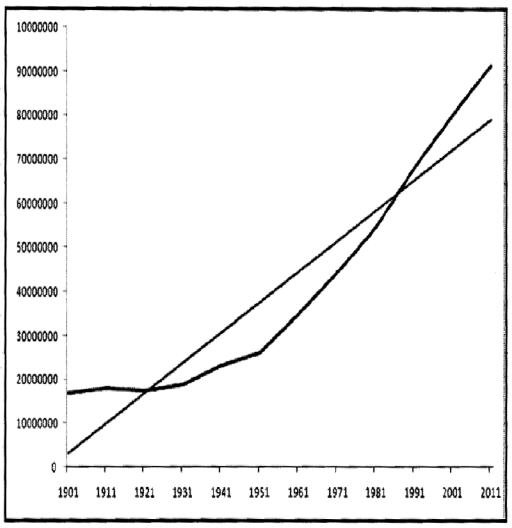
T – 2.3

Decadal Variation of Population – West Bengal

CENSUS (1901-2011) WEST BENGAL			
Year	Population	Decadal Growth(%)	Change in Growth (%)
2011	91,276,115	13.84	-3.93
2001	80,176,197	17.77	-6.96
1991	68,077,965	24.73	1.56
1981	54,580,647	23.17	-3.70
1971	44,312,011	26.87	-5.93
1961	34,926,279	32.80	19.58
1951	26,299,980	13.22	-9.71
1941	23,229,552	22.93	14.79
1931	18,897,036	8.14	11.05
1921	17,474,348	-2.91	-9.16
1911	17,998,769	6.25	-
1901	16,940,088	-	-

T – 2.4

Year	Population	Decadal Growth (%)	Change in Growth (%)
2011	**91,276,115**	**13.84**	**-3.93**
2001	**80,176,197**	**17.77**	**-6.96**
1991	**68,077,965**	**24.73**	**1.56**
1981	**54,580,647**	**23,17**	**-3.7**
1971	**44,312,011**	**26.87**	**-5.93**
1961	**34,926,279**	**32.8**	**19.58**
1951	**26,299,980**	**13.22**	**14.79**
1941	**23,229,592**	**22.93**	**14.79**
1931	**18,897,036**	**8.14**	**11.05**
1921	**17,474,348**	**-2.91**	**-9.16**
1911	**17,998,769**	**6.25**	**-**
1901	**16,940,008**	**-**	**-**



Decadal Growth Curve

2.4 Calculation table for absolute growth, index of growth and decadal growth

YEAR	POPULATION	ABSOLUTE GROWTH	INDEX OF GROWTH (%)	DECADAL GROWTH (%)
1951	102333	–	–	–
1961	222948	120615	217.86	117.86
1971	384859	161911	172.62	72.62
1981	671081	286222	174.37	74.37
1991	1062771	391690	158.36	58.36
2001	1437354	374583	135.24	35.24

Formula used

(i) Absolute Growth = $Ca-Pa$

(ii) Index of Growth = $\frac{Ca}{Pa}\times 100$

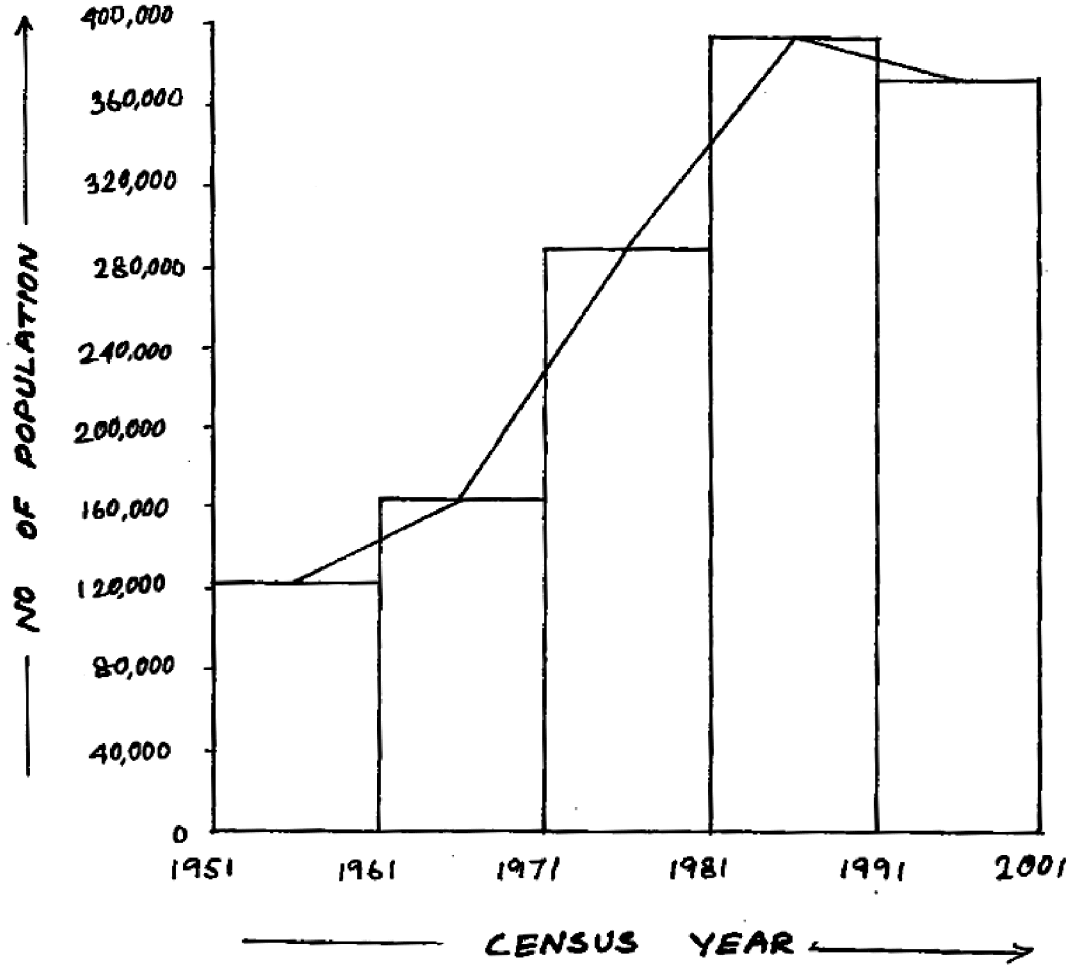
(iii) Decadal Growth (%) = $\frac{Ca-Pa}{Pa}\times 100$

[Where – Ca = Population of current census year

Pa = Population of Previous census year]

ABSOLUTE GROWTH

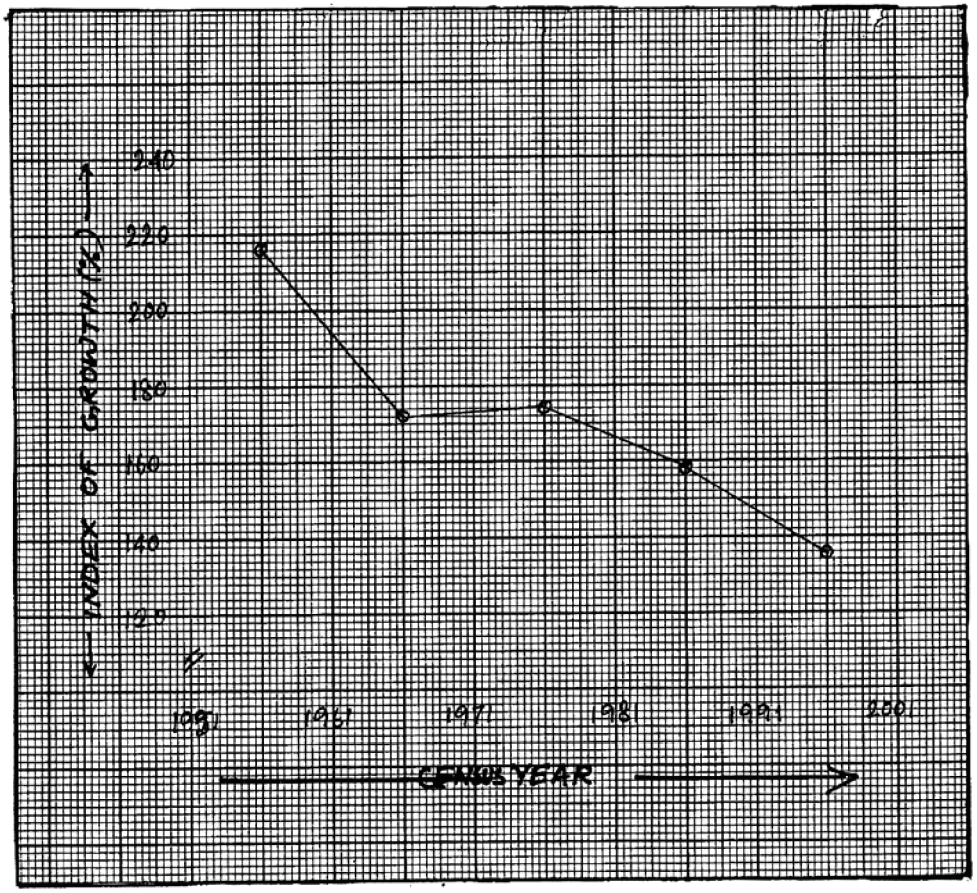
OF TOTAL POPULATION (1951 - 2001)



VERTICAL SCALE - 1cm to 40,000 persons

Index of Growth

of total population (1951-2001)

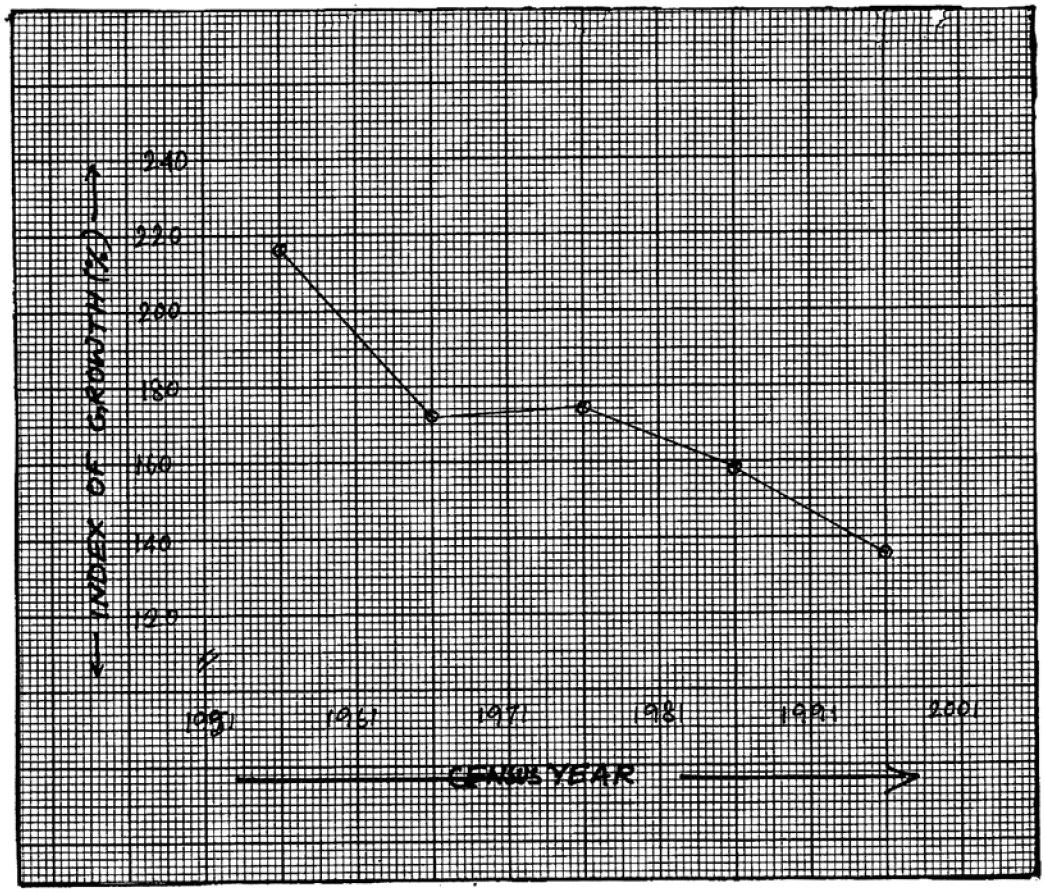


Vertical Scale

1 cm to 20 unit

DECADAL GROWTH RATE

of total population (1951-2001)



Vertical Scale
1 cm to 20 unit

Interpretation :

Based on the census data of different census year and using the above formulas the absolute growth have been calculated. And based on above calculation a bar graph of absolute growth and two line graph of inden of growth and decadal growth rate have been drawn.

We found what from the census year 1951 the population is gradually increased and upto the census year 1991. After 1991 there is a decline of population which have been proved from the 2001 census data.

CALCULATION TABLE FOR DECADAL GROWTH AND INDEX OF GROWTH OF RURAL AND URBAN POPULATION :

Census year	URBAN			RURAL		
	Population	Decadal Growth	Index of Growth	Population	Decadal Growth	Index of Growth
1901	75908	–	–	1246578	–	–
1911	83483	9.979	109.98	1261590	1.204	101.20
1921	87885	5.273	105.27	1136296	–9.931	90.07
1931	91808	4.464	104.46	1278869	12.547	112.55
1941	120449	31.197	131.20	1520081	18.861	118.86
1951	134927	12.020	112.02	1580832	3.996	104.00
1961	195464	44.866	144.87	2094546	32.496	132.50
1971	248425	27.095	127.09	2691779	28.514	128.51
1981	346018	39.285	139.28	3351534	24.510	124.5
1991	494347	42.867	142.87	4245802	26.682	126.68
2001	732734	48.223	148.22	5133853	20.916	120.91
2011	1400692	91.159	191.16	5703115	11.088	111.09

2.5 Summary

The decadal growth rate of population is a very important measure of an economy and for this reason the study helps to understand the spatio-temporal changes in a country or region.

Unit-3 ❑ Types of age-sex pyramids : Graphical representation and analysis

Structure

3.1 Objective

- The learners will acquaint themselves with the different types of age-sex pyramids and its graphical representation.

3.2 Introduction

Population is a dynamic entity. It changes every day and has typical composition. Age and sex composition is one of the major characteristics of the population. ***A population pyramid, also called an "age-sex-pyramid", is a graphical illustration that shows the distribution of various age groups in a population (typically that of a country or gegion of the world), which forms the shape of a pyramid when the population is growing.*** **Population pyramids** are used by demographers as a tool for understanding the make-up of a given **population**, whether a city, country, region, or the world. It is a graphic profile of the **population's** residents.

3.3 Population Pyramid

Population Pyramids are ideal for detecting changes or differences in population patterns. Multiple **Population Pyramids** can be used to compare patterns across nations or selected **population** groups. The shape of a **Population** gives valuable information for human resource planners. If the population pyramid shows higher proportion of children - then planner can think of controlling birth rate. If it is lower

then, they can consider increasing birth rate. If the proportion of old age population is high, means the societly need lots of take care persons and health facility for them. Besides, they may think of bringing the working age people from somewhere else.

Population pyramid is a graphical representation of age and sex composition of the given population. It tells what is the proportion of male and female in different age groups. In other words it is just like a 'family picture' (of the people living in given territory). It usually represented by Bar graph showing the age-sex structure of a population. It consists of two sets of horizontal bar graphs (one for each sex, with males on the left and females on the right), with the member of persons in each age groups along the horizontal axis, and ages along the vertical axis. ***The first population pyrmid was published in 1874 in a statistical atlas of the United States. For a population with high birth and death rates, it is wide at the bottom (young ages) and narrow at the top (old ages), hence the term pyramid. Its shape varies however.*** The population pyramid gives an immediate picture of the demographic regime and the history of a country over a long period. When fertility declines and life expectancy increases, as is the case during the demographic transition, the population pyramid changes shape to resemble a cylinder, or even a spinning top if fertility falls below replacement level. Annual fluctuations in birth and death rates are visible in the pyramid, leaving a durable imprint of the crises experienced by a country, such as famine or war, or of temporary surges in the birth rate, such as the baby boom.

Types of Population Pyramids

Each country will have diferent or unique population pyramids. However, population pyramids will be defined as the following : ***stationary, expansive or constrictive. These types have been indentified by the fertility and mortality rates of a country.***

"Stationary" pyramid

A pyramid can be described as stationary if the precentages of population (age and sex) remains constant over time. Stationary population is when a population contains equal birth rates and death rates.

"Expansive" pyramid

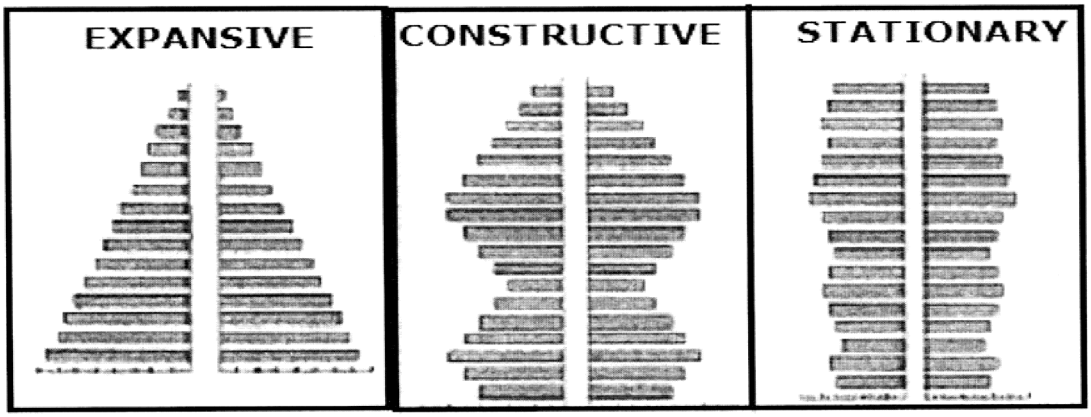
A population pyramid that is very wide at the younger ages, characteristic of countries with high birth rate and low life expectancy. the population is said to be fast-growing, and the size of each birth cohort gets larger than the size of the previous year.

"Constructive" pyramid

A population pyramid that is narrowed at the bottom. The population is generally

older on average, as the country has long life expectancy, a low death rate, but also a low birth rate. However, the perecentage of younger population are extremely low, this can cause issues with dependency ratio of the population. This pyramid is more common when immigrants are factored out. This is a typical pattern for a very developed country, a high level of education, easy access to and incentive to use birth control, good health care, and few negative environmental factors.

Different Types of Population Pyramids



Method of representing Age-Sex Pyramid

In a population pyramid, the size of the population under investigation is depicted on the horizontal axis, and age is aligened on the vertical axis. The result is a series of bars stacked on top of one another, each representing an age category (typically in 5-year age groups), with the youngest age group represented by the bottom bar and the oldest age grouop by the uppermost bar. The horizontal length of each bar represents the number of individuals in the specific age group for the population depicted. The age groups that correspond to each bar are displayed along the central axis or along one side or both sides of the graph. Often the years of birth for each age category are also displayed on the graph. To maintain proportyonality, the age groups are the same size (e.g., 1-year, 5-year, or 10-year age groups), and the bars are all of equal height. Population pyramids intended for comparison should be drawn to the same scale and should depict the same age categories, ***The population pyramid can be used to represent additional characteristics of a population, such as marital status, race, or geographic location. In this case the bar for each age-sex group is further subdivided to represent the additional categories.***

3.4 Interpreting population pyramids

The sape of the population pyramid efficiently communicates considerable

information about the age-sex structure of a specific population. A borad-based pyramid indicates that people in the younger age categories make up a relatively large proportion of the population, and a narrow or pointed top indicates that older people make up a relatively small proportion of the population. In the older age groups of many populations, the number of females is much greater than the number of males; this is reflected in the shape of the pyramid, such that the bars on the right side of the central axis (the female side) are longer than those on the left (male) side. The median age of the population would be the age group (bar) represented by the point on the vertical axis that equally divides the area within the pyramid (equal areas within the pyramid fall above and below the age represented by the bar).

Table 1.1: Distribution of Population by sex and by age group in West Bengal as per Census 2011

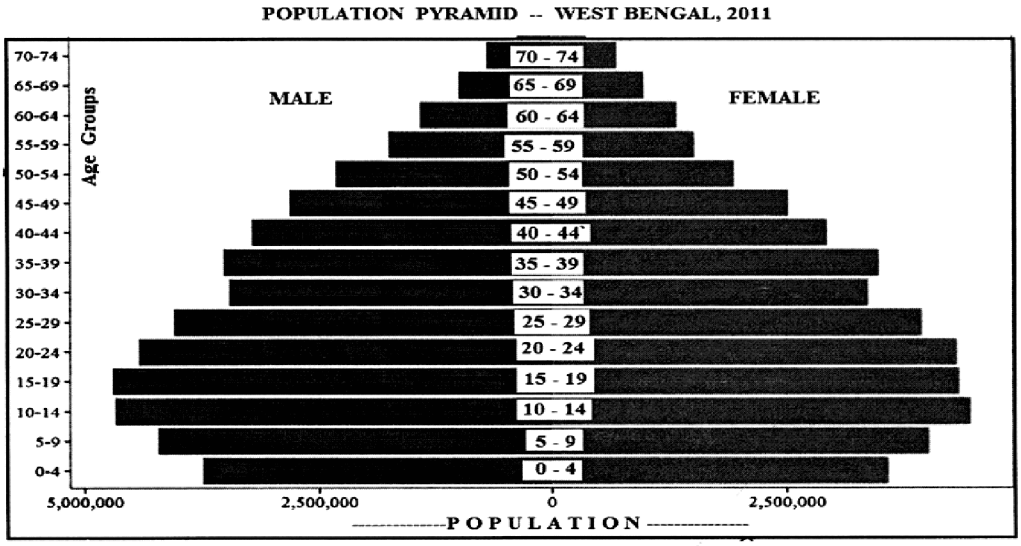
Age		Rural			Urban			Total	
	Male	Female	Total	Male	Female	Total	Male	Female	Total
0–4	2770164	2662779	543294	973698	926502	1900200	3743862	3589281	733143
5–9	3085358	2963285	6048643	1131405	1067761	2199166	4216763	4031046	8247809
10–14	3401028	3276066	6677094	1276478	1202951	2479429	4677506	4479017	9156523
15–19	3359604	3083091	6442695	1342721	1272615	2615336	4702325	4355706	9058031
20–24	3062672	2980049	6042721	1359958	1355643	2715601	4422630	4335692	8758322
25–29	2723495	2631501	5354996	1321409	1321504	2642913	4044904	3953005	7997909
30–34	2255183	2197087	4452270	1209476	1179844	2389320	3464659	3376931	6841590
35–39	2306077	2290624	4596701	1217284	1198661	2415945	3523361	3489285	7012646
40–44	2100597	1924134	4024731	1119007	1009322	2128329	3219604	2933456	6153060
45–49	1807534	1614943	3422477	1006678	906564	1913242	2814212	2521507	5335719
50–54	1469321	1224592	2693913	847911	716056	1563967	2317232	1940648	4257880
55–59	1086200	962669	2048869	660703	559078	1219781	1746903	1521747	3268650
60–64	881516	864930	1746446	524885	474123	999008	1406401	1339053	2745454
65–69	625865	654549	1280414	365415	337164	702579	991280	991713	1982993
70–74	421658	455273	8769311	265223	248453	513679	686881	703726	1390607
75–79	217061	238972	456033	143155	140579	283734	360216	379551	739767
80+	245245	291610	536855	161291	185415	346706	406536	477025	883561
Age not stated	26367	22014	48381	37385	26685	64070	63752	48699	112451
All ages	31844945	30338168	62183113	14964082	14128920	29093002	46809027	44467088	91276115

Source : Census 2011 (WB)

POPULATION PYRAMID – WEST BENGAL, 2011

3.5 Method of drawing population pyramid :

1. Take a column in the mid position at least 2 cm wide for spacing the age group.
2. Then consider the number of population and their respective categories (Male, Female and Rural, Urban etc).



3. Then calculate the % of respective category with respect to total population, or Rural and Urban population.
4. Then draw horizontal bar graph for Rural and Urban (if available) otherwise for Male and Female.
5. Then, again divide the respective bar for Male and Female Rural and Urban. So, finish the pyramid diagram of the particular state or district.

3.5 Summary

The age-sex pyramids used by demographers is an important tool for understanding the populations of a city, country or region.

Unit-4 ❑ Nearest Neighbour Analysis from SOI (R.F. - 1:50,000) Topographical Maps

Structure

4.1 Objective

- The learner will know about Nearest Neighbour Analysis which is an important technique used settlement geography and urban studies.

4.2 Introduction

NNA or Nearest Neighbour Analysis is a significant spatial analysis of point pattern. **Nearest neighbom analysis examines the distances between each point and the closest point to it, and then compares these to expected values for a random sample of points from a CSR (complete spatial randomness) pattern,** CSR is generated by means of two assumptions : 1) that all places are equally likely to be the recipient of a case (event) and 2) all cases are located independently of one another. In settlement geography, this method is very much popular to determine and analysis the point pattern in any region.

4.3 Principles of NNA

Nearest neighbour index (NNI) was originally devised by plant ecologist CLARKE and EVANS. It measures the deviation of any spatial pattern of the distributions of point from randomness. Generally in sattlement geography, **NNA** is widely used.

$$\textbf{NEAREST NEIGHBOUR INDEX} = \frac{Observedvalue(\overline{D})}{Expectedvalue(D_r)}$$

$\overline{(D)}$ = *Summation of nearest distance from each settlement/ Total number of settlements.*

$$D_r = 2 * \sqrt{(N / A)}$$

Where, N= Total number pf settlements. A= Area of the region.

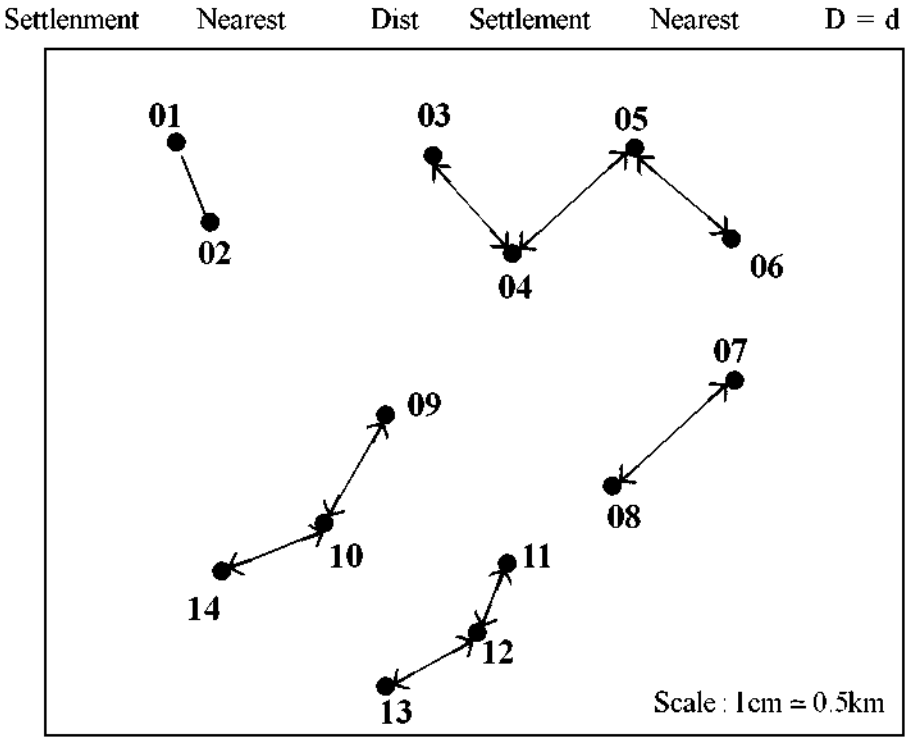
(A) The mean nearest neighbor distance

$$\bar{d} = \frac{\sum_{i=1}^{N} d_i}{N} \quad \text{..... (i)}$$

where N is the number of points. d_i is the nearest neighbor distance for point i.

For getting NND (Nearest Neighbour Distance) take any grid (here ane grid has been selected from 73/F/14) and mark the centroid of the respective settlement.

Topographical Map No- 73/F/14



ID	Neighbour	Km	ID	Neighbour	Dist (Km)
01	02	1.35	14	10	1.50
02	01	1.35			
03	04	1.20			
04	03	1.20			
05	04	1.50			
06	05	2.00			
07	08	1.50			
08	07	1.50		$\sum d = 20.6$	
09	10	1.40			
10	09	1.40		$\bar{d} = 20.6/14 = 1.47$	
11.	12	1.20			
12	11	1.20			
13	12	2.30			

Same process has been applied for the following calculation for Map no - 73/F/14 - Grid No - A_2

CALCULATION OF NEAREST NEIGHBOUR INDEX

NEAREST DISTANCE FROM EACH SETTLEMENT

MAP NO. 73F/14(GRID A2)

T-4.2

Settlement I.D.	Nearest Distance (cm)	Settlement I.D.	Nearest Distance (cm)	Settlement I.D.	Nearest Distance (cm)
1	2.2	17	1.3	33	0.7
2	1.0	18	2.3	34	0.7
3	1.0	19	2.2	35	0.9
4	1.7	20	1.9	36	1.5
5	1.7	21	1.7	37	1.8
6	4.0	22	2.0	38	1.8
7	1.6	23	1.7	39	2.7
8	2.1	24	2.4	40	1.5
9	2.1	25	1.3	41	1.2
10	1.8	26	1.7	42	1.2
11	1.5	27	2.0	43	2.1
12	12	1.5	28	1.7	44
13	1.8	29	1.9	45	0.6
14	1.6	30	2.3	46	0.6
15	1.3	31	2.1	47	2.2
16	2.2	32	1.0	–	–

$\sum$ **d** (Nearest distance from each settlement) = 80.0

Mean of d = 80/47 = 1.7021 (Mean Nearest Neighbour distance)

$$D_r = 2 * \sqrt{(47/316.48)}$$

$$= 2 * \sqrt{0.1485}$$

$$= 2 * 0.3854$$

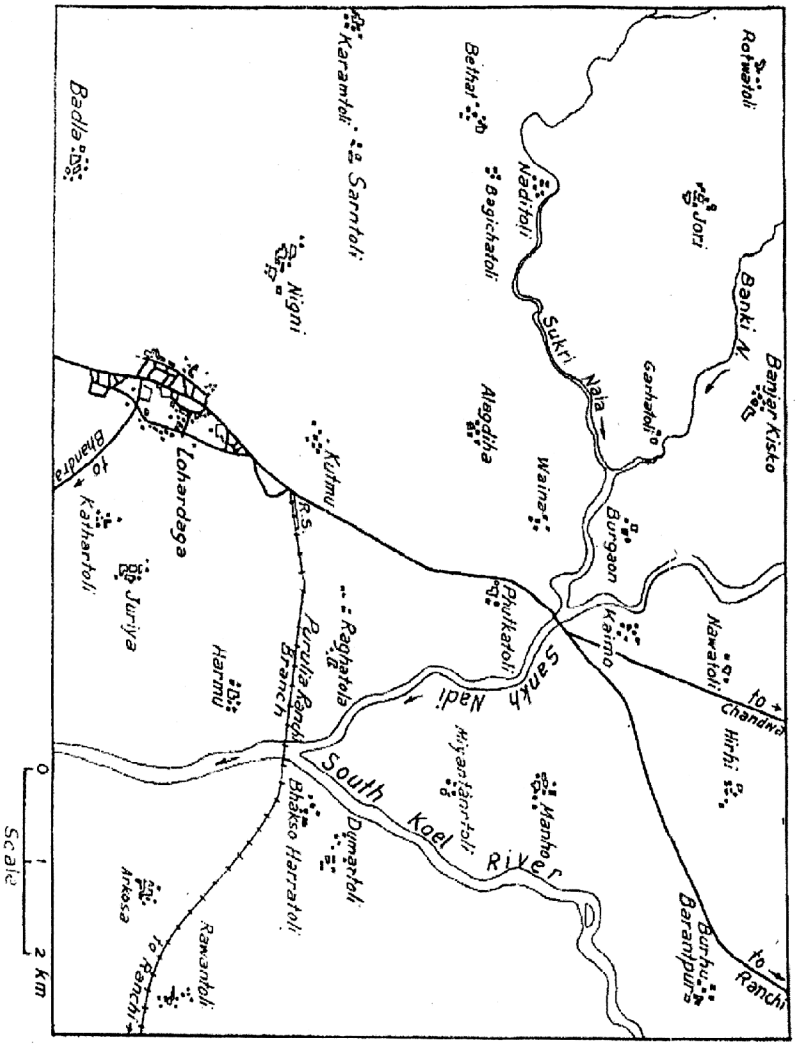
$$= 0.7707$$

$$\textbf{NNI} = \overline{D} / D_r = 2.2085$$

4.3 Interpretation

The given data is taken from map no. 73f/14, grid A3. Here the nearest distance

from each settlement is 80.0 cm. NNA measures the deviation of any spatial pattern of the distribution of points from randomness. ***The value of Nearest Neighbour Index vary between 0 to 2.149. The NNI of the given data is 2.2085. A value of 0 indicates complete clustering and values more than 0 indicates randomness. Therefore a value of 2.2085 indicate that the settlement are spaced out or dispersed and not nucleated.*** Settlements are randomly distributed throughout the chosen grid of the map. The settlement pattern shows uniformly distributed dispersed settlement.



4.4 Summary

The NNA provides a numerical value that described the extent to which a set of points are clustered or uniformly spaced.

Unit-5 ❑ Choropleth mapping based on population data

Structure

5.1 Objective

Choropleth is derived from the Greek word choros (place), adn plethein (to fill). Choropleth mapping is a common technique for representing enumeration data. These are maps where enumeration units, such as states or countries, are shaded a particular colour depending on that unit's data value.

5.2 The choropleth technique is defined by the International Cartographic Association as follows

"A method of cartographic representation which employs distinctive colour or shading applied to areas other than those bounded by isolines. These are usually statistical or administrative areas." Major concerns of the cartographer regarding Choropleth are :– Data classification, Area symbolization and Legend design.

Table - T- 5.1 - Population of West Bengal

Sl. No.	District	Area Sq.Km	Population2001	Population2011		Decennial Growth Rate (%)				Population Density Per Sq. Km				
			P	M	F	P	M	F	1991 – 2001	2001 – 2011	2001	2011
1	2	3	4	5	6	7	8	9	10	11	12	13
	West Bengal	88,752	80176197	41465985	38710212	91347736	46927389	44420347	17.77	13.93	903	1029
1	Darjiling	3,149	1609172	830644	778528	1842034	934796	907238	23.79	14.47	511	585
2	Jalpaiguri	6,227	3401173	1751145	1650028	3869675	1980068	1889607	21.45	13.77	546	621
3	Koch Bihar	3,387	2479155	1272094	1207061	2822780	1453590	1369190	14.19	13.86	732	833
4	Uttar Dinajpur	3,140	2441794	1259737	1182057	3000849	1550219	1450630	28.72	22.90	778	956
5	Dakshin Dinajpur	2,219	1503178	770335	732843	1670931	855104	815827	22.15	11.16	677	753
6	Maldah	3,733	3290468	1689406	1601062	3997907	2061593	1936377	24.78	21.50	881	1071
7	Murshidabad	5,324	5866596	3005000	2861569	7102430	3629595	3472835	23.76	21.07	1102	1334
8	Birbhum	4,545	3015422	1546633	1468789	3502387	1791017	1711370	17.99	16.15	663	771
9	Barddhman	7,024	6895514	3588376	3307138	7723663	3975356	3748307	13.96	12.01	982	1100
10	Nadia	3,927	4604827	2366853	2237974	5168488	2655056	2513432	19.54	12.24	1173	1316
11	North Twenty Four Parganas	4,094	8934286	4638756	4295530	10082852	5172138	4910714	22.69	12.86	2182	2463
12	Hugli	3,149	5041976	2589625	2452351	5520389	2819100	2701289	15.77	9.49	1601	1753
13	Bankura	6,882	3192695	1636002	1556693	3596292	1840504	1755788	13.82	12.64	464	523
14	Purulya	6,259	2536516	1298078	1238438	2927965	1497656	1430309	14.02	15.43	405	468
15	Haora	1,467	4273099	2241898	2031201	4841638	2502453	2339185	14.57	13.31	2913	3300
16	Kolkata	185	4572876	2500040	2072836	4486679	2362662	2124017	3.93	1.88	24718	24252
17	South Twenty Four Parganas	9,960	6906689	3564993	3341696	8153176	4182758	3970418	20.85	18.05	693	819
18	Paschim Medinipur*	9,345	5193411	2648048	2545363	5943300	3032630	2910670	15.76	14.44	556	636
19	Purba Medinipur*	4,736	4417377	2268322	2149055	5094238	2631094	2463144	14.87	15.32	933	1076

5.3 Methods of drawing choropleth

From the above table total population distribution of the northern part of West Bengal will be showing with the help of choropleth map. It is very much significant for representing such discrete data. The steps for preparation of choropleth map are as follows :

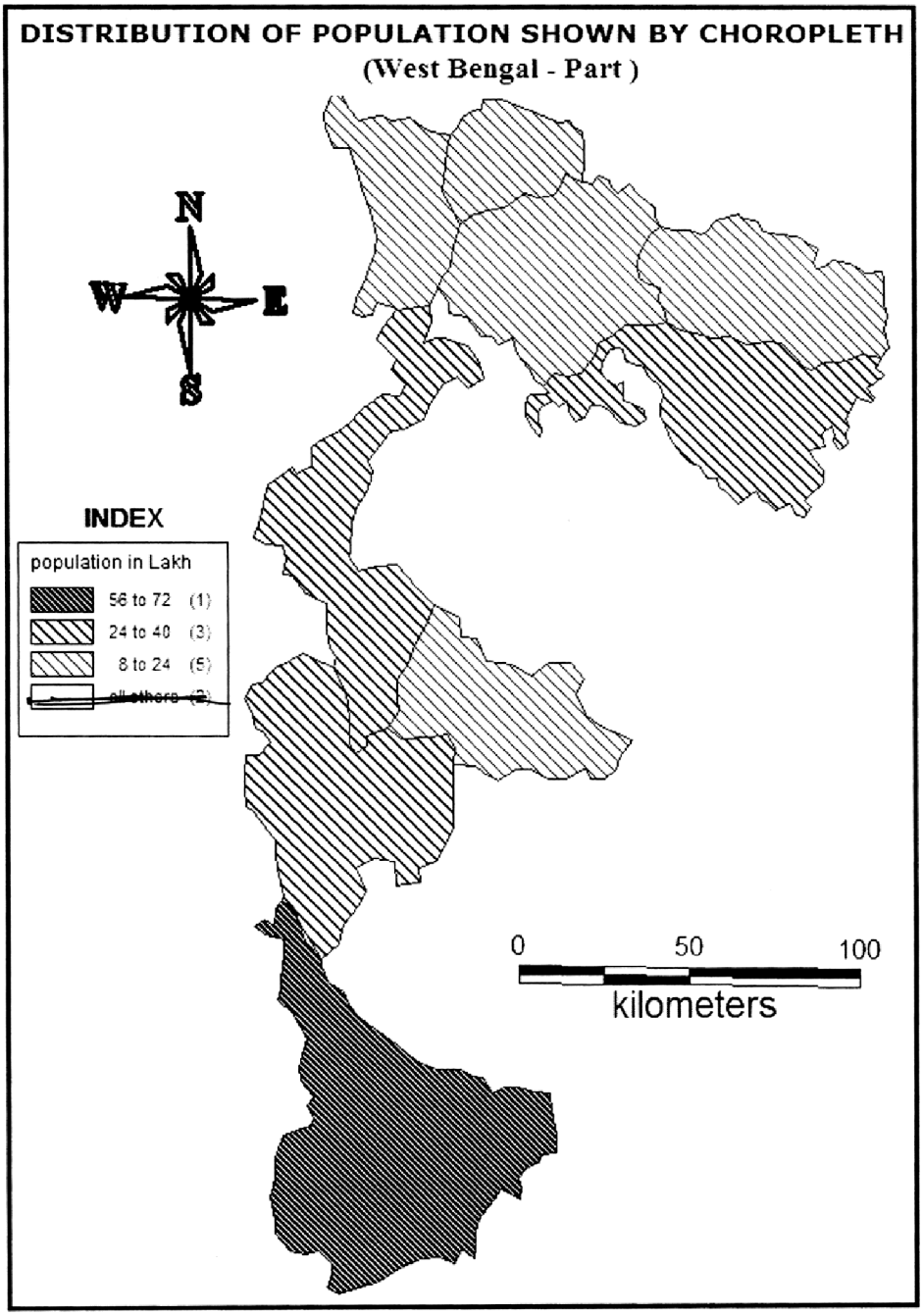
a) Distribution or any type of production etc. are to be selected in relation to the respective space. The regional boundary, here, is to be considered for this purpose.

b) After that, processing and calculation of the respective data will be taken into consideration. After proper verification about the range and nature of data, classes are to be made for next step following this-(1+3.3 log of N)-rule of Thumb.

c) After the confirmation of classes gradational shades or colours are selected for final composition of the map.

d) Then complete the map with proper title, scale, discrete legend/index for final shape of the choropleth map.

After the completion of map, proper interpretation regarding even or odd distribution and its causes are to be explain and zones are also to mentioned for better management of resources.



Here for calculation purpose district serial numbers 01 to 07 from the given table have taken into consideration.

5.4 Summary

Choropleth mapping in a very important technique to visualize data differences and gives an aggregate summary of a geographic chatacteristics with spatial enumeration.

Unit-6 ❑ Variation in occupational structure by proportional divided circles

Structure

6.1 Objective

- The learners will learn about the construction of proportional divided circles.

6.2 Introduction

Occupational structure is distribution of occupations in society, classified according to skill level, economic function, or social status. The occupational structure is shaped by factors such as the structure of the economy, technology, bureaucracy, the labor-market segmentation, the primary labour market and the secondary labour market, and by status and prestige. Demand mobility takes place over time and is not caused by individuals ascending or descending in social class or status, but rather by changes in the occupational structure of the economy.

6.3 Occupational Distribution

Occupational distribution of population refers to the proportion of total working **population** employed in different broad areas of the economy, The proportion of workers engaged in various occupations can highlight the traits of the social-economic development of a region.

Occupational distribution of population helps us to understand the following:

1) Rate at which the population grows.
2) Then number of people employed
3) Productivity of the working population.

4) Industries which are becoming unimportant.

5) Number of people that construct the labor force.

6.4 Steps for processing and calculating the data for occupational structure

1. Fix the categories of data for occupational structure and properly tabulate it.
2. After considering the data, add all categories to confirm that it is 100% or representing total number of population involved in all categories.
3. For the preparation of pie diagram for the respective spatial unit the total number of working population will represent the radius of circle and the variation of working population will create the proportional pie diagram.
4. **Total working population = 100 360^0**

Primary Sector = 45 = $(45/100)x360^0 = 162^0$

Secondary Sector = 25 = $(25/100)x360^0 = 90^0$

Tertiary Sector = 30 = $(30/100)x360^0 = 108^0$

So, total number of working population will be sub-divided into segments of the circle, that is $162^0 + 90^0 + 108^0 = 360^0$

Data Matrix

Table - 6.1. - Districtwise number of female main and marginal workers as per population census 2011

District	Cutivatiors	Agricultura Labourers	Household Industry Workers	Other Workers	Cultivators	Agricultural Labourers	Household Industry Workers	Other Workers
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Burdwan	12,333	127330	29,676	158,722	10,720	145,344	30,083	103,960
Birbhum	9,699	48,573	15,770	53,154	20,977	179,053	21,978	39,085
Bankura	11,765	83,272	9,409	49,859	20,977	179,053	21,978	39,085
Purba Medinipur	14,681	27,063	21,530	63,862	34,581	131,540	40,311	60,683
Paschim Medinipur	32,480	126,084	18,499	81,278	52,967	304,964	65,568	69,961
Howrah	4,524	7,611	49,756	129,941	4,888	13,521	49,023	62,411
Hooghly	10,017	82,765	19,890	138,540	10,133	84,665	25,314	67,938
Purulia	21,634	39,592	19,657	39,283	43,631	212,875	28,860	42,856
24 Prganas (N)	12,337	38,993	36,418	336,975	9,214	50,645	37,910	103,943
24 Prganas (S)	20,289	31,639	33,173	169,870	32,943	119,661	74,016	126,332
Kolkata	2,711	2,136	11,976	265,330	5,172	1,479	8,663	85,807
Nadia	6,608	25,267	52,300	106,514	5,079	17,567	30,188	44,994
Murshidabad	9,528	29,778	218,076	102,077	8,020	32,674	130,420	73,667
Uttar Dinajpur	15,231	55,528	11,856	52,958	16,906	80,835	13,423	29,262
Dakshin Dinajpur	13,885	45,288	10,110	29,609	11,427	63,054	9,886	20,168
Malda	13,009	41,409	73,501	71,144	16,042	95,016	87,264	54,001
Jalpaiguri	13,870	45,756	4,781	167,176	18,461	83,331	7,660	80,514
Darjeeling	10,352	8,715	2,654	114,793	11,132	16,958	3,258	36,278
Coochbehar	28,415	59,613	10,084	38,397	33,637	74,837	11,240	26,446
West Bengal	2633	926412	649089	2169482	353279	1809616	700352	1168710

Table : 6.2 : ***Districtwise number of male main worlers and marginal workers as per population census 2011.***

District	Cutivatiors	Agricultura Labourers	Household Industry Workers	Other Workers	Cultivators	Agricultural Labourers	Household Industry Workers	Other Workers
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Burdwan	284,860	476,956	51662	1,025,203	34,253	223,552	13,.137	183,460
Birbhum	181,962	302,096	18,272	283,318	28,244	159,244	5,733	50,901
Bankura	226,414	219,998	22,070	293,606	50,567	165,051	7,929	65,187
Purba Medinipur	227,108	250,205	36,537	486,353	68,845	293,496	20,465	133,087
Paschim Medinipur	403,904	363,115	39,114	446,074	82,917	311,038	19,421	91,775
Howrah	284,860	476,956	51,662	1,025,203	34,253	223,552	13,137	183,460
Hooghly	181,962	302,096	18,272	283,318	28,244	159,244	5,733	50,901
Purulia	226,414	219,998	22,070	293,606	50,567	165,051	7,929	65,187
24 Prganas (N)	227,108	250,205	36,537	486,353	68,845	293,496	20,465	133,087
24 Prganas (S)	403,904	365,115	39,114	446,074	82,917	311,038	19,421	91,775
Kolkata	4,702	6,767	41,487	1,241,310	3,454	2,006	6,312	106,428
Nadia	286,621	436,313	78,186	604,071	10,434	76,987	8,464	53,014
Murshidabad	336,521	597,348	97,611	631,144	27,007	182,494	19,900	93,642
Uttar Dinajpur	211,115	238,667	10,091	228,896	14,125	67,298	2,608	26,827
Dakshin Dinajpur	156,797	129,402	10,375	131,004	11,167	42,188	1,919	16,708
Malda	206,232	281,043	24,882	339,775	19,899	128,291	11,040	75,299
Jalpaiguri	167,234	166,901	13,334	574,713	15,367	53,684	4,315	95,831
Darjeeling	41,632	23,372	7,863	323,474	13,062	16,996	2,804	50,383
Coochbehar	283,599	212,822	16,063	237,364	19,146	44,603	3,206	28,505
West Bengal	3940399	4943086	869039	11925755	559642	2509728	245644	1722754

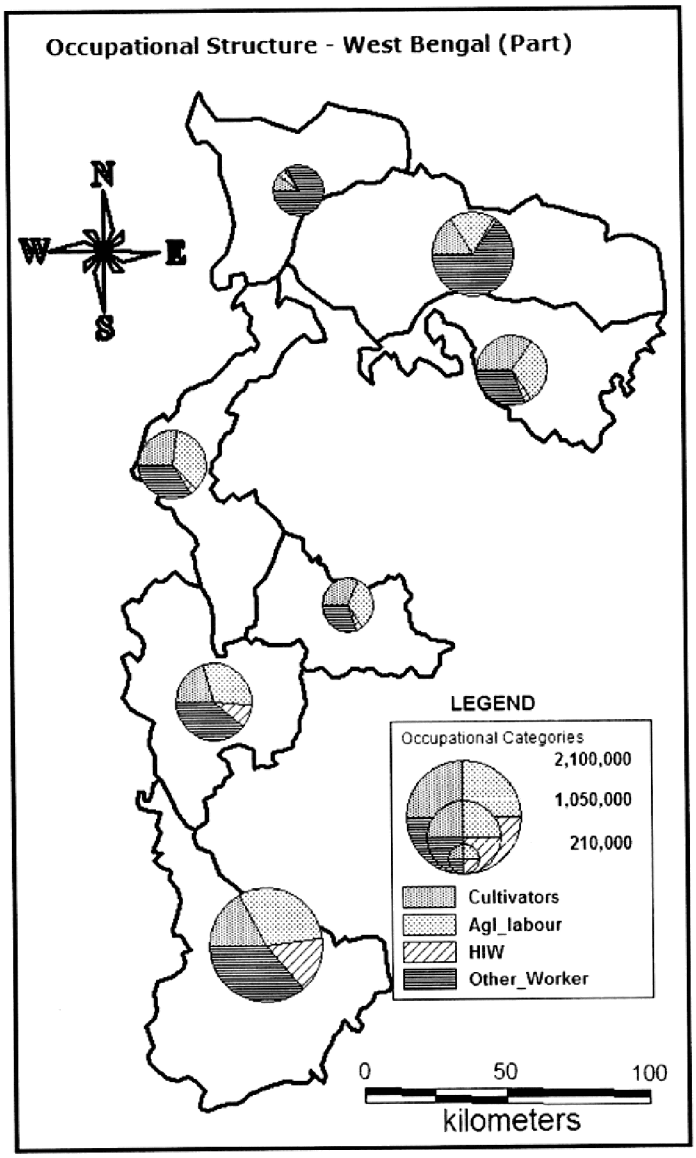
T – 6.3 DATA MATRIX

Broad Religious Composition (2011) of West bengal (District Wise)

District	Hindu	Muslim	Others
Uttar Dinajpur	51.72	47.36	0.91
South 24-Parganas	65.86	33.24	0.90
Puruliya	83.42	7.12	9.45
North 24-Parganas	75.23	24.22	0.55
Nadia	73.75	25.41	0.83
Murshidabad	35.92	63.67	0.40
Medinipur	85.58	11.33	3.09
Malda	49.28	49.72	1.00
Kolkata	77.68	20.27	2.05
Koch Bihar	75.50	24.24	0.26
Jalpaiguri	83.30	10.85	5.84
Hugli	83.63	15.14	1.23
Haora	74.98	24.44	0.58
Darjeeling	76.92	5.31	17.78
Dakshin Dinajpur	74.01	24.02	1.97
Birbhum	64.49	35.8	
Barddhman	78.89	19.78	1.32
Bankura	84.35	7.51	8.14
West Bengal	**72.47**	**25.25**	2.28

Source-Censusindia.com

VARIATION OF OCCUPATIONAL STRUCTURE
Shown By Divided Proportional Circle



T – 6.4 Broad Occupational Categories of West Bengal (North Districts)

Sl No.	Dist(s)	Cultivation	Agricultural labours	Household Workers	Industrial	Other Workers
01	Darjeeling	51984	32087	10517		438267
02	Jalpaiguri	181104	212657	18115		741889
03	Koch bihar	312014	272435	26147		275761
04	Uttar Dinajpur	226346	294195	21947		281854
05	Dakhin Dinajpur	170682	174690	20485		160613
06	Malda	219241	322452	100383		410919
07	Murshidabad	346103	627126	315687		733221

Source-WB District Census, 2011

Map showing occupational structure of northern districts of West Bengal with the help of the above data.

6.4 Summary

It is a very useful method in statistical analysis and knowing about the proportional circles gives us an overview of categorising data into broad occupational categories.

Unit-7 ❑ Time Series Analysis of Industrial Production (India and West Bengal)

Structure

7.1 Objective

- The learners will learn about the different methods of Time-Series Analysis.

7.1 Introduction

Time Series is a sequence of well-defined data points measured at consistent time intervals over a period of time. Data collected on anad-hoc basis or irregularly does not form a time series. ***Time series analysis is the use of statistical methods to analyze time series data and extract meaningful statistics and characteristics about the data.***

7.3 Objective of Time series analysis –

i) Data compression-provide compact descrioption of the data.

ii) Explanatory-seasonal factors- relationships with other variables (temperature, humidity, pollution, etc),

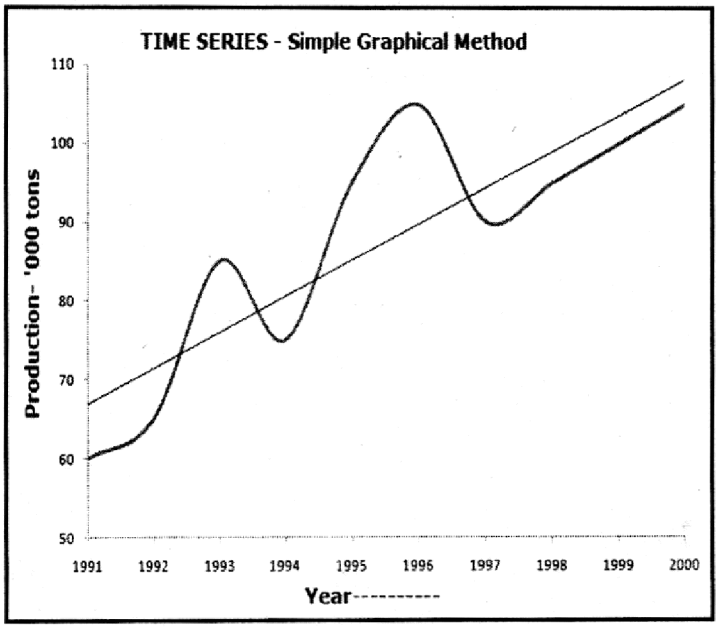
iii) Signal processing - extracting a signal in the presence of noise,

iv) Prediction-use the model to predict future values of the time series.

T - 7.1

Time Series Analysis by Simple Graphical Method

Year	*Production '000 Tons*
1991	*60*
1992	*65*
1993	*85*
1994	*75*
1995	*95*
1996	*105*
1997	*90*
1998	*95*
1999	*100*
2000	*105*



T - 7.2

Year	Production of Steel (million tons)
1994	20
1995	22
1996	30
1997	28
1998	32
1999	25
2000	29
2001	35
2002	40
2003	32

7.4 Semi Average Method

This method is as simple and relatively objective as the free hand method. The data is divided in two equal halves and the arithmetic mean of the two sets of values of Y is plotted against the center of the relative time span. If the number of observations is even the division into halves will be straight forward; however, if the number of observations is odd, then the middle most item, i.e., *(n+12)th*, is dropped. The two points so obtained are joined through a straight line which shows the trend. The trend values of *Y*, i.e., γ, can then be read from the graph corresponding to each time period.

Since the arithmetic mean is greatly affected by extreme values, it is subjected to misleading values, and hence the trend obtained by plotting by means might be distored. However, if extreme values are not above noted two methods, consider the following working example.

Example :

Measure the trend by the method of semi-averages by using the table given below. Also write the equation of the trend line with origin at **1984-85.**

T-7.3

Year	Value in Million
1984 - 85	18.6
1985 - 86	22.6
1986 - 87	38.1
1987 - 88	40.9
1988 - 89	41.4
1989 - 90	40.1
1990 - 91	46.6
1991 - 92	60.7
1992 - 93	57.2
1993 - 94	53.4

Table for Semi-Average Method

T-7.4

Year	Values	Semi-Totals	Semi-Average	Trend Values
1984 - 85	18.6			28.664 - 3.656 = 25.008
1985 - 86	22.6			32.32 - 3.656 = 28.664
1986 - 87	38.1	161.6	32.32	32.32
1987 - 88	40.9			32.32 + 3.656 = 35.976
1988 - 89	41.4			35.976 + 3.656 = 39.632
1989 - 90	40.1			39.632 + 3.656 = 43.288
1990 - 91	46.6			43.288 + 3.656 = 46.944
1991 -92	60.7	253	5.6	50.6
1992 - 93	57.2			50.60 + 3.656 = 54.256
1993 - 94	53.4			54.256 + 3.656 = 57.912

Trend for 1991 - 92 = 50.60
Trend for 1986 - 87 = 32.32
Increase in trend in 5 years = 18.28
Increase in trend in 1 year = 3.656

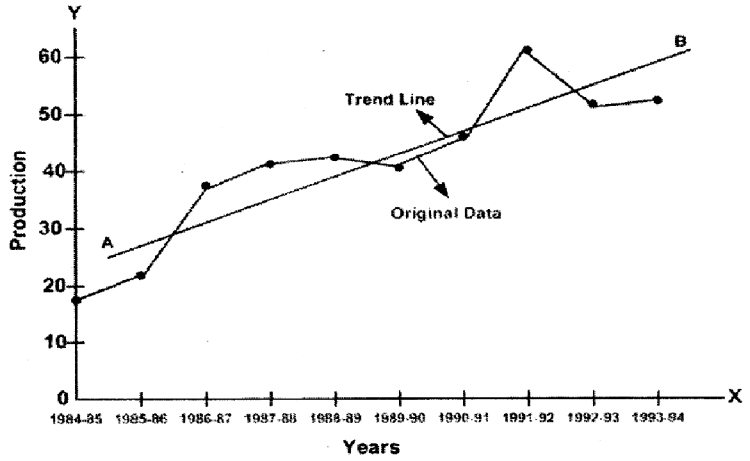
The trend for one year is 3.656. This is called the slope of the trend line and is denoted by ***b***. Thus, ***b = 3.656.*** The trend for **1987 – 88** is calculated by adding **3.656** to **32.32** and similar calculations are done for the subsequent years. The trend for **1985 – 86** is less than the trend for **1986 – 87.** Thus the trend for **1985 – 86** is **32.32 – 3.656 = 28.664.** The trend for the year **1984 – 85 = 25.008.** This is called the interecept because **1984 – 85** is the origin. The intercept is the value of Y, when **X=0.** The interecept is donoted by a. The equation of the trend line is $\Upsilon = a + bX = 25.008 + 3.656X (1984 - 85 = 0)$, where Υ, shows the trend values. This equation can be used to calculate the trend values of the time series. It can also be used for forecasting the future values of the variable.

7.5 Method of Least Squares

If a straight line is fitted to the data it will serve as a satisfactory trend, perhaps the most accurate method of fitting is that of least squares. This method is designed to accomplish two results.

(i) The sum of the vertical deviations from the straight line must equal to zero.

(ii) The sum of the squares of all deviations must be less than the sum of the squares for any other conceivable straight line.



There will be many straight lines which can meet the first condition. Among all different lines, only line will satisfy the second condition. It is because of this second condition that this method is known as the method of least squares. It may be mentioned that a line fitted to satisfy the second condition, will automatically satisfy the first condition.

The formula for a straight-line trend can most simply be expressed as— $Y_c=a+bX$

Where X represents time variable, Yc is the dependent variable for which trend values are to be calculated and a and b are the constants of the straight tine to be found by the method of least squares.

Constant is the Y-ntercept. This is the difference between the point of the orgin (O) and the point of the trend line an Y-axis intersect. It shows the value of Y when X = 0, constant b indicates the slope which is the change in Y for each unit change in X.

Let us assume that we are given observations of Y for n number of years. If we wish to find the values of constants a and b in such a manner that the two conditions laid down above are satisfied by the fitted equation.

Mathematical reasoning suggests thatm to obtain the values of constants a and b according to the Principle of Least Squares, we have to solve simulataneously the following two equations.

$$\sum Y = na + b\sum Y \quad .....(i)$$

$$\sum XY = a\sum X + b\sum X^2 \quad .....(ii)$$

Solution of the two normal equations yield the values for 'a' and 'b'

Least Squares Long Method : It makes use of the above mentioned two normal equations without attempting to shift the time variable to convenient mid-year. This method is illustrated by the following example.

Illustration : Fit aliner trend curve by the least-squares method to the following data:

Year	Production (Kg.)	Year	Production (Kg.)
2001	3	2006	10
2002	5	2007	11
2003	6	2008	12
2004	6	2009	13
2005	8	2010	15

Solution : The first year 2001 is assumed to be 0, 2002 would become 1,2003 would be 2 and so on. The various steps are outlined in the following table.

T - 7.5

Year	Production			
t-X	Y	X (t)	XY	X^2
1	**2**	**3**	**4**	**5**
2001	3	0	0	0
2002	5	1	5	1
2003	6	2	12	4
2004	6	3	18	9
2006	10	5	50	25
2007	11	6	66	36
2008	12	7	84	49
2009	13	8	104	64
2010	15	9	135	11
Total	**89**	**45**	**506**	**285**

The above table yields the following values for various terms mentioned below:
$n = 10$, $\Sigma X = 45$, $\Sigma X^2 = 285$, $\Sigma Y = 89$, and $\Sigma XY = 506$

Substituting these values in the two normal equations, we obtain

89 = 10a + 45b ... (i)

506 = 45a + 285b ... (ii)

Multiplying equation (i) by 9 and equation (ii) by 2, we obtain

801 = 90a + 405b ... (iiii)

1012 = 90a + 570b ... (iv)

Subtracting equation (iii) from equation (iv), we obtain

211 = 165b or b = 211/165 = 1.28 = (b)

Substituting the value of b equation (i), we obtain

$89 = 10a + 45 \times 1.28$

$89 = 10a + 57.60$

$10a = 89 - 57.6$

$10a = 31.4$

$a = 31.4/10 = 3.14 = (a)$

Substituting these values of a and b in the linear equation, we obtain the following trend line

$Y_c = 3.14 + 1.28X$

Inserting various values of X in this equation, we obtain the trend values as below :

Year	Observed	Y	b x X	Y_c (Col. 3 plus Col. 4)
1	2	3	4	5
2001	3	3.14	1.28×0	3.14
2002	5	3.14	1.28×1	4.42
2003	6	3.14	1.28×2	5.70
2004	6	3.14	1.28×3	6.98
2005	8	3.14	1.28×4	8.26
2006	10	3.14	1.28×5	9.54
2007	11	3.14	1.28×6	10.82
2008	12	3.14	1.28×7	12.10
2009	13	3.14	1.28×8	13.38
2010	15	3.14	1.28×9	14.66

Year = t (Time) = X, Yc = a +bx (Yc = a + bt) = 3.14 + 1.28 x t(X)

Table - 7.6

7.6 Simple Graphical Method

The representation of Geographical data (here industrial production) by simple graphical method on a plain graph paper is a very simple method to observe and interpret the trend of production. It is very easy to draw and simple.

Example : Year wise Production of Finished Products of a Jute industry in Hooghly District.

T – 7.7

Year	Production (touns)
1991	60
1992	65
1993	85
1994	75
1995	95
1996	105
1997	90
1998	95
1999	100
2000	105

Semi Average Method :

This is also a Simple Method for represent the Goegraphical data. In this method the whole time series data are divided into two equal parts and averages of each part are calculates. When the numbers of observations even. There will be exactly two equal parts, but in case of odd number, the central value is usually neglected when the division is made. This is represented in the following example.

T – 7.8

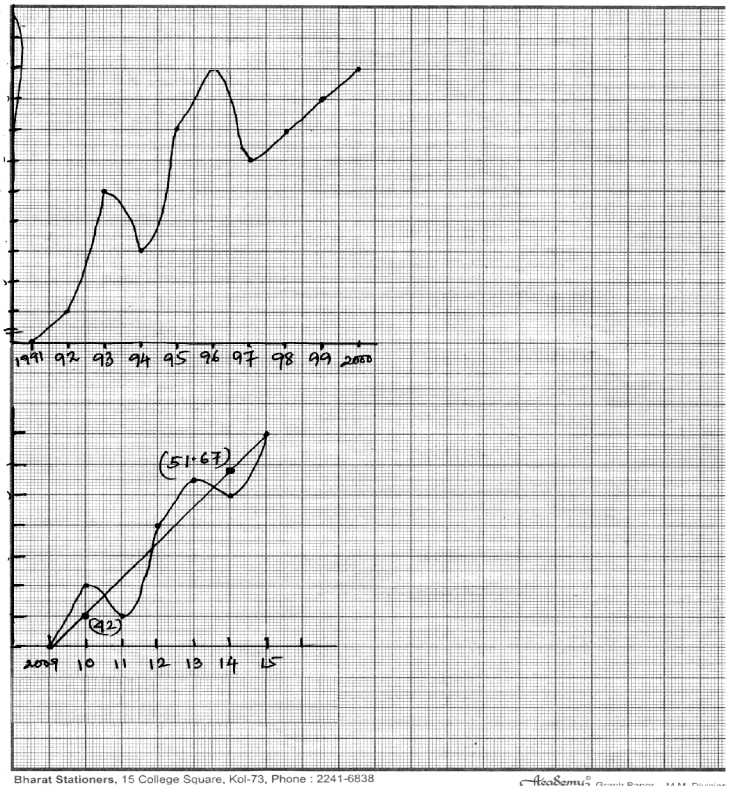
Year	**Production '000 tons**	**Semi Average**
2009	40	$\overline{x_1} = \frac{40+44+42}{3}$
2010	44	$= 42$
2011	42	
2012	48	$\overline{x_2} = \frac{51+50+54}{3}$

2013 2014 2015	51 50 54	= 51.67

From this curve and trend line we care also froject the production of the following years. here we will project the production of 2017 and 2019 respectively from the graphical representation.

Simple Graphical Method

Trend Line of Semi Average



Moving Average Method – Representation

This is also a Simpliest Method to observe the trends values. The purpose of the moving average method is to smooth out cydical, seasonal and irregular variations of the time series data in order to isolate the trend.

Example : Following table shows the annual production of an industrial unit.

T– 7.9

Year	2004	2005	2006	2007	2008	2009	2010	2011	2012
Production '000 tons	6.4	4.3	4.3	3.4	4.4	5.4	3.4	2.4	1.4

Calculate and represent 5 yearly moving averages.

T– 7.10

Year	Production '000 tons	5-Year Miving Total	5-Year Moving Average
2004	6.4	–	–
2005	4.3	–	–
2006	4.3	22.8	4.56
2007	3.4	21.8	4.36
2008	4.4	20.9	4.18
2009	5.4	19.0	3.80
2010	3.4	16.0	3.20
2011	2.4	–	–
2012	1.4	–	–

First Moving Total = 6.4+4.3+4.3+3.4+4.4 = 22.8

First Moving Average = 22.8/5 = 4.56

In this Process all others are calculated.

See graph for representation.

Example – B – 4 Yearly Moving Average

T – 7.11

Year (1)	Production '000 tons	4-year Moving Total (3)	2- item Moving Total (3) (4)	4-year centered Moving Agerage Col (4)/8
2001	506			
2002	620	2835	5752	32.32 - 3.656 = 719.00
2003	1036	2917	5910	738.75
2004	673	2993	6066	758.25
2005	588	3073	6211	776.38
2006	696	3138	6351	793.87
2007	1116	3213	6503	812.87
2008	738	3290	6653	831.67
2009	663	3363	6806	850.75
2010	773	3443	6968	871
2011	1189	3525	7122	890.25
2012	818	3597	7122	910.12
2013	745	3684		
2014	845			
2015	1276			

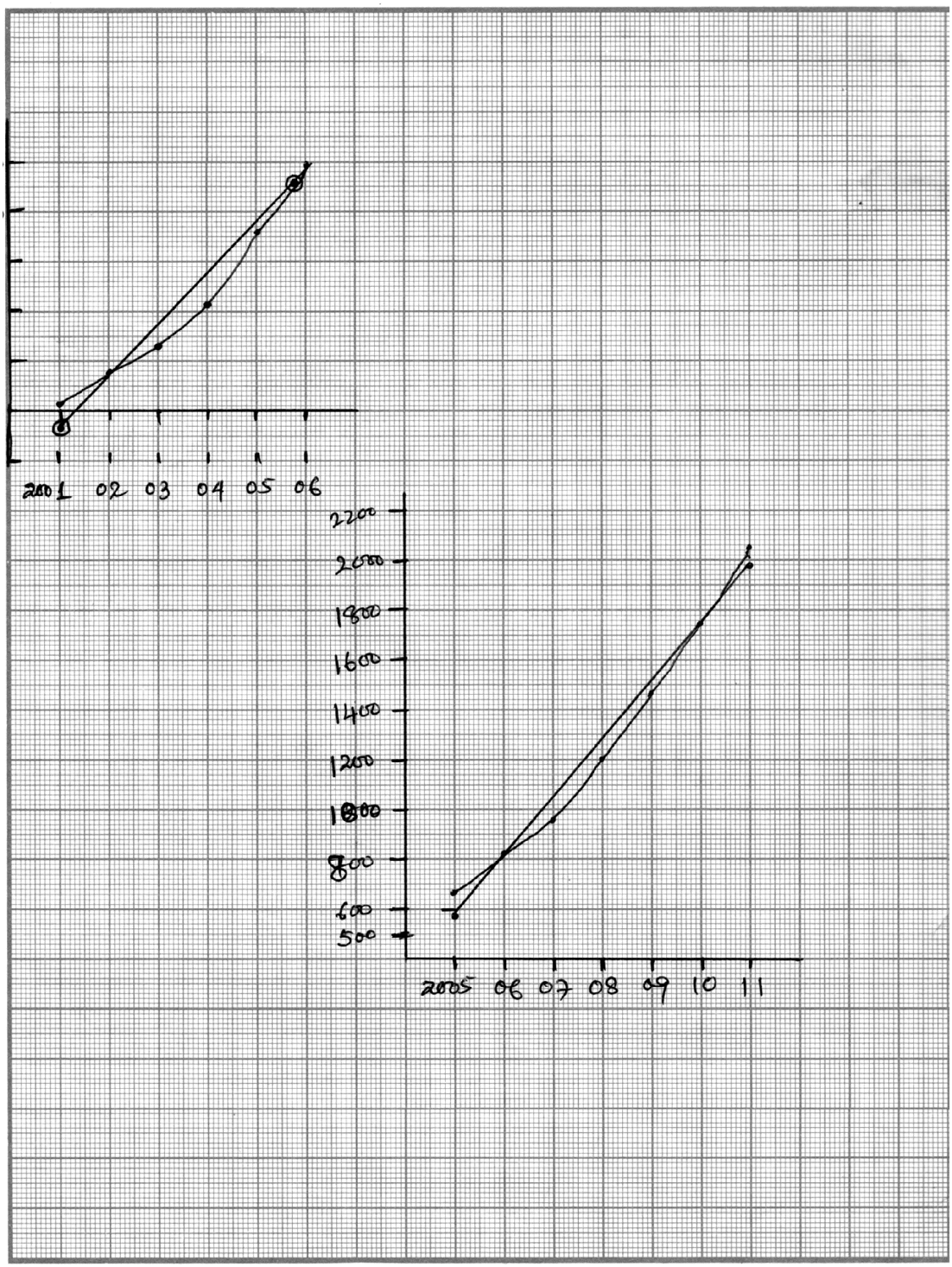
Method of Least Square

The method of least square is the most objective and widly used method in determining the trend in s time series data. From the following example it will be represented.

There are linear and non-liner trend for this method. When the trend is linear the equation may be represented by y = a + bt which can be solved by the so-called normal equations.

$$\sum y = na + b\sum t \quad ...(i)$$

$$\sum yt = a\sum t + b\sum t^2 \quad ...(ii)$$


01
02
03
04
05
06
2200
2000
1800
1600
1400
1200
1000
800
600
500
2005
06
07
08
09
10
11

T - 7.12

Example : When the number of years even

Year (X)	Production (Y)		t^2	yt = (y x t)
2001	101	= –5	25	–505
2002	107	–3	9	–321
2003	113	–1	1	–113
2004	121	1	1	121
2005	136	3	9	408
2006	148	5	25	740
N =6	726	0	70	330

Form normal equation of even years.

$\sum y = an + b\sum t \quad or \quad 726 = 6a + b \times 0$

$or \quad 726 = 6a = a \quad = \frac{726}{6} = 121$

$$a = 121$$

$\sum yt = a\sum t + b\sum t^2 \; or \; 330 = a \times 0 + b \times 70$

$or \quad 330 = 70b \quad or \quad b = \frac{330}{70} = 4.71$

$$b = 4.71$$

The trend equation is

Y = 121 + 4.71t with origin at the mind point of 1953 and 1954 and $\left(t = \frac{1}{2}\right)$

So For estimation of production in the year of 2011 will be

$y = 121 + 4.71 \times 15 = 191.65$

For the year 2011 the production is 191.65 thousand tons.

Where n = number of paired observations. The normal equation are obtained by multtiplying $y = a + bt$ by the coeffocoemts of a and b, *i.e.*, by 1 t throught and summing up.

Case –I – When the number of years is odd. In this case $\Sigma t = 0$ and the two normal equations take the form.

$$\Sigma y = na$$
$$\Sigma yt = b\Sigma t^2$$

Hence $a = \left[\frac{\Sigma y}{n}\right]$ and $b = \left[\frac{\Sigma yt}{\Sigma t^2}\right]$

So, a and b will be calculated from the above expressions.

Case – 2 – When the number of years is even.

In this case the origin is placed in the mindway between the two middle years and the unit is taken to be $\frac{1}{2}$ year instade of one year. In this situation we have again $\Sigma t = 0$ and Hence – $a = \left[\frac{\Sigma y}{n}\right]$ and $b = \left[\frac{\Sigma yt}{\Sigma t^2}\right]$

Example : Number of years odd.

Year (X)	Production (Y)	t = Year - Yo 2008 (Yo)	t^2	yt	Result
2005	672	-3	9	-2016	$\Sigma t = 0$
2006	824	-2	4	-1648	$\Sigma t^2 = 28$
2007	967	-1	1	-967	$\Sigma yt = 6520$
2008	1204 (Yo)	0	0	0	
2009	1464	+1	1	1464	n = 7
2010	1758	+2	4	3516	$\Sigma y = 8946$
2011	2057	+3	9	6171	

From normal equation :

$$8946 = 7a + b \times 0$$
$$or \quad 8946 = 7a \quad or \quad a = 1278 \; (a)$$
$$6520 = a \times 0 + b \times 28$$
$$or \quad 6520 = 28b$$
$$or \quad b = \frac{6520}{28} = 232.9 = (b)$$

Estimate the prodiction for the year 2015

Now the actual trend equation is

$y = 1278 + 232.9t$

For estimation of the year 2015

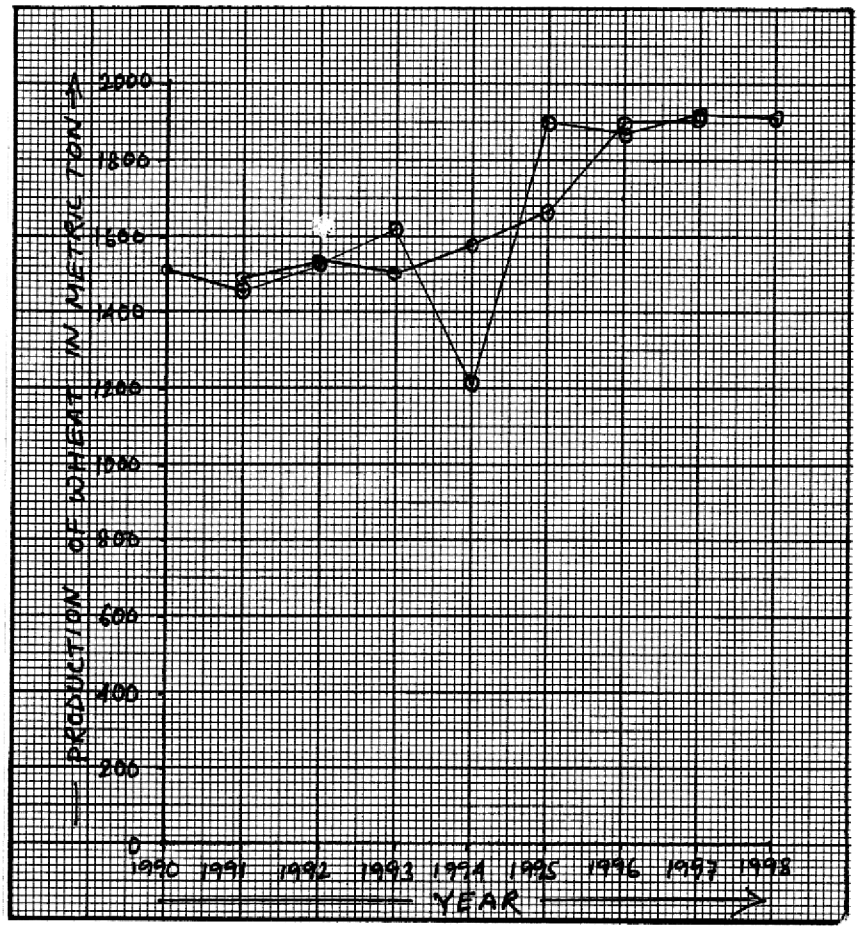
$t = 7$

Y2015 = 1278+232.9 × 7

= 2908.3 ('000 tons)

Year	Production of wheat (Metricton)	Moving Average		
		3-Year	4-Year	5-Year
1990	1510	-	-	-
1991	1450			-
1992	1532		5126	1462.8
1993	1621		1451	1540.8
1994	1210		1563.5	1626
1995	1990		1651.75	1750.6
1996	1876		1902.75	1764.2
1997	1921			-
1998	1914			-

Moving Average Trend (3-Year)
Showing
The production of wheat (1990-98)

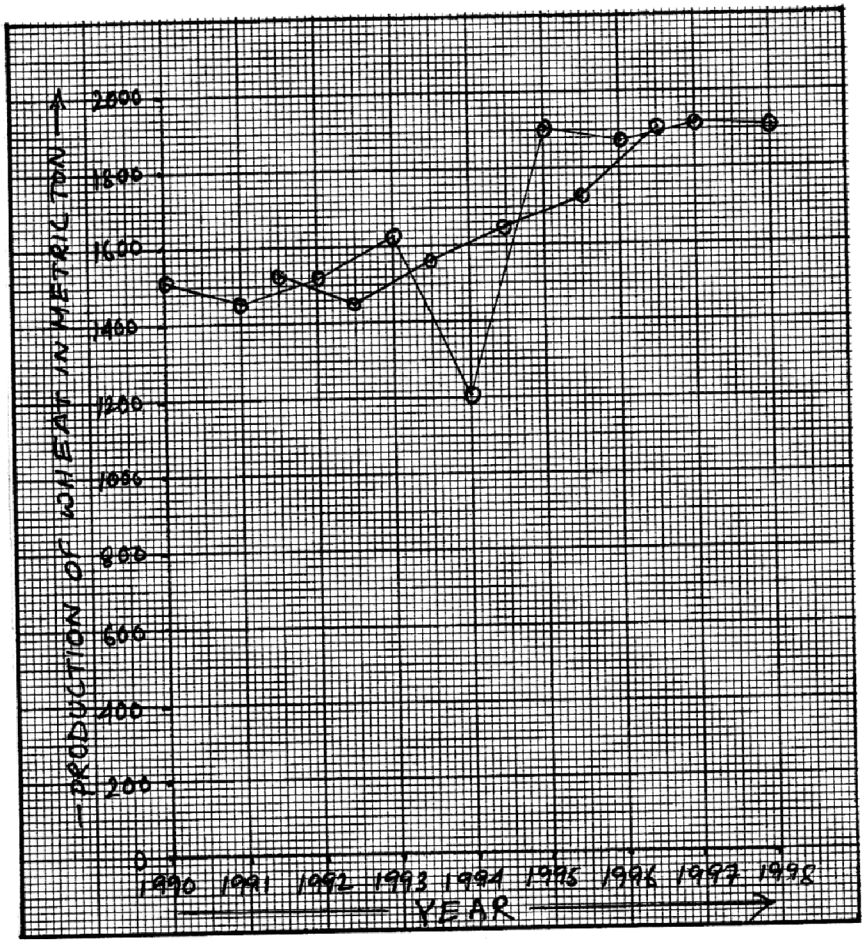


Legend

Actual Trend

Moving Average Trend (3-year)

4 Year Moving Average Trend
Showing
The production of wheat (1990-98)

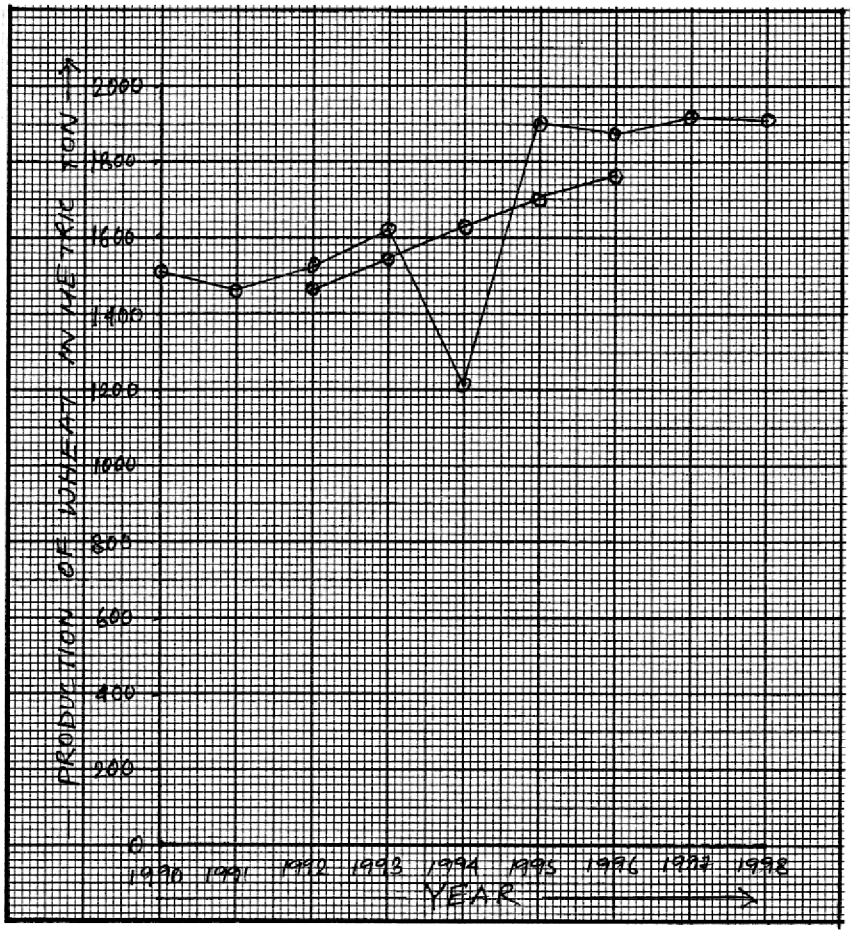


Legend

Actual Trend

Moving Average Trend (3-year)

5 Year Moving Average Trend
Showing
The Production of wheat (1990-98)



Legend

Actual Trend

Moving Average Trend (3-year)

7.7 Summary and Conclusion

- Time series Analysis helps us understand what are the underlying forces leading to a praticular trend in the time series data points and helps us in forecasting and monitoring the data points by fiting appropriate models to it.
- The biggest advantage of using time series analysis is that it can be used to understand the past as well as predict the future.
- Further, time series analysis is based on past data plotted against time which is rather readily available in most areas of study.

Unit-8 ❑ Transport Network Analysis by Shortest Path Matrix Method

Structure

8.1 Objective

- The learners with learn about Shortest Path Matrix and the method of calculating it.

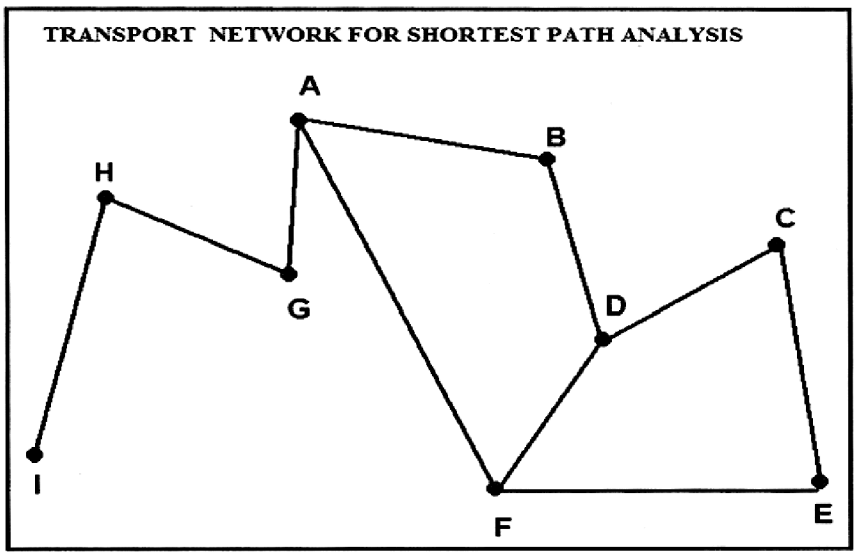
8.2 Introduction

Any network is to be represented as a matrix with rows as set of origin and the columns as the set of destinations. Here, a sample road network has been represented from the Topographical map, 79/B/13 with RF - 1: 50.000 for this purpose. Nine (09) nodes are represented here with their connected path or link and hence network has been created.

8.3 Shortest Path Matrix

In shortest path method for transport network analysis, Nodes and Links have been associated with their connectivity status.

From one node to other how many nodes are to be crossed for easy approach, this should be the consideration for this method. In this way total matrix will be prepared which will originally a diagonal matrix after replacing the same in the reverse columns, it will finnaly be a square matrix.



Then, Mark the highest and Lowest number of Nodes in the matrix **which is called Associated Number.**

Then Add the respective rows for getting Shimbel Index; **Lowest the Shimbel Index, greater will be the accessibility.**

Following Table and Diagram representing the Transport Network Analysis by Shortest Path Analysis.

TRANSPORT NETWORK FOR SHORTEST PATH ANALYSIS

Following this NEtwork Acessibility Matrix has been prepared and from this result shortest path analysis has been done with the help of Associated Number and Shimbel Index; A to I indicating Nodes and A------B indicating Link.

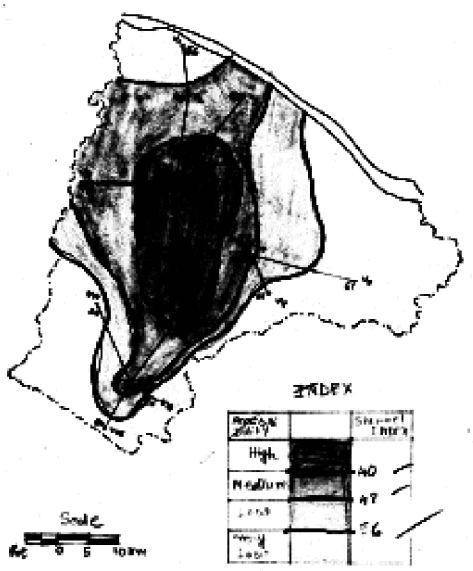
Node	A	B	C	D	E	F	G	H	I	Associated Number	Shimbel Index
A	00	01	03	02	02	01	01	02	03	**03**	15
B	01	00	02	01	03	02	02	03	04	04	18
C	03	02	00	01	01	02	04	05	06	06	24
D	02	01	01	00	02	01	03	04	05	**05**	19
E	02	01	01	02	00	01	03	04	05	**05**	19
F	01	02	02	01	01	00	02	03	04	04	16
G	01	04	04	03	03	02	00	01	02	04	20
H	02	05	05	04	04	03	01	00	01	**05**	25
I	03	06	06	05	05	04	02	01	00	**06**	32

Transport Network Analysis by Shortest Path Matrix

**** -----From A to C cand I indicating least Associated Number (03) and hence most Accessible**

**** ---- From -D to I, E to I, H to B, C and from I to B, C indicating Large Associated Number and hence Least Accessible.**

**** ----- A is most Accessible from all nodes indicated by smallest Shimbel Index (15)**



Calculation Table For Shimbel Index :

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}	V_{12}	V_{13}	V_{14}	V_{15}	Total
V_1	0	1	3	2	3	3	4	5	5	3	4	5	6	6	6	56
V_2	1	0	2	1	2	2	3	4	4	2	3	4	5	5	5	43
V_3	3	2	0	1	2	2	3	4	4	2	3	4	5	5	5	45
V_4	2	1	1	0	1	1	2	3	3	1	2	3	4	4	4	32 (L)
V_5	3	2	2	1	0	2	3	4	4	2	3	4	5	5	5	45
V_6	3	2	2	1	2	0	1	2	2	2	3	4	5	5	5	39
V_7	4	3	3	2	3	1	0	1	1	3	4	5	6	6	6	48
V_8	5	4	4	3	4	2	1	0	2	4	5	6	7	7	7	61 (H)
V_9	5	3	4	3	4	2	1	2	0	4	5	6	7	7	7	61
V_{10}	3	2	2	1	2	2	2	4	4	0	1	2	3	3	3	34
V_{11}	4	3	3	2	3	3	4	5	5	1	0	1	2	2	2	40
V_{12}	5	4	4	3	4	4	5	6	6	2	1	0	1	1	1	47
V_{13}	6	5	5	4	5	5	6	7	7	3	2	1	0	2	2	60
V_{14}	6	5	5	4	5	5	6	7	7	3	2	1	2	0	2	60
V_{15}	6	5	5	4	5	5	6	7	7	3	2	1	2	2	0	60

8.3 Summary

It measures the shortest topologic distance and is the ultimate goal of most transportation planning.

NOTES