

B.Sc. GEOGRAPHY LAB MANUAL

3rd Semester



Prepared By
Pure & Applied Science Dept.
Geography

MIDNAPORE CITY COLLEGE



MIDNAPORE CITY COLLEGE
Department of Pure and Applied Sciences
Course Module for Bachelor of Science (Honours)
Major in Geography
(CCFUP), 2023 & NEP, 2020
Semester: III
Course Type: Major-3P
Course Code: GEOHMJ03
Human Geography (Practical)

MJ-3P: Human Geography (Practical)

Credits 01

Course objectives

The meaning of the term 'development' and Human Development has been refashioned multiple times post-World War II. Initially, development was conceived as merely the economic growth of the country. The belief was that the fruits of economic growth would trickle down to the poorest of the poor. Development theory moved away from traditional ideas of gathering wealth towards a more subjective understanding of development. This unit discusses the concepts of Human Development and various Human Development Indices which are important for assessing Human development resources of a nation.

Course outcome

After completing the course, the students will be able to understand 1. Human Equity, Empowerment, Sustainability and Productivity of human development. 2. Methods to represent human data through various indices and cartograms and thematic maps

Course contents:

1. Different approaches to measure Human Development: Income approach, welfare approach, Basic needs approach and capabilities approach.
 2. Methods of studying human development: Human Development Index and Gender Development Index, Components of HDI and GDI, Methods to measure Poverty Index (Pv1 and Pv2), Measuring methods
 3. Concept of GEM and methods to measure GEM, Gender Equality and Inequality Indices, Women's Economic Opportunity Index (WEOI), Gender Gap Measure (GGM) index.
-

Different approaches to measure human development

Human development is a multidimensional concept that goes beyond mere economic growth to encompass overall well-being and quality of life. Various approaches have been proposed to measure human development, each emphasizing different aspects. The four major approaches are:

1. Income Approach:

This approach equates human development with **economic growth** and focuses on **per capita income** as a proxy for development. Emphasizes **gross national income (GNI)** or **GDP per capita**. Assumes that a higher income leads to a better standard of living. Often used in traditional development economics. Ignores inequality, distribution of income, and non-economic aspects like health, education, and freedom. Doesn't reflect actual quality of life or human well-being.

2. Welfare Approach:

This approach links human development with **levels of consumption and access to goods and services**, assuming these contribute to individual welfare. Focuses on **consumption levels**, **public services** (health, education, housing), and **social security**. Welfare indicators include **life expectancy**, **literacy**, **access to clean water**, etc. May include **social safety nets** and **subsidies**. Still largely quantitative and economic in nature. Does not fully account for personal freedom or individual capabilities.

3. Basic Needs Approach:

Development is achieved when individuals have access to **minimum requirements** for a dignified life. Focuses on **satisfaction of basic human needs**: food, shelter, clothing, education, and healthcare. Developed by the International Labour Organization (ILO) in the 1970s. Emphasizes **equity** and **redistribution** to meet the needs of the poor. Can be seen as paternalistic. May overlook long-term development by focusing on immediate needs. Doesn't measure personal freedom or potential.

4. Capabilistic Approach (Capability Approach):

Proposed by **Amartya Sen**, this approach defines human development in terms of **expanding people's capabilities and freedoms** to lead lives they value. Focuses on what people are **able to be and do** (functionings), not just what they have. Emphasizes **freedom**, **choice**, and **agency**. Basis for the **Human Development Index (HDI)** developed by UNDP. Living a long and healthy life. Being educated. Participating in community life. Holistic and people-centered. Incorporates both **qualitative and quantitative** dimensions. Abstract and sometimes difficult to measure. Requires subjective judgment in choosing capabilities.

Method of studying human development: Human development Index, Gender Development Index, Components of HDI and GDI

To study human development, researchers and policymakers commonly use composite indices that quantify and compare well-being across regions and populations. The **Human Development Index (HDI)** and the **Gender Development Index (GDI)** are two key methods developed by the United Nations Development Programme (UNDP) to measure human development.

1. Human Development Index (HDI)

The HDI is a composite index that measures average achievement in three basic dimensions of human development:

- A long and healthy life
- Access to knowledge
- A decent standard of living

Components of HDI:

Dimension	Indicator	Measurement
Health	Life expectancy at birth	Number of years
Education	a. Mean years of schooling (for adults) b. Expected years of schooling (for children)	Average years
Standard of Living	Gross National Income (GNI) per capita (PPP, US\$)	Income (logarithmically transformed)

Calculation:

Each dimension index is calculated using the formula:

$$\text{Dimension Index} = (\text{Actual Value} - \text{Minimum Value}) / (\text{Maximum Value} - \text{Minimum Value})$$

The HDI is the **geometric mean** of the normalized indices of all three dimensions:

$$\text{HDI} = (\text{Health Index} \times \text{Education Index} \times \text{Income Index})^{(1/3)}$$

2. Gender Development Index (GDI)

The GDI measures gender inequalities in the same three basic dimensions as HDI. It compares the HDI values for **females and males** to assess gender gaps in human development.

Calculation:

- First, calculate HDI separately for males and females.
- Then, compute the ratio of female to male HDI.
- **GDI = Female HDI / Male HDI**

A GDI of **1.0** indicates gender equality; a value **<1.0** indicates inequality against females.

Components of GDI:

It uses the same components as HDI but disaggregated by gender:

Dimension	Female Indicator	Male Indicator
Health	Female life expectancy at birth	Male life expectancy at birth
Education	Mean & expected years of schooling (females)	Mean & expected years (males)
Income	Female estimated earned income	Male estimated earned income

Method to measure poverty index: PV1 and PV2

The **Poverty Index** is used to measure the level and intensity of poverty in a population. The terms **PV1** and **PV2** likely refer to **poverty measures** aligned with the **Foster-Greer-Thorbecke (FGT) poverty indices**, which are widely used in poverty analysis. These are sometimes notated as **P α** where $\alpha = 0, 1, 2$ — corresponding to:

- **P0 (Headcount Ratio)**: Measures the proportion of the population below the poverty line.
- **P1 (Poverty Gap Index)**: Measures the **depth** of poverty.
- **P2 (Squared Poverty Gap Index)**: Measures the **severity** of poverty.

Thus, **PV1** and **PV2** likely correspond to:

PV1 = Poverty Gap Index (P1)

- Formula:

$$P_1 = \frac{1}{N} \sum_{i=1}^q \left(\frac{z - y_i}{z} \right)$$

Where:

- N = total population
- q = number of people below the poverty line
- z = poverty line
- y_i = income (or consumption) of poor individual i
- **Interpretation**: Average shortfall of the poor from the poverty line, expressed as a percentage.

PV2 = Squared Poverty Gap Index (P2)

- Formula:

$$P_2 = \frac{1}{N} \sum_{i=1}^q \left(\frac{z - y_i}{z} \right)^2$$

- Interpretation: Gives greater weight to the poorest among the poor. It accounts for both the depth and inequality of poverty.

Concept of GEM and Methods to measure GEM: Gender equality and inequality indices

The **Gender Empowerment Measure (GEM)** is a composite index developed by the **United Nations Development Programme (UNDP)** to evaluate the **extent of gender equality** in terms of **economic and political participation and decision-making power**. It complements the **Gender-related Development Index (GDI)**, which measures gender disparities in development achievements.

Concept of GEM (Gender Empowerment Measure)

GEM assesses whether women and men are able to actively participate in economic and political life and take part in decision-making processes. It focuses on **agency**, not just well-being.

Purpose:

- Highlight gender inequalities in key areas of power.
- Measure the relative empowerment of women compared to men.
- Encourage policies to reduce gender disparities in leadership and income.

Dimensions of GEM:

GEM incorporates **three dimensions**, each focusing on different aspects of empowerment:

Dimension	Indicator
Political participation and decision-making power	Share of parliamentary seats held by women
Economic participation and decision-making power	Share of women among legislators, senior officials, and managers
Power over economic resources	Female share of income (estimated earned income of females vs. males)

Methods to Measure GEM:

1. Collecting Gender-disaggregated Data:

- Data on income, occupation, and parliamentary representation disaggregated by sex.
- Sources: National statistical agencies, international databases (e.g., World Bank, UNDP).

2. Standardization of Indicators:

- Values are normalized to ensure comparability.
- Example: Income is adjusted using purchasing power parity (PPP).

3. Index Construction (UNDP Methodology):

- For each indicator, scores are computed and aggregated.
- The indicators are scaled (typically between 0 and 1).
- Final GEM score is calculated using an **equally weighted average** of the three dimensions.

4. Interpretation:

- A GEM value close to **1** indicates higher gender empowerment and equality.
- A GEM value closer to **0** indicates greater gender inequality in power and opportunities.

◆ Related Gender Equality & Inequality Indices

Index	Developer	Focus Area
GDI (Gender-related Development Index)	UNDP	Adjusts HDI for gender disparities
GII (Gender Inequality Index)	UNDP (since 2010)	Measures inequality in reproductive health, empowerment, and labor
GEI (Gender Equality Index)	EIGE (EU)	Multidimensional approach across work, money, knowledge, time, etc.
WEF Gender Gap Index	World Economic Forum	Gender gaps in health, education, economy, politics

Women's Economic Opportunity Index (WEOI), Gender Gap Measure (GGM) Index

1. Women's Economic Opportunity Index (WEOI)

Published by: *The Economist Intelligence Unit (EIU)*

Objective:

Measures the enabling environment for women's economic participation across countries.

Key Features:

- Focuses specifically on **economic opportunities** for women.
- Examines **laws, regulations, practices, and attitudes** that influence women's ability to participate in the workforce.

Key Indicators (grouped into 5 categories):

1. **Labour policy and practice**
2. **Access to finance**
3. **Education and training**
4. **Legal and social status**
5. **General business environment**

Scoring:

Countries are scored from **0 to 100**, with 100 being the most favourable environment for women's economic opportunity.

2. Gender Gap Measure (GGM)

Also known as the **Global Gender Gap Index**

Published by: *World Economic Forum (WEF)*

Objective:

Measures **gender-based disparities** across multiple dimensions, not just economic.

Key Features:

- Broader scope than WEOI.
- Focuses on **outcomes rather than resources/input**.

Key Pillars (4 dimensions):

1. **Economic Participation and Opportunity**
2. **Educational Attainment**
3. **Health and Survival**
4. **Political Empowerment**

Scoring:

Ranges from **0 (inequality)** to **1 (equality)** — the closer to 1, the smaller the gender gap.

□ 1. Human Development Index (HDI)

◆ Formula (as per UNDP method)

HDI is the geometric mean of three normalized indices:

$$\text{HDI} = (I_{\text{Health}} \times I_{\text{Education}} \times I_{\text{Income}})^{1/3}$$

Where:

- I_{Health} = Life Expectancy Index
- $I_{\text{Education}}$ = Education Index
- I_{Income} = Income Index

◆ Each component is calculated as:

$$\text{Index} = \frac{\text{Actual value} - \text{Minimum value}}{\text{Maximum value} - \text{Minimum value}}$$

✓ Assumed Maximum & Minimum Values:

Indicator	Max	Min
Life Expectancy	85	20
Mean years of schooling	15	0
Expected years of schooling	18	0
GNI per capita (PPP \$)	75,000	100

$$I_{\text{Education}} = \frac{\text{MYSI} + \text{EYSI}}{2}$$

$$I_{\text{Income}} = \frac{\log(\text{GNIpc}) - \log(100)}{\log(75000) - \log(100)}$$

✓ Example HDI Calculation

Given:

- Life Expectancy = 70 years
- Mean Years of Schooling = 10
- Expected Years of Schooling = 14
- GNI per capita = 20,000 (PPP USD)

1. Health Index

$$I_{Health} = \frac{70 - 20}{85 - 20} = \frac{50}{65} \approx 0.769$$

2. Education Index

$$MYSI = \frac{10}{15} = 0.667, \quad EYSI = \frac{14}{18} = 0.778 \Rightarrow I_{Education} = \frac{0.667 + 0.778}{2} \approx 0.722$$

3. Income Index

$$I_{Income} = \frac{\log(20000) - \log(100)}{\log(75000) - \log(100)} \approx \frac{4.301 - 2}{4.875 - 2} \approx \frac{2.301}{2.875} \approx 0.800$$

Final HDI:

$$HDI = (0.769 \times 0.722 \times 0.800)^{1/3} \approx (0.445)^{1/3} \approx 0.763$$

□ 2. Gender-related Development Index (GDI)

◆ Definition:

GDI adjusts HDI for gender inequalities. It compares female and male HDIs.

◆ Formula:

$$GDI = \frac{HDI_{female}}{HDI_{male}}$$

- If GDI = 1, gender equality in human development.
- GDI < 1 means females are disadvantaged.

✔ Example GDI Calculation

Let's say:

- HDI (Female) = 0.700
- HDI (Male) = 0.800

$$GDI = \frac{0.700}{0.800} = 0.875$$

→ This indicates a gender gap in human development.

□ 3. Gender Empowerment Measure (GEM) (Older UNDP indicator, now replaced by GII)

◆ Formula Overview (Simplified):

GEM is calculated using an equally weighted average of:

- Female share in parliamentary representation
- Female share in professional & managerial positions
- Female earned income share relative to male

$$GEM = \frac{A + B + C}{3}$$

Where:

- A = Female % of parliamentary seats
- B = Female % in economic participation positions
- C = Female income / Male income (adjusted for population)

✔ Example GEM Calculation

Let's say:

- Women hold 20% of parliamentary seats → A = 0.20
- Women make up 30% of managerial/professional roles → B = 0.30
- Women earn 60% of what men earn → C = 0.60

$$GEM = \frac{0.20 + 0.30 + 0.60}{3} = 0.367$$

→ Indicates relatively low empowerment.

1. PV1 (P1) – Poverty Gap Index:

It measures the average shortfall of the poor from the poverty line as a proportion of the poverty line. It reflects the depth of poverty.

Formula:

$$P_1 = \frac{1}{N} \sum_{i=1}^q \left(\frac{z - y_i}{z} \right)$$

Where:

- N : total population
- q : number of poor people (i.e., those with $y_i < z$)
- z : poverty line income
- y_i : income of poor individual i

2. PV2 (P2) – Squared Poverty Gap Index:

Also called the severity of poverty, this considers inequality among the poor by squaring the poverty gaps.

Formula:

$$P_2 = \frac{1}{N} \sum_{i=1}^q \left(\frac{z - y_i}{z} \right)^2$$

✓ Example Calculation:

Suppose:

- Poverty line, $z = \text{₹}1000$
- Population: $N = 5$
- Incomes of 5 individuals: ₹300, ₹600, ₹800, ₹1200, ₹1500

Step 1: Identify the Poor (i.e., income < ₹1000):

Poor individuals: ₹300, ₹600, ₹800

So, $q = 3$

Step 2: Calculate Poverty Gap Index (P1 / PV1)

$$P_1 = \frac{1}{5} \left(\frac{1000 - 300}{1000} + \frac{1000 - 600}{1000} + \frac{1000 - 800}{1000} \right) = \frac{1}{5} (0.7 + 0.4 + 0.2) = \frac{1.3}{5} = 0.26$$

So, PV1 = 0.26 or 26%

Step 3: Calculate Squared Poverty Gap Index (P2 / PV2)

$$P_2 = \frac{1}{5} \left(\left(\frac{1000 - 300}{1000} \right)^2 + \left(\frac{1000 - 600}{1000} \right)^2 + \left(\frac{1000 - 800}{1000} \right)^2 \right)$$
$$= \frac{1}{5} (0.49 + 0.16 + 0.04) = \frac{0.69}{5} = 0.138$$

So, PV2 = 0.138 or 13.8%

Suggested Readings

1. Papalia, D.E., Olds, S.W. and Feldman, R.D. (2006). Human development.9th Ed. New Delhi: Tata McGraw- Hill. 2. Journal of Human Development and Capabilities, Published by Taylor & Francis (Routledge), Print ISSN: 1945-2829
3. Klasen, Stephan (2017): Working Paper UNDP's gender-related measures: Current problems and proposals for fixing them
4. Economist Intelligence Unit, 2012, Women's Economic Opportunity Index 2012, EIU, August 23.

MIDNAPORE CITY COLLEGE
Department of Pure and Applied Sciences
Laboratory Manual for Bachelor of Science (Honours)
Major in Geography
(CCFUP), 2023 & NEP, 2020
Semester – III
Course Type: Major - 4
Course Code: GEOHMJ04
Course Title: Statistical Methods

PREFACE TO THE FIRST EDITION

This is the first edition of Lab Manual for BSc Honours Major in Geography (Third Semester). Hope this edition will help you during practical. This edition mainly tried to cover the whole syllabus. Some hard topics are not present here that will be guided by responsive teachers at the time of practical.

ACKNOWLEDGEMENT

We are really thankful to our students, teachers, and non-teaching staffs to make this effort little bit complete. Mainly thanks to Director and Principal Sir to motivate for making this lab manual.

MJ-4P: Statistical methods (Practical)

Credits 04

Course Objective

This course is designed to introduce students to the fundamentals of statistics. Spread with five sub-modules, this course is for the beginners to get an acquaintance with basic statistical methods. The objectives of this course are-

- 1. To build fundamental knowledge of statistical analysis.*
- 2. To get an adequate understanding of distribution of data and its implications to further bivariate analysis.*

Course Learning Outcomes

Upon completion of this course, the students will be able to -

- 1. Learn the basic concepts of variables, vectors, random sampling process.*
- 2. Learn the distribution of data and its applications.*
- 3. Gain a working knowledge of describing statistical data.*

Course contents:

1. Fundamental concepts: variable, vector, concepts of population and sample, random sampling; data type in statistics: nominal, ordinal, interval, and ratio scale data, types of sampling.

2. Frequency distribution and its implication, construction of cumulative frequency distribution curves; extraction of samples from a data matrix in excel using random, systematic and stratified method of sampling.
 3. The shape of the distribution of data - skewness and kurtosis; normal distribution- properties of normal distribution. Examples of normal distribution; Representation of data distribution through boxplot. Data standardization in statistics.
 4. Measures of central tendency and dispersion: mean, mode, median, standard deviation.
 5. Measures of association: Pearson's correlation and Spearman's rank correlation.
-
-

1. Fundamental concepts: variable, vector, concepts of population and sample, random sampling; data type in statistics: nominal, ordinal, interval, and ratio scale data, types of sampling.

Statistics in Geography

Statistics is the branch of mathematics that deals with the collection, analysis, interpretation, presentation, and organization of numerical data. It is a key tool in scientific research, business, economics, and social sciences, including geography. Statistics can be broadly divided into two main categories: descriptive statistics and inferential statistics. Descriptive statistics involve the analysis and presentation of data using summary statistics, graphs, and charts. Inferential statistics involve making inferences and predictions about a population based on a sample of data.

In geography, statistics are used to understand spatial patterns and relationships in various aspects of human and physical geography. Geographers use statistics to analyze and interpret data related to population distribution, land use, natural resources, climate, and other geographic phenomena. Statistical methods such as regression analysis, spatial analysis, and geographic information systems (GIS) are used to explore relationships between different variables and to identify patterns and trends.

Statistics are also used in geography to make predictions and inform decision-making. For example, geographers use statistical models to predict changes in population growth, land use patterns, and climate change impacts. Statistics can also be used to evaluate the effectiveness of policies and interventions aimed at addressing spatial problems such as urban sprawl, environmental degradation, and social inequality.

Overall, statistics play a crucial role in geography as they provide quantitative data and insights that help geographers understand and analyze spatial patterns and relationships, make predictions, inform decision-making, and evaluate policies and interventions.

Variables and vectors

Variable: A variable is any characteristic, number, or quantity that can be measured or counted. It can vary from one individual or observation to another.

Types of Variables:

1. **Quantitative (Numerical)** – Represent measurable quantities.
 - *Example:* Age, height, income, test scores.
2. **Qualitative (Categorical)** – Represent categories or labels.
 - *Example:* Gender, religion, occupation, blood type.

Examples:

- A student's **age**: 18, 19, 20, ...
- A person's **gender**: male, female, other
- A household's **monthly income**: ₹15,000, ₹30,000, ₹50,000, ...

Vector: A vector is a data structure (usually a one-dimensional array) that stores a sequence of values, often representing observations of a single variable.

Example 1: Vector of a Quantitative Variable

- Variable: *Height (in cm)*
- Vector: $H = [160, 170, 165, 175, 168]$ (*Each value is the height of a person.*)

Example 2: Vector of a Categorical Variable

- Variable: *Education level*
- Vector: $E = ["Graduate", "Postgraduate", "Graduate", "PhD", "Diploma"]$

Discrete and continuous data**Discrete data**

Discrete data is a type of data that consists of values or observations that can only take on specific, separate, and distinct numerical values. These values are often integers, but they can also be non-numeric categories or labels.

Discrete data is different from continuous data, which can take on any value within a range. For example, the height of a person is continuous data because it can take on any value within a range, such as 5'6.5" or 5'6.75". In contrast, the number of siblings a person has is discrete data because it can only take on specific values, such as 0, 1, 2, 3, and so on.

Some examples of discrete data include:

- Number of students in a classroom
- Number of cars in a parking lot

- Number of goals scored by a soccer team in a game
- Number of children in a family
- Number of pets in a household

Discrete data is often analyzed using statistical methods that are specific to this type of data, such as frequency distributions, histograms, and measures of central tendency like the mode or median. These statistical methods help to summarize and understand the patterns and relationships in discrete data.

Continuous data

Continuous data is a type of data that can take on any numerical value within a range, such as decimals or fractions. Continuous data is different from discrete data, which can only take on specific, separate, and distinct numerical values.

Examples of continuous data include:

- Height and weight of individuals
- Temperature readings
- Length of time between two events
- Amount of rainfall in a day
- Speed of a car

Continuous data is often analyzed using statistical methods such as descriptive statistics, correlation analysis, and regression analysis. These statistical methods help to summarize and understand the patterns and relationships in continuous data. Graphical methods such as histograms, box plots, and scatterplots are also commonly used to visualize and analyze continuous data.

It is important to note that while continuous data can theoretically take on any numerical value within a range, in practice it may be limited by the precision of measurement tools and methods. For example, the height of a person may be measured in centimeters, but the measuring tool may only be precise to the nearest millimeter, resulting in discrete data.

Population and samples

Population

In statistics, population is the entire set of items from which you draw data for a statistical study. It can be a group of individuals, a set of items, etc. It makes up the data pool for a study.

Generally, population refers to the people who live in a particular area at a specific time. But in statistics, population refers to data on your study of interest. It can be a group of individuals, objects, events, organizations, etc. You use populations to draw conclusions.



Population

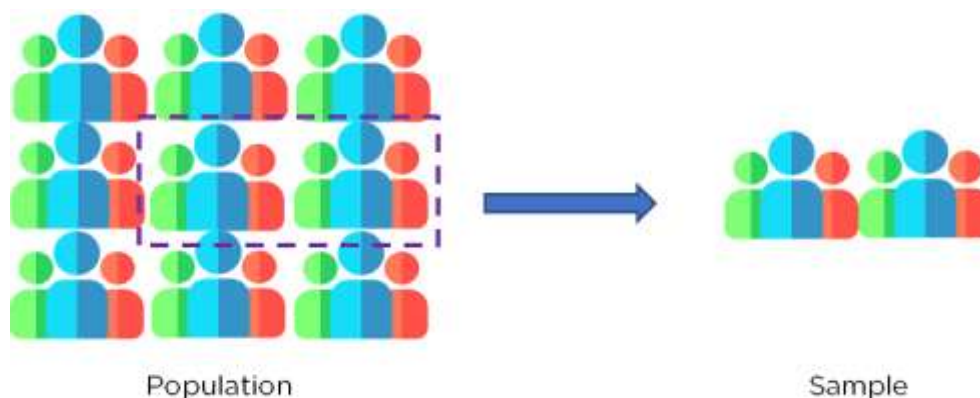
Sample

A sample is defined as a smaller and more manageable representation of a larger group. A subset of a larger population that contains characteristics of that population. A sample is used in statistical testing when the population size is too large for all members or observations to be included in the test.

The sample is an unbiased subset of the population that best represents the whole data.

To overcome the restraints of a population, you can sometimes collect data from a subset of your population and then consider it as the general norm. You collect the subset information from the groups who have taken part in the study, making the data reliable. The results obtained for different groups who took part in the study can be extrapolated to generalize for the population.

The process of collecting data from a small subsection of the population and then using it to generalize over the entire set is called Sampling.



Samples are used when:

- The population is too large to collect data.
- The data collected is not reliable.

- The population is hypothetical and is unlimited in size. Take the example of a study that documents the results of a new medical procedure. It is unknown how the procedure will affect people across the globe, so a test group is used to find out how people react to it.

A sample should generally:

- Satisfy all different variations present in the population as well as a well-defined selection criterion.
- Be utterly unbiased on the properties of the objects being selected.
- Be random to choose the objects of study fairly.

Population and Sample Formulas

1. Population Parameter:

- Mean: $\mu = (\Sigma X) / N$, where ΣX is the sum of all values in the population and N is the size of the population
- Standard Deviation: $\sigma = \sqrt{[(\Sigma(X-\mu)^2) / N]}$, where X is a value in the population, μ is the population mean, and N is the size of the population

2. Sample Statistic:

- Mean: $\bar{x} = (\Sigma x) / n$, where Σx is the sum of all values in the sample and n is the size of the sample
- Standard Deviation: $s = \sqrt{[(\Sigma(x-\bar{x})^2) / (n-1)]}$, where x is a value in the sample and \bar{x} is the sample mean

Note that the formulas for the population parameter and sample statistic are similar, but they use different notation and have slightly different calculations. The population parameter uses the entire population, while the sample statistic uses a subset (i.e., sample) of the population.

Data types in statistics (Scales of measurement)

Nominal, ordinal, interval, and ratio scale data

In statistics, there are four scales of measurement, which are:

1. Nominal Scale: This is the lowest level of measurement and involves assigning names or labels to items or variables. Examples include gender, race, religion, and nationality. Nominal scales do not have any numerical value, and the categories are not ordered.

Nominal scale is the lowest level of measurement in statistics. It involves assigning names or labels to items or variables without any numerical value. The categories in a nominal scale are not ordered, and

each category is distinct and separate. Nominal scale data is often qualitative or categorical, and examples include gender (male or female), race (white, black, Asian, etc.), religion (Christian, Muslim, Hindu, etc.), and marital status (single, married, divorced, etc.). In nominal scale data, you can only count the frequency of occurrence of each category or group, but you cannot perform arithmetic operations such as addition or subtraction.

2. Ordinal Scale: This scale involves ranking items or variables in a specific order. Examples include educational level (elementary, high school, college, graduate), or rating scales such as Likert scales. Ordinal scales have a specific order, but the distance between the categories is not meaningful.

Ordinal scale is a level of measurement in statistics that involves ranking items or variables in a specific order. The categories or responses in an ordinal scale have a specific order, but the distance between them is not necessarily equal or meaningful. Ordinal scale data is often qualitative or categorical and can be expressed in the form of rating scales, surveys, or questionnaires. Examples of ordinal scale data include educational level (elementary, high school, college, graduate), level of satisfaction (very satisfied, somewhat satisfied, neutral, somewhat dissatisfied, very dissatisfied), or quality ratings (poor, fair, good, very good, excellent). In ordinal scale data, you can perform arithmetic operations such as counting, ranking, and calculating percentages, but you cannot perform operations such as addition, subtraction, multiplication, or division.

3. Interval Scale: This scale involves measuring variables with equal distances between them, but the scale does not have a true zero point. Examples include temperature (in Celsius or Fahrenheit), time, and IQ scores. On an interval scale, 0 does not indicate a total absence of the variable being measured.

Interval scale is a level of measurement in statistics that involves measuring variables with equal distances between them, but the scale does not have a true zero point. The distance between the intervals or categories is meaningful, and the values on an interval scale can be expressed as numerical values. Examples of interval scale data include temperature (in Celsius or Fahrenheit), time, and IQ scores. In interval scale data, you can perform arithmetic operations such as addition and subtraction, but you cannot perform operations such as multiplication or division. The concept of zero in interval scale data does not indicate a total absence of the variable being measured, but rather a point on the scale that is arbitrary and does not have a true value of zero.

4. Ratio Scale: This is the highest level of measurement and involves measuring variables that have a true zero point. Examples include height, weight, and income. On a ratio scale, 0 indicates a complete absence of the variable being measured, and measurements are meaningful in terms of ratios.

Ratio scale is a level of measurement in statistics that involves measuring variables that have a true zero point. The values on a ratio scale can be expressed as numerical values, and the distance between the intervals or categories is meaningful. Examples of ratio scale data include height, weight, distance, and income. In ratio scale data, you can perform all four arithmetic operations (addition, subtraction, multiplication, and division) because the scale has a true zero point, indicating a complete absence of the variable being measured. In ratio scale data, the concept of zero is absolute and meaningful, and you can calculate ratios between values. For example, if one person's height is twice as high as another person's height, you can say that the first person's height is a ratio of 2:1 compared to the second person's height.

Types of sampling

In statistics, **sampling** is the process of selecting a subset (sample) from a larger population to make inferences about the whole population. There are two main categories of sampling methods:

I. Probability Sampling

In probability sampling, each member of the population has a known, non-zero chance of being selected.

1. Simple Random Sampling

- Every individual has an equal chance of being selected.
- **Example:** Choosing 10 students randomly from a class of 50 using a lottery method.

2. Systematic Sampling

- Selecting every k th individual from a list after a random start.
- **Example:** Selecting every 5th household from a list of 1,000 households.

3. Stratified Sampling

- Population is divided into subgroups (*strata*) and samples are taken from each stratum.
- **Example:** Dividing a college population into male and female students and randomly selecting 50 from each group.

4. Cluster Sampling

- Population is divided into clusters, and a few clusters are randomly selected for full observation.
- **Example:** Randomly selecting 3 villages out of 20 and surveying all households in those 3 villages.

II. Non-Probability Sampling

In non-probability sampling, not all members have a chance of being selected. Often used when probability sampling is impractical.

1. Convenience Sampling

- Sampling those who are easiest to reach.
- **Example:** Surveying students in a nearby canteen because they are easily accessible.

2. Judgmental or Purposive Sampling

- The researcher selects samples based on their judgment about who is most useful.
- **Example:** Interviewing only village leaders to study rural governance.

3. Snowball Sampling

- Existing study subjects recruit future subjects from among their acquaintances.
- **Example:** Studying drug users by asking each participant to refer others they know.

4. Quota Sampling

- Researcher ensures a specific number (quota) of participants from different subgroups.
- **Example:** Interviewing 50 urban and 50 rural residents, without random selection.

Sampling Type	Category	Key Feature	Example
Simple Random Sampling	Probability	Equal chance for all	Lottery method
Systematic Sampling	Probability	Every k th element	Every 10th person on a list
Stratified Sampling	Probability	Sample from subgroups (strata)	50 men and 50 women
Cluster Sampling	Probability	Sample entire selected clusters	3 schools from a district
Convenience Sampling	Non-Probability	Easy access	Surveying in a café
Judgmental Sampling	Non-Probability	Researcher decides who to sample	Experts only
Snowball Sampling	Non-Probability	Participants recruit others	Hidden populations
Quota Sampling	Non-Probability	Fixed quota from each subgroup	20 youth, 20 elderly

2. Frequency distribution and its implication, construction of cumulative frequency distribution curves; extraction of samples from a data matrix in excel using random, systematic and stratified method of sampling.

Frequency distribution and its implications

You are given, below the production of paddy (metric-ton) of 90 farms of a region. Arrange the data in frequency distribution table with 10 equal classes. [**Univariate Frequency Distribution, Continuous Data, after two decimal place, Inclusive Method**]

1.10	1.13	1.44	1.44	1.27	1.17	1.98	1.36	1.30
1.27	1.24	1.73	1.51	1.12	1.42	1.03	1.58	1.46
1.40	1.21	1.62	1.31	1.55	1.33	1.04	1.48	1.20
1.60	1.70	1.09	1.49	1.86	1.95	1.51	1.82	1.42
1.29	1.54	1.38	1.87	1.41	1.77	1.15	1.57	1.07
1.65	1.36	1.67	1.41	1.55	1.22	1.69	1.67	1.34
1.45	1.39	1.25	1.26	1.75	1.57	1.53	1.37	1.59
1.19	1.52	1.56	1.32	1.81	1.40	1.47	1.38	1.62
1.76	1.28	1.92	1.46	1.46	1.35	1.16	1.42	1.78
1.68	1.47	1.37	1.35	1.47	1.43	1.66	1.56	1.48

Step-I: Range = Highest Value - Lowest Value = 1.98 - 1.03 = 0.95

Step-II: Selection of no. of classes, [Sturge's rule (C)] = $1+3.332 \log N$ (N = Total Frequency)
 = $1+3.332 \log 90$
 = 7.5115 (8 approx., but 10 as per question)

Step-III: Determination of class interval (i) = $\frac{\text{Range}}{\text{No. of Class}} = \frac{0.95}{10} = 0.095 = (0.10 \text{ approx})$

Frequency Distribution Table (production of paddy, metric-ton)

Class Limit Paddy production (Metric ton)	Class Boundary Paddy production (Metric ton)	Class Mark (x_i) Paddy production (Metric ton)	Class Width(w_i) Paddy production (Metric ton)	Tally Marks	Frequency(f_i) (No. of Farms)	Remarks
1.01-1.10	1.005-1.105	1.055	0.10		5	*INCLUSIVE METHOD IS APPLIED TO FORM CLASSES.
1.11-1.20	1.105-1.205	1.155	0.10		7	
1.21-1.30	1.205-1.305	1.255	0.10		10	
1.31-1.40	1.305-1.405	1.355	0.10		15	
1.41-1.50	1.405-1.505	1.455	0.10		18	
1.51-1.60	1.505-1.605	1.555	0.10		14	
1.61-1.70	1.605-1.705	1.655	0.10		9	
1.71-1.80	1.705-1.805	1.755	0.10		5	
1.81-1.90	1.805-1.905	1.855	0.10		4	
1.91-2.00	1.905-2.005	1.955	0.10		3	
					$\Sigma f_i / N = 90$	

Note:

- Univariate frequency distribution is given.
- Inclusive method is applied. Class limit & class boundary is not same. [Stated Class Limit \neq Class Boundary]
- In this question data is given to two decimal point, so class limit must also be shown two decimal point & class boundary is shown by three decimal points; in case of inclusive method.
- It is also calculated by exclusive method that is much easier than previous. It is just example; students are directed to go for *exclusive method as it is continuous data.*

Histograms and frequency distribution curves

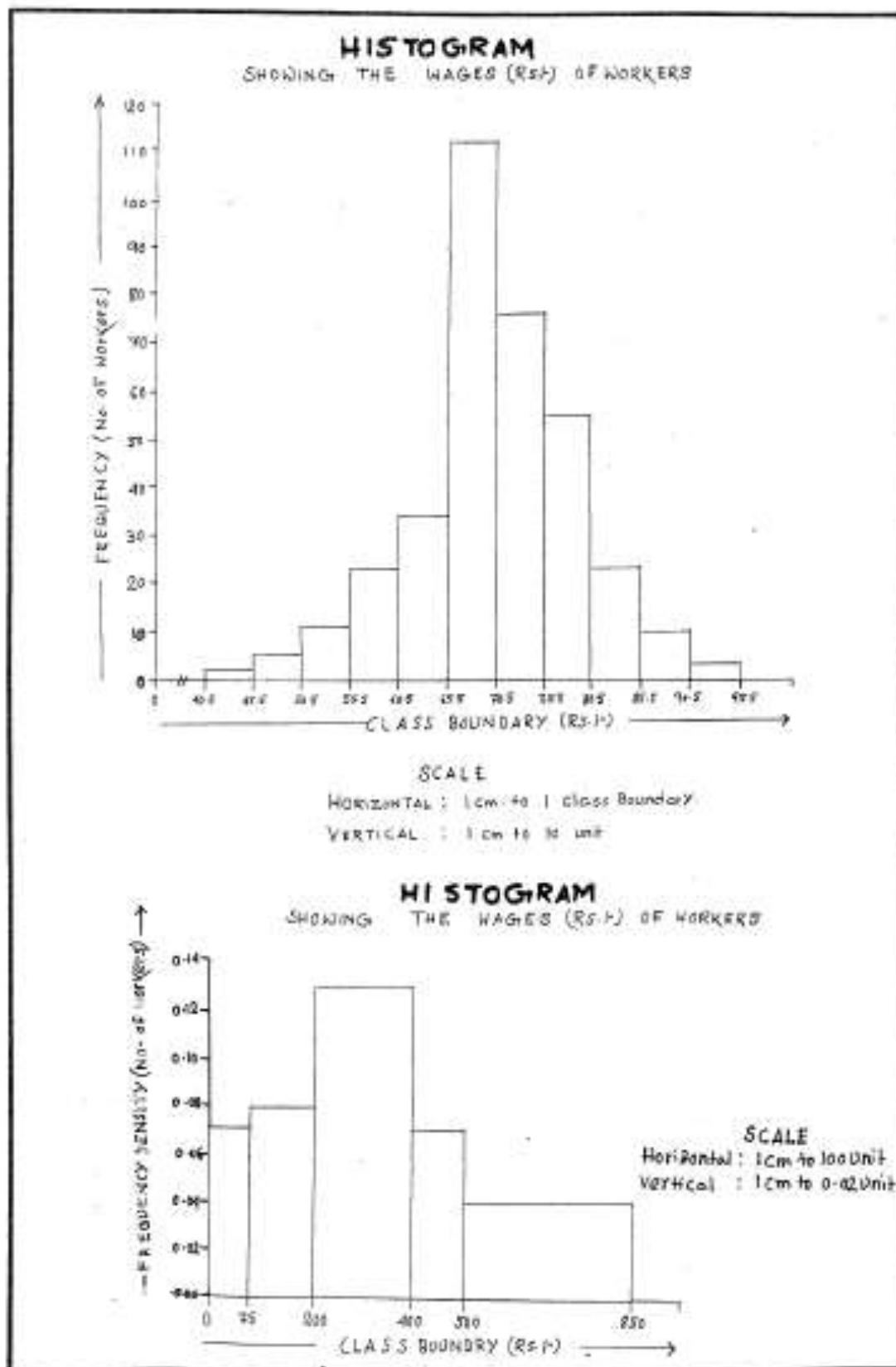
With the help of the following data draw Histogram, Frequency polygon & Frequency curve, with suitable scale.

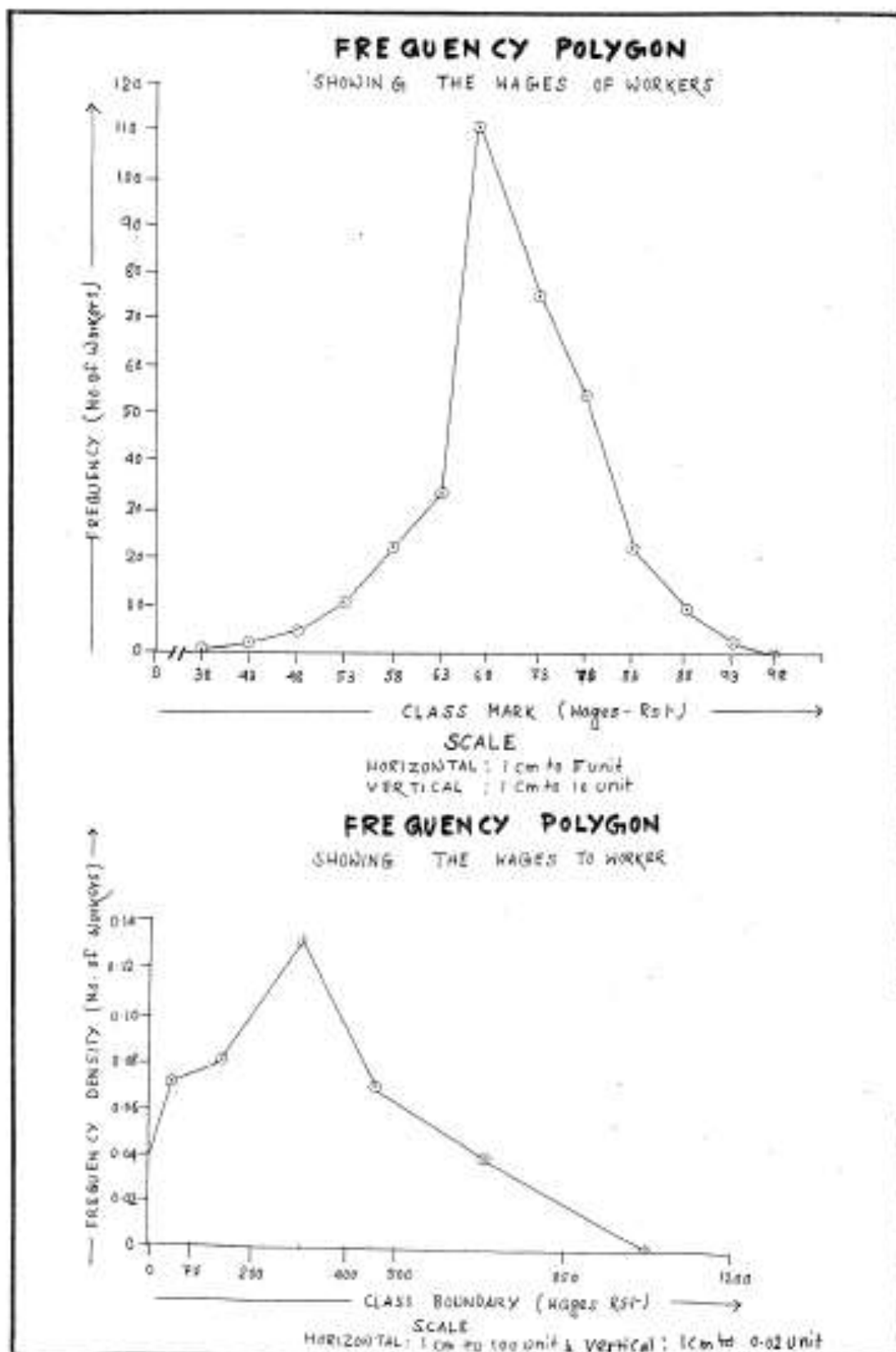
Frequency Distribution [Wages (Rs/-) of Worker]

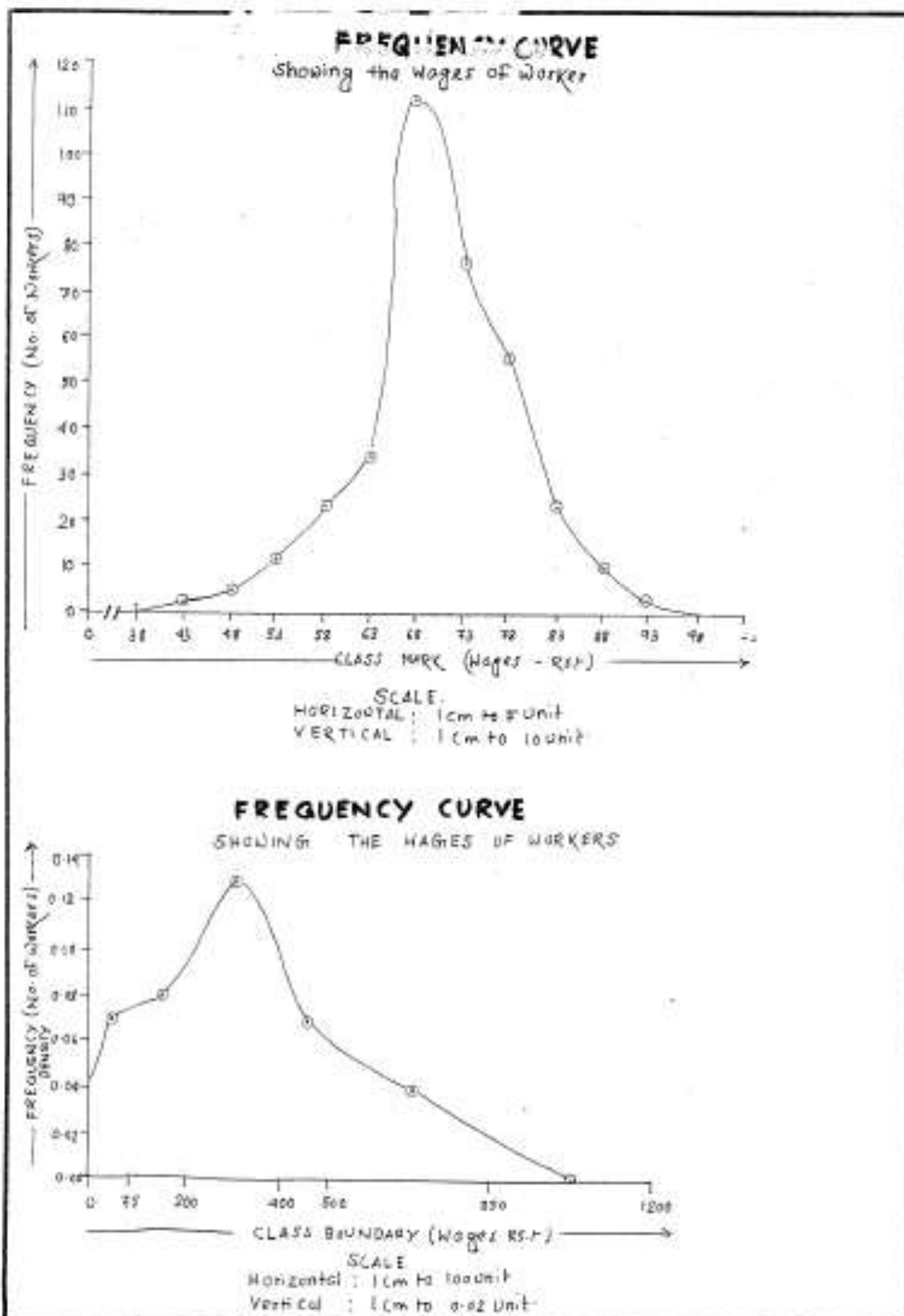
Class Boundary (Rs/-)	Frequency (f _i) No. of Workers	Class Marks (x _i)	Class Width (w _i)	Remarks
40.50-45.50	2	43	5	Equal Class Boundary
45.50 -50.50	5	48	5	
50.50 -55.50	11	53	5	
55.50- 60.50	23	58	5	
60.50-65.50	34	63	5	
65.50-70.50	112	68	5	
70.50 -75.50	76	73	5	
75.50 - 80.50	55	78	5	
80.50-85.50	23	83	5	
85.50-90.50	10	88	5	
90.50- 95.50	3	93	5	

Frequency Distribution [Wages (Rs/-) of Worker]

Class Boundary (Rs/-)	Frequency (f _i) No. of Workers	Class Marks (x _i)	Class Width (w _i)	Remarks	Class Boundary (Rs/-)
0-75	5	75	37.5	0.07	Unequal Class Boundary
75-200	10	125	137.5	0.08	
200-400	26	200	300.0	0.13	
400-500	7	100	450.0	0.07	
500-850	6	350	675	0.04	







Extraction of samples from a data matrix in Excel

Using random, systematic and stratified methods of sampling

To extract samples from a data matrix in Excel using random, systematic, and stratified sampling, you can leverage Excel's built-in functions and the Data Analysis add-in. Random sampling uses the =RAND() function, systematic sampling selects every nth row, and stratified sampling divides the data into subgroups and samples from each.

1. Random Sampling:

- **Step 1:** In a new column, use the =RAND() function in each row to generate a random number. For example, in cell B2, type =RAND(), press Enter, and then copy this formula down the column.
- **Step 2:** Sort your data by the random number column (A-Z or Z-A) to get a random sample.

2. Systematic Sampling:

- **Step 1:** Determine the sampling interval (n). For example, if you want to sample every 5th row, n = 5.
- **Step 2:** Use a formula to identify the rows to be sampled. For example, if n = 5, in cell C2, type =IF(MOD(ROW()-1,5)=0, "Sample", "") and drag it down. The "Sample" value will appear in rows 0, 5, 10, 15, etc.
- **Step 3:** Filter the column with the formula results to extract the sampled rows.

3. Stratified Sampling:

- **Step 1:** Divide your data into strata (groups) based on a specific characteristic (e.g., age, gender, etc.).
- **Step 2:** Determine the desired sample size for each stratum.
- **Step 3:** Use Excel's Data Analysis add-in (if installed) to perform stratified sampling:
 - Go to Data > Data Analysis > Sampling.
 - Choose "Stratified Random" and specify the strata column, input range, sample size, and output range.
- **Step 4:** Alternatively, you can manually sample within each stratum using the random sampling method described above.

Example:

Imagine you have a list of employees and want to sample 20% for an interview, stratifying by department.

1. **Random Sampling:** You could use =RAND() in a new column, sort by that, and select the top 20% of the sorted list.
2. **Systematic Sampling:** You could select every 5th employee.
3. **Stratified Sampling:** You could sample a specific number of employees from each department to reflect the department's proportion in the company.

❖ **Following data set is female literacy rate (%) of 25 blocks of Purba Medinipur district, 2011.**

1. Prepare a sample [n = 5 (20% of data set)] from the given data set through applying random sampling techniques.
2. Derive the sample arithmetic mean.

SI. No.	C.D.Block	Female Literacy Rate (%)	SI. No.	C.D.Block	Female Literacy Rate (%)
1	Tamluk	83.74	14	Pataspur-II	80.53
2	Sahid Matangini	80.89	15	Bhagabanpur-I	82.50
3	Paskura-I	78.07	16	Egra-I	78.72
4	Kolaghat	78.37	17	Egra-II	79.45
5	Moyna	80.24	18	Khejuri-I	84.36
6	Nandakumar	80.07	19	Khejuri-II	79.80
7	Chandipur	82.93	20	Bhagabanpur-II	86.29
8	Mahisadal	80.84	21	Ramnagar-I	81.72
9	Nandigram-I	80.71	22	Ramnagar-II	83.37
10	Nandigram-II	84.88	23	Contai-I	83.73
11	Sutahata	80.09	24	Deshapran	87.02
12	Haldia	81.97	25	Contai-III	84.75
13	Pataspur-I	79.90	-	-	-

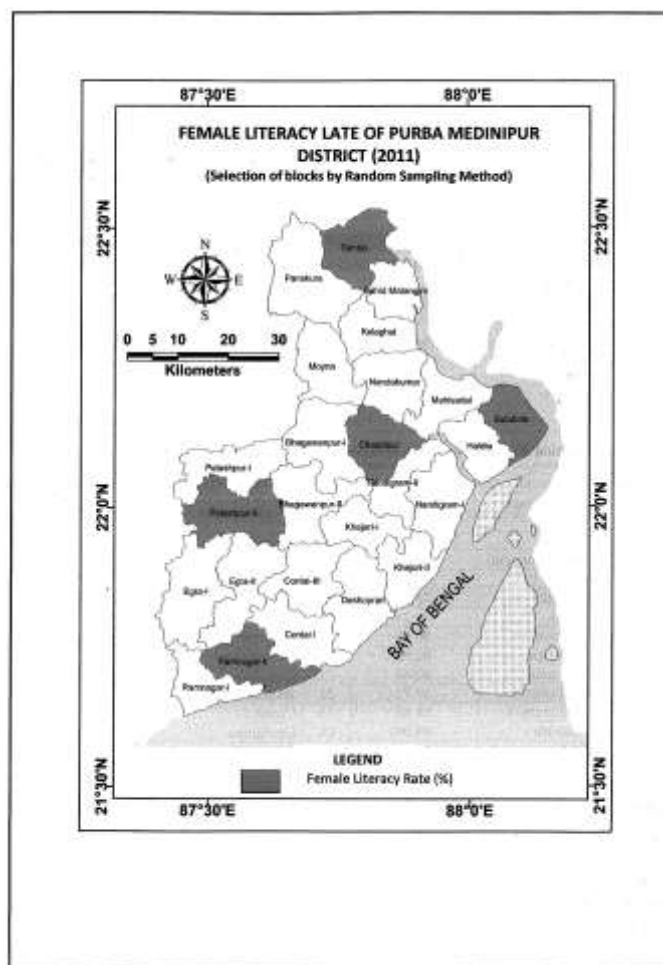
Computation Table for Selection of Samples from Population by Random Sampling

SI. No.	C.D.Block	Female Literacy Rate (%)	SI. No.	C.D.Block	Female Literacy Rate (%)
1	Tamluk	83.74	14	Pataspur-II	80.53
2	Sahid Matangini	80.89	15	Bhagabanpur-I	82.50
3	Paskura-I	78.07	16	Egra-I	78.72
4	Kolaghat	78.37	17	Egra-II	79.45
5	Moyna	80.24	18	Khejuri-I	84.36
6	Nandakumar	80.07	19	Khejuri-II	79.80
7	Chandipur	82.93	20	Bhagabanpur-II	86.29
8	Mahisadal	80.84	21	Ramnagar-I	81.72
9	Nandigram-I	80.71	22	Ramnagar-II	83.37
10	Nandigram-II	84.88	23	Contai-I	83.73
11	Sutahata	80.09	24	Deshapran	87.02
12	Haldia	81.97	25	Contai-III	84.75

1. Selections of blocks by random sampling are **Tamluk (83.74%)**; **Chandipur (82.93%)**; **Sutahata (80.09%)**; **Pataspur-II (80.53%)** & **Ramnagar-II (83.37%)**.
2. Sample Mean = $\frac{\sum X}{n} = \frac{410.66\%}{5} = 82.13\%$, Population Mean = $\frac{\sum X}{n} = \frac{2044.94\%}{25} = 81.80\%$

Interpretation

Here Simple random sampling techniques are applied for selection of samples. When homogeneous data but no. of population is quite low are given then this method is applied. There is no significant difference between population mean & sample mean (82.13% - 81.80%) = 0.33%. So it is assumed that the selection of sample is very much correct.



Following data set is female literacy rate (%) of 25 blocks of Purba Medinipur district, 2011.

1. Prepare a sample [n = 6 (24% of data set)] from the given data set through applying systematic sampling techniques without replacement after arranging all the data items in array and selecting every 5th item starting from first one .
2. Derive the sample arithmetic mean.

SI. No.	C.D.Block	Female Literacy Rate (%)	SI. No.	C.D.Block	Female Literacy Rate (%)
1	Tamluk	83.74	14	Pataspur-II	80.53
2	Sahid Matangini	80.89	15	Bhagabanpur-I	82.50
3	Paskura-I	78.07	16	Egra-I	78.72
4	Kolaghat	78.37	17	Egra-II	79.45
5	Moyna	80.24	18	Khejuri-I	84.36
6	Nandakumar	80.07	19	Khejuri-II	79.80
7	Chandipur	82.93	20	Bhagabanpur-II	86.29
8	Mahisadal	80.84	21	Ramnagar-I	81.72
9	Nandigram-I	80.71	22	Ramnagar-II	83.37
10	Nandigram-II	84.88	23	Contai-I	83.73
11	Sutahata	80.09	24	Deshapran	87.02

12	Haldia	81.97	25	Contai-III	84.75
13	Pataspur-I	79.90	-	-	-

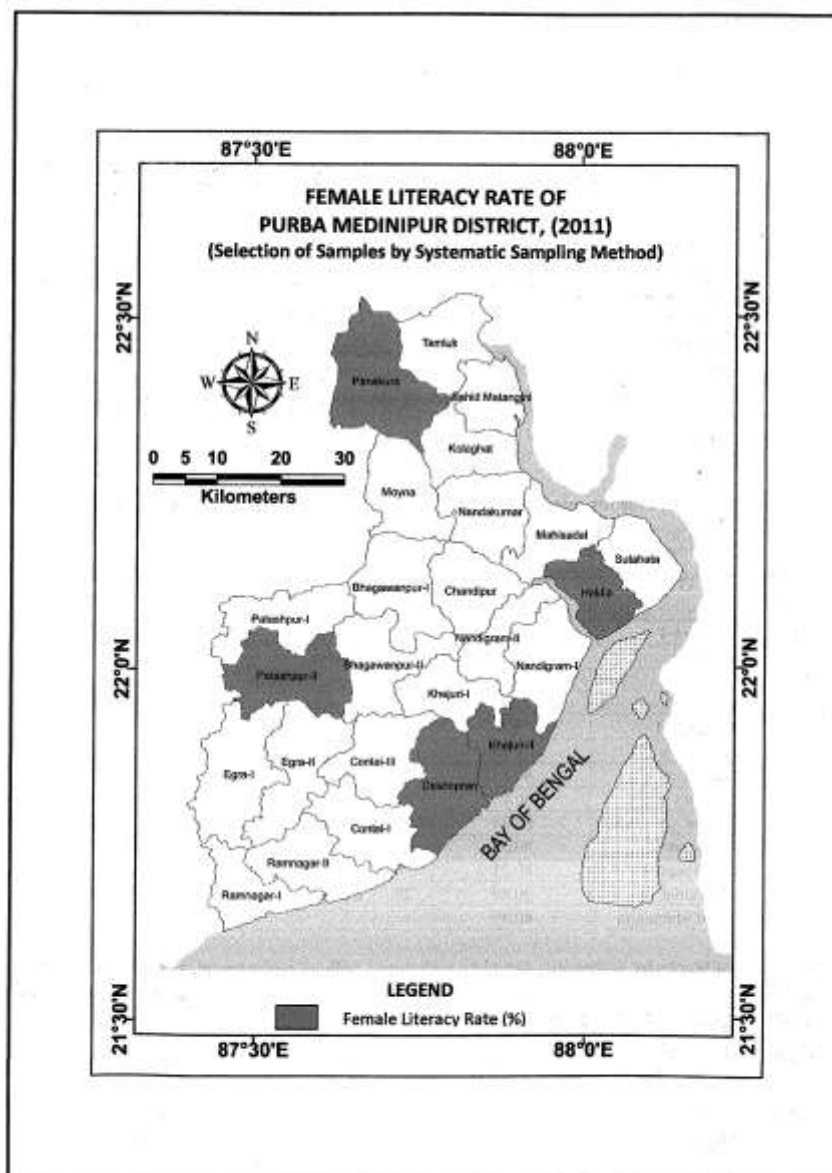
**Computation Table for Selection of Samples from Population by Systematic Sampling
(Ascending Order)**

SI. No.	C.D.Block	Female Literacy Rate (%)	SI. No.	C.D.Block	Female Literacy Rate (%)
1	Paskura-I	78.07	14	Ramnagar-I	81.72
2	Kolaghat	78.37	15	Haldia	81.97
3	Egra-I	78.72	16	Bhagabanpur-I	82.50
4	Egra-II	79.45	17	Chandipur	82.93
5	Khejuri-II	79.80	18	Ramnagar-II	83.37
6	Pataspur-I	79.90	19	Contai-I	83.73
7	Nandakumar	80.07	20	Tamluk	83.74
8	Sutahata	80.09	21	Khejuri-I	84.36
9	Moyna	80.24	22	Contai-III	84.75
10	Pataspur-II	80.53	23	Nandigram-II	84.88
11	Nandigram-I	80.71	24	Bhagabanpur-II	86.29
12	Mahisadal	80.84	25	Deshapran	87.02
13	Sahid Matangini	80.89	-	-	-

1. Selections of blocks by systematic sampling techniques without replacement, selecting every 5th item starting from first one is **Paskura-I (78.07%)**, **Khejuri-II (79.80)**, **Pataspur-II (80.53%)**, **Haldia (81.97%)**, **Tamluk(83.74%)** & **Deshapran (87.02%)**.
2. Sample Mean = $\frac{\sum X}{n} = \frac{491.13\%}{6} = 81.86\%$, Population Mean = $\frac{\sum X}{n} = \frac{2044.94\%}{25} = 81.80\%$

Interpretation

Here Systematic sampling technique is applied for selection of samples from population. When homogeneous data, but no. of population is quite more than this method is applied. There is no significant difference between population mean & sample mean (81.86% - 81.80%) = 0.06 %. So it is assumed that the selection of sample is very much correct.



Following data set is female literacy rate (%) of 25 blocks of Purba Medinipur district, 2011.

1. Prepare a sample [n = 5 (20% of data set)] from the given data set through applying stratified sampling techniques.
2. Derive the sample arithmetic mean.

SI. No.	C.D.Block	Female Literacy Rate (%)	SI. No.	C.D.Block	Female Literacy Rate (%)
1	Tamluk	83.74	14	Pataspur-II	80.53
2	Sahid Matangini	80.89	15	Bhagabanpur-I	82.50
3	Paskura-I	78.07	16	Egra-I	78.72
4	Kolaghat	78.37	17	Egra-II	79.45
5	Moyna	80.24	18	Khejuri-I	84.36
6	Nandakumar	80.07	19	Khejuri-II	79.80

7	Chandipur	82.93	20	Bhagabanpur-II	86.29
8	Mahisadal	80.84	21	Ramnagar-I	81.72
9	Nandigram-I	80.71	22	Ramnagar-II	83.37
10	Nandigram-II	84.88	23	Contai-I	83.73
11	Sutahata	80.09	24	Deshapran	87.02
12	Haldia	81.97	25	Contai-III	84.75
13	Pataspur-I	79.90	-	-	-

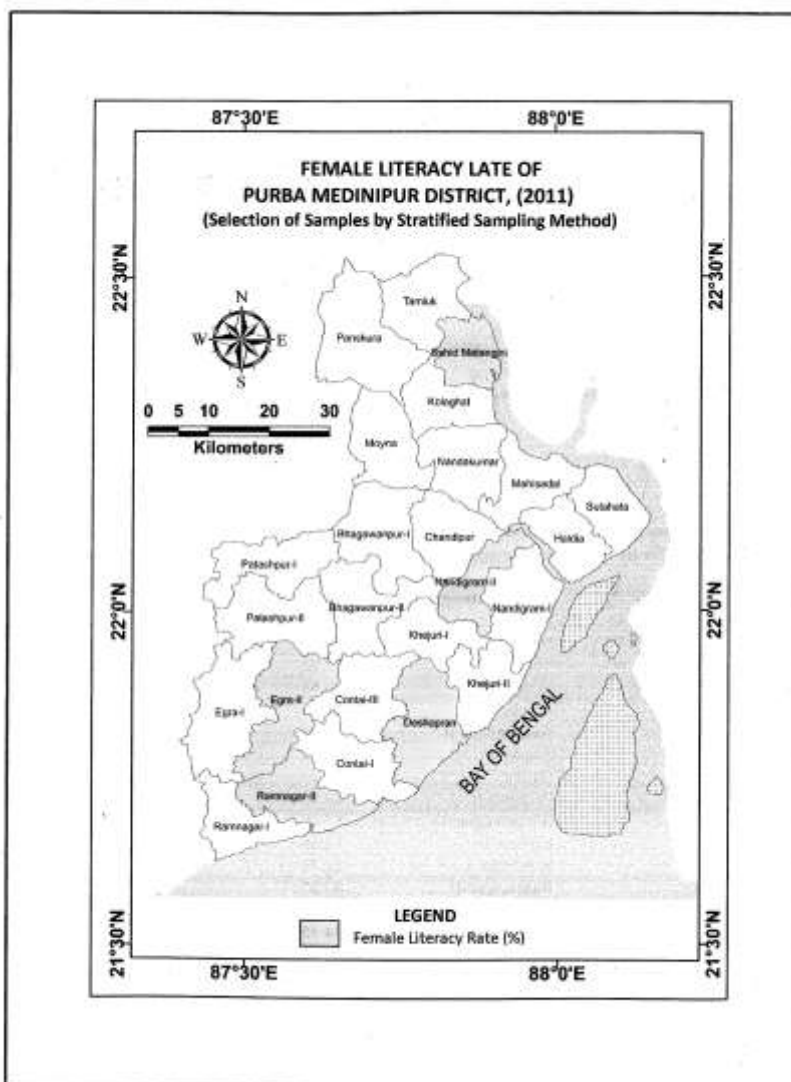
Computation Table for Selection of Samples from Population by Stratified Sampling

Female Literacy Rate (%)					
Low Education Rate (Below 80 %)		Medium Education Rate (80% - 85 %)		Medium Education Rate (Above 85%)	
Paskura-I	78.07	Tamluk	83.74	Bhagabanpur-II	86.29
Egra-I	78.72	Sahid Matangini	80.89	Deshapran	87.02
Egra-II	79.45	Moyna	80.24		
Khejuri-II	79.80	Nandakumar	80.07		
Pataspur-I	79.90	Chandipur	82.93		
Kolaghat	78.37	Mahisadal	80.84		
		Nandigram-I	80.71		
		Nandigram-II	84.88		
		Sutahata	80.09		
		Haldia	81.97		
		Pataspur-II	80.53		
		Bhagabanpur-I	82.50		
		Khejuri-I	84.36		
		Ramnagar-I	81.72		
		Ramnagar-II	83.37		
		Contai-I	83.73		
		Contai-III	84.75		

1. Selections of blocks from different stratification layer are **Egra-II (79.45%)**, **Sahid Matangini (80.89%)**, **Nandigram-II (84.88%)**, **Ramnagar-II (83.37%)** & **Deshapran (87.02%)**.
2. Sample Mean = $\frac{\sum X}{n} = \frac{415.61\%}{5} = 81.86\%$, Population Mean = $\frac{\sum X}{n} = \frac{2044.94\%}{25} = 81.80\%$

Interpretation

Here Stratified sampling techniques are applied for selection of samples. When heterogeneous data set is given then to reduce heterogeneity, some stratification layering has been done, after that samples are collected from each layer, and then this method is applied. There is no significant difference between population mean & sample mean (83.12% - 81.80%) = 1.32%. So it is assumed that the selection of sample is very much correct.



3. The shape of the distribution of data - skewness and kurtosis; normal distribution- properties of normal distribution. Examples of normal distribution; Representation of data distribution through boxplot. Data standardization in statistics.

The shape of the distribution of data - skewness and kurtosis

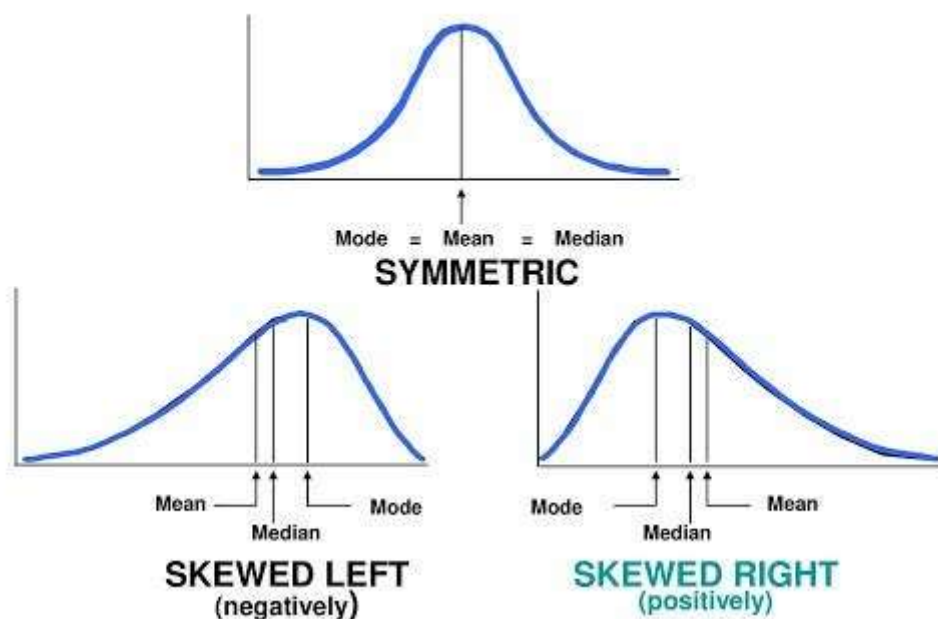
Skewness and kurtosis are statistical measures that describe the shape of a data distribution. Skewness measures the asymmetry of the distribution, while kurtosis measures the "tailedness" or peakedness of the distribution. These measures are crucial in data analysis to understand data patterns and identify outliers.

Skewness:

Skewness measures the degree of asymmetry in a distribution. It indicates whether the data is concentrated more on one side of the mean than the other. **Types:**

1. **Symmetrical:** The distribution is balanced, with the left and right sides of the distribution mirroring each other. The mean, median, and mode are all equal in a symmetrical distribution.
2. **Positively Skewed (Right-Skewed):** The tail of the distribution is longer on the right side, indicating a higher concentration of data on the left side and some extreme values on the right.
3. **Negatively Skewed (Left-Skewed):** The tail of the distribution is longer on the left side, indicating a higher concentration of data on the right side and some extreme values on the left.

Describe the shape, center, and spread of a distribution... for shape, see below...



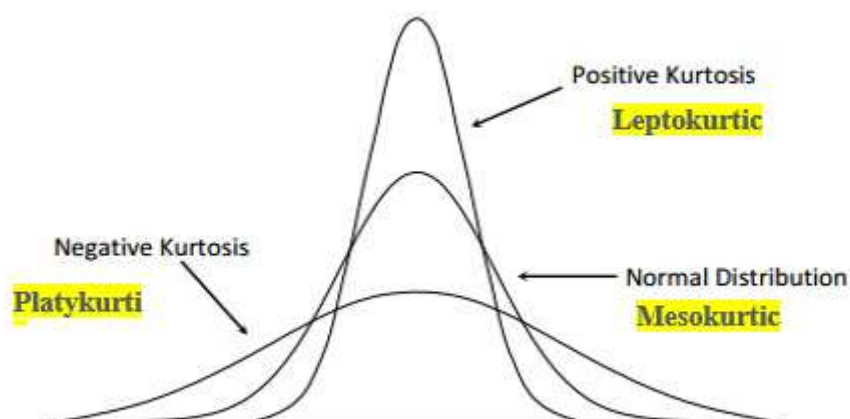
- **Interpretation:**

- A skewness value of 0 indicates a symmetrical distribution.
- A positive skewness value (greater than 0) indicates a right-skewed distribution.
- A negative skewness value (less than 0) indicates a left-skewed distribution.

Kurtosis:

Kurtosis measures the "tailedness" of a distribution, indicating the presence and extent of outliers. It also describes the peakedness or flatness of the distribution. **Types:**

1. **Mesokurtic:** The distribution is similar to a normal distribution, with a moderate peak and tails.
2. **Leptokurtic:** The distribution has a sharp peak and heavy tails, indicating a greater concentration of data around the mean and more outliers.
3. **Platykurtic:** The distribution has a flat peak and light tails, indicating a more even distribution of data and fewer outliers.



Interpretation:

- Kurtosis of 3 (or excess kurtosis of 0) is considered mesokurtic, indicating a normal distribution.
- Kurtosis greater than 3 (positive excess kurtosis) indicates a leptokurtic distribution.
- Kurtosis less than 3 (negative excess kurtosis) indicates a platykurtic distribution.

Importance: Skewness and kurtosis are valuable tools in data analysis for:

- **Identifying data patterns:** Understanding the shape of a distribution helps in identifying trends and patterns in the data.
- **Detecting outliers:** Kurtosis can be used to identify outliers, which are extreme values that can significantly impact data analysis.
- **Choosing appropriate statistical tests and models:** The shape of a distribution can influence the choice of statistical tests and models used in analysis.
- **Data transformation and normalization:** Understanding skewness and kurtosis can help in transforming or normalizing data for better analysis and modeling.

Normal distribution- properties of normal distribution.

Examples of normal distribution;

Normal Distribution

A **normal distribution** is a **bell-shaped**, symmetric probability distribution that is commonly seen in nature, social sciences, and statistics. It is also called the **Gaussian distribution**.

Properties of a Normal Distribution

Property	Description
1. Symmetry	The curve is symmetric about the mean .
2. Bell-shaped	The highest point is at the mean, and the tails taper off equally.
3. Mean = Median = Mode	All central measures are equal.
4. Asymptotic tails	The tails approach the x-axis but never touch it.
5. Total area under the curve = 1	Represents total probability (100%).
6. Empirical Rule (68-95-99.7 Rule)	In a normal distribution: • ~68% of data falls within 1 SD • ~95% within 2 SD • ~99.7% within 3 SD of the mean.

Empirical Rule Illustration (Assume Mean = 0, SD = 1)

Range	Percent of Data
$\mu \pm 1\sigma$ (-1 to +1)	~68%
$\mu \pm 2\sigma$ (-2 to +2)	~95%
$\mu \pm 3\sigma$ (-3 to +3)	~99.7%

Examples of Normal Distribution

Context	Example
Natural sciences	Heights of adult humans in a population
Education	IQ scores of students
Social science	Age at first marriage (in large populations)
Medical research	Blood pressure or cholesterol levels
Manufacturing	Measurement errors in production (like length of machine-made screws)

Representation of data distribution through boxplot.

Boxplot (Box-and-Whisker Plot)

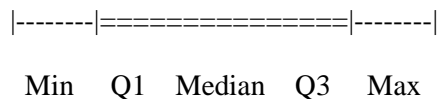
A **boxplot** is a graphical representation that shows the **spread, central tendency, and skewness** of a dataset using five summary statistics. It is widely used in **exploratory data analysis**.

Five-Number Summary in a Boxplot

1. **Minimum** – the smallest value (excluding outliers)
2. **Q1 (1st Quartile)** – 25% of data falls below this

3. **Median (Q2)** – middle value (50% below, 50% above)
4. **Q3 (3rd Quartile)** – 75% of data falls below this
5. **Maximum** – the largest value (excluding outliers)

Boxplot Components



- **Box:** Shows the interquartile range ($IQR = Q3 - Q1$)
- **Line inside the box:** Median (Q2)
- **Whiskers:** Extend from Q1 to minimum and Q3 to maximum (excluding outliers)
- **Dots/asterisks:** Represent outliers (values beyond $1.5 \times IQR$)

What a Boxplot Shows

Feature	Interpretation
Spread (IQR)	The width of the box indicates variability
Skewness	If the median is not centered, the data is skewed
Outliers	Shown as individual points outside whiskers
Symmetry	If whiskers are of equal length, data is symmetric

Example Interpretation

Dataset: Test scores of 20 students

Scores: [45, 48, 50, 52, 54, 55, 56, 58, 60, 62, 64, 65, 66, 68, 70, 72, 74, 75, 78, 80]

Summary	Value
Minimum	45
Q1	54
Median	62
Q3	70

Maximum	80
---------	----

Boxplot Interpretation:

- The box spans from 54 to 70 (IQR = 16).
- Median is 62 (centered → roughly symmetrical).
- No outliers.
- The data is moderately spread and balanced.

Use of Boxplots in Geography/Social Science

- Comparing income levels across regions
- Analyzing rainfall variation across seasons
- Showing literacy rates by gender in different districts

Data standardization in statistics.

Data standardization (also called z-score normalization) is a process of rescaling variables so they have a mean of 0 and a standard deviation of 1. It allows variables with different units or scales to be compared meaningfully.

Why Standardize Data?

Purpose	Explanation
Comparability	Puts all variables on the same scale for fair comparison.
Preprocessing for analysis	Required in many statistical models (e.g., regression, PCA, clustering).
Handling different units	Useful when variables are measured in different units (e.g., income in ₹, height in cm).

Formula for Standardization (Z-score)

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- X = original value
- μ = mean of the data
- σ = standard deviation of the data
- Z = standardized value (z-score)

Example

Suppose we have heights of five students in cm:

$X = [160, 165, 170, 175, 180]$

- Mean (μ) = 170
- Standard Deviation (σ) = 7.9 (approx)

Standardizing 180 cm:

$$Z = \frac{180 - 170}{7.9} \approx 1.27$$

So, 180 cm is **1.27 standard deviations above the mean.**

Key Characteristics of Standardized Data

Statistic	Value
Mean	0
SD	1
Unit	Dimensionless (unitless)

When to Use Standardization

- Before performing **Principal Component Analysis (PCA)**
- In **machine learning algorithms** sensitive to scale (e.g., K-means, SVM, logistic regression)
- For **comparing scores** (e.g., test marks across different subjects)

Real-life Example (Social Sciences/Geography)

- **Income data** across regions: Standardize before comparing because urban and rural incomes vary in range and scale.
- **Climate variables** like temperature and rainfall: Standardize to perform multivariate climate zone classification.

4. Measures of central tendency and dispersion: mean, mode, median, standard deviation

Central tendency and dispersion

MEASURES OF CENTRAL TENDENCY
(MEAN VALUES)

Measures	Name	Symbol	*Ungroup Data	*Group Data
MEAN \bar{X}	Arithmetic Mean	AM (Simple Arithmetic Mean)	$AM = \frac{\sum Xi}{n}$ $x_i = \text{variables}$ $n = \text{No. of observations}$	$AM = \frac{\sum fi.Xi}{N}$ $f_i = \text{Frequency}$ $x_i = \text{Class Mark}$ $N = \text{Total frequency}$
		AM (Weighted Arithmetic Mean)	$\frac{w_1x_1 + w_2x_2 + \dots + w_nx_n}{w_1 + w_2 + \dots + w_n}$ $WAM = \frac{\sum wx}{\sum w}$	-----
	Geometric Mean	GM	$GM = \text{Antilog } \frac{\sum \text{Log } xi}{n}$ $x_i = \text{variables}$ $n = \text{No. of observations}$	$GM = \text{Antilog } \frac{\sum fi.\text{Log } xi}{\sum fi}$ $f_i = \text{Frequency}$ $x_i = \text{Class Mark}$ $\sum fi = \text{Total frequency}$
	Harmonic Mean	HM	$HM = \frac{n}{\sum (\frac{1}{xi})}$ $x_i = \text{variables}$ $n = \text{No. of observations}$	$HM = \frac{\sum fi}{\sum (\frac{fi}{xi})}$ $f_i = \text{Frequency}$ $x_i = \text{Class Mark}$
For any distribution $AM > GM > HM$ Ungroup series (Solved example): $23.90 > 22.07 > 20.39$ Group series (Solved example): $70.50 > 70.16 > 69.83$ *All calculations are solved by <i>Algebraic Method</i> . It is easier one. Same formula for equal & unequal class boundary.				
Empirical relation between Mean , Median & Mode $\text{Mean} - \text{Mode} = 3\text{Mean} - 3\text{Median}$ $\text{Mean} - 3\text{Mean} + 3\text{Median} = \text{Mode}$ $3 \text{ Median} - 3\text{Mean} = \text{Mode}$ $\text{Mode} = 3\text{Median} - 2\text{Mean}$				

Arithmetic Mean
[Ungroup & Group Series]

Ungroup Data		Group Data						
Data set (x_i)	Computation	Class Boundary Score (%)	Frequency No. of Students (f_i)	Class Mark (x_i)	($f_i x_i$)	Computation		
12	Ungroup Data Algebraic Method Arithmetic Mean $AM = \frac{\sum xi}{n}$ $= \frac{239}{10}$ $= 23.90$	55-60	7	57.5	402.5	Group Data Algebraic Method Arithmetic Mean $AM = \frac{\sum fi.Xi}{N}$ $= \frac{7050}{100}$ $= 70.50$		
14		60-65	10	62.5	625			
16		65-70	34	67.5	2295			
15		70-75	28	72.5	2030			
23		75-80	10	77.5	775			
28		80-85	8	82.5	660			
43		85-90	3	87.5	262.5			
34		$\sum f_i / N = 100$			$\sum f_i x_i = 7050$			
29		Same formula for equal & unequal class						
25								
$\sum x_i = 239$								

**Weighted Arithmetic Mean
[Ungroup Series]**

Price (Rs/-) Per Table (x)	No. of Table Sold (f/w)	fx or wx	Computation
36	14	504	$\text{Weighted Mean} = \frac{\sum wx \text{ or } \sum fx}{\sum w \text{ or } \sum f} = \frac{1628}{40} = 40.70$
40	11	440	
44	9	396	
48	6	288	
$\sum x = 168$	$\sum f \text{ or } \sum w = 40$	$\sum fx \text{ or } \sum wx = 1628$	$\text{Simple Mean} = \frac{\sum X}{n} = \frac{168}{4} = 42.00$

• Note: It is applicable only for ungroup series, when group series is given then frequencies of values are themselves treated as weights & the formula is same as like group mean.

**Geometric Mean
[Ungroup & Group Series]**

Ungroup Data			Group Data						
Data set (xi %)	Log (xi)	Computation	Class Boundary Score (%)	Frequency No. of Students (fi)	Class Mark (xi)	Log (xi)	(fi Log xi)	Computation	
12	1.07918	Ungroup Data Algebraic Method Geometric Mean $\text{Antilog} \frac{\sum \text{Log } xi}{n}$ $\text{Antilog} \frac{13.43969}{10}$ *Antilog 1.343969 = 22.07%	55-60	7	57.5	1.75967	12.31769	Group Data Algebraic Method Geometric Mean $\text{Antilog} \frac{\sum fi \text{Log } xi}{\sum fi}$ $\text{Antilog} \frac{184.6124}{100}$ *Antilog 1.8461284 = 70.17%	
14	1.14613		60-65	10	62.5	1.79588	17.95880		
16	1.20412		65-70	34	67.5	1.82930	62.19620		
15	1.17609		70-75	28	72.5	1.86034	52.08952		
23	1.36173		75-80	10	77.5	1.88930	18.89300		
28	1.44716		80-85	8	82.5	1.91645	15.33160		
43	1.63347		85-90	3	87.5	1.94201	5.82603		
34	1.53148		$\sum fi / N = 100$				$\sum (fi \text{Log } xi) = 184.61284$		
29	1.46239		*[Shift → log = Antilog]						
25	1.39794								
$\sum xi = 239$	$\sum \text{Log } xi = 13.43969$								

**Harmonic Mean
[Ungroup & Group Series]**

Ungroup Data			Group Data						
Data set (xi %)	$\frac{1}{xi}$	Computation	Class Boundary Score (%)	Frequency No. of Students (fi)	Class Mark (xi)	$\frac{fi}{xi}$	Computation		
12	0.08333	Ungroup Data Algebraic Method Harmonic Mean $\text{HM} = \frac{n}{\sum (\frac{1}{xi})}$ $= \frac{10}{0.49025}$ = 20.39%	55-60	7	57.5	0.12174	Group Data Algebraic Method Harmonic Mean $\text{HM} = \frac{\sum fi}{\sum (\frac{fi}{xi})}$ $\text{HM} = \frac{100}{1.43194}$ = 69.83%		
14	0.07143		60-65	10	62.5	0.16000			
16	0.06250		65-70	34	67.5	0.50370			
15	0.06666		70-75	28	72.5	0.38621			
23	0.04348		75-80	10	77.5	0.12903			
28	0.03571		80-85	8	82.5	0.09697			
43	0.02325		85-90	3	87.5	0.03429			
34	0.02941		$\sum fi / N = 100$					$\sum \frac{fi}{xi} = 1.43194$	
29	0.03448		*[Shift → log = Antilog]						
25	0.04000								
$\sum xi = 239$	$\sum \frac{1}{xi} = 0.49025$								

**MEASURES OF CENTRAL TENDENCY
(FRACTILES VALUES)**

Measures	Symbol	Name	Ungroup Data	Group Data
Median	\bar{X}	-----	$\frac{N+1}{2}$ th item	$L + \frac{\frac{N}{2} - F}{f} \times w$
Quartile	Q ₁	First/ Lower Quartile	$\frac{1 \times (N+1)}{4}$ th item	$L + \frac{\frac{1 \times N}{4} - F}{f} \times w$
	Q ₂	Second Quartile	$\frac{N+1}{2}$ th item	$L + \frac{\frac{N}{2} - F}{f} \times w$
	Q ₃	Third/ Upper Quartile	$\frac{3 \times (N+1)}{4}$ th item	$L + \frac{\frac{3 \times N}{4} - F}{f} \times w$
Decile	D ₁	First/ Lower Decile	$\frac{1 \times (N+1)}{10}$ th item	$L + \frac{\frac{1 \times N}{10} - F}{f} \times w$
	D ₂	Second Decile	$\frac{2 \times (N+1)}{10}$ th item	$L + \frac{\frac{2 \times N}{10} - F}{f} \times w$
	D ₃	Third Decile	$\frac{3 \times (N+1)}{10}$ th item	$L + \frac{\frac{3 \times N}{10} - F}{f} \times w$
	D ₄	Fourth Decile	$\frac{4 \times (N+1)}{10}$ th item	$L + \frac{\frac{4 \times N}{10} - F}{f} \times w$
	D ₅	Fifth Decile	$\frac{5 \times (N+1)}{10}$ th item	$L + \frac{\frac{5 \times N}{10} - F}{f} \times w$
	-----	-----	-----	-----
	D ₉	Ninth Decile/ Upper Decile	$\frac{9 \times (N+1)}{10}$ th item	$L + \frac{\frac{9 \times N}{10} - F}{f} \times w$
Percentile	P ₁	First Percentile/ Lower Percentile	$\frac{1 \times (N+1)}{100}$ th item	$L + \frac{\frac{1 \times N}{100} - F}{f} \times w$
	-----	-----	-----	-----
	P ₅₀	Fifth Percentile	$\frac{50 \times (N+1)}{100}$ th item	$L + \frac{\frac{50 \times N}{100} - F}{f} \times w$
	-----	-----	-----	-----
P ₉₉	Ninety nine Percentile/ Upper Percentile	$\frac{99 \times (N+1)}{100}$ th item	$L + \frac{\frac{99 \times N}{100} - F}{f} \times w$	
Description			x = Position N = No. of observation	L= Lower boundary of fractile Class f= Frequency of fractile class w = Interval of fractile class N = Total frequency F = Cumulative frequency upto fractile class

Note:

- All partition values are called together Fractile.
- In any distribution the value of $\bar{X} = Q_2 = D_5 = P_{50}$ is same.

Computation of Median
[Ungroup Series, Even & Odd Number]

Data set (x_i)	Array (Ascending Order)		Computation	Data set (x_i)	Array (Ascending Order)		Computation
	Rank	(x_i)			Rank	(x_i)	
12	1.	12	$N = 10$ (Even No.) $\bar{X} = \frac{N+1}{2}$ th item $= \frac{10+1}{2}$ $= 5.5$ Value = $\frac{23+25}{2}$ $= 24$	12	1.	12	$N = 9$ (Odd No.) $\bar{X} = \frac{N+1}{2}$ th item $= \frac{9+1}{2}$ $= 5$ Value = The value of Sl.no.5 is 23
14	2.	14		14	2.	14	
16	3.	15		16	3.	15	
15	4.	16		15	4.	16	
23	5.	23		23	5.	23	
28	6.	25		28	6.	28	
43	7.	28		43	7.	29	
34	8.	29		34	8.	34	
29	9.	34		29	9.	43	
25	10.	43					

Computation of Quartiles
[Ungroup Series]

Temperature ($^{\circ}\text{C}$) (x_i)	Array Temperature ($^{\circ}\text{C}$) [Ascending Order]		Computation
	Rank	(x_i)	
16.1	1.	16.1	$Q_1 = \frac{1 \times (N+1)}{4}$ th item $Q_1 = \frac{1 \times (12+1)}{4} = 3.25$ th item $= 3^{\text{RD}}$ Value + (4 TH Value - 3 RD Value) x 0.25 $= 17.9^{\circ} + (18.1^{\circ} - 17.9^{\circ}) \times 0.25$ $= 17.95^{\circ}\text{C}$ $Q_3 = \frac{3 \times (N+1)}{4}$ th item $Q_3 = \frac{3 \times (12+1)}{4} = 9.75$ th item $= 9^{\text{TH}}$ Value + (10 TH Value - 9 TH Value) x 0.75 $= 25.9^{\circ} + (26.7^{\circ} - 25.9^{\circ}) \times 0.75$ $= 26.5^{\circ}\text{C}$
16.5	2.	16.5	
17.9	3.	17.9	
20.4	4.	18.1	
23.4	5.	20.4	
25.9	6.	21.2	
27.9	7.	23.4	
27.4	8.	24.0	
26.7	9.	25.9	
24.0	10.	26.7	
21.2	11.	27.4	
18.1	12.	27.9	

**MEASURES OF CENTRAL TENDENCY
(MODAL VALUES)**

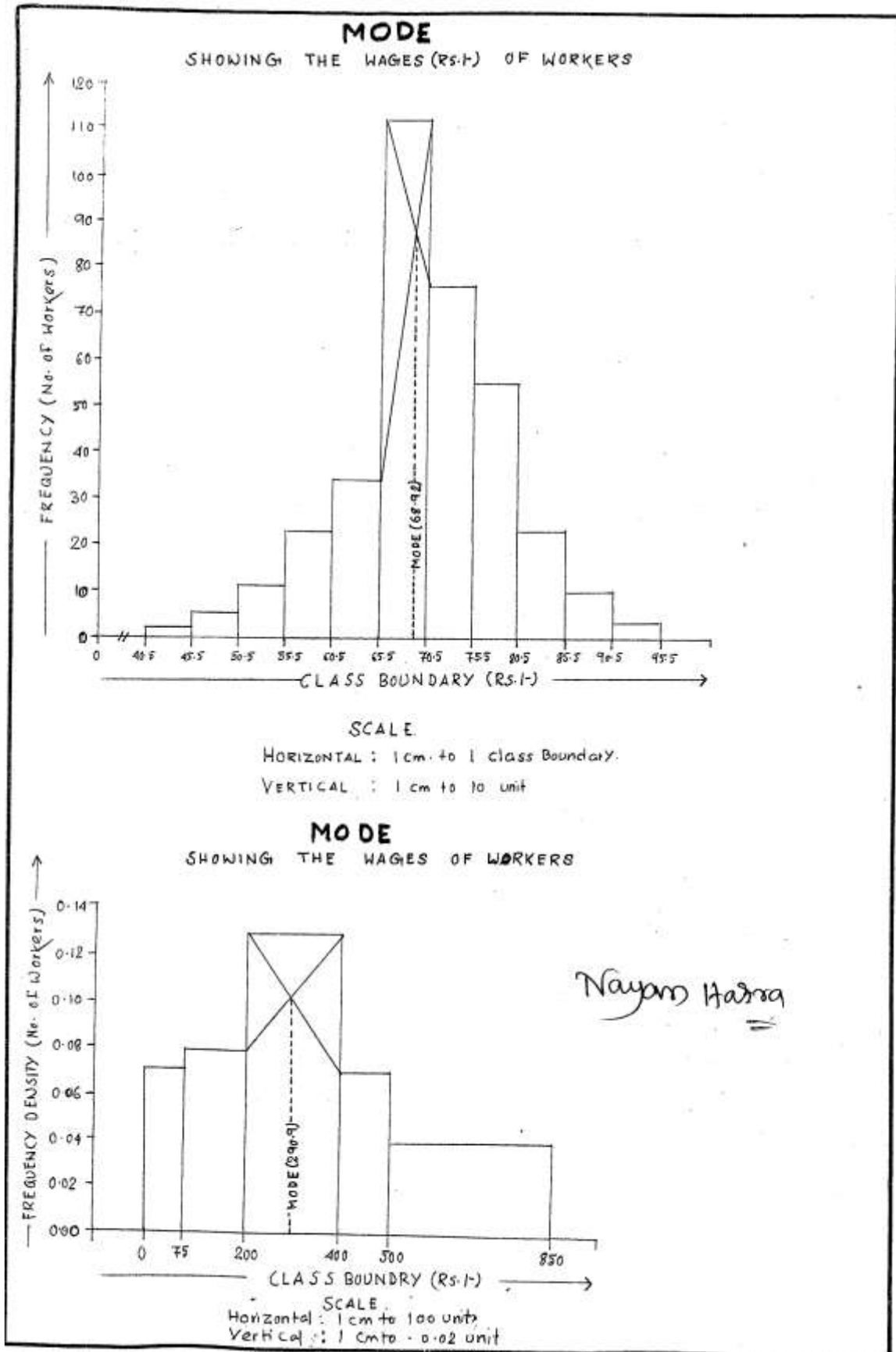
Measures	Symbol	Ungroup Data	Group Data
Mode	\bar{X}	Mode is found by inspection only.	$L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$
			<p>L_1 = Lower class boundary of modal class. Δ_1 = Difference between modal & pre modal class. Δ_2 = Difference between modal & post modal class. i = Width of the modal class.</p> <p>-----</p> <p>*Same formula for equal class & unequal class, but in case of unequal class mode is found by frequency densities except absolute frequency. *In case of group frequency if highest value /density occur in more than one class, then mode is also more than one. *If highest value lies in between two consecutive classes than mode is middle most value of these consecutive class.</p>

**Computation of Modal values
[Ungroup Series & Group Series with Equal Class Boundary]**

Ungroup Data series		Group Data series			
Data set (x_i)	Mode	Class Boundary (Rs/-)	Frequency (f_i) No. of Workers	Class Width (w_i)	Computation
12	Occurrences are 12=2 13=2 14=3 15=4 16=1	40.50 - 45.50	2	5	$\bar{X} = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$ $= 65.50 + \frac{112-34}{(112-34)+(112-76)} \times 5$ $= 68.92$
14		45.50 - 50.50	5	5	
16		50.50 - 55.50	11	5	
15		55.50 - 60.50	23	5	
13		60.50 - 65.50	34	5	
14		65.50 - 70.50	112 (MC)	5	
12	Therefore mode = 15	70.50 - 75.50	76	5	
15	*If all occurrences are same value then no mode whatsoever.	75.50 - 80.50	55	5	
15		80.50 - 85.50	23	5	
14		85.50 - 90.50	10	5	
13		90.50 - 95.50	3	5	

**Computation of Modal values
[Group Series with Unequal Class Boundary]**

Class Boundary (Rs/-)	Frequency (f_i) No. of Workers	Class Width (w_i)	Frequency Density (fd_i) $\frac{f_i}{w_i}$	Computation
0-75	5	75	0.07	$\bar{X} = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$ $= 200 + \frac{0.13-0.08}{(0.13-0.08)+(0.13-0.07)} \times 200$ $= 290.9$
75-200	10	125	0.08	
200-400	26	200	0.13(Mfd)	
400-500	7	100	0.07	
500-850	6	350	0.04	



MEASURES OF DISPERSION
[ABSOLUTE MEASURES & RELATIVE MEASURES]

Measures	Absolute Measures		Relative Measures
	Ungroup Series	Group Series	
Range	(Highest value - Lowest value)	UCB-LCB	Co-efficient of Range $\frac{\text{Highest Value} - \text{Lowest Value}}{\text{Highest Value} + \text{Lowest Value}}$
Quartile Deviation/ Semi Inter-quartile Range (QD)	$\frac{Q_3 - Q_1}{2}$ Q ₃ = Third Quartile Q ₁ = First Quartile Use same formula for equal & unequal class. Use group formula for group calculations & ungroup formula for ungroup calculation for Q ₃ & Q ₁ .		Co-efficient of Quartile Deviation $\frac{Q_3 - Q_1}{Q_3 + Q_1} \dots \dots \dots (1)$ $\frac{\text{Quartile Deviation}}{\text{Median}} \times 100 \dots \dots \dots (2)$
Mean Deviation about mean (MD \bar{x})	$\frac{1}{n} \sum x_i - \bar{x} $ x _i = Variables \bar{x} = Mean n = No. of observations	$\frac{1}{N} \sum f_i x_i - \bar{x} $ x _i = Class Mark \bar{x} = Mean f _i = Frequency N = Total frequency	Co-efficient of Mean Deviation $\frac{\text{Mean Deviation about mean}}{\text{Mean or median}} \times 100$
Standard Deviation (SD) σ	$\sqrt{\frac{1}{n} \sum (x - \bar{x})^2}$ x _i = Variables \bar{x} = Mean n = No. of observations	$\sqrt{\frac{1}{N} \sum f(x - \bar{x})^2}$ x _i = Class Mark \bar{x} = Mean f _i = Frequency N = Total frequency	Co-efficient of Variation (CV) [Variability] $\frac{\text{Standard Deviation } (\sigma)}{\text{Mean } (\bar{x})} \times 100$
	Use same formula for equal & unequal class interval.		
Variance (σ^2)	$\frac{\sum (x - \bar{x})^2}{n}$ x _i = Variables \bar{x} = Mean n = No. of observations	$\frac{\sum f(x - \bar{x})^2}{\sum f_i}$ x _i = Class Mark \bar{x} = Mean f _i = Frequency $\sum f_i$ = Total frequency	
	Use same formula for equal & unequal class interval.		

*** NOTE:**

- Use same formula for equal & unequal class.
- Use group formula for group calculations & ungroup formula for ungroup calculation for Q₃, Q₁, σ , \bar{x} etc.
- When **all frequencies** are taken into account for calculation of mean or standard deviation, is called *population mean* or *population standard deviation*, when calculation is done with the help of selection of **sample** then it is called *sample mean* or *sample standard deviation*.

In any distribution,

- Standard deviation (σ) \geq Mean Deviation.
- Standard deviation (σ) < Variance
- Range = 6 σ
- Quartile Deviation = $\frac{2\sigma}{3}$, Mean Deviation = $\frac{4\sigma}{5}$
- Z- Score = $\frac{x_i - \bar{x}}{\sigma}$ [When group data is given use group formula, when ungroup data is given use ungroup formula]
- Standard Error/Standard Error of Mean = $\frac{\sigma}{\sqrt{n}}$ (Ungroup Series, n = No. of observation) $\frac{\sigma}{\sqrt{N}}$ (Group Series, N = Total frequency)
- In a normal distribution 68.27% of the frequencies lie within a range ($\bar{x} \pm 1\sigma$), 95.45% within ($\bar{x} \pm 2\sigma$) and 99.73 % within ($\bar{x} \pm 3\sigma$).

Range
[Ungroup & Group Series]

Ungroup Data		Group Data			
Data set (x_i)	Computation	Class Boundary Score (%)	Frequency No. of Students (f_i)	Range (UCB-LCB)	
12	Highest value (43) Lowest value (12) Therefore, Range = (43-12) = 31	55-60	7	5	
14		60-65	10	5	
16		65-70	34	5	
15		70-75	28	5	
23		75-80	10	5	
28		80-85	8	5	
43		85-90	3	5	
34				$\Sigma f_i / N = 100$	
29			Same formula for equal & unequal class & unequal class boundary		
25					
$\Sigma x_i = 239$					

Computation of Quartile Deviation or Semi Inter Quartile Range (QD)
[Ungroup Series]

Temperature (°C) (x_i)	Array Temperature (°C) [Ascending Order]		Computation
	Rank	(x_i)	
16.1	1.	16.1	$Q_1 = \frac{1 \times (N+1)}{4} \text{th item } Q_1 = \frac{1 \times (12+1)}{4} = 3.25 \text{th item}$ $= 3^{\text{RD}} \text{ Value} + (4^{\text{TH}} \text{ Value} - 3^{\text{RD}} \text{ Value}) \times 0.25$ $= 17.9^\circ + (18.1^\circ - 17.9^\circ) \times 0.25$ $= 17.95^\circ\text{C}$ $Q_3 = \frac{3 \times (N+1)}{4} \text{th item } Q_3 = \frac{3 \times (12+1)}{4} = 9.75 \text{th item}$ $= 9^{\text{TH}} \text{ Value} + (10^{\text{TH}} \text{ Value} - 9^{\text{TH}} \text{ Value}) \times 0.75$ $= 25.9^\circ + (26.7^\circ - 25.9^\circ) \times 0.75$ $= 26.5^\circ\text{C}$ $QD = \frac{Q_3 - Q_1}{2} = \frac{26.5 - 17.95}{2} = 4.275 \text{ (4.28 Approx)}$
16.5	2.	16.5	
17.9	3.	17.9	
20.4	4.	18.1	
23.4	5.	20.4	
25.9	6.	21.2	
27.9	7.	23.4	
27.4	8.	24.0	
26.7	9.	25.9	
24.0	10.	26.7	
21.2	11.	27.4	
18.1	12.	27.9	

Computation of Quartile Deviation or Semi Inter Quartile Range (QD)

[Group Series]

Class Boundary (Rs/-)	Frequency (f _i) No. of Workers	Class Width (w _i)	Cumulative Frequency		Computation
			Less Than	(f _c <)	
40.50 - 45.50	2	5	40.50	0	$Q_1 = L + \frac{\frac{xN}{4} - F}{f} \times W = 65.50 + \frac{\frac{354}{4} - 75}{112} \times 5 = 66.01$ $Q_3 = L + \frac{\frac{xN}{4} - F}{f} \times W = 75.50 + \frac{\frac{1062}{4} - 263}{55} \times 5 = 75.73$ $QD = \frac{Q_3 - Q_1}{2} = \frac{75.73 - 66.01}{2} = 4.86$
45.50 - 50.50	5	5	45.50	2	
50.50 - 55.50	11	5	50.50	7	
55.50 - 60.50	23	5	55.50	18	
60.50 - 65.50	34	5	60.50	41	
65.50 - 70.50	112	5	65.50	75	
70.50 - 75.50	76	5	70.50	187	
75.50 - 80.50	55	5	75.50	263	
80.50 - 85.50	23	5	80.50	318	
85.50 - 90.50	10	5	85.50	341	
90.50 - 95.50	3	5	90.50	351	
	$\Sigma f_i / N = 354$		95.50	354	

Computation of Mean Deviation about Mean (MD \bar{X})

[Ungroup Series]

Data set (x _i %)	Computation	x _i - \bar{x}	Computation	
12	Arithmetic Mean $AM = \frac{\Sigma x_i}{n}$ $= \frac{239}{10}$ $= 23.90$	11.9	Mean Deviation about Mean = $\frac{1}{n} \Sigma x_i - \bar{x} $ $= \frac{1 \times 79.0}{10}$ $= 7.90$	
14		9.9		
16		7.9		
15		8.9		
23		0.9		
28		4.1		
43		19.1		
34		10.1		
29		5.1		
25		1.1		
$\Sigma x_i = 239$				$\Sigma x_i - \bar{x} = 79.0$

Computation of Mean Deviation about Mean (MD \bar{X})

[Group Series]

Class Boundary Score (%)	Frequency No. of Students (f _i)	Class Mark (x _i)	(f _i x _i)	\bar{x}	x _i - \bar{x}	f _i x _i - \bar{x}	Computation
55-60	7	57.5	402.5	Arithmetic Mean $AM = \frac{\Sigma f_i x_i}{N}$ $= \frac{7050}{100}$ $= 70.50$	13	91	Mean Deviation about Mean $\frac{1}{N} \Sigma f_i x_i - \bar{x} $ $= \frac{1 \times 546}{100}$ $= 5.46$
60-65	10	62.5	625		8	80	
65-70	34	67.5	2295		3	102	
70-75	28	72.5	2030		2	56	
75-80	10	77.5	775		7	70	
80-85	8	82.5	660		12	96	
85-90	3	87.5	262.5		17	51	
	$\Sigma f_i / N = 100$		$\Sigma f_i x_i = 7050$			$\Sigma f_i x_i - \bar{x} = 546$	

**Computation of Standard Deviation (σ)/ Population Standard Deviation
[Ungroup Series]**

Data set (x_i %)	Computation	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	Computation
12	Arithmetic Mean $AM = \frac{\sum x_i}{n}$ $= \frac{239}{10}$ $= 23.90$	-11.9	141.61	$\sigma = \sqrt{\frac{1}{n} \sum (x - \bar{x})^2}$ $\sigma = \sqrt{1 \times \frac{892.90}{10}}$ $= 9.45$ $SE = \frac{\sigma}{\sqrt{n}} = \frac{9.45}{\sqrt{10}} = 2.99$
14		-9.9	98.01	
16		-7.9	62.41	
15		-8.9	79.21	
23		0.9	0.81	
28		4.1	16.81	
43		19.1	364.81	
34		10.1	102.01	
29		5.1	26.01	
25		1.1	1.21	
$\sum x_i = 239$			$\sum (x_i - \bar{x})^2 = 892.90$	

Note:

- When sample is selected (say 20% or 30%) from the distribution and calculation is done with those data set is called sample standard deviation in case of ungroup series.

**Computation of Standard Deviation (σ)/ Population Standard Deviation
[Group Series]**

Class Boundary Score (%)	Frequency No. of Students (f_i)	Class Mark (x_i)	$(f_i x_i)$	\bar{x}	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$	Computation
55-60	7	57.5	402.5	Arithmetic Mean $= \frac{\sum f_i x_i}{N}$ $= \frac{7050}{100}$ $= 70.50$	-13	169	1183	$\sigma = \sqrt{\frac{1}{N} \sum f (x - \bar{x})^2}$ $= \sqrt{1 \times \frac{4750}{100}}$ $= 6.89$ $SE = \frac{\sigma}{\sqrt{N}} = \frac{6.89}{\sqrt{100}} = 0.69$
60-65	10	62.5	625		-8	64	640	
65-70	34	67.5	2295		-3	9	306	
70-75	28	72.5	2030		2	4	112	
75-80	10	77.5	775		7	49	490	
80-85	8	82.5	660		12	144	1152	
85-90	3	87.5	262.5		17	289	867	
	$\sum f_i / N = 100$		$\sum f_i x_i = 7050$				$\sum f_i (x_i - \bar{x})^2 = 4750$	

Note:

- When all the data is taken into account for calculation of standard deviation of a distribution is called population standard deviation in case of group series.

**Computation of Standard Deviation (σ)/ Sample Standard Deviation
[Group Series, 20% sample size taken into account]**

Class Boundary Score (%)	Frequency No. of Students (f_i)	Class Mark (x_i)	$(f_i x_i)$	\bar{x}	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$	Computation
55-60	3	57.5	172.5	Arithmetic Mean $= \frac{\sum f_i x_i}{N}$ $= \frac{1420}{20}$ $= 71.0$	-13.5	182.25	546.75	$\sigma = \sqrt{\frac{1}{N} \sum f (x - \bar{x})^2}$ $= \sqrt{1 \times \frac{1505.0}{20}}$ $= 75.25$
60-65	2	62.5	125.0		-8.5	72.25	144.5	
65-70	4	67.5	270.0		-3.5	12.25	49.0	
70-75	5	72.5	362.5		1.5	2.25	11.25	
75-80	2	77.5	155.0		6.5	42.25	84.5	
80-85	3	82.5	247.5		11.5	132.25	396.75	
85-90	1	87.5	87.5		16.5	272.25	272.25	
	$\sum f_i / N = 20$		$\sum f_i x_i = 1420$				$\sum f_i (x_i - \bar{x})^2 = 1505.0$	

Note:

- When sample is selected (say 20% or 30%) from the distribution and calculation is done with those data set is called sample standard deviation in case of group series.

Computation of Variance (σ^2)
[Ungroup Series]

Data set (x_i %)	Computation	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	Computation
12	Arithmetic Mean $AM = \frac{\sum x_i}{n}$ $= \frac{239}{10}$ $= 23.90$	-11.9	141.61	$(\sigma^2) = \frac{\sum(x - \bar{x})^2}{n}$ $= \frac{892.90}{10}$ $= 89.29$
14		-9.9	98.01	
16		-7.9	62.41	
15		-8.9	79.21	
23		0.9	0.81	
28		4.1	16.81	
43		19.1	364.81	
34		10.1	102.01	
29		5.1	26.01	
25		1.1	1.21	
$\sum x_i = 239$			$\sum(x_i - \bar{x})^2 = 892.90$	

Computation of Variance (σ^2)
[Group Series]

Class Boundary Score (%)	Frequency No. of Students (f_i)	Class Mark (x_i)	$(f_i x_i)$	\bar{x}	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^2$	Computation
55-60	7	57.5	402.5	Arithmetic Mean $= \frac{\sum f_i x_i}{N}$ $= \frac{7050}{100}$ $= 70.50$	-13	169	1183	$(\sigma^2) = \frac{\sum f(x - \bar{x})^2}{\sum f_i}$ $= \frac{4750}{100}$ $= 47.50$
60-65	10	62.5	625		-8	64	640	
65-70	34	67.5	2295		-3	9	306	
70-75	28	72.5	2030		2	4	112	
75-80	10	77.5	775		7	49	490	
80-85	8	82.5	660		12	144	1152	
85-90	3	87.5	262.5		17	289	867	
	$\sum f_i / N = 100$		$\sum f_i x_i = 7050$				$\sum f_i(x_i - \bar{x})^2 = 4750$	

You are given below the annual rainfall of two stations of 10 years. Compare the variability of rainfall of two stations with the help of CV.

Computation of Co-efficient of Variance (CV)
[Ungroup Series]

Annual Rainfall at Station-1 (cm) (x_i)	Mean & Spread	Annual Rainfall at Station-2 (cm) (x_i)	Mean & Spread	Remarks
100	$\bar{x} = 102$ $\sigma = 56.13$ $CV = \frac{(\sigma)}{(\bar{x})} \times 100$ $= \frac{56.13}{102} \times 100$ $= 55.03\%$	95	$\bar{x} = 102$ $\sigma = 19.89$ $CV = \frac{(\sigma)}{(\bar{x})} \times 100$ $= \frac{19.89}{102} \times 100$ $= 19.50\%$	<ul style="list-style-type: none"> 10 year total rainfall same at both stations, i.e; = 1020 cm 10 year mean rainfall same at both station = 102 cm Standard deviation of rainfall at station-1 is larger than at station-2. Variability of rainfall at station-1 is greater than at station-2. Therefore, rainfall is more consistent, less variable and more reliable at station-2
150		100		
120		105		
60		120		
0		100		
90		110		
70		115		
200		50		
150		115		
80		110		
$\sum x_i = 1020$		$\sum x_i = 1020$		

Note:

- CV is more useful in ungroup data for comparison.
- But it is calculated in group data also. Basic formula is same. When ungroup data is given \bar{X} , σ is calculated with the help of ungroup formula but When group data is given \bar{X} , σ is calculated with the help of group formula.

5. Measures of association: Pearson's correlation and Spearman's rank correlation

Measures of association help us understand the strength and direction of a relationship between two variables. Two commonly used correlation methods are:

1. Pearson's Correlation Coefficient (r)

Definition: Pearson's **r** measures the **linear relationship** between **two continuous (interval/ratio)** variables.

Formula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Key Features:

Aspect	Details
Type of data	Continuous (numeric)
Relationship	Linear
Range	-1 to +1
r = +1	Perfect positive correlation
r = -1	Perfect negative correlation
r = 0	No linear correlation

Example:

- **Variables:** Study time (hours) vs. test score
- **Result:** **r = 0.85** → Strong positive correlation

2. Spearman's Rank Correlation Coefficient (ρ)

Definition: Spearman's **ρ** measures the **monotonic relationship** (not necessarily linear) between **two ranked (ordinal or continuous)** variables.

Formula (when there are no tied ranks):

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

Where:

- d_i = difference in ranks of each pair
- n = number of observations

Key Features:

Aspect	Details
Type of data	Ordinal or non-normally distributed continuous
Relationship	Monotonic (increasing or decreasing)
Range	-1 to +1
Use case	When data has ranks, ties , or is non-linear

Example:

- **Variables:** Ranks in Math and Physics
- **Result:** $\rho = 0.90$ → Strong agreement in rankings

Comparison Table: Pearson vs. Spearman

Feature	Pearson's r	Spearman's ρ
Data type	Continuous	Ordinal or Continuous
Measures	Linear relationship	Monotonic relationship
Sensitive to outliers	Yes	Less sensitive
Suitable when	Data is normally distributed	Data is skewed or not normal
Example	Hours studied vs. exam score	Rank in English vs. rank in History

Application in Geography/Social Science:

- **Pearson's r:** Correlation between **population density and air pollution**.
- **Spearman's ρ :** Rank correlation between **development index and literacy ranking** of districts.

REGRESSION & CORRELATION

MEASURES	FORMULA
Scatter Diagram & Regression line [Least square method]	$\hat{Y} = a + bX$ $\hat{Y} = \text{Predicted dependent variable}$ $a = \text{Intercept}$ $b = \text{On slope}$ $\Sigma Y = na + b \Sigma X \dots\dots\dots(1)$ $\Sigma XY = a \Sigma X + b \Sigma X^2 \dots\dots\dots(2)$
Scatter Diagram & Regression line [Regression co-efficient method]	$\hat{Y} = a + bX$ $\hat{Y} = \text{Predicted dependent variable}$ $a = \text{Intercept}$ $b = \text{On slope}$ $b = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}} \dots\dots\dots(1)$ $a = \bar{Y} - b\bar{X} \dots\dots\dots(2)$
Scatter Diagram & Regression line [Correlation co-efficient method]	$\hat{Y} = a + bX$ $\hat{Y} = \text{Predicted dependent variable}$ $a = \text{Intercept}$ $b = \text{On slope}$ $b = r \cdot \frac{\sigma_y}{\sigma_x} \dots\dots\dots(1)$ $a = \bar{Y} - b\bar{X} \dots\dots\dots(2)$ $r = \frac{\frac{1}{n} \Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\frac{1}{n} \Sigma (X - \bar{X})^2} \cdot \sqrt{\frac{1}{n} \Sigma (Y - \bar{Y})^2}} \dots\dots\dots(3)$
Product moment correlation co-efficient [Normal Method]	$r_{xy} = \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[n \Sigma x^2 - (\Sigma x)^2][n \Sigma y^2 - (\Sigma y)^2]}}$
Product moment correlation co-efficient [Co-variance Method]	$r_{xy} = \frac{\text{Covariance } xy}{\sigma_x \sigma_y} \quad (\text{or}) \quad r_{xy} = \frac{\frac{1}{n} \Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\frac{1}{n} \Sigma (X - \bar{X})^2} \cdot \sqrt{\frac{1}{n} \Sigma (Y - \bar{Y})^2}}$ $\text{Cov}_{xy} = \frac{1}{n} \Sigma (X - \bar{X})(Y - \bar{Y})$ $\sigma_x = \sqrt{\frac{1}{n} \Sigma (x - \bar{x})^2}$ $\sigma_y = \sqrt{\frac{1}{n} \Sigma (y - \bar{y})^2}$
Probable error of correlation co-efficient	$(PE_r) = 0.6745 \times \frac{1 - r^2}{\sqrt{n}}$ $r = \text{correlation coefficient}$ $n = \text{No. of observation}$ $\left[\frac{r}{PE_r} < 6 = \text{Insignificant} \quad \& \quad \frac{r}{PE_r} \geq 6 = \text{Significant} \right]$
Test of significance of Correlation Co-efficient	$\text{Student's 't'} = r \sqrt{\frac{n-2}{1-r^2}}$ $n = \text{No. of observation}$ $r = \text{Correlation Co-efficient}$

The following data shows rainfall and rainy days of a year:

Months	J	F	M	A	M	J	Ju	A	S	O	N	D
Rainy days	0.8	1.3	2.0	3.0	6.0	12.6	16.4	17.0	18.8	6.9	1.2	0.4
Rainfall (cm)	1.4	1.6	2.6	5.1	10.3	28.0	32.7	31.4	29.4	13.4	1.7	0.7

1. Draw a scatter diagram with regression line on the basis of above data.
2. What will be the rainfall (cm) if the rainy days are 5 & 15?
3. Interpret the nature of variables with the help of product moment correlation co-efficient method.
4. Calculate the probable error of the correlation coefficient to justify the nature of correlation.

5. Find out the standard error of estimate for the model.
6. Justify the relationship exists between the rainy days and rainfall with the application of test of significance.
7. Compute the percentage of variance in rainfall explained by rainy days.
8. Calculate the co-efficient of determination. (R^2)

Correlation coefficient method

Months	Rainy Days (X)	Rainfall (cm) (Y)	\bar{X}	(X- \bar{X})	(X- \bar{X}) ²	\bar{Y}	(Y- \bar{Y})	(Y- \bar{Y}) ²	(X- \bar{X}).(Y- \bar{Y})	COMPUTATION	
J	0.8	1.4	Mean = $\frac{\Sigma X}{n}$ = $\frac{86.4}{12}$ = 7.20	-6.4	40.96	Mean = $\frac{\Sigma Y}{n}$ = $\frac{158.3}{12}$ = 13.19	-11.79	139.00	75.46	$\sigma_x = \sqrt{\frac{1}{n} \Sigma (x - \bar{x})^2}$	
F	1.3	1.6		-5.9	34.81		-11.59	134.33	68.38		$\sigma_x = \sqrt{1 \times \frac{548.62}{12}}$ = 6.76153 r.days
M	2.0	2.6		-5.2	27.04		-10.59	112.14	55.07		
A	3.0	5.1		-4.2	17.64		-8.09	65.45	33.98		
M	6.0	10.3		-1.2	1.44		-2.89	8.35	3.47	$\sigma_y = \sqrt{\frac{1}{n} \Sigma (Y - \bar{Y})^2}$	
J	12.6	28.0		5.4	29.16		14.81	219.34	79.97		$\sigma_y = \sqrt{1 \times \frac{1941.68}{12}}$ = 12.72032 cm
J	16.4	32.7		9.2	84.64		19.51	380.64	179.49		
A	17.0	31.4		9.8	96.04		18.21	331.60	178.46		
S	18.8	29.4		11.6	134.56		16.21	262.76	188.04	REGRESSION EQUATION $\hat{Y} = a + bX$ Solving equation is $\hat{Y} = -0.14388 + 1.85216X$	
O	6.9	13.4		-0.3	0.09		0.21	0.04	0.06		
N	1.2	1.7		-6.0	36.00		-11.49	132.02	68.94		
D	0.4	0.7		-6.8	46.24		-12.49	156.00	84.93		
n = 12	ΣX 86.4	ΣY 158.3			$\Sigma (X-\bar{X})^2$ 548.62			$\Sigma (Y-\bar{Y})^2$ 1941.68	$\Sigma (X-\bar{X}).(Y-\bar{Y})$ 1016.25		

Computation of Regression Equation

In the method of least square the regression equation is $\hat{Y} = a + bX$

To solve a & b, the normal equation is

$$b = r \cdot \frac{\sigma_y}{\sigma_x} \dots\dots\dots(1)$$

$$a = \bar{Y} - b\bar{X} \dots\dots\dots(2)$$

$$r = \frac{\frac{1}{n} \Sigma (X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{1}{n} \Sigma (X-\bar{X})^2} \cdot \sqrt{\frac{1}{n} \Sigma (Y-\bar{Y})^2}} \dots\dots\dots(3)$$

\hat{Y} = Predicted dependent variable

a = Intercept

b = On slope

X = Independent Variable

r = Correlation co-efficient

σ_y = Standard deviation of y variables

σ_x = Standard deviation of x variables

\bar{Y} = Mean of y variables

\bar{X} = Mean of x variables

Putting the value n = 12, $\Sigma (X-\bar{X}).(Y-\bar{Y}) = 1016.25$, $\Sigma (X-\bar{X})^2 = 548.62$, $\Sigma (Y-\bar{Y})^2 = 1941.68$ in equation... (3), then we get,

$$r = \frac{\frac{1}{n} \Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{1}{n} \Sigma(X-\bar{X})^2} \cdot \sqrt{\frac{1}{n} \Sigma(Y-\bar{Y})^2}}$$

$$r = \frac{\frac{1}{12} \times 1016.25}{\frac{1}{12} \times \sqrt{548.62} \times \sqrt{1941.68}}$$

$$r = \frac{1016.25}{23.42264 \times 44.06450}$$

$$r = \frac{1016.25}{1032.10692}$$

$$r = 0.98452$$

Therefore, $b = r \cdot \frac{\sigma_y}{\sigma_x}$

$$b = 0.948452 \times \frac{12.72032}{6.76153}$$

$$b = 0.948452 \times 1.88128$$

$$b = 1.85216$$

$$a = \bar{Y} - b\bar{X}$$

$$a = 13.19167 - 1.85216 \times 7.2$$

$$a = -0.14388$$

So the regression equation is $\hat{Y} = a + bX$ [$\hat{Y} = -0.14388 + 1.85216X$]

For drawing of regression line, When $X = 4$ then, $\hat{Y} = a + bX$
 $\hat{Y} = -0.14388 + 1.85216 \times 4$
 $\hat{Y} = 7.26476 \text{ cm}$

When $X = 16$ then, $\hat{Y} = a + bX$
 $\hat{Y} = -0.14388 + 1.85216 \times 16$
 $\hat{Y} = 29.49068 \text{ cm}$

➤ What will be the rainfall (cm) if the rainy days are 5 & 15?

This problem is solved by the least square method.

Computation of Regression Equation

Months	Rainy Days (X)	Rainfall (cm) (Y)	X ²	XY	Solution
J	0.8	1.4	0.64	1.12	$\hat{Y} = a + bX$ Solving These Equations: $a = -0.14373$ & $b = 1.85214$ Therefore the Regression Equation is : $[\hat{Y} = -0.14373 + 1.85214X]$
F	1.3	1.6	1.69	2.08	
M	2.0	2.6	4.00	5.20	
A	3.0	5.1	9.00	15.30	
M	6.0	10.3	36.00	61.80	
J	12.6	28.0	158.76	352.80	
J	16.4	32.7	268.96	536.28	
A	17.0	31.4	289.00	533.80	
S	18.8	29.4	353.44	552.72	
O	6.9	13.4	47.61	92.46	
N	1.2	1.7	1.44	2.04	
D	0.4	0.7	0.16	0.28	
n= 12	$\Sigma X = 86.4$	$\Sigma Y = 158.3$	$\Sigma X^2 = 1170.70$	$\Sigma XY = 2155.88$	

So the regression equation is $\hat{Y} = a + bX$ [$\hat{Y} = -0.14373 + 1.85214X$]

If the rainy days are 5 then expected rainfall (cm) is (When X = 5) then,

$$\begin{aligned} \hat{Y} &= a + bX \\ \hat{Y} &= -0.14373 + 1.85214 \times 5 \\ \hat{Y} &= 9.11697 \text{ cm} \end{aligned}$$

If the rainy days are 15 then expected rainfall (cm) is (When X = 15) then,

$$\begin{aligned} \hat{Y} &= a + bX \\ \hat{Y} &= -0.14373 + 1.85214 \times 15 \\ \hat{Y} &= 27.63837 \text{ cm} \end{aligned}$$

MIDNAPORE CITY COLLEGE
Department of Pure and Applied Sciences
Laboratory Manual for Bachelor of Science (Honours)
Major in Geography
(CCFUP), 2023 & NEP, 2020
Semester – III
Course Type: SEC - 3
Course Code: GEOSEC03
Course Title: Computer Programming (R/Python) (Practical)

PREFACE TO THE FIRST EDITION

This is the first edition of Lab Manual for BSc Honours Major in Geography (Third Semester). Hope this edition will help you during practical. This edition mainly tried to cover the whole syllabus. Some hard topics are not present here that will be guided by responsive teachers at the time of practical.

ACKNOWLEDGEMENT

We are really thankful to our students, teachers, and non-teaching staffs to make this effort little bit complete. Mainly thanks to Director and Principal Sir to motivate for making this lab manual.

SEC 3: Computer Programming (R/Python)

Credits 03

SEC3P: Computer Programming (R/Python) - Practical Full Marks: 50

Course Objective

This course provides an introduction to computer programming using the R or Python languages (whichever feasible based on the infrastructure). The course is designed to escalate the skill sets of the students to bridge the gap towards their future academic endeavors in Geospatial Science or any related application areas in the Earth, Atmosphere and Planetary Sciences as well as in the areas of Human and Population Geography and Developmental Studies. The objectives of the course are -

- 1. To enhance the analytical skills using the first programming language.*
- 2. To take the leverage of the analytical platforms to analyze and comprehend the geographic data.*
- 3. To prepare the students for doing better research*

Course Learning Outcomes

Upon completion of this course, students will be able to -

1. Learn the basics of computer programming using R/Python
2. Learn the data structures that these two platforms can deal with, writing functions, data filtering and summary methods
3. Learn how to build graphics and regression models using R and Python.

Course Outline:

1. Basics of computer programming: understanding the interface of programming languages such as R and Python
2. Data structure in R/Python: array, vector, matrix, data frame, importing data in R/Python.
3. Functions in R/Python: the basic syntax of statistical functions, writing and using functions in R/Python
4. Data filtering and summary methods in R/Python: conditional functions, loop functions such as for and while loop for iterative calculation. Executing statistical tests in R/Python.
5. Building graphics in R/Python: creating histogram, scatterplot, boxplot, line plot; building a regression model, visualisation data using R/Python.

Basics of computer programming

▪ **Computer Programming**

Computer programming is the process of writing code (sets of instructions) to perform tasks on a computer. It uses **programming languages** to communicate with the computer.

▪ **Basic Programming Concepts**

Concept	Description
Syntax	Rules that define the structure of code in a programming language.
Variables	Containers to store data values (e.g., numbers, text).
Data Types	Types of data, such as integers (<code>int</code>), floating point numbers (<code>float</code>), strings (<code>str</code>), and booleans (<code>bool</code>).
Operators	Symbols used to perform operations (+, -, *, /, etc.).
Control Structures	Code blocks that manage decision-making and repetition (e.g., <code>if</code> , <code>else</code> , <code>for</code> , <code>while</code>).
Functions	Reusable blocks of code that perform a specific task.
Input/Output	Getting data from users (input) and displaying results (output).
Loops	Repeat actions using <code>for</code> or <code>while</code> loops.
Comments	Notes in the code for explanation, ignored by the computer.

▪ Popular Programming Languages

- **Python** – Easy to read, widely used in data science and automation.
- **C/C++** – Offers control over system resources, used in systems programming.
- **Java** – Object-oriented and platform-independent.
- **JavaScript** – Widely used for web development.

▪ Example in Python

```
python

# This is a comment
name = input("Enter your name: ") # Taking user input
print("Hello, " + name + "!")     # Displaying output
```

Understanding the interfaces of the Python programming language

Understanding the interfaces of the Python programming language involves exploring the different ways you can interact with and use Python to write, run, and manage code. Here's a structured overview of the main interfaces:

➤ Command-Line Interface (CLI)

- Python can be run from the terminal or command prompt.
- You can type `python` or `python3` to enter an interactive mode (REPL).
- Scripts can be run directly with `python script.py`.

➤ Interactive Shells

- **Python REPL**: Basic interface in the terminal.
- **IPython**: Enhanced interactive shell with features like auto-completion and magic commands.
- **Jupyter Notebook**: Web-based interface used heavily in data science and education.

➤ Integrated Development Environments (IDEs)

- **IDLE**: Comes bundled with Python; simple and beginner-friendly.
- **PyCharm**: Full-featured professional IDE.
- **VS Code**: Lightweight, extensible editor with Python support.
- Other IDEs: Spyder, Thonny, Atom, Sublime Text.

➤ GUI-based Interfaces

- You can build Graphical User Interfaces (GUIs) using libraries like:
 - Tkinter (standard)
 - PyQt / PySide
 - Kivy

➤ Web Interfaces

- Python runs on the backend of web applications using frameworks like:
 - Django
 - Flask
 - FastAPI

➤ APIs and External Interfaces

- Python can interface with:
 - Databases (via SQLAlchemy, sqlite3, psycopg2, etc.)
 - External applications through APIs (e.g., REST APIs with requests)
 - Other languages using bindings (e.g., Python-C API, pybind11)

Notebook and Cloud Interfaces

- Google Colab: Cloud-based notebooks for Python with free GPU access.
- Kaggle Notebooks: For data science competitions and datasets.





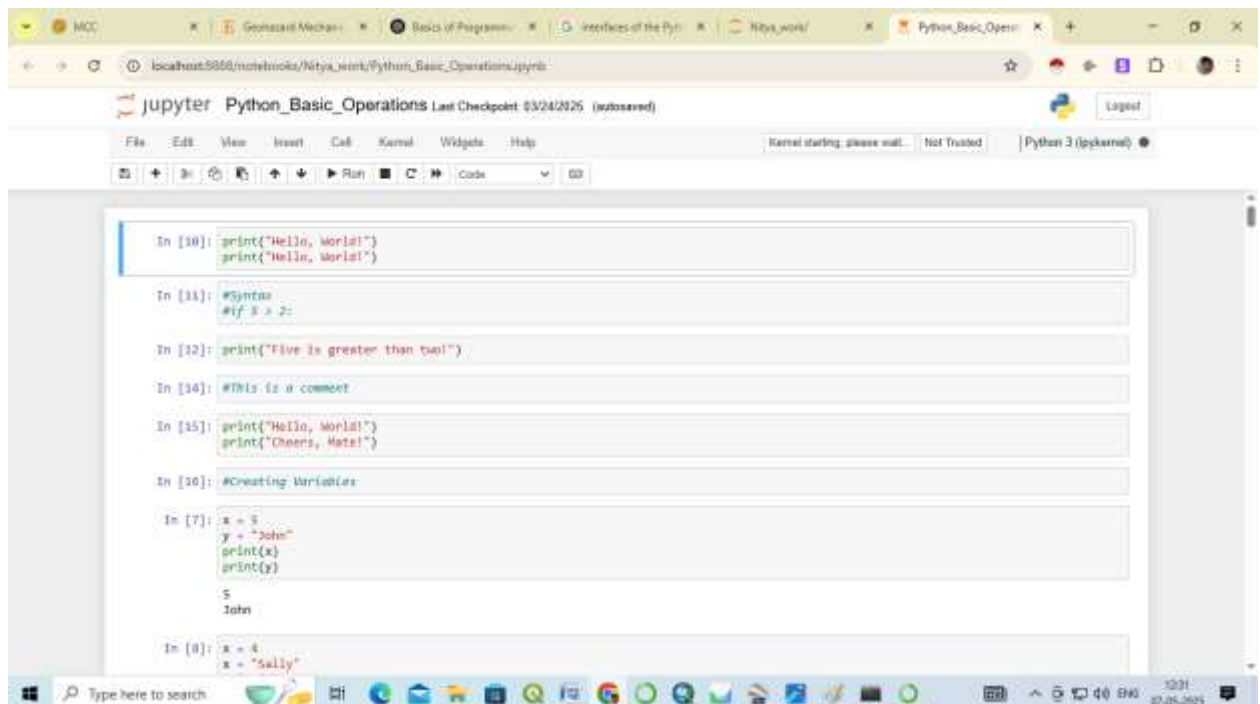
Quit Logout

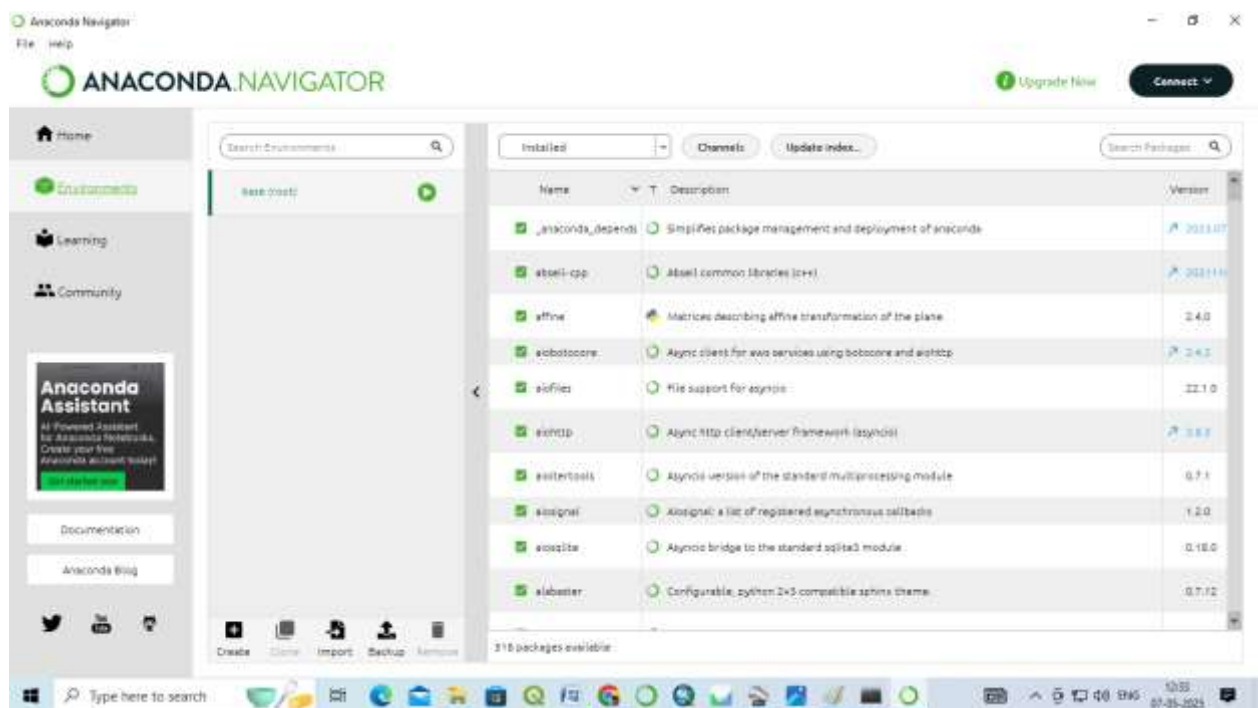
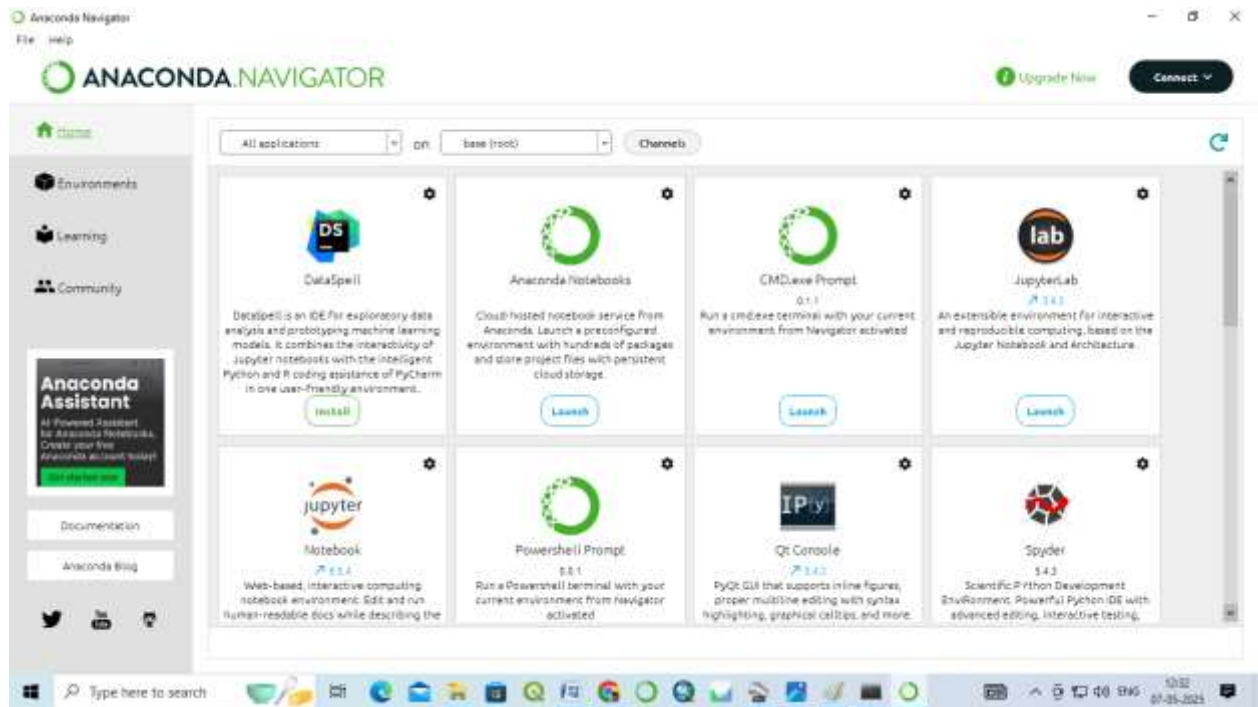
Files Running Clusters

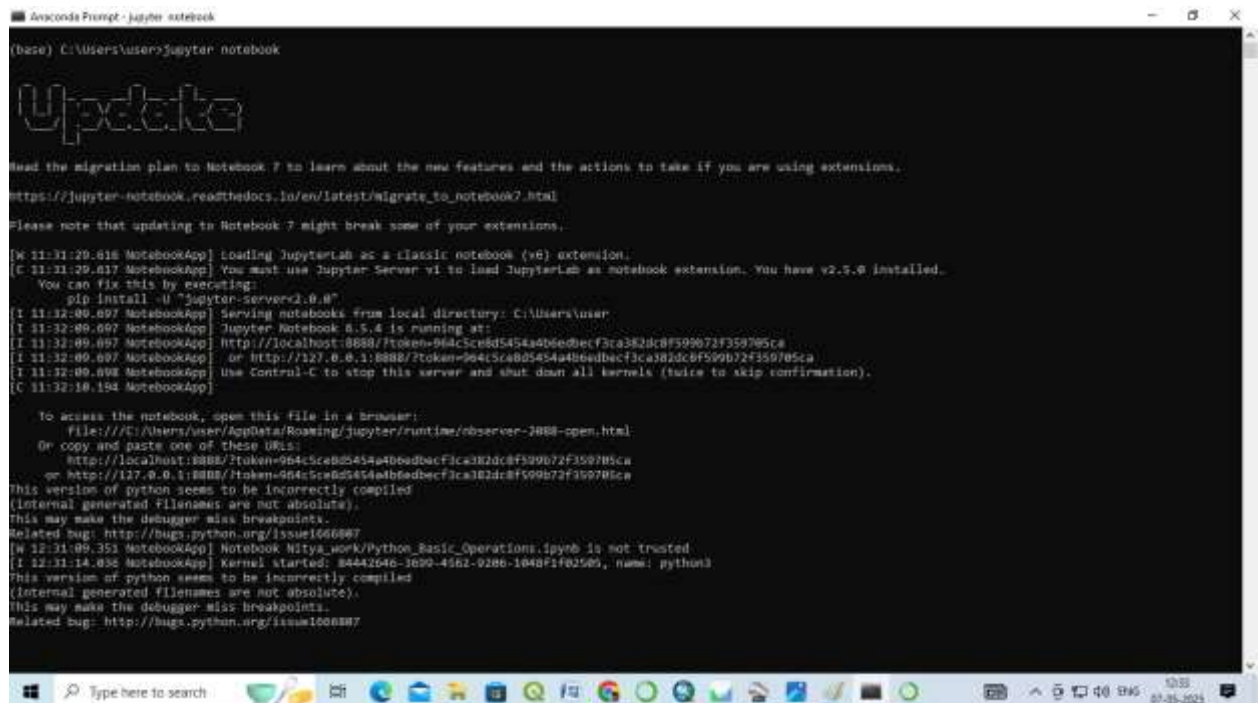
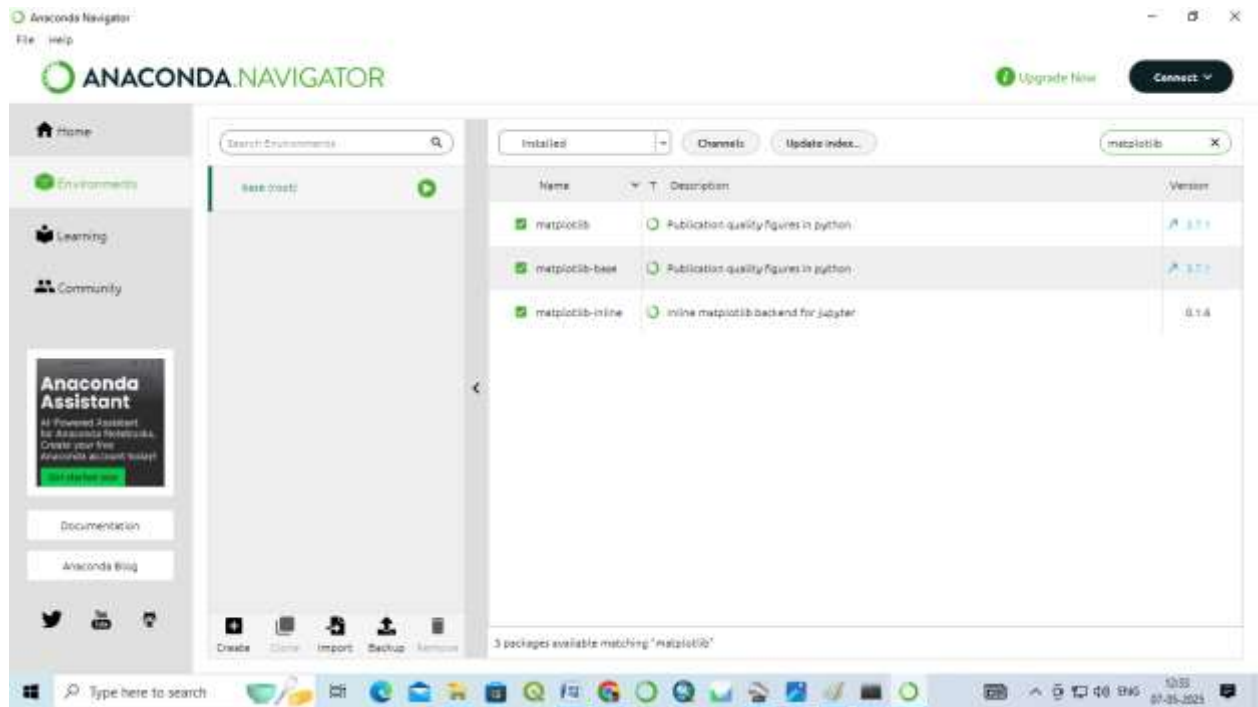
Select items to perform actions on them

Upload New ↕

<input type="checkbox"/>	Name	Last Modified	File size
<input type="checkbox"/>	3D Objects	3 years ago	
<input type="checkbox"/>	anaconda3	2 years ago	
<input type="checkbox"/>	Contacts	3 years ago	
<input type="checkbox"/>	Desktop	7 minutes ago	
<input type="checkbox"/>	Documents	9 days ago	
<input type="checkbox"/>	Downloads	22 minutes ago	
<input type="checkbox"/>	Favorites	3 years ago	
<input type="checkbox"/>	Links	3 years ago	
<input type="checkbox"/>	Machine Learning	a month ago	







Data structures in python

Python provides several powerful and versatile built-in data structures, each designed for specific types of data storage and manipulation. Here's an overview of the core data structures in Python:

◆ 1. List

- **Ordered, mutable, allows duplicates.**
- Used to store a sequence of items.

```
python Copy Edit  
  
fruits = ['apple', 'banana', 'cherry']  
fruits.append('mango')
```

◆ 2. Tuple

- **Ordered, immutable, allows duplicates.**
- Often used for fixed collections of items.

```
python Copy Edit  
  
coordinates = (10.5, 20.7)
```

◆ 3. Set

- **Unordered, mutable, no duplicates.**
- Useful for membership testing and removing duplicates.

```
python Copy Edit  
  
unique_numbers = {1, 2, 3, 2}
```

◆ 4. Dictionary (dict)

- **Unordered** (as of Python 3.6+, insertion order is preserved), key-value pairs.
- Keys must be **unique** and **immutable**.

```
python Copy Edit  
  
student = {'name': 'Alice', 'age': 21}  
student['grade'] = 'A'
```

◆ 5. String

- Technically not a container type, but it behaves like a sequence.
- **Immutable** sequence of Unicode characters.

```
python
```

[Copy](#)[Edit](#)

```
message = "Hello, World!"
```

◆ 6. Arrays (via `array` module)

- Like lists but for **homogeneous numeric data**.
- More memory-efficient than lists for large numeric datasets.

```
python
```

[Copy](#)[Edit](#)

```
import array  
arr = array.array('i', [1, 2, 3])
```

a. Vector (1D array-like)

- Typically represented using a list, NumPy array, or pandas Series.

Using list:

```
python
```

[Copy](#)[Edit](#)

```
vector = [1, 2, 3, 4]
```

Using NumPy:

```
python
```

[Copy](#)[Edit](#)

```
import numpy as np  
vector_np = np.array([1, 2, 3, 4])
```

Using pandas:

```
python Copy Edit  
  
import pandas as pd  
vector_series = pd.Series([1, 2, 3, 4])
```

b. Matrix (2D array)

- Represented using nested lists or NumPy arrays.

Using NumPy:

```
python Copy Edit  
  
matrix = np.array([[1, 2], [3, 4]])
```

c. DataFrame

- A table-like structure with rows and columns, ideal for labeled data. Provided by pandas.

Example:

```
python Copy Edit  
  
data = {'Name': ['Alice', 'Bob'], 'Age': [25, 30]}  
df = pd.DataFrame(data)
```

2. Importing Data into Python

a. From CSV:

```
python Copy Edit  
  
df = pd.read_csv('filename.csv')
```

b. From Excel:

```
python Copy Edit  
  
df = pd.read_excel('filename.xlsx')
```

c. From a JSON file:

```
python Copy Edit  
  
df = pd.read_json('filename.json')
```

c. From a JSON file:

```
python Copy Edit  
  
df = pd.read_json('filename.json')
```

d. From a SQL database:

```
python Copy Edit  
  
import sqlite3  
conn = sqlite3.connect('mydatabase.db')  
df = pd.read_sql_query("SELECT * FROM my_table", conn)
```

Functions in python

Basic Syntax of a Function

```
python Copy Edit  
  
def function_name(parameters):  
    # function body  
    return result
```

Example

```
python Copy Edit  
  
def greet(name):  
    return f"Hello, {name}!"  
  
print(greet("Alice")) # Output: Hello, Alice!
```

Key Concepts

- `def` : Keyword used to define a function.
- `parameters` : Variables passed to the function.
- `return` : Keyword used to send back a result.

Types of Functions

1. **Built-in functions:** Like `print()`, `len()`, `type()`.
2. **User-defined functions:** Created using the `def` keyword.
3. **Lambda functions:** Anonymous, single-expression functions.

Example:

```
python

square = lambda x: x * x
print(square(5)) # Output: 25
```



1. Writing a Function

To define a function, use the `def` keyword:

```
python

def greet(name):
    print(f"Hello, {name}!")
```

This function takes one parameter (name) and prints a greeting.

2. Calling a Function

To use (or "call") the function:

```
python

greet("Alice")
```

Output:

```
Hello, Alice!
```

3. Returning a Value

Functions can return results using the return keyword:

```
python Copy Edit  
  
def add(a, b):  
    return a + b
```

Call it like this:

```
python Copy Edit  
  
result = add(3, 4)  
print(result) # Output: 7
```

4. Default Parameters

You can set default values for parameters:

```
python Copy Edit  
  
def greet(name="Guest"):  
    print(f"Hello, {name}!")
```

5. Keyword Arguments

Call functions using keyword arguments for clarity:

```
python Copy Edit  
  
greet(name="Sam")
```

Data Filtering and Summery Methods in Python

1. Data Filtering Methods

Assume you have a DataFrame named df:

```
python Copy Edit  
  
import pandas as pd  
  
# Sample DataFrame  
data = {  
    'Name': ['Alice', 'Bob', 'Charlie', 'David', 'Eve'],  
    'Age': [25, 30, 35, 40, 29],  
    'City': ['Kolkata', 'Delhi', 'Mumbai', 'Chennai', 'Kolkata'],  
    'Score': [85, 90, 78, 92, 88]  
}  
df = pd.DataFrame(data)
```

Filter rows by condition:

```
python Copy Edit  
  
df[df['Age'] > 30]
```

Filter by multiple conditions:

```
python Copy Edit  
  
df[(df['Age'] > 30) & (df['City'] == 'Kolkata')]
```

Filter by specific values (isin):

```
python Copy Edit  
  
df[df['City'].isin(['Kolkata', 'Delhi'])]
```

Filter using query:

```
python Copy Edit  
  
df.query('Age > 30 and City == "Kolkata"')
```

2. Data Summary Methods

General summary:

```
python
```

[Copy](#)[Edit](#)

```
df.describe()
```

Summary of specific columns:

```
python
```

[Copy](#)[Edit](#)

```
df['Score'].mean()      # Mean
df['Score'].median()   # Median
df['Score'].std()      # Standard deviation
df['Score'].min(), df['Score'].max() # Min & Max
```

Count of unique values:

```
python
```

[Copy](#)[Edit](#)

```
df['City'].value_counts()
```

Group-wise summary:

```
python
```

[Copy](#)[Edit](#)

```
df.groupby('City')['Score'].mean()
```

Correlation between numeric variables:

```
python
```

[Copy](#)[Edit](#)

```
df.corr(numeric_only=True)
```

Executing Statistical Tests in Python

1. One-sample t-test

```
python Copy Edit  
  
from scipy import stats  
  
data = [2.3, 2.5, 3.1, 2.8, 2.9]  
t_stat, p_value = stats.ttest_1samp(data, popmean=3.0)  
print(f"T-statistic: {t_stat}, P-value: {p_value}")
```

2. Two-sample (independent) t-test

```
python Copy Edit  
  
group1 = [2.3, 2.9, 3.1]  
group2 = [3.0, 3.4, 3.5]  
t_stat, p_value = stats.ttest_ind(group1, group2)  
print(f"T-statistic: {t_stat}, P-value: {p_value}")
```

3. Paired t-test

```
python Copy Edit  
  
before = [2.3, 2.9, 3.1]  
after = [2.7, 3.0, 3.2]  
t_stat, p_value = stats.ttest_rel(before, after)  
print(f"T-statistic: {t_stat}, P-value: {p_value}")
```

4. Chi-square test (for independence)

```
python Copy Edit  
  
import numpy as np  
obs = np.array([[10, 20], [20, 40]])  
chi2, p, dof, expected = stats.chi2_contingency(obs)  
print(f"Chi2: {chi2}, P-value: {p}")
```

5. ANOVA (one-way)

python

Copy Edit

```
group1 = [2.1, 2.5, 2.9]
group2 = [3.1, 3.4, 3.3]
group3 = [4.1, 4.2, 4.0]
f_stat, p_value = stats.f_oneway(group1, group2, group3)
print(f"F-statistic: {f_stat}, P-value: {p_value}")
```

Building Graphics in Python

5/7/25, 3:55 PM

Python_Basic_Operations - Jupyter Notebook

```
In [10]: print("Hello, World!")
print("Hello, World!")
```

```
In [11]: #Syntax
#if 5 > 2:
```

```
In [12]: print("Five is greater than two!")
```

```
In [14]: #This is a comment
```

```
In [15]: print("Hello, World!")
print("Cheers, Mate!")
```

```
In [16]: #Creating Variables
```

```
In [7]: x = 5
y = "John"
print(x)
print(y)
```

```
5
John
```

```
In [8]: x = 4
x = "Sally"
print(x)
```

```
Sally
```

```
In [11]: x = str(3) # x will be '3'
y = int(3) # y will be 3
z = float(3) # z will be 3.0
print(y)
```

```
3
```

```
In [12]: x = 5
y = "John"
print(type(x))
print(type(y))
```

```
<class 'int'>
<class 'str'>
```

```
In [14]: x = "John"
# is the same as
x = 'John'
print(x)
```

```
John
```

5/7/25, 3:57 PM

Python_Basic_Operations - Jupyter Notebook

```
In [17]: #Case-Sensitive
a = 4
A = "Sally"
#A will not overwrite a
print(a)
```

4

```
In [18]: #Variable Names
myvar = "John"
my_var = "John"
_my_var = "John"
myVar = "John"
MYVAR = "John"
myvar2 = "John"
print(myvar)
```

John

```
In [21]: #Illegal variable names
2myvar = "John"
my-var = "John"
my var = "John"
print(my-var)
```

Cell In[21], line 2

2myvar = "John"

^

SyntaxError: invalid decimal literal

```
In [22]: #Multiple Variables
x, y, z = "Orange", "Banana", "Cherry"
print(x)
print(y)
print(z)
```

Orange
Banana
Cherry

```
In [23]: x = y = z = "Orange"
print(x)
print(y)
print(z)
```

Orange
Orange
Orange

5/7/25, 3:57 PM

Python_Basic_Operations - Jupyter Notebook

```
In [24]: fruits = ["apple", "banana", "cherry"]
x, y, z = fruits
print(x)
print(y)
print(z)
```

```
apple
banana
cherry
```

```
In [25]: #Output Variables
x = "Python is awesome"
print(x)
```

```
Python is awesome
```

```
In [26]: x = "Python"
y = "is"
z = "awesome"
print(x, y, z)
```

```
Python is awesome
```

```
In [27]: #mathematical operator
x = 5
y = 10
print(x + y)
```

```
15
```

```
In [28]: x = 5
y = "John"
print(x + y)
```

```
-
TypeError                                Traceback (most recent call last)
```

```
Cell In[28], line 3
      1 x = 5
      2 y = "John"
----> 3 print(x + y)
```

```
TypeError: unsupported operand type(s) for +: 'int' and 'str'
```

```
In [29]: x = 5
y = "John"
print(x, y)
```

```
5 John
```

5/7/25, 3:57 PM

Python_Basic_Operations - Jupyter Notebook

```
In [ ]: Text Type: str
Numeric Types: int, float, complex
Sequence Types: list, tuple, range
Mapping Type: dict
Set Types: set, frozenset
Boolean Type: bool
Binary Types: bytes, bytearray, memoryview
None Type: NoneType
```

```
In [ ]: #Example                                #Data Type
x = "Hello World"                             str
x = 20                                         int
x = 20.5                                       float
x = 1j                                         complex
x = ["apple", "banana", "cherry"]            list
x = ("apple", "banana", "cherry")             tuple
x = range(6)                                   range
x = {"name": "John", "age": 36}                dict
x = {"apple", "banana", "cherry"}             set
x = frozenset({"apple", "banana", "cherry"})  frozenset
x = True                                       bool
x = b"Hello"                                  bytes
x = bytearray(5)                             bytearray
x = memoryview(bytes(5))                     memoryview
x = None                                       NoneType
```

```
In [30]: #Identification of Data type
x = "Hello World"
print(type(x))
```

```
<class 'str'>
```

```
In [31]: #Identification of Data type
x = {"name": "John", "age": 36}
print(type(x))
```

```
<class 'dict'>
```

```
In [32]: #Python Numbers
x = 1    # int
y = 2.8  # float
z = 1j   # complex
```

```
In [33]: print(type(x))
print(type(y))
print(type(z))
```

```
<class 'int'>
<class 'float'>
<class 'complex'>
```

5/7/25, 3:57 PM

Python_Basic_Operations - Jupyter Notebook

```
In [34]: #Int
x = 1
y = 35656222554887711
z = -3255522
print(type(x))
print(type(y))
print(type(z))
```

```
<class 'int'>
<class 'int'>
<class 'int'>
```

```
In [35]: #Float
x = 1.10
y = 1.0
z = -35.59
```

```
print(type(x))
print(type(y))
print(type(z))
```

```
<class 'float'>
<class 'float'>
<class 'float'>
```

```
In [36]: #Floats another type
x = 35e3
y = 12E4
z = -87.7e100
```

```
print(type(x))
print(type(y))
print(type(z))
```

```
<class 'float'>
<class 'float'>
<class 'float'>
```

```
In [37]: #Complex
x = 3+5j
y = 5j
z = -5j
```

```
print(type(x))
print(type(y))
print(type(z))
```

```
<class 'complex'>
<class 'complex'>
<class 'complex'>
```

5/7/25, 3:57 PM

Python_Basic_Operations - Jupyter Notebook

```
In [38]: #Type Conversion
x = 1 # int
y = 2.8 # float
z = 1j # complex

#convert from int to float:
a = float(x)

#convert from float to int:
b = int(y)

#convert from int to complex:
c = complex(x)

print(a)
print(b)
print(c)

print(type(a))
print(type(b))
print(type(c))
```

```
1.0
2
(1+0j)
<class 'float'>
<class 'int'>
<class 'complex'>
```

```
In [39]: #Python Casting
#Integers
x = int(1) # x will be 1
y = int(2.8) # y will be 2
z = int("3") # z will be 3

#Floats
x = float(1) # x will be 1.0
y = float(2.8) # y will be 2.8
z = float("3") # z will be 3.0
w = float("4.2") # w will be 4.2

#Strings
x = str("s1") # x will be 's1'
y = str(2) # y will be '2'
z = str(3.0) # z will be '3.0'
```

```
In [40]: #Python Strings
print("Hello")
print('Hello')
```

```
Hello
Hello
```

5/7/25, 3:58 PM

Python_Basic_Operations - Jupyter Notebook

```
In [41]: #Python Strings
#Quotes Inside Quotes
print("It's alright")
print("He is called 'Johnny'")
print('He is called "Johnny"')
```

```
It's alright
He is called 'Johnny'
He is called "Johnny"
```

```
In [42]: #Python Strings
#Multiline Strings
a = """Lorem ipsum dolor sit amet,
consectetur adipiscing elit,
sed do eiusmod tempor incididunt
ut labore et dolore magna aliqua."""
print(a)
```

```
Lorem ipsum dolor sit amet,
consectetur adipiscing elit,
sed do eiusmod tempor incididunt
ut labore et dolore magna aliqua.
```

```
In [ ]: #Python Operators

#Python Arithmetic Operators
```

Operator	Name	Example
+	Addition	x + y
-	Subtraction	x - y
*	Multiplication	x * y
/	Division	x / y
%	Modulus	x % y
**	Exponentiation	x ** y
//	Floor division	x // y

```
In [ ]: #Python Assignment Operators
```

Operator	Example	Same As
=	x = 5	x = 5
+=	x += 3	x = x + 3
-=	x -= 3	x = x - 3
*=	x *= 3	x = x * 3
/=	x /= 3	x = x / 3
%=	x %= 3	x = x % 3
//=	x //= 3	x = x // 3
**=	x **= 3	x = x ** 3
&=	x &= 3	x = x & 3
=	x = 3	x = x 3
^=	x ^= 3	x = x ^ 3
>>=	x >>= 3	x = x >> 3
<<=	x <<= 3	x = x << 3
:=	print(x := 3)	x = 3

5/7/25, 3:58 PM

Python_Basic_Operations - Jupyter Notebook

```
In [6]: #Python Comparison Operators
Operator   Name      Example
==         Equal     x == y
!=         Not equal x != y
>          Greater than x > y
<          Less than  x < y
>=         Greater than or equal to x >= y
<=         Less than or equal to x <= y
```

```
In [20]: #Python Lists
mylist = ["apple", "banana", "cherry"]
print(mylist)

['apple', 'banana', 'cherry']
```

```
In [19]: thislist= ["apple", "banana", "cherry"]
print(thislist)

['apple', 'banana', 'cherry']
```

```
In [21]: thislist = ["apple", "banana", "cherry", "apple", "cherry"]
print(thislist)

['apple', 'banana', 'cherry', 'apple', 'cherry']
```

```
In [26]: #List Items - Data Types
#String, int and boolean data types
list1 = ["apple", "banana", "cherry"]
list2 = [1, 5, 7, 9, 3]
list3 = [True, False, False]
print(list3)

[True, False, False]
```

```
In [23]: list1 = ["abc", 34, True, 40, "male"]
print(list1)

['abc', 34, True, 40, 'male']
```

```
In [27]: #Checking data types
mylist = ["apple", "banana", "cherry"]
print(type(mylist))

<class 'list'>
```

```
In [39]: #Access Items
thislist = ["apple", "banana", "cherry"]
print(thislist[1])

banana
```

```
In [57]: #Negative Indexing
thislist = ["apple", "banana", "cherry"]
print(thislist[-1])

cherry
```

localhost:8888/notebooks/Nitya_work/Python_Basic_Operations.ipynb

8/16

5/7/25, 3:58 PM

Python_Basic_Operations - Jupyter Notebook

```
In [40]: #Return the third, fourth, and fifth item
thislist = ["apple", "banana", "cherry", "orange", "kiwi", "melon", "mango"]
print(thislist[2:5])
```

```
['cherry', 'orange', 'kiwi']
```

```
In [41]: #Add List Items
thislist = ["apple", "banana", "cherry"]
thislist.append("orange")
print(thislist)
```

```
['apple', 'banana', 'cherry', 'orange']
```

```
In [42]: #Remove Specified Item
thislist = ["apple", "banana", "cherry"]
thislist.remove("banana")
print(thislist)
```

```
['apple', 'cherry']
```

```
In [43]: #Change Item Value
thislist = ["apple", "banana", "cherry"]
thislist[1] = "blackcurrant"
print(thislist)
```

```
['apple', 'blackcurrant', 'cherry']
```

```
In [44]: #Sort Lists
thislist = ["orange", "mango", "kiwi", "pineapple", "banana"]
thislist.sort()
print(thislist)
```

```
['banana', 'kiwi', 'mango', 'orange', 'pineapple']
```

```
In [45]: thislist = [100, 50, 65, 82, 23]
thislist.sort()
print(thislist)
```

```
[23, 50, 65, 82, 100]
```

```
In [46]: #Sort Descending
thislist = ["orange", "mango", "kiwi", "pineapple", "banana"]
thislist.sort(reverse = True)
print(thislist)
```

```
['pineapple', 'orange', 'mango', 'kiwi', 'banana']
```

```
In [47]: thislist = [100, 50, 65, 82, 23]
thislist.sort(reverse = True)
print(thislist)
```

```
[100, 82, 65, 50, 23]
```

5/7/25, 3:59 PM

Python_Basic_Operations - Jupyter Notebook

```
In [49]: #Python Tuples
mytuple = ("apple", "banana", "cherry")
print(mytuple)

('apple', 'banana', 'cherry')
```

```
In [50]: thistuple = ("apple", "banana", "cherry")
print(thistuple)

('apple', 'banana', 'cherry')
```

```
In [51]: thistuple = ("apple", "banana", "cherry", "apple", "cherry")
print(thistuple)

('apple', 'banana', 'cherry', 'apple', 'cherry')
```

```
In [52]: #Tuple Length
thistuple = ("apple", "banana", "cherry")
print(len(thistuple))

3
```

```
In [53]: thistuple = ("apple",)
print(type(thistuple))

#NOT a tuple
thistuple = ("apple")
print(type(thistuple))

<class 'tuple'>
<class 'str'>
```

```
In [55]: tuple1 = ("apple", "banana", "cherry")
tuple2 = (1, 5, 7, 9, 3)
tuple3 = (True, False, False)
print(tuple1)
print(tuple2)
print(tuple3)

('apple', 'banana', 'cherry')
(1, 5, 7, 9, 3)
(True, False, False)
```

```
In [56]: tuple1 = ("abc", 34, True, 40, "male")
print(tuple1)

('abc', 34, True, 40, 'male')
```

```
In [58]: #Access Tuple Items and Negative Indexing
thistuple = ("apple", "banana", "cherry")
print(thistuple[1])

thistuple = ("apple", "banana", "cherry")
print(thistuple[-1])

banana
cherry
```

localhost:8888/notebooks/Nitya_work/Python_Basic_Operations.ipynb

10/16

5/7/25, 3:59 PM

Python_Basic_Operations - Jupyter Notebook

```
In [62]: #Add and remove Items (Tuples)
thistuple = ("apple", "banana", "cherry")
y = list(thistuple)
y.append("orange")
thistuple = tuple(y)
print(thistuple)

thistuple = ("apple", "banana", "cherry")
y = list(thistuple)
y.remove("apple")
thistuple = tuple(y)
print(thistuple)

('apple', 'banana', 'cherry', 'orange')
('banana', 'cherry')
```

```
In [63]: #Python Sets
thisset = {"apple", "banana", "cherry"}
print(thisset)

{'banana', 'apple', 'cherry'}
```

```
In [64]: thisset = {"apple", "banana", "cherry", True, 1, 2}
print(thisset)

{True, 2, 'cherry', 'banana', 'apple'}
```

```
In [65]: thisset = {"apple", "banana", "cherry"}
print(len(thisset))

3
```

```
In [66]: set1 = {"apple", "banana", "cherry"}
set2 = {1, 5, 7, 9, 3}
set3 = {True, False, False}
print(set1)
print(set2)
print(set3)

{'banana', 'apple', 'cherry'}
{1, 3, 5, 7, 9}
{False, True}
```

```
In [67]: myset = {"apple", "banana", "cherry"}
print(type(myset))

<class 'set'>
```

5/7/25, 3:59 PM

Python_Basic_Operations - Jupyter Notebook

```
In [68]: #Access the set
thisset = {"apple", "banana", "cherry"}

for x in thisset:
    print(x)
```

```
banana
apple
cherry
```

```
In [70]: #Add and remove set
thisset = {"apple", "banana", "cherry"}

thisset.add("orange")

print(thisset)

thisset = {"apple", "banana", "cherry"}

thisset.remove("banana")

print(thisset)
```

```
{'banana', 'apple', 'cherry', 'orange'}
{'apple', 'cherry'}
```

```
In [71]: #Python Dictionaries
thisdict = {
    "brand": "Ford",
    "model": "Mustang",
    "year": 1964}
print(thisdict)
```

```
{'brand': 'Ford', 'model': 'Mustang', 'year': 1964}
```

```
In [72]: #Python Datetime
import datetime

x = datetime.datetime.now()
print(x)
```

```
2025-03-24 15:55:42.470136
```

```
In [73]: import datetime

x = datetime.datetime.now()

print(x.year)
print(x.strftime("%A"))
```

```
2025
Monday
```

5/7/25, 4:07 PM

Python_Basic_Operations - Jupyter Notebook

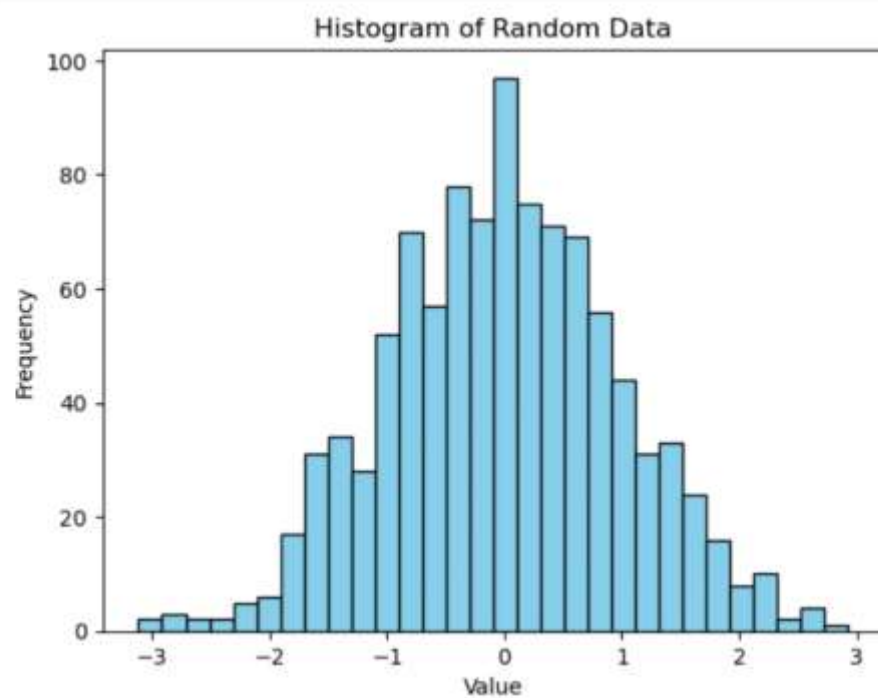
```
In [11]: import matplotlib.pyplot as plt
import numpy as np

# Generate some data
data = np.random.randn(1000) # 1000 values from a normal distribution

# Create histogram
plt.hist(data, bins=30, color='skyblue', edgecolor='black')

# Add labels and title
plt.title('Histogram of Random Data')
plt.xlabel('Value')
plt.ylabel('Frequency')

# Show plot
plt.show()
```



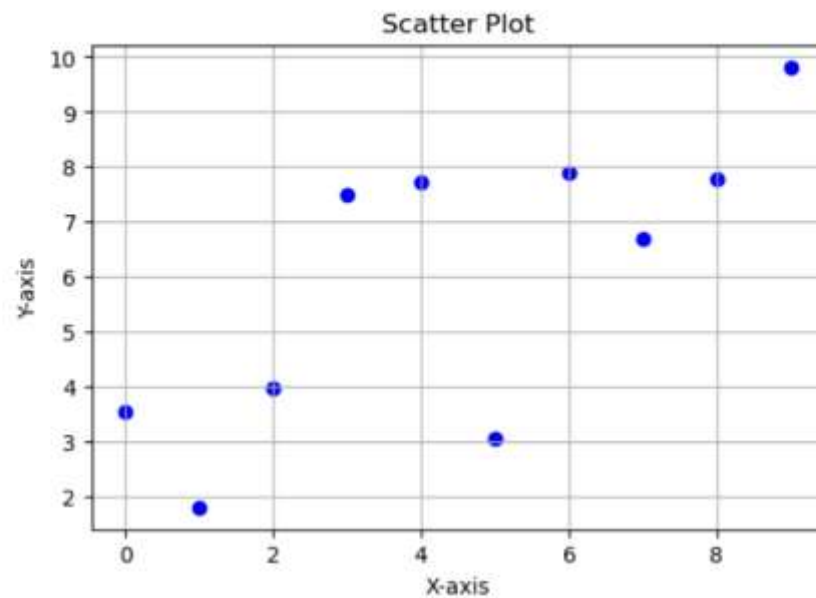
5/7/25, 4:07 PM

Python_Basic_Operations - Jupyter Notebook

```
In [3]: import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd

# Generate random data
np.random.seed(0)
x = np.arange(10)
y = x + np.random.randn(10) * 2
data = pd.DataFrame({
    'Group': np.repeat(['A', 'B', 'C'], 10),
    'Value': np.random.randn(30)})

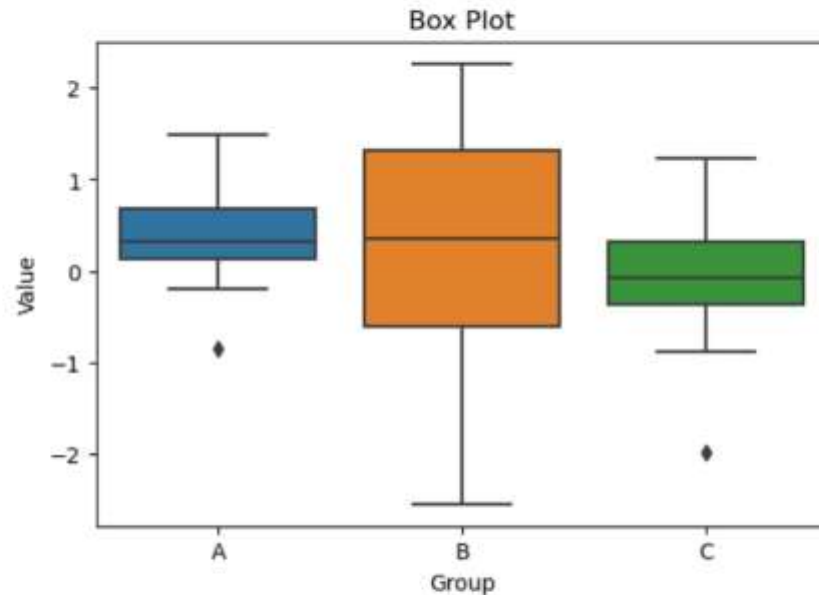
plt.figure(figsize=(6,4))
plt.scatter(x, y, color='blue', marker='o')
plt.title("Scatter Plot")
plt.xlabel("X-axis")
plt.ylabel("Y-axis")
plt.grid(True)
plt.show()
```



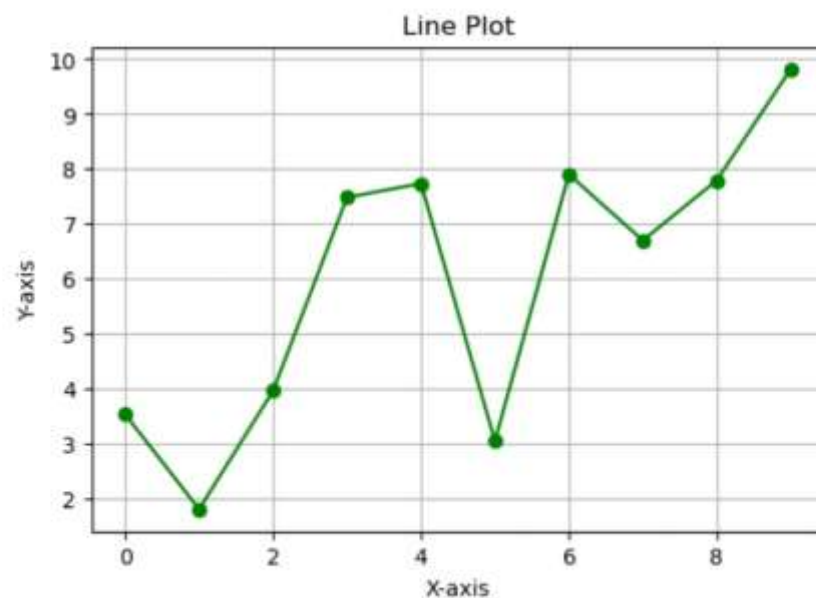
5/7/25, 4:08 PM

Python_Basic_Operations - Jupyter Notebook

```
In [4]: plt.figure(figsize=(6,4))
sns.boxplot(x='Group', y='Value', data=data)
plt.title("Box Plot")
plt.show()
```



```
In [5]: plt.figure(figsize=(6,4))
plt.plot(x, y, marker='o', linestyle='-', color='green')
plt.title("Line Plot")
plt.xlabel("X-axis")
plt.ylabel("Y-axis")
plt.grid(True)
plt.show()
```



Building a Regression Model in Python

5/7/25, 4:08 PM

Python_Basic_Operations - Jupyter Notebook

```
In [6]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Generate synthetic dataset
# For example: y = 3x + noise
np.random.seed(0)
X = 2 * np.random.rand(100, 1)
y = 3 * X + np.random.randn(100, 1)

# Split into training and testing data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, ran

# Create the model
model = LinearRegression()
model.fit(X_train, y_train)

# Predict
y_pred = model.predict(X_test)

# Evaluate
print("Coefficients:", model.coef_)
print("Intercept:", model.intercept_)
print("Mean Squared Error:", mean_squared_error(y_test, y_pred))
print("R2 Score:", r2_score(y_test, y_pred))

# Plotting
plt.scatter(X_test, y_test, color='blue', label='Actual')
plt.plot(X_test, y_pred, color='red', label='Predicted')
plt.title('Linear Regression Example')
plt.xlabel('X')
plt.ylabel('y')
plt.legend()
plt.show()

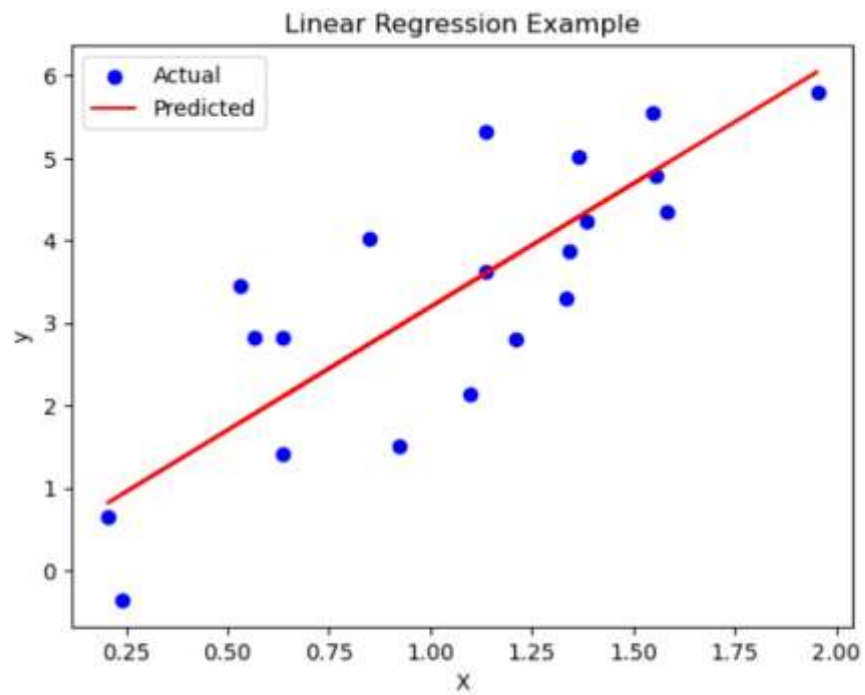
Coefficients: [[2.9902591]]
Intercept: [0.20634019]
Mean Squared Error: 0.9177532469714288
R2 Score: 0.6521157503858557
```

localhost:8888/notebooks/fitya_work/Python_Basic_Operations.ipynb

16/17

5/7/25, 4:08 PM

Python_Basic_Operations - Jupyter Notebook



In []:

Suggested Readings:

1. Tilman M. Davies. 2016. The book of R: A first course in programming and statistics. No Starch Press, US.
 2. Andy Field, Jeremy Miles, and Zoe Field. 2022. Discovering statistics using R. SAGE Publications India Pvt Ltd. India.
 3. Jared P. Lander. 2018. R for everyone: Advanced analytics and graphics. 2nd Edition. Pearson Education. India.
 4. Bharti Motwani. 2019. Data analytics with R. Wiley. India.
 5. Sandip Rakshit. 2017. R programming for beginners. McGraw Hill Education. India.
 6. Eric Matthes. 2016. Python crash course. No Starch Press, US.
 7. Paul Barry. 2016. Head-first Python, 2nd Edition. O'Reilly.
 8. William McKinney. 2017. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media.
 9. David Landup. 2021. Data Visualization in Python with Pandas and Matplotlib. Stack Abuse.
 10. Log2Base2 courses on python (<https://log2base2.com/courses/python>).
 11. Coursera. Data Analysis with Python (<https://www.coursera.org/learn/data-analysis-with-python>)
 12. Free codecamp. Data analysis with python (<https://www.freecodecamp.org/learn/data-analysis-with-python/>)
 13. Coursera. Data analysis with R-programming (<https://www.coursera.org/learn/data-analysis-r?B>).
-

DISCLAIMER

This self-learning material is based on different books,
journals and web sources.