Chapter # 05
Data Analytics

Q.1. What is a model in computer science, and how does it help in real life? Give examples to explain.

Ans. A model is a simple way to understand or solve a big problem by focusing on the most important parts and ignoring extra details. Think of it like using a map to find your way in a city instead of looking at every building and road. A model helps you make decisions faster and easier.

How Models Work:

A model has three main parts:

1. **Input:** The information you start with (like data or facts).
2. **Process:** The steps or rules used to solve the problem.
3. **Output:** The final result or answer.

Example 1: Study Plan Model

- **Input:** A student has exams in Math, Science, and English.
- **Process:** They decide to spend more time on weak subjects (e.g., Math) and less on strong ones (e.g., English).
- **Output:** A plan that helps them study better and pass exams.

Example 2: Farmer's Planting Model

- **Input:** Weather data (like rain patterns) and soil quality.
- **Process:** The farmer uses this data to guess the best time to plant crops.
- **Output:** A good harvest because they planted at the right time.

Why Are Models Important?

1. **Make Decisions Easier:** Models help you choose the best action by showing possible outcomes.
   Example: A farmer avoids planting during heavy rain.
2. **Save Time and Effort:** They ignore unimportant details.
   Example: A study plan focuses on key subjects, not every book in the library.
3. **Predict the Future:** Models use data to guess what might happen.
   Example: Weather models predict rain, so you bring an umbrella.

Where Are Models Used?

1. **Weather Forecasting:**
   - Meteorologists collect data (temperature, wind speed) and use models to guess if it will rain or be sunny.
2. **Finance:**
   - Banks use models to guess if stock prices will rise or fall.
3. **Science:**
   - Scientists use models to test ideas, like how a new medicine works in the body.

Q.2. What are the basic statistical concepts used in data analysis? Explain the measures of central tendency, with suitable examples. / Discuss the concept of measure of tendency with example.

Ans. **Data analytics** is the process of studying **data** to find useful information, patterns, or trends that help in decision-making. To analyze data properly, we use some basic statistical concepts.

Basic Statistical Concepts:

Statistics is a branch of mathematics that helps us summarize and understand large sets of data. With the help of statistics, we can draw meaningful conclusions easily.

Statistical tools like **mean, median, mode, variance, standard deviation, and probability** help us understand data more clearly. They play an important role in making better decisions in fields like **education, healthcare, business, and science**.

Measures of Central Tendency:

Measures of central tendency help us find the **middle** or **average** value in a group of numbers. There are **three main types:**

1. Mean (Average):

The **mean** is the sum of all values divided by the number of values. It shows the **average value** in a data set.

Formula:

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Total number of values}}$$

Example:

Scores: 50, 60, 70, 80, 90

$$\text{Mean} = \frac{50+60+70+80+90}{5} = 70$$

2. Mode:

The **mode** is the value that **appears most often** in a data set. A data set can have **no mode, one mode**, or **more than one mode.**

Example:

Scores = 50, 60, 70, 70, 90 → Mode = **70**

Scores = 50, 60, 70, 70, 60, 90 → Modes = **60** and **70**

The mode helps us identify the most frequent or common value in the data.

3. Median:

The **median** is the **middle value** when the data is arranged in **ascending order** (smallest to largest).

If there are:

- **Odd** number of values → the median is the middle one.
- **Even** number of values → the median is the **average of the two middle values.**
- **Example (Odd):** 50, 60, 70, 80, 90 → Median = **70**
- **Example (Even):** 50, 60, 70, 80 → Median $= (60+70)/2 = 65$, the median is 65

The median helps us understand the middle point of the data.

Q.3. Explain the measures of dispersion, and introduction to probability with suitable examples.

Ans. Measures of Dispersion:

Dispersion tells us **how much the data is spread out**. It helps us see whether the values in a dataset are close to the average (mean) or far away from it.

Two common ways to measure dispersion are:

- Variance
- Standard Deviation

1. Variance (σ^2):

Variance shows **how far each number in the data is from the mean.** If the variance is small, the values are close to the mean. If it is large, the values are spread out.

Formula:

$$\text{Variance } (\sigma^2) = \frac{Sum\ of\ (x_i - \mu)^2}{N}$$

Where:

- x_i = each value
- μ = mean
- N = number of values

Example – Variance for Class A

- **Scores:** 50, 52, 55, 57, 60
- **Step 1: Find the Mean**

$$(\mu) = \frac{50+52+55+57+60}{5} = \frac{274}{5}$$

$$= 54.8$$

- **Step 2: Find the squared differences from the mean**

x_i	$x_i - \mu$	$(x_i - \mu)^2$
50	-4.8	23.04
52	−2.8	7.84
55	0.2	0.04
57	2.2	4.84
60	5.2	27.04

Step 3: Find Variance

$$\text{Variance } (\sigma^2) = \frac{23.04+7.84+0.04+4.84+27.04}{5}$$

$$= \frac{62.8}{5} = 12.56$$

Example – Class B:

Scores: 30, 45, 55, 75, 90

Step 1: Mean

$$\mu = \frac{30+45+55+75+90}{5}$$

$$= \frac{295}{5} = 59$$

Step 2: Squared differences

x_i	$x_i - \mu$	$(x_i - \mu)^2$
30	-29	841
45	-14	196
55	-4	16
75	16	256
90	31	961

Step 3: Variance

$$\sigma^2 = \frac{841+196+16+256+961}{5}$$

$$= \frac{2270}{5} = 454$$

- Variance of Class A: 12.56
- Variance of Class B: 454

So, Class B has a much higher variance than Class A. That means the marks in Class B are **more spread out**.

2. Standard Deviation (σ):

Standard deviation tells us **how far the numbers are from the mean**. It is just the **square root of variance**. It is easier to understand than variance because it is in the same unit as the data.

Formula:

$$\text{Standard Deviation } \sigma = \sqrt{\text{Variance}}$$

Example:

Class A:

Standard Deviation $= \sqrt{12.56} = 3.55$

Class B:

Standard Deviation $= \sqrt{456} = 21.26$

This shows Class A's values are closer to the mean, while Class B's values are more spread out.

Introduction to Probability:

Probability is the study of how likely an event is to happen. It is used to **predict future outcomes.**

Formula:

$$\text{Probability} = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

Example: Tossing a coin

- Total outcomes: Head, Tail = 2
- Favorable outcome (e.g., Head) = 1

Probability of Head $= \frac{1}{2} = 0.5 = 50\%$

Tidbits:
Statistics can help us understand patterns in data, leading to better decision-making in various fields, from healthcare to marketing!

Where is Probability Used?

Probability is used in many areas like:

- Weather forecasting (e.g., 80% chance of rain)
- Business decisions (e.g., product success)
- Games and sports
- Medical testing (e.g., chance of illness).

Do You Know?
Statistics is used in many everyday activities, such as predicting weather patterns or analyzing sports performance.

Class Activity

Instructions: You will analyze a small dataset, calculate measures of central tendency (mean, median, mode), and measures of dispersion (variance and standard deviation).

1. **Collect Data:** Imagine you are surveying your classmates about the number of hours they spend on homework in a week. Gather data from 10 classmates. Record the number of hours (you can use any reasonable numbers, e.g., between 0 to 20 hours).
2. **Calculate Measures of Central Tendency:**
   - **Mean:** Calculate the average number of hours spend on homework.
   - **Median:** Determine the middle value when the hours are arranged in order.
   - **Mode:** Identify which number appear most frequently in your data.
3. **Calculate Measures of Dispersion:**
   - **Variance:** Use the formula to calculate variance based on your data.
   - **Standard Deviation:** Calculate the standard deviation using the variance you found.
4. **Reflect:** Write a brief reflection (3-4 sentences) about what these calculations tell you regarding your classmates' study habits.

Answer.

1. **Collect Data (Sample Data):**

   Number of hours spent on homework by 10 classmates:

   [5, 8, 7, 6, 10, 12, 5, 8, 9, 6]
2. **Measures of Central Tendency**
   - Mean = (5 + 8 + 7 + 6 + 10 + 12 + 5 + 8 + 9 + 6) / 10
     = 76 / 10
     = 7.6 hours
   - Median: Arrange the data: [5, 5, 6, 6, 7, 8, 8, 9, 10, 12]
     Middle two values = 7 and 8
     Median = (7 + 8) / 2 = 7.5 hours

- Mode: 5, 6, and 8 each appear twice → Multiple modes
  Mode = 5, 6, and 8 hours

3. **Measures of Dispersion**
   - **Step 1:** Find the mean = 7.6
   - **Step 2:** Calculate squared differences from the mean:

x	$(x - \mu)$	$(x - \mu)^2$
5	5 - 7.6 = -2.6	6.76
8	8 - 7.6 = 0.4	0.16
7	7 - 7.6 = -0.6	0.36
6	6 - 7.6 = -1.6	2.56
10	10 - 7.6 = 2.4	5.76
12	12 - 7.6 = 4.4	19.36
5	-2.6	6.76
8	0.4	0.16
9	1.4	1.96
6	-1.6	2.56

   Total of squared differences = 46.4
   Variance = 46.4 / 10 = 4.64
   Standard Deviation = ≈ 2.15

4. **Reflection:**

Most classmates spend between 5 to 10 hours on homework, with an average of 7.6 hours per week. Modes of 5, 6, and 8 show common but varied study habits. A standard deviation of 2.15 suggests that study times are fairly consistent across students.

Q.4. Explain data collection methods. / What are the different methods of data collection, and how do they work?

Ans. **Data collection and preparation** are the most important steps in any research or analysis. The quality and correctness of data directly affect the accuracy of the results. This process involves collecting the right data, preparing it by cleaning and transforming it, and then handling any missing or incorrect values.

Data collection is the process of gathering relevant information needed to answer a research question, solve a problem, or make informed decisions. The choice of method depends on:

- The research objective
- The type of data needed (quantitative or qualitative)
- Available time and resources

There are **three main methods** of data collection:

1. **Surveys :**

Surveys are a structured way to collect data by asking a fixed set of questions from a group of people (sample). They help collect **large amounts of data quickly**.

Key Features:

- Questions are predefined (same for everyone)
- Useful for collecting **opinions, preferences, habits, or behaviors**
- Can be conducted through:
  - Online forms (e.g., Google Forms)
  - Phone interviews
  - Face-to-face interviews
  - Printed questionnaires

Advantages:

- Fast and cost-effective
- Can reach a large number of people
- Easy to analyze (quantitative data)

Limitations:

- Responses may be biased or dishonest
- May not allow deep explanations

Example:

A local grocery store in Islamabad wants to know which products customers prefer. They make a short 5-question survey and give it to 50 customers. Responses help them stock the most demanded products.

Survey Questions:

1. Which product categories do you buy most often?
2. Are there any products you would like to see more often?
3. How often do you shop at this grocery store?
4. What influences your purchasing decisions? (price, quality, etc.)
5. Any comments or suggestions?

2. Observations:

Observation means watching people, things, or events carefully in their natural setting to collect information, without asking questions or making any changes.

Key Features:

- No direct interaction with subjects
- Useful for studying **habits, movements, and real-time behaviors**
- Can be done openly or discreetly
- Can be manual (human observer) or automatic (cameras, sensors)

Advantages:

- Realistic and natural data
- Useful when participants cannot explain their behavior
- Captures non-verbal actions

Limitations:

- Time-consuming
- Observer bias possible
- Cannot collect deep opinions or reasons behind behavior

Example:

A restaurant observes which tables customers choose during lunch for one week. Based on results, they rearrange tables to improve comfort and flow of movement.

3. Experiments :

Experiments involve **changing one or more variables** to observe their effect on another variable. This method is used to test **cause-and-effect relationships**.

Key Features:

- Used in scientific, educational, or technical research
- One group is exposed to the change (experimental group), and one is not (control group)
- Conducted in a **controlled environment** to avoid errors

Advantages:

- Produces accurate and reliable results
- Helps establish clear relationships between variables
- Data is often quantitative and easy to analyze

Limitations:

- Requires careful planning
- Not always practical in real-life settings
- May involve ethical concerns in some studies

Example:

A teacher wants to see if providing printed notes helps students perform better in exams. The teacher splits the class into two groups: one group gets printed notes, and the other only gets lectures. After a month, both groups take the same test, and the teacher compares their results to see if the printed notes helped improve performance.

Q.5. What is data preparation, and why is it important for analysis? Explain how missing or incorrect data is handled. Support your answer with examples.

Ans. **Data Preparation:**

After collecting data, the next important step is to **prepare it for analysis**.

This means:

- **Cleaning the data** to remove mistakes or wrong entries
- **Organizing it** in a proper and meaningful way
- **Changing it into a format** that is easy to study

Sometimes, some data may be **missing or incorrect**. In such cases, researchers can use **special techniques** like:

- **Interpolation** (estimating missing values)
- **Statistical adjustments** (making corrections using math or patterns)

Example:

If a survey has missing answers, those missing values can be **estimated** using the rest of the available answers.

Proper data preparation helps make the analysis **accurate and trustworthy**.

Data Cleaning and Transformation:

Raw data (the original data collected) is often not ready to use.

It may have:

- Errors
- Missing information
- Wrong format

To make it useful, we need to **clean and transform** the data.

Data Cleaning:

Data cleaning means **fixing or removing problems** in the data.

These problems can be:

- **Incorrect entries** (like spelling mistakes)
- **Missing values** (like blank marks)
- **Duplicate data** (same entry repeated)

If we don't clean the data, the final results of the analysis will be **wrong or confusing.**

Example:

A school collects data about student scores. But there are some mistakes:

Name	Scores	Class	Section
Ali	84	10	A
Ali	90	10	A
Sara		10	A

Here:

- "Alie" is a spelling mistake (should be "Ali")
- Sara's score is missing

After cleaning, the data looks like this:

Name	Grade	Class	Section
Ali	84	10	A
Ali	90	10	A
Sara	87	10	A

What was fixed:

- "Alie" corrected to "Ali"
- Sara's missing score was **filled with the average (87)**

Data Transformation:

After cleaning, we sometimes need to **change the data into a better format** for analysis.

This is called **Data Transformation**.

This can include:

- **Making new columns**
- **Rearranging the data**
- **Changing formats** to make it easier to read or understand

Example:

After fixing student grade records, the school might:

- Group the data by **class, section,** or **average marks**
- This makes it easier to compare and study the data

Handling Missing Data:

Sometimes, some data is **missing or not recorded**.

There are **three ways** to handle missing data:

1. **Imputation:**

This means **filling the missing value** using the average or similar data.

Example:

Sara's marks are missing. The school calculates the class average (87) and assigns that value to Sara.

This keeps the dataset complete and gives a fair estimate.

2. **Flagging:**

Instead of filling the value, the missing data is **left empty but marked**.

This tells the analyst that a value is missing, but nothing is changed.

Example:

The school adds a **note or symbol** in Sara's record to show her marks are missing.

This keeps the analysis honest and transparent.

3. **Removal:**

If **only a few records** have missing data, and they are not important for the final results, they can be **removed** from the study.

Example:

If Sara is the only one with missing marks, the school might decide to **remove her record** from some parts of the analysis.

But this also means losing useful information about Sara.

By cleaning and transforming the data, and dealing with any missing information, researchers can ensure that their analysis is accurate and meaningful.

Q.6. Explain the role and importance of statistical models in solving real-world problems.

Ans. Statistical models help us understand real-life problems using data. These models show how different things are related and help us make better decisions.

For example, if we want to guess how much money a person will spend on groceries next month, we can look at past spending data. A statistical model can then help us **predict future expenses**.

Importance of Statistical Models:

- They are used to **predict future events**.
- Help us **understand relationships** between different factors.
- Make **better and faster decisions** using past data.

Examples of Use:

- In **business**, to predict sales.
- In **education**, to check student performance.
- In **health**, to track disease spread.
- In **banks**, to check who should get a loan.

Steps to build a statistical model:

1. **Define the Problem:**
   First, we must know what we want to predict. **Example:** grocery costs.
2. **Collect Data:**
   Gather useful information like past spending, number of people in the family, etc.
3. **Choose an Algorithm:**
   Select a method (like linear or logistic regression) to make the model.
4. **Train the Model:**
   Teach the model using the collected data so it can learn and make predictions.
5. **Evaluate the Model:**
   Test the model with new data to check if it gives correct results.

There are different types of models:

- **Linear regression** is used to predict values (like income).
- **Logistic regression** is used to predict categories (like pass/fail).
- **Clustering** is used to group similar items (like students with similar marks).

In short, statistical models are **useful tools** that make problem-solving easier, faster, and more accurate by using real data.

Q.7. Describe the steps involved in building a basic statistical model (e.g., linear regression). Include details on data collection, model training, and evaluation.

Ans. **Linear regression** is a type of statistical model used to study the relationship between two numerical variables: one **independent variable** (cause) and one **dependent variable** (effect).

It helps us **predict** the future value of the dependent variable based on the independent variable by fitting a straight line through the data.

Linear Regression Formula:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Where:

- Y = Dependent variable (what we want to predict)
- X = Independent variable (what we use to predict)
- β_0 = Intercept (starting value when X = 0)
- β_1 = Slope (how much Y changes with each unit increase in X)
- ε = Error term (difference between actual and predicted values)

Tidbits:

To improve your statistical model, consider these suggestions:

1. Use more data points for better accuracy.
2. Include relevant factors such as like family size or special events that may affect spending's.
3. Regularly update your model with new data to keep it relevant.
4. Test your predictions against actual spending to refine your approach.

Steps in Building a Linear Regression Model:

To build a basic statistical model such as **linear regression**, we follow several important steps. Each step has a specific purpose and helps improve the model's accuracy.

Step 1: Define the Problem

This means clearly understanding what we are trying to predict or solve.
Before building a model, we must know the question we want to answer.
Example: If we want to predict daily income from a fruit stall, we must find what affects that income. In this case, the number of customers is the key factor.

Step 2: Collect the Data

Data collection means gathering information related to the problem.
We collect data on the important factors (variables). This data helps the model learn and make predictions.

Example Data:

Customers	Earnings (Rs.)
10	500
15	700
20	900
25	1100
30	1300

Here, **customers** are the independent variable (X), and **earnings** are the dependent variable (Y).

Step 3: Choose the Model (Linear Regression)

Select an appropriate method that best fits the type of data.

We use **linear regression** when we want to predict a numerical value using another variable. It shows a straight-line relationship between two variables.

Formula:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Step 4: Train the Model

Training means teaching the model by using the collected data. We use the collected data to calculate the slope (β_1) and intercept (β_0).

From data:

Customers (X)	Earnings (Y)
10	500
15	700
20	900
25	1100
30	1300

- Increase in earnings = Rs. 200
- Increase in customers = 5

Step 1: Find the Slope (β_1)

We notice that:

When the number of customers increases by 5, the earnings increase by Rs. 200.

So,

- β_1 = Change in Earnings / Change in Customers = 200 / 5 = 40 (slope)

This means: **Every extra customer adds Rs. 40 to the earnings.**

Step 2: Find the Intercept (β_0)

Use any data point, for **Example:**

When **Customers** = 10, **Earnings** = Rs. 500

Use the formula:

$$Y = \beta_0 + \beta_1 x$$

Substitute values:

$500 = \beta_0 + (40 \times 10) \quad \Rightarrow \quad 500 = \beta_0 + 400 \quad \Rightarrow \quad \beta_0 = 500 - 400 = 100$

This means: **Even if no customers come, you will still earn Rs. 100** (maybe from regular or fixed earnings).

Final Equation:

Earnings = 100 + 40 × Customers

This equation is now your **trained model.** We can use it to **predict future earnings** based on the number of customers.

Step 5: Make Predictions

Prediction means using the model to estimate future outcomes.

We put new values into the equation to predict results.

Example:

If 22 customers visit, *predicted earnings:*

Earnings = 100 + 40 × 22 = 980 Rs.

Step 6: Evaluate the Model

Evaluation means checking how accurate the model's predictions are.

We compare the model's predictions with real values. This shows how well the model works.

Example:

If actual earnings with 28 customers = Rs. 1250,
Predicted earnings = 100 + (40 × 28) = Rs. 1220
Error = 1250 – 1220 = Rs. 30
Since the error is small, the model is considered **accurate**.

Q.8. How can we improve the accuracy and performance of a statistical model?

Ans. To improve a statistical model, consider the following suggestions:

1. **Use more data points** to make the model more accurate and reliable.
2. **Include relevant factors** like family size, income, or special events that may affect the outcome.
3. **Regularly update the model** with new data to keep it current and useful.
4. **Test predictions against real outcomes** to identify and correct any errors in the model.

Q.9. Explain Logistic Regression & Clustering statistical modeling techniques.

Ans. Statistical modeling techniques help us understand data, make predictions, and find patterns. Some important techniques include **logistic regression** and **clustering**.

1. Logistic Regression:

Logistic regression is a method used to predict an outcome that can be either "yes" or "no" (a **binary outcome**). Instead of giving a specific number, it gives a **probability** between 0 and 1, which helps us understand how likely an event is to happen.

Example:

We can use logistic regression to **predict** the chance of a student **passing** an exam based on the number of **study hours**.

Steps to Use Logistic Regression:

1. **Collect Data:**
   You gather data on students' study hours and whether they passed or failed.

Hours Studied	Passed (1) / Failed (0)
1	0
2	0
3	0
4	1
5	1
6	1

2. **Build the Model:**
   In logistic regression, we create a model using the **logistic function**:

$$P(\text{Pass}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \times \text{Hours})}}$$

   Where:
   - **P(Pass)** is the probability of passing.
   - **e** is the base of the natural logarithm.
   - β_0 is the intercept (log-odds of passing when no hours are studied).
   - β_1 is the coefficient (how the log-odds change for each additional hour studied).
3. **Interpret the Model:**
   After fitting the model to the data, we get values for β_0 and β_1.
   Suppose $\beta_0 = 1$ and $\beta_1 = 0.6$. This means:
   - For each hour studied, the chance of passing increases.
   - If a student studies for **5 hours**, we calculate the passing probability:

$$P(\text{Pass}) = \frac{1}{1 + e^{-(1 + 0.6 \times 5)}} \approx 0.83$$

   This means there is an **83% chance** that the student will pass after studying for 5 hours.

4. **Make Predictions:**

Let's predict for a student who studies for **4 hours:**

$$P(\text{Pass}) = \frac{1}{1+e^{-(1+0.6\times4)}} \approx 0.73$$

This gives a **73% chance** of passing the exam.

2. Clustering Techniques:

Clustering is a statistical method used to group similar items or individuals based on common characteristics. It helps us identify patterns and create meaningful groups from data without using labels.

Example – Clustering Students by Performance:

Suppose we have the scores of five students in **Math** and **English**. Using clustering, we can divide them into two groups:

- **Group 1**: Students who are strong in **Math** (e.g., *Basim* and *Umer*)
- **Group 2**: Students who are strong in **English** (e.g., *Anie* and *Tallat*)

This helps teachers or school staff plan targeted support for different groups of students.

How Clustering Works:

1. **Define the Problem**: Decide what you want to group.
   Example: Group 5 students based on their math and English scores.
2. **Collect Data**: Gather the information (scores) for each student.
   Example:

Name	Math	English
Basim	80	40
Umer	75	35
Anie	45	85
Tallat	50	90
Maliha	70	45

3. **Choose Clusters:** Decide how many groups you want.
   Example: We choose 2 clusters (math-strong and English-strong).
4. **Find Cluster Centers:** Pick starting points (e.g., average scores) for each group.
   Example:
   - Cluster 1 (math-strong): Start with (Math = 80, English = 40).
   - Cluster 2 (English-strong): Start with (Math = 45, English = 85).
5. **Assign Students to Clusters:**
   Calculate how close each student is to the cluster centers.
   Example:
   - Basim, Umer, and Maliha are closer to the math-strong cluster.
   - Anie and Tallat are closer to the English-strong cluster.
6. **Update Cluster Centers:**
   Recalculate the average scores for each group.
   Example:
   - Math-strong group: New average = (80 + 75 + 70) ÷ 3 = 75 (Math), (40 + 35 + 45) ÷ 3 = 40 (English).
   - English-strong group: New average = (45 + 50) ÷ 2 = 47.5 (Math), (85 + 90) ÷ 2 = 87.5 (English).
7. **Repeat Until Stable:**
   Recheck distances and reassign students if needed.
   Example: After updating centers, Basim, Umer, and Maliha stay in the math-strong group. Anie and Tallat stay in the English-strong group. The clusters are now stable (no more changes).

Final Result:

- Cluster 1 (Math-Strong): Basim, Umer, Maliha.
- Cluster 2 (English-Strong): Anie, Tallat.

Why Clustering Matters:

This helps in:

- Better understanding of data
- Easier decision-making
- Identifying hidden patterns

It's especially useful when there are no predefined labels, as clustering helps discover natural groupings.

Example: Schools can group students based on performance data to provide targeted support (e.g., Cluster 2 gets extra help in math).

Real-life analogy: It's like sorting toys into boxes by color or size, there are no strict rules, just putting similar items together.

K-means Clustering:

K-means is one of the most widely used clustering techniques. Here's how it works:

- First, we decide how many clusters (groups) we want. For example, 2 clusters.
- The algorithm calculates the **distance** between each data point (e.g., student scores) and the cluster centers.
- It then groups the data points based on how close they are to each cluster center.

The algorithm repeats this process, updating cluster centers and reassigning data until the clusters become stable (no more changes).

This method is useful in many fields such as education, marketing, health, and research for analyzing data and making smart decisions.

Q.10. How we evaluate models. Also explain the importance of ethical considerations while using data.

Ans. Model Evaluation:

After building a statistical model, it is important to check how well the model is working. This process of checking how well a model performs and understanding the results it provides is called **Model Evaluation.**

It helps us improve the model and make better decisions based on its results.

> **Tidbits:**
> Always visualize your data before building models and test your results on new data for better accuracy.

1. Performance Metrics:

Performance metrics are tools used to measure the model's accuracy and reliability. There are two common types:

- **Error Metrics:**
  These show the difference between the predicted value and the actual (real) value.
  Example: If a model predicts grocery expenses to be **Rs. 8,000**, but the actual cost is **Rs. 10,000**, the **error** is **Rs. 2,000**.
- **Accuracy Metrics:**
  These tell us how many predictions made by the model were correct.
  Example: If a model correctly predicts that **9 out of 10** students will pass an exam, the **accuracy** is **90%.**

Type	What It Measures	Example
Error Metrics	Difference between predicted and actual values	Predicted bill = 8,000, Actual bill = 10,000 → Error = **2,000 rupees**
Accuracy Metrics	How often predictions are correct	Model predicts pass/fail, and is correct **80 out of 100 times** → Accuracy = **80%**

2. *Interpreting Outputs:*

After evaluating a model, we must understand what the results mean. This helps in making correct decisions.

Example:

If a model shows that students who study more hours score higher in exams, we can suggest that students increase their study time to improve their performance.

Ethical Considerations in Modeling:

While building and using models, it is important to follow ethical rules. This means we should be fair, honest, and respect the privacy of people whose data is used.

1. **Fairness and Bias:**

A good model should treat all individuals equally and should not be biased toward or against any group.

Example:

A model used to approve bank loans should give equal chances to all applicants, regardless of gender, race, or background.

2. **Data Privacy:**

The personal data used in building models must be kept safe and not shared without permission.

Example:

If a company collects data from customers, it must protect that data and not use or sell it without the customer's approval.

Q.11. Discuss the types of data visualizations and their uses.

Ans. Introduction:

Data visualization means showing data in the form of pictures like graphs, charts, or diagrams. It helps us understand the data more easily and quickly. With the help of these visuals, we can see patterns, trends, and comparisons without reading long numbers or tables.

Types of Data Visualizations:

1. **Bar Chart:**

A **bar chart** is a type of data visualization used to **compare different categories**.

- Each **bar** represents a category (like a product, subject, or student).
- The **height or length** of the bar shows the value (like sales, scores, or quantities).

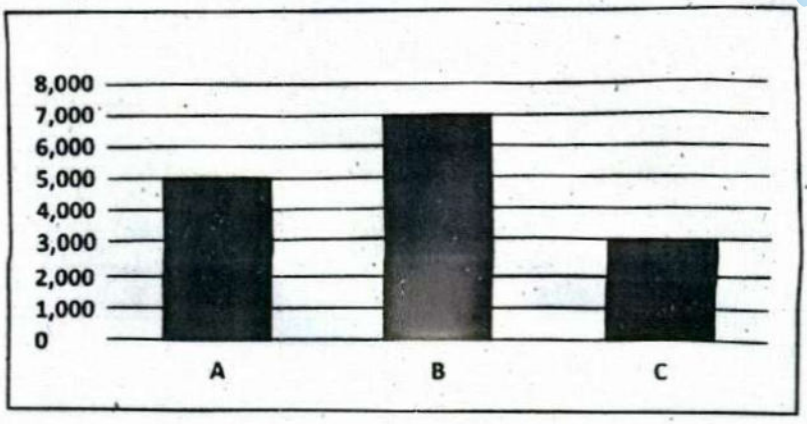
Use: Bar charts are useful to compare different things like sales of products or number of students in different sections.

Example:

Imagine you have sales data for three products:

Product	Sales (in Rs.)
A	5,000
B	7,000
C	3,000

A bar chart would show three bars, one for each product and their heights would represent how much each product sold. This makes it easy to see which product sold the most.



2. Line Graph:

A line graph shows changes over time. It connects points with lines, so we can easily see if the values are going up or down.

- **Use:** Good for showing trends like temperature changes, monthly expenses, or growth over time.

Example:

If you check the temperature every day for a week, a line graph will show how the temperature went higher or lower each day.

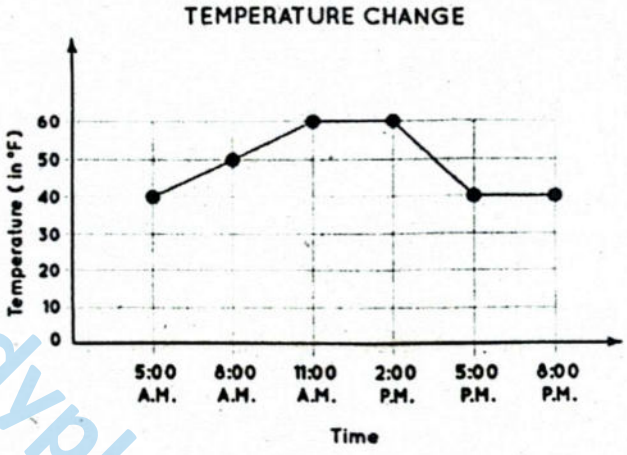


Figure: A line graph showing variation of temperature over time

3. Histogram:

Histograms are used to show how data is spread out. It looks like a bar chart but it groups numbers into ranges (called bins) and show how many values fall into each range.

- **Use:** Helpful to study the distribution of marks, heights, weights, etc.

Example:

If you want to check how students did in a math exam, a histogram can show how many students got scores in different ranges, like 60–70, 70–80, and so on.

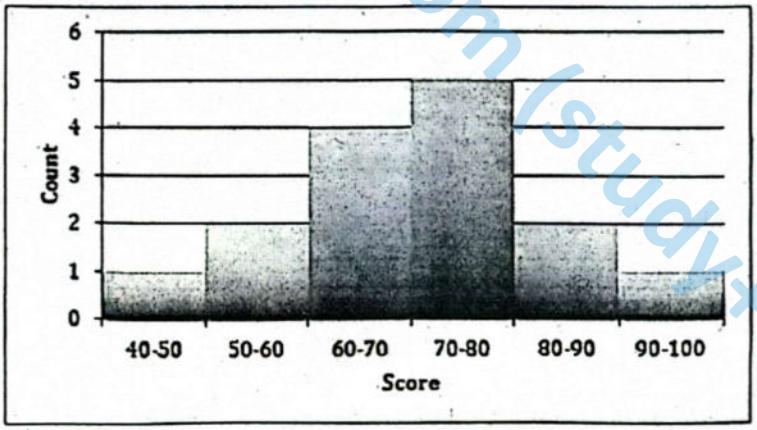


Figure: Example of a Histogram showing the distribution of exam scores

4. Scatterplot:

A scatterplot shows the relationship between two things (variables). Each dot on the graph stands for one piece of data and shows the values for both things.

- **Use:** Useful to find if two things are related, like study time and marks.

Example:

A scatterplot can help you see if there's a link between how many hours a student studied and their Test Grade.

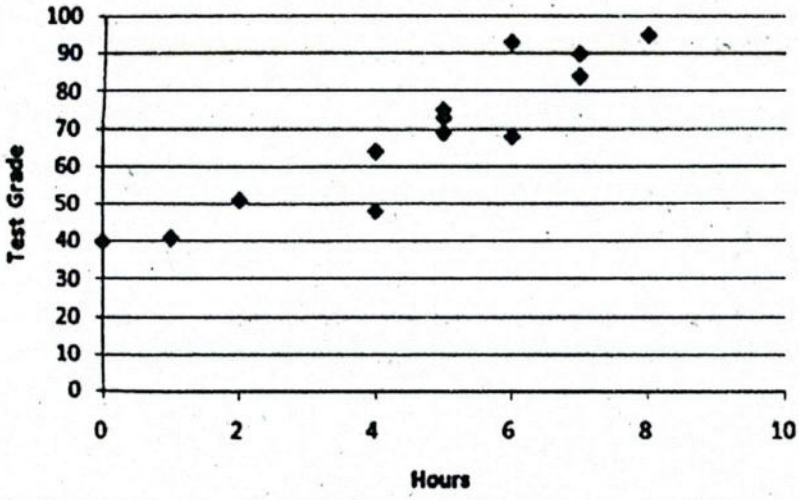


Figure: Scatterplot showing the relationship between hours studied and exam scores

5. Boxplot (Whisker Plot):

Boxplots (also called whisker plots) are used to show how data is spread out. It shows the middle value (median), the range where most values fall (quartiles), and any unusual values (outliers). It basically shows a summary of data.

- **Use:** Best to compare data of different groups like different classes or sections.

Example:

You can use a boxplot to compare exam scores from 3 different classes and see which class did better overall.

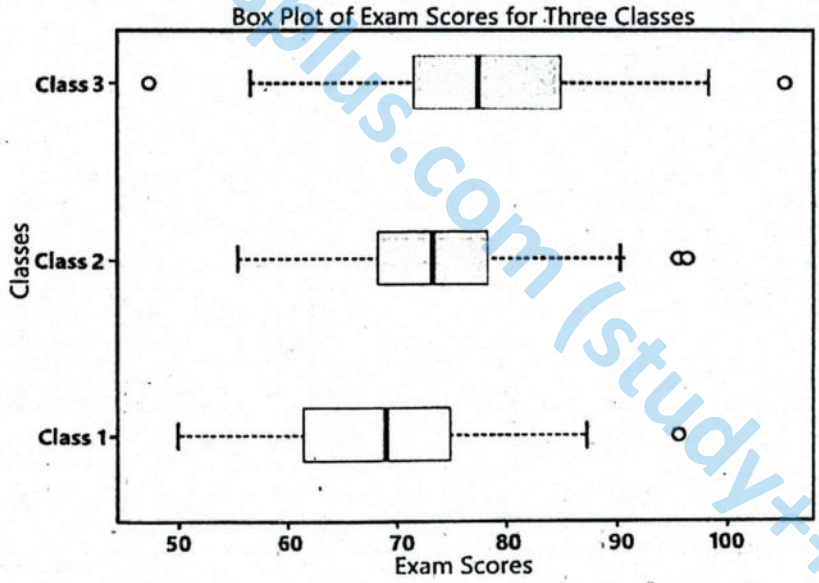


Figure: A Boxplot, showing class scores performance of three classes

Q.12. Explain the tools used for data visualization and describe the steps to create and interpret visualizations using Excel or Google Sheets.

Tools for Data Visualization:

Data visualization helps us understand large amounts of information by turning numbers into easy-to-understand charts and graphs. Many tools are available to create these visualizations, but some of the simplest and most common tools are **Microsoft Excel** and **Google Sheets**. These tools are easy to use and widely available on most computers and devices.

Both Excel and Google Sheets allow you to enter data and quickly create different types of visualizations such as bar **charts, line graphs, and pie charts.**

For example, if you run a small business and want to track the number of products sold each month, you can enter the monthly sales data into these tools and generate a bar chart to easily see which month had the highest sales.

Steps to Create and Interpret Visualizations in Excel or Google Sheets:

1. **Enter Your Data:**
   Begin by typing your data into the spreadsheet. For example, list the months (January, February, etc.) in one column and the sales figures in the next column.

Month	Sales
January	120
February	150
March	100
April	200

2. **Select the Data:**
   Use your mouse to highlight the cells containing the data you want to visualize.
3. **Choose a Chart Type**
   - Click on the **Insert** tab.
   - Select the chart type you want (e.g., **Bar Chart**, **Line Graph**, or **Pie Chart**).
4. **Customize the Chart**
   - Add axis labels: e.g., label the **x-axis** with months and the **y-axis** with sales figures.
   - Add a chart title, adjust colors, or insert data labels to make your chart more informative.
5. **Understanding Statistical Representations**
   - When you create a chart, make sure you understand:
   - What each axis represents.
   - The trend shown (e.g., increasing, decreasing, fluctuating).
   - The highest and lowest values.
   - Patterns that can help you make business decisions.

For example, a bar chart may show which month had the highest or lowest sales, or a line graph can reveal whether sales are increasing or decreasing over time. Reading the chart correctly **helps** you make **better decisions** based on the data.

Conclusion:

Using tools like Excel and Google Sheets for data visualization makes it easy to organize and display information clearly. By following simple steps, anyone can create meaningful charts and graphs that help in understanding and analyzing data effectively.

Exercise

Multiple Choice Questions

1. **An example of a basic statistical model:**
   a) Linear Regression
   b) Neural Networks
   c) Decision Trees
   d) Support Vector Machines
2. **The activity involved in experimental design in data science:**
   a) Creating visualizations
   b) Collecting and analyzing data systematically
   c) Writing code for machine learning
   d) Building databases
3. ***A commonly used tool for creating data visualizations:***
   a) MS Excel
   b) Python (Matplotlib)
   c) Tableau
   d) All of the above

4. **The meaning of the slope in a linear regression model:**
   a) The intercept of the model
   b) The change in the dependent variable for a unit change in the independent variable
   c) The error term
   d) The mean of the data
5. **An example of a real-world application of statistical models:**
   a) Predicting house prices
   b) Creating social media posts
   c) Designing websites
   d) Writing essays
6. **Option not considered a benefit of data visualization:**
   a) Identifying trends and patterns
   b) Communicating insights effectively
   c) Making data more complex
   d) Summarizing large datasets
7. **A primary goal of K-Means Clustering:**
   a) To classify data into predefined categories
   b) To group data into clusters based on similarity
   c) To predict continuous outcomes
   d) To reduce the dimensionality of data
8. **The meaning of "K" in K-Means Clustering:**
   a) Number of features in the dataset
   b) Number of clusters to be formed
   c) Number of iterations required for convergence
   d) Number of data points in the dataset

Correct Answers

Q. No	Correct Option
1.	a) Linear Regression
2.	b) Collecting and analyzing data systematically
3.	d) All of the above
4.	b) The change in the dependent variable for a unit change in the independent variable
5.	a) Predicting house prices
6.	c) Making data more complex
7.	b) To group data into clusters based on similarity
8.	b) Number of clusters to be formed

Short Answer Questions

Q.1. What is the importance of building statistical models in real-world applications?

Ans.
- Statistical models help us **understand and analyze data** in a meaningful way.
- They are used to **identify patterns and make predictions** about future events.

These models are helpful in **real-life fields** like business, health, education, and economics.
- **Example:** A shopkeeper can use a model to **predict future income** based on past sales data.

Q.2. Name one basic statistical model used for predicting outcomes and explain its purpose.

Ans.
- One basic statistical model is **Linear Regression.**
- Its purpose is to **predict the value of one variable** (dependent) based on the value of another variable (independent) by fitting a straight line to the data.
- **Example:** It can be used to predict a student's marks based on the number of hours studied.

Q.3. List two types of data visualizations and describe when you would use each.

Ans.
- **Bar Chart:** *A bar chart is used to compare different categories.* ***For example****, it is useful when* comparing sales of various products in a shop.
- **Line Graph:** A line graph is used to show changes or trends over time. **For example**, it can be used to track temperature changes during a week.
- Both types help in understanding data in a simple and visual way.

Q.4. How does visualizing data help in understanding descriptive statistics?

Ans.

- Data visualization shows numbers in the form of **graphs and charts** like bar charts, histograms, and boxplots.
- It helps us easily understand **key statistics** such as mean, median, mode, and range.
- By looking at a chart, we can quickly see **patterns, trends, and differences** in the data.
- It makes the information **clear, simple, and easy to analyze** without going through complex numbers.

Long Question Answers

Q.1. Explain the role and importance of statistical models in solving real-world problems.

Ans. See answer of chapter question # 6

Q.2. Describe the steps involved in building a basic statistical model (e.g., linear regression). Include details on data collection, model training, and evaluation.

Ans. See answer of chapter question # 7

Q.3. Discuss the types of data visualizations and their uses.

Ans. See answer of chapter question # 11

Q.4. Explain data collection methods.

Ans. See answer of chapter question # 4

Q.5. Discuss the concept of measure of tendency with example.

Ans. See answer of chapter question # 2

SLO Based Questions

Q.1. What is meant by Measures of Central Tendency? Name them.

Ans. Measures of central tendency are values that represent the **center or average** of a data set. They help us understand the typical value in the data.

The three main types are:

1. **Mean** (average)
2. **Median** (middle value)
3. **Mode** (most frequent value)

Q.2. Define Mean. How is it calculated?

Ans. **Mean** is the **average** of all the values in a data set. It is found by **adding all the values** and then **dividing by the number of values.**

Formula:

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Total number of values}}$$

Q.3. What is Mode? Give an example.

Ans. **Mode** is the value that **occurs most often** in a data set.

Example:

Data: 4, 5, 6, 6, 7

Mode = 6 (because it appears twice)

Q.4. Define Median. How is it calculated?

Ans. **Median** is the **middle value** of the data when arranged in order.

- If the number of values is **odd**, the middle value is the median.
- If the number of values is **even**, the average of the two middle values is the median.

Q.5. What is meant by Dispersion? Why is it important?

Ans. **Dispersion** means the **spread of values** in a dataset. It shows how far the values are from the mean.
It is important because it helps us understand **how consistent or varied** the data is.

Q.6. Define Variance. What does it tell us?

Ans. **Variance** measures how much each value differs from the mean.
It tells us **how spread out** the values are. A **higher variance** means more spread; a **lower variance** means values are close to the mean.

Q.7. What is Standard Deviation? How is it related to Variance?

Ans. **Standard deviation** shows how much the values differ from the mean in the same unit.
It is the **square root of the variance.**
Formula:

$$\text{Standard Deviation } \sigma = \sqrt{Variance}$$

Q.8. Compare the standard deviation of two classes if: Class A = 3.55, Class B = 21.26

Ans. Class A has a **smaller standard deviation** (3.55), which means its values are **closer to the mean.**
Class B has a **larger standard deviation** (21.26), meaning its values are **more spread out.**

Q.9. What is Probability? Write its formula.

Ans. **Probability** is the **chance or likelihood** that a certain event will happen.
Formula:

$$\text{Probability} = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

Q.10. Give one real-life example of probability.

Ans. **Example:** When tossing a coin, the probability of getting **heads** is:

$$\text{Probability of Head} = \frac{1}{2} = 0.5 = 50\%$$

Q.11. Write two uses of probability in real life.

Ans.
1. **Weather forecasting** (e.g., 70% chance of rain)
2. **Business decisions** (e.g., success chances of a product)

Q.12. What is the difference between Variance and Standard Deviation?

Ans.

Aspect	Variance	Standard Deviation
Definition	Measures the average squared distance from the mean	Measures the average distance from the mean
Formula	$\text{Variance } (\sigma^2) = \frac{Sum\ of\ (x_i - \mu)^2}{N}$	$\text{Standard Deviation } \sigma = \sqrt{Variance}$
Units	Square of the original units (e.g., cm^2, $marks^2$)	Same as original units (e.g., cm, marks)
Interpretation	Harder to interpret directly	Easier to understand and compare
Relationship	Square of the standard deviation	Square root of the variance

Q.13. What is data collection? Name three common methods.

Ans. Data collection is the process of gathering information relevant to a study or research. Common methods include:

- **Surveys** – Asking structured questions.
- **Observations** – Watching subjects without interference.
- **Experiments** – Testing variables under controlled conditions.

Q.14. Explain surveys as a method of data collection with example.

Ans. Surveys involve asking a fixed set of questions to collect structured data from people.
Example: A grocery store in Islamabad distributes a 5-question survey to 50 customers to learn their product preferences and uses the feedback to improve stock.

Q.15. What is observation in data collection? Give an example.

Ans. Observation means watching or monitoring people or events in their natural setting without interference.
Example: A restaurant notes which tables are used most at lunch to improve customer seating arrangements

Q.16. Define experiments as a data collection method with one example.

Ans. Experiments involve changing one or more variables to study their effect on another, under controlled conditions.
Example: A teacher gives printed notes to one group of students and compares their test results with a group that received only lectures.

Q.17. What is data preparation and why is it necessary?

Ans. Data preparation means cleaning, organizing, and converting collected data into a usable format. It ensures accuracy and improves the quality of analysis and results.

Q.18. What is data cleaning? Support your answer with an example.

Ans. Data cleaning removes errors such as spelling mistakes, missing values, or duplicates.
Example: "Alie" is corrected to "Ali" and Sara's missing score is filled with the class average (87) to complete the dataset.

Q.19. What is data transformation? Give a suitable example.

Ans. Data transformation changes clean data into a better format for analysis, like creating new columns or summaries.
Example: Grouping student grades by class or section helps in easier comparison and reporting.

Q.20. Write any three techniques to handle missing data.

Ans.
- **Imputation:** Estimate missing values using averages or patterns.
- **Flagging:** Leave the value empty but mark it as missing.
- **Removal:** Delete missing records if they are few and not critical to analysis.

Q.21. Why is it important to handle missing data before analysis?

Ans. Missing data can lead to inaccurate or incomplete analysis. Handling it ensures that the results are reliable, valid, and reflective of the true situation.

Q.22. Differentiate between data cleaning and data transformation.

Ans.

Aspect	Data Cleaning	Data Transformation
Definition	Process of correcting or removing errors in data	Process of converting clean data into a usable format
Purpose	To ensure accuracy and remove issues like duplicates or typos	To prepare data for analysis by changing its structure or format
Example	Fixing name "Alie" to "Ali", filling missing score	Grouping student data by class average or section

Q.23. What is a statistical model, and why is it important in real life?

Ans. A statistical model is a mathematical tool used to understand the relationship between different variables in data. It helps us analyze patterns, make predictions, and support decision-making.
Importance:
- Helps in predicting future outcomes.
- Useful in fields like business, healthcare, and education.
- **Example:** A shopkeeper can use a statistical model to estimate future sales based on past customer data.

Q.24. What is logistic regression and where is it used?

Ans. Logistic regression is a statistical method used to predict outcomes that are categorical, such as "yes" or "no".
Example: Predicting whether a student will pass or fail an exam based on hours studied.
It gives the probability of an event occurring between 0 and 1, instead of predicting exact values.

Q.25. What is clustering in data science? Give one example.

Ans. Clustering is a technique used to group similar items together based on common features.
Purpose: It helps in understanding patterns and organizing data into meaningful categories.
Example: Grouping students based on their performance in Math and English using their scores.

Q.26. What is K-means clustering? How does it work?

Ans. K-means is a popular clustering method where the user selects the number of clusters (groups), and the algorithm assigns data points based on similarity.

Example: Students with high math scores go into one cluster, and those with high English scores into another.

The model uses distances between scores to group similar students together.

Q.27. What are performance metrics in model evaluation? Give two types.

Ans. Performance metrics help measure how well a statistical model works.

Types:

- **Error Metrics** – Show how much the predicted value differs from the actual value.
- **Accuracy Metrics** – Show how often the model's predictions are correct.

These metrics help in improving model quality.

Q.28. What does interpreting a model mean? Why is it useful?

Ans. Interpreting a model means understanding the results and drawing useful conclusions.

Purpose: It helps in making better decisions based on the model's output.

Example: If a model shows that more study hours improve grades, teachers can advise students accordingly.

Q.29. What are ethical considerations in statistical modeling?

Ans. Ethical considerations are rules to ensure fairness, honesty, and privacy in model building.

Key Points:

- **Fairness:** Models should not favor or harm any group unfairly.
- **Data Privacy:** Personal data must be kept secure and not misused.

This ensures trust and responsible use of data

Q.30. How can you improve a statistical model?

Ans. To improve a model:

- Use more data points for better accuracy.
- Include important factors like family size or seasonal trends.
- Regularly update the model with new data.
- Compare predictions with actual outcomes to refine it

Q.31. What is linear regression and how is it used?

Ans. Linear regression is a statistical method used to study the relationship between two numerical variables.

It fits a straight line to data to help predict the dependent variable (Y) using the independent variable (X).

Example: Predicting daily income based on the number of customers.

Q.32. What is data visualization and why is it important?

Ans. Data visualization is the process of showing data in the form of charts, graphs, or pictures.

It helps us understand large amounts of information quickly by making it easy to see patterns, trends, and comparisons.

This is important because it makes complex data simple and helps in better decision-making.

Q.33. Describe any three types of data visualizations and their uses.

Ans.

- **Bar Chart:** Used to compare different categories. For example, it can show sales of different products.
- **Line Graph:** Shows changes over time, like temperature changes during a week.
- **Histogram:** Shows how data is spread in different ranges, like exam scores of students.

Each type helps in understanding data in a specific way.

Q.34. What is a scatterplot and how is it useful?

Ans. A scatterplot is a graph that shows the relationship between two variables by plotting points on a chart. Each point shows a pair of values. It is useful to see if two things are related, for example, the number of hours studied and the exam marks of students.

Q.35. How can Microsoft Excel and Google Sheets be used for data visualization?

Ans. Excel and Google Sheets are tools that help create charts easily. You enter data like months and sales figures, select the data, choose a chart type (bar chart, line graph), and add labels to make the chart clear. These tools help turn numbers into simple visual forms for easy understanding.

Q.36. Write the steps to create a chart in Excel or Google Sheets.

Ans. Enter the data in columns (for example, months and sales).
Select the data using the mouse.
Click on the "Insert" tab and choose the chart type (bar chart, line graph).
Add labels to the chart for clarity, such as months on the x-axis and sales on the y-axis.
These steps help create clear and easy-to-understand visualizations.

Important MCQs

Basic Statistical Concepts

1. Which of the following is NOT a measure of central tendency?
A) Mean B) Median
C) Mode D) Variance

2. The mean of the data set 50, 60, 70, 80, 90 is:
A) 60 B) 70
C) 80 D) 90

3. If the scores are 50, 60, 70, 70, 90, the mode is:
A) 50 B) 60
C) 70 D) 90

4. If the scores are 50, 60, 70, 70, 60, 90, the modes are:
A) 50 and 60 B) 60 and 70
C) 70 and 90 D) 60 and 90

5. The median of 50, 60, 70, 80 is:
A) 60 B) 65
C) 70 D) 75

6. Measures of dispersion help us understand:
A) The most common value
B) The middle value
C) How spread out the data is
D) The average value

7. Variance is calculated as:
A) The sum of all values
B) The average of squared deviations from the mean
C) The square root of the mean
D) The difference between max and min

8. A higher variance means:
A) Data is more spread out
B) Data is closer to the mean
C) Data is always positive
D) Data is always negative

9. Standard deviation is:
A) The square of the variance
B) The square root of the variance
C) The mean of the data
D) The mode of the data

10. If Class A has a standard deviation of 3.55 and Class B has 21.26, which class has more spread out scores?
A) Class A B) Class B
C) Both are equal D) Cannot be determined

Introduction to Probability

11. Probability is the study of:
A) Data collection
B) How likely an event is to happen
C) Data cleaning
D) Data transformation

12. The probability of getting heads in a fair coin flip is:
A) 0 B) 1
C) 0.5 D) 2

13. Probability can be used in:
A) Weather prediction B) Business decisions
C) Sports D) All of the above

14. Probability is calculated as:
A) Total number of outcomes / Number of favorable outcomes
B) Number of favorable outcomes / Total number of outcomes
C) Mean / Median
D) Variance / Standard deviation

Data Collection and Preparation

15. Which of the following is NOT a data collection method?
A) Surveys B) Observations
C) Experiments D) Data cleaning

16. Surveys are best for:
A) Collecting large amounts of structured data
B) Observing natural behavior
C) Testing cause and effect
D) Data transformation

17. Observation as a data collection method involves:
A) Manipulating variables
B) Watching subjects in their natural environment
C) Sending online forms
D) Removing errors

18. Experiments are useful for:
A) Collecting opinions
B) Observing natural behavior
C) Determining cause and effect
D) Cleaning data

19. Data preparation includes:
A) Cleaning and organizing data
B) Collecting data
C) Making predictions
D) Calculating probability

20. If survey responses are incomplete, you should:
A) Ignore the data
B) Estimate missing values or adjust statistically
C) Only use complete responses
D) Discard the survey

Data Cleaning and Transformation

21. Data cleaning involves:
A) Adding more data
B) Correcting errors and removing duplicates
C) Making predictions
D) Calculating mean

22. An example of data cleaning is:
A) Correcting misspelled names
B) Collecting survey responses
C) Observing customer behavior
D) Calculating variance

23. Data transformation is:
A) Changing data into a suitable format
B) Removing all data
C) Ignoring missing values
D) Printing data

24. Creating new columns in a dataset is part of:
A) Data collection B) Data transformation
C) Data cleaning D) Probability calculation

Handling Missing Data

25. Imputation means:
A) Removing all missing data
B) Estimating missing values using existing data
C) Ignoring missing data
D) Duplicating data

26. Flagging missing data means:
A) Estimating missing values
B) Marking missing data for transparency
C) Removing all data
D) Printing missing values

27. If only a few records have missing data, you can:
A) Remove those records
B) Remove all data
C) Ignore the problem
D) Duplicate the records

Building Statistical Models

28. Statistical modeling is used to:
A) Make sense of data and predict outcomes
B) Clean data
C) Transform data
D) Collect data

29. The first step in building a statistical model is:
A) Collecting data
B) Defining the problem
C) Evaluating the model
D) Choosing an algorithm

30. Which is an example of a statistical model?
A) Linear regression B) Data cleaning
C) Data transformation D) Data collection

31. Training a model means:
A) Teaching the model using data
B) Removing errors
C) Collecting new data
D) Calculating mean

32. Evaluating a model involves:
A) Testing the model with new data
B) Collecting more data
C) Cleaning the data
D) Calculating variance

33. Which of the following is NOT a step in model development?
A) Define the problem B) Collect data
C) Ignore errors D) Evaluate the model

34. A real-world application of statistical models is:
A) Predicting grocery expenses
B) Making a phone call
C) Writing an essay
D) Drawing a picture

35. Which of the following is a benefit of using statistics in daily life?
A) Understanding patterns in data
B) Making random guesses
C) Ignoring data
D) Increasing errors

ANSWER KEY

1.	D	2.	B	3.	C	4.	B	5.	B	6.	C	7.	B	8.	A	9.	B	10.	B
11.	B	12.	C	13.	D	14.	B	15.	D	16.	A	17.	B	18.	C	19.	A	20.	B
21.	B	22.	A	23.	A	24.	B	25.	B	26.	B	27.	A	28.	B	29.	B	30.	A
31.	A	32.	A	33.	C	34.	A	35.	A										