CBSE Class 10 Data Science Question Paper with Solutions(Memory Based)

Time Allowed :3 Hours	**Maximum Marks :**70	**Total questions :**37

General Instructions

Read the following instructions very carefully and strictly follow them:

1. Please check that this question paper contains 23 printed pages.
2. Q.P. Code given on the right hand side of the question paper should be written on the title page of the answer-book by the candidate.
3. Please check that this question paper contains 37 questions.
4. 15 minute time has been allotted to read this question paper. The question paper will be distributed at 10.15 a.m. From 10.15 a.m. to 10.30 a.m., the candidates will read the question paper only and will not write any answer on the answer-book during this period.

1. Which Python library is primarily used for Data Manipulation and Data Frames?

Solution:

Concept: Python provides several libraries for data analysis and manipulation. One of the most widely used libraries for handling structured data is Pandas.

Answer: The Python library primarily used for **data manipulation and working with data frames** is **Pandas**.

Explanation:

- **Pandas** is an open-source Python library designed for data analysis and data handling.
- It provides powerful data structures such as:
  - **Series:** One-dimensional labeled array.

  - **DataFrame:** Two-dimensional table-like structure (rows and columns).

- It is widely used in:
  - Data cleaning and preprocessing
  - Data analysis
  - Handling missing data
  - Reading/writing data (CSV, Excel, SQL)

Example:

```
import pandas as pd

data = {"Name": ["Aman", "Riya"], "Marks": [85, 90]}
df = pd.DataFrame(data)
print(df)
```

Conclusion: Pandas is the most commonly used Python library for data manipulation and working with data frames in data science and analytics.

> **Quick Tip**
>
> **Remember:** DataFrames in Python → Use **Pandas**.

---

2. Identify the process of dividing a dataset into 100 equal parts.

Solution:

Concept: In statistics, datasets are often divided into equal parts to understand distribution and relative position of data values. These divisions are known as quantiles.

Answer: The process of dividing a dataset into **100 equal parts** is called **Percentiles**.

Explanation:

- **Percentiles** divide data into 100 equal parts.

- Each percentile represents the percentage of data below a specific value.
- For example:
  - 25th percentile → 25% of data lies below it.
  - 50th percentile → Median of the dataset.
  - 90th percentile → 90% of data lies below it.

Related Terms:

- **Quartiles:** Divide data into 4 equal parts.
- **Deciles:** Divide data into 10 equal parts.
- **Percentiles:** Divide data into 100 equal parts.

Conclusion: When a dataset is divided into 100 equal parts to analyze distribution, the process is called calculating percentiles.

Quick Tip

Remember: 4 parts → Quartiles 10 parts → Deciles 100 parts → Percentiles

---

3. Name the type of join used to combine rows from two tables based on a related column between them.

Solution:

Concept: In relational databases, joins are used to combine data from multiple tables using a common column. Different types of joins serve different purposes depending on how the data should be merged.

Answer: The join used to combine rows from two tables based on a related column between them is called an **INNER JOIN**.

Explanation:

- An **INNER JOIN** returns only the rows that have matching values in both tables.

- It uses a common column (such as an ID or key) to relate the tables.
- Rows without matching values are excluded from the result.

Example:

```
SELECT *
FROM Students
INNER JOIN Marks
ON Students.ID = Marks.ID;
```

Result: Only students who have matching marks records will appear in the output.

Other Types of Joins (for reference):

- **LEFT JOIN:** All records from left table + matching from right.
- **RIGHT JOIN:** All records from right table + matching from left.
- **FULL JOIN:** All records from both tables.

Conclusion: An INNER JOIN is the standard method used to combine rows from two tables using a related column where matching values exist in both tables.

Quick Tip

Remember: Matching rows from both tables → INNER JOIN.

---

4. In a dataset, if the mean is significantly higher than the median, what does it indicate about the distribution?

Solution:

Concept: The relationship between mean and median helps identify the **skewness** of a data distribution. Skewness indicates whether data is stretched more toward the left or right side.

Answer: If the **mean is significantly higher than the median**, the distribution is **positively skewed (right-skewed)**.

Explanation:

- In a **positively skewed distribution**, a few very large values pull the mean toward the right.
- The median remains less affected because it depends on the middle value.
- Therefore:

$$\text{Mean} > \text{Median} > \text{Mode} \quad \text{(usually)}$$

Characteristics of Positively Skewed Distribution:

- Long tail on the right side
- Presence of high-value outliers
- Common in income distributions, exam scores with few toppers, etc.

Comparison for Clarity:

- Mean $>$ Median $\rightarrow$ Positively skewed
- Mean $<$ Median $\rightarrow$ Negatively skewed
- Mean $\approx$ Median $\rightarrow$ Symmetrical distribution

Conclusion: When the mean is much higher than the median, it indicates a positively skewed distribution caused by extreme high values pulling the average upward.

Quick Tip

Remember: Mean pulled right $\rightarrow$ Positive skew. Mean pulled left $\rightarrow$ Negative skew.

---

5. Define Encryption in the context of data security.

Solution:

Concept: In data security, protecting sensitive information from unauthorized access is essential. Encryption is a key technique used to ensure data privacy and confidentiality.

Definition: **Encryption** is the process of converting **readable data (plaintext)** into an **encoded form (ciphertext)** so that only authorized users can access or understand it.

Explanation:

- Encryption uses mathematical algorithms and keys to transform data.
- The encrypted data appears meaningless to anyone without the correct key.
- To read the original data, it must be converted back using **decryption**.

Where Encryption is Used:

- Online banking and digital payments
- Secure messaging apps
- Password protection
- Data storage and cloud services
- Secure websites (HTTPS)

Types of Encryption:

- **Symmetric Encryption:** Same key for encryption and decryption.
- **Asymmetric Encryption:** Uses public and private keys.

Conclusion: Encryption is a fundamental data security technique that protects sensitive information by converting it into an unreadable format, ensuring that only authorized users can access it.

Quick Tip

Remember: Encryption = Plaintext $\rightarrow$ Ciphertext (for security).

---

6. State whether Selection Bias occurs during data collection or during model deployment.

Solution:

Concept: Bias in data science refers to systematic errors that affect the fairness or accuracy of analysis and models. Selection bias is a common type of bias related to how data is gathered.

Answer: Selection Bias occurs during data collection.

Explanation:

- Selection bias arises when the data collected is **not representative of the entire population**.
- This happens due to improper sampling methods or limited data sources.
- As a result, the model trained on such data may produce biased or inaccurate results.

Example: If a survey about smartphone usage is conducted only among college students, the data will not represent older age groups. This leads to selection bias during the **data collection stage**.

Why Not During Deployment? While bias effects may appear during model deployment, the **origin of selection bias is in the sampling or data collection phase**.

Conclusion: Selection bias occurs during the data collection process when the sample chosen does not accurately represent the target population.

Quick Tip

Remember: Selection Bias = Problem in sampling/data collection.

---

7. Differentiate between Structured and Unstructured data with one example each.

Solution:

Concept: Data can be classified based on how it is organized and stored. The two main types are structured and unstructured data.

Structured Data: Structured data is data that is **organized in a fixed format** and stored in rows and columns. It is easy to store, search, and analyze using databases and spreadsheets.

Features:

- Well-defined structure
- Stored in tables (rows and columns)
- Easy to analyze using SQL or Excel

Example: Student records stored in a table with columns like Name, Roll No., and Marks.

Unstructured Data: Unstructured data is data that **does not follow a fixed format** or predefined structure. It is more complex and harder to analyze directly.

Features:

- No fixed format
- Harder to store and process
- Requires advanced tools like AI or NLP for analysis

Example: Photos, videos, social media posts, or emails.

Difference Between Structured and Unstructured Data:

Feature	Structured Data	Unstructured Data
Format	Organized	Not organized
Storage	Tables, databases	Media files, text, etc.
Analysis	Easy	Complex
Example	Student database	Images or videos

Conclusion: Structured data is organized and easy to analyze, while unstructured data lacks a fixed format and requires advanced methods for processing.

Quick Tip

Remember: Tables = Structured data. Images/Videos = Unstructured data.

8. Why is the Median considered a more robust measure of central tendency than the Mean when outliers are present?

Solution:

Concept: Measures of central tendency describe the center of a dataset. However, their reliability depends on how sensitive they are to extreme values (outliers).

Answer: The **median** is considered more robust than the mean because it is **less affected by outliers or extreme values**.

Explanation:

- The **mean** is calculated by adding all values and dividing by the total number of observations. A single very large or very small value can significantly shift the mean.
- The **median** is the middle value when data is arranged in order. It depends only on the position of values, not their magnitude.
- Therefore, extreme values do not strongly influence the median.

Example: Consider the dataset:

$$2,\ 3,\ 4,\ 5,\ 100$$

- Mean $= \frac{2+3+4+5+100}{5} = 22.8$ (heavily affected by 100)
- Median $= 4$ (represents the central value better)

Key Insight:

- Mean $\rightarrow$ Sensitive to outliers
- Median $\rightarrow$ Resistant to outliers

Conclusion: Because the median depends on the order of data rather than extreme values, it provides a more reliable measure of central tendency in datasets containing outliers.

Quick Tip

Remember: Outliers present $\rightarrow$ Use Median, not Mean.

9. Explain the concept of the Central Limit Theorem and its significance in data analysis.

Solution:

Concept: The Central Limit Theorem (CLT) is one of the most important results in statistics. It explains how the distribution of sample means behaves when multiple samples are drawn from a population.

Definition: The **Central Limit Theorem** states that **the distribution of the sample mean approaches a normal (bell-shaped) distribution as the sample size increases**, regardless of the original population distribution (provided the sample size is sufficiently large, usually $n \geq 30$).

Key Points:

- Applies to sample means, not individual data points.
- Works even if the original data is skewed or non-normal.
- Larger sample sizes produce distributions closer to normal.
- The mean of the sampling distribution equals the population mean:

$$\mu_{\bar{x}} = \mu$$

- The standard deviation of the sampling distribution (standard error) is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Intuition: If we repeatedly take samples from any population and calculate their averages, the pattern of those averages will form a normal distribution, even if the original data is not normally distributed.

Significance in Data Analysis:

- **Foundation of Inferential Statistics:** Enables estimation of population parameters from samples.

- **Confidence Intervals:** Used to calculate reliable intervals for population means.
- **Hypothesis Testing:** Allows use of normal-based tests like Z-tests and t-tests.
- **Real-world Applications:** Used in quality control, surveys, finance, and machine learning.
- **Simplifies Analysis:** Makes analysis easier by allowing normal approximation.

Example: Even if individual incomes in a city are highly skewed, the average income calculated from many random samples will follow an approximately normal distribution.

Conclusion: The Central Limit Theorem explains why normal distributions appear frequently in statistics and enables reliable analysis of population characteristics using sample data, making it fundamental to modern data analysis.

Quick Tip

Remember: Large samples $\rightarrow$ Sample means become normally distributed (CLT).

---

10. What is Data Merging? Mention one scenario where it is required.

Solution:

Concept: In data analysis, information is often stored in multiple datasets or tables. To perform meaningful analysis, these datasets may need to be combined.

Definition: Data Merging is the process of **combining two or more datasets into a single dataset** based on a common key or related field.

Explanation:

- It is commonly used in databases, spreadsheets, and data analysis tools.
- Merging helps integrate related information from different sources.
- It is often performed using common identifiers such as ID, name, or date.

Scenario Where Data Merging is Required: Suppose a school has:

- One table containing **student details** (Name, Roll No.)
- Another table containing **marks** (Roll No., Marks)

To generate a report showing both names and marks together, the two tables must be **merged using Roll No. as the common key**.

Other Common Use Cases:

- Combining sales and customer data
- Joining survey results with demographic data
- Integrating data from multiple departments

Conclusion: Data merging is an essential data preparation step that combines related datasets into a unified view, enabling more comprehensive analysis and reporting.

Quick Tip

Remember: Data merging = Combining datasets using a common key.

---

11. Explain the four steps of the Statistical Problem-Solving Process in detail.

Solution:

Concept: The statistical problem-solving process is a systematic approach used to collect, analyze, and interpret data in order to make informed decisions. It helps transform real-world problems into data-driven solutions.

The Four Steps of the Statistical Problem-Solving Process:

1. Define the Problem (Formulate the Question): This is the first and most important step, where the objective of the study is clearly defined.

Key Activities:

- Identify the problem or research question.
- Determine what information is needed.

• Define variables and target population.

Example: A company wants to know whether customer satisfaction has improved after introducing a new product.

2. Collect the Data: Once the problem is defined, relevant data must be gathered using appropriate methods.

Key Activities:

• Choose data collection methods (surveys, experiments, observations).

• Decide between primary and secondary data.

• Ensure data accuracy and reliability.

Example: Conducting an online survey to collect customer feedback.

3. Analyze the Data: In this step, collected data is organized and examined to identify patterns and insights.

Key Activities:

• Cleaning and preparing data.

• Using statistical tools (mean, median, graphs, charts).

• Applying statistical models if required.

Example: Calculating average satisfaction scores and plotting trends over time.

4. Interpret and Communicate Results: The final step involves drawing conclusions and presenting findings clearly.

Key Activities:

• Interpret statistical results in context.

• Make data-driven decisions.

• Present findings using reports, charts, or presentations.

Example: Reporting that satisfaction increased by 15% and recommending expansion of the product line.

Importance of This Process:

- Ensures structured decision-making
- Reduces bias and errors
- Supports evidence-based conclusions
- Widely used in business, healthcare, research, and data science

Conclusion: The statistical problem-solving process involves defining the problem, collecting relevant data, analyzing it systematically, and interpreting the results to make informed decisions. This structured approach ensures accurate and meaningful data analysis.

Quick Tip

Remember: Define → Collect → Analyze → Interpret (4-step statistical cycle)

---

12. Define and explain the following types of Biases: Recall Bias, Survivor Bias

Solution:

Concept: Bias refers to systematic errors in data collection or analysis that lead to incorrect conclusions. Understanding different types of bias is important for accurate data interpretation.

1. Recall Bias:

Definition: Recall Bias occurs when participants **do not remember past events accurately**, leading to incorrect or incomplete data.

Explanation:

- Common in surveys, interviews, and retrospective studies.
- People may forget details or unintentionally distort memories.
- This results in unreliable self-reported data.

Example: In a health survey, participants may not accurately remember how often they exercised or what they ate last year, leading to inaccurate data.

2. Survivor Bias:

Definition: Survivor Bias occurs when analysis focuses only on **successful or surviving cases** while ignoring those that failed or were excluded.

Explanation:

- Leads to overly optimistic or misleading conclusions.
- Happens when incomplete data is analyzed.
- Important failures or missing cases are overlooked.

Example: Studying only successful startups to identify success factors while ignoring failed startups leads to survivor bias.

Key Difference:

Feature	Recall Bias	Survivor Bias
Cause	Memory errors	Ignoring failed cases
Occurs In	Surveys/interviews	Data analysis/selection
Effect	Inaccurate reporting	Overly positive conclusions

Conclusion: Recall bias arises from inaccurate memory during data collection, while survivor bias results from analyzing only successful outcomes and ignoring failures, both of which can distort statistical conclusions.

Quick Tip

Remember: Recall Bias → Memory errors. Survivor Bias → Ignoring failures.

---

13. Compare CSV, Spreadsheets, and SQL as data storage formats. Which one is best for handling massive, interrelated datasets?

Solution:

Concept: Different data storage formats are used depending on the size, complexity, and structure of the data. CSV files, spreadsheets, and SQL databases each serve different purposes.

1. CSV (Comma-Separated Values):

Description: A CSV file is a simple text-based format where data is stored in rows and columns separated by commas.

Features:

- Lightweight and easy to create
- No built-in data relationships
- Limited scalability
- Easily readable by many tools

Use Case: Small datasets, data exchange between systems.

2. Spreadsheets (Excel, Google Sheets):

Description: Spreadsheets store data in tabular format with built-in tools for calculations and visualization.

Features:

- User-friendly interface
- Supports formulas, charts, and formatting
- Limited handling of very large datasets
- Weak support for relationships between tables

Use Case: Business reports, small to medium data analysis.

3. SQL Databases:

Description: SQL databases (e.g., MySQL, PostgreSQL) store data in structured tables with defined relationships using keys.

Features:

- Handles very large datasets efficiently
- Supports relationships using primary and foreign keys
- Enables complex queries and joins
- High scalability and reliability

Use Case: Enterprise systems, banking, e-commerce platforms.

Comparison Table:

Feature	CSV	Spreadsheets	SQL
Structure	Simple text	Tabular with UI	Relational database
Scalability	Low	Medium	Very High
Data Relationships	None	Limited	Strong (joins)
Ease of Use	Easy	Very easy	Requires SQL knowledge
Best For	Small data	Medium data	Large, complex data

Answer to the Question: The best option for handling **massive, interrelated datasets** is **SQL databases**, because they support relationships, scalability, and complex querying.

Conclusion: While CSV and spreadsheets are useful for small to moderate datasets, SQL databases are the most suitable choice for managing large-scale, interconnected data efficiently.

Quick Tip

Remember: Small data → CSV Medium data → Spreadsheets Massive relational data → SQL
