FREE EDITION · NOTES + 3 SAMPLE MCQS

CUET · COMPUTER SCIENCE · CLASS XII · CODE 308

Understanding Data

CUET unit: Understanding Data

By **UniDrill** · NCERT-grounded study material

WWW.UNIDRILL.IN

📌 Snapshot

- Data is the core concept — what it is, why it matters, and how it is collected, stored, and processed for decision-making.
- Data is classified into structured and unstructured types, with metadata as a descriptor for unstructured data — a distinction that appears regularly in CUET objective questions.
- The data processing cycle (Input → Processing → Output) and its real-world applications (ATM, admit card generation, train ticketing) are central models tested in assertion-reason and case-based items.
- Five key statistical techniques — Mean, Median, Mode, Range, and Standard Deviation — are explained with worked examples; CUET frequently tests formula application and selection of the appropriate technique for a given scenario.
- Python is a programming tool for data analysis with dedicated libraries, linking this topic to later Python data-processing topics.

📖 Detailed Notes

2.1 Core concepts

- **What is data?** Data is a collection of characters, numbers, and other symbols that represent values of some situations or variables. "Data" is plural; the singular form is "datum". Data stored electronically can be processed faster and more easily than manual processing. (NCERT §7.1, p. 98)
- **Importance of data:** Humans rely on data for decision-making. Large amounts of data, when processed by a computer, reveal hidden patterns not otherwise visible. Examples include ATM transactions (bank debiting the withdrawn amount), meteorological offices monitoring satellite data for cyclones, businesses using dynamic pricing based on demand-supply data, and electronic voting machines accumulating votes for quick result declaration. (NCERT §7.1.1, p. 98–99)
- **Types of data — Structured:** Data organised in a well-defined format, stored in a tabular (rows and columns) structure where each column is an attribute/characteristic/variable and each row is one observation. Examples: inventory tables, fee records, ATM withdrawal records. (NCERT §7.1.2(A), p. 99–100)

- **Types of data — Unstructured:** Data that is not in a traditional row-and-column format. Examples include newspaper pages, emails, web pages with multimedia content, text documents, business reports, audio/video files, social media messages. Unstructured data is sometimes described using metadata. (NCERT §7.1.2(B), p. 100–101)
- **Metadata:** Data about data. For an email, metadata includes subject, recipient, main body, attachment details. For an image file, metadata includes image size (KB/MB), image type (JPEG, PNG), and image resolution. (NCERT §7.1.2(B), p. 101)
- **Data collection:** The process of identifying already available data or collecting it from appropriate sources. Data may exist as physical diary/register entries (requiring digitisation into a spreadsheet), already in digital format (e.g., CSV file), or not yet recorded (requiring software development using Python with MySQL or CSV storage). (NCERT §7.2, p. 101)
- **Data storage:** The process of storing data on storage devices so it can be retrieved later. Common digital storage devices include Hard Disk Drive (HDD), Solid State Drive (SSD), CD/DVD, Tape Drive, Pen Drive, and Memory Card. File processing has limitations that can be overcome through Database Management System (DBMS). (NCERT §7.3, p. 102)
- **Data processing:** Converting raw data (numbers/text/image) into information (tables/charts/text). The data process cycle has three stages: Input (data collection, data preparation, data entry) → Processing (store, retrieve, classify, update) → Output (reports, results, processing system). Figure 7.1 and Figure 7.2 illustrate these steps with real-world examples (competitive exam admit card, bank ATM withdrawals, train ticket issuance). (NCERT §7.4, p. 102–103)
- **Statistical techniques — overview:** Summarisation methods applied on tabular data for easy comprehension. The five techniques covered are Mean, Median, Mode (measures of central tendency) and Range, Standard Deviation (measures of variability/dispersion). (NCERT §7.5, p. 103–104)
- **Mean:** The average of numeric values of an attribute. Given n values $x_1, x_2, \cdots, x_?$, mean = (sum of all values) / n. Mean is not suitable when outliers are present; outliers should be removed before calculating mean. Example: heights [90,102,110,115,85,90,100,110,110] → mean = 912/9 = 101.33 cm. (NCERT §7.5.1(A), p. 104)
- **Median:** The middle value when all values are sorted in ascending or descending order. For odd number of values, median is the value at the middle position. For even number of values, median is the average of the two middle values. Median represents the central value at which data is equally divided into two halves. Example: sorted heights [85,90,90,100,102,110,110,110,115] → median = 102 cm (position 5). (NCERT §7.5.1(B), p. 104–105)
- **Mode:** The value that appears most frequently in the data. Computed on the basis of frequency of occurrence. A dataset has no mode if each value occurs only once;

it may have multiple modes. Mode can be found for numeric as well as non-numeric data. Example: in height data, mode = 110 (frequency 3). (NCERT §7.5.1(C), p. 105)

- **Range:** The difference between the maximum and minimum values in the data (M – S). Range can be calculated only for numerical data and indicates coverage/spread of data values. However, any outlier in the data badly influences the result since range is based on only the two extreme values. Example: max height 115 cm, min height 85 cm → range = 30 cm. (NCERT §7.5.2(A), p. 105–106)
- **Standard deviation:** Refers to differences within a group/set of data. Unlike Range, it considers all data values. Calculated as the positive square root of the average of squared differences of each value from the mean: $\sigma = \sqrt{[\Sigma (x_i - \bar{x})^2 / n]}$. Smaller σ means data are less spread; larger σ means more spread. Example: for the nine student heights, $\sigma = \sqrt{(938/9)} = \sqrt{104.22} \approx 10.2$ cm. (NCERT §7.5.2(B), p. 106)
- **Python for data analysis:** Python has libraries specially built for data processing and analysis, making it a preferred programming tool for efficient analysis of large data volumes. These will be covered in following chapters. (NCERT §7.5, p. 107)
- **Why summarisation matters (NCERT §7.5, p. 103-104).** Raw tabular data — even small datasets — quickly becomes hard to interpret. A single number (mean, median, mode) or a measure of spread (range, σ) compresses many values into a comprehensible summary, enabling decisions without scanning every record.
- **Sensitivity of mean to outliers (NCERT §7.5.1(A), p. 104).** Median, by contrast, is order-based and ignores the magnitudes of extreme values — making it the robust choice for income, real-estate prices, exam scores with outliers, etc.
- **Why σ uses squared differences (NCERT §7.5.2(B), p. 106).** Squaring ensures that positive and negative deviations from the mean do not cancel out; taking the square root at the end restores the original measurement unit (e.g., cm, not cm²).
- **CSV vs Excel (NCERT §7.2, p. 101).** Both are tabular formats, but CSV is plain text (cross-platform, language-agnostic) while Excel files are proprietary binary formats. Pandas can read both with `read_csv()` and `read_excel()` respectively.

2.2 Definitions to memorise

Term	Definition	Page
Data	A collection of characters, numbers, and other symbols that represent values of some situations or variables	98
Datum	Singular form of data	98
Structured data	Data organised in a well-defined tabular format (rows and columns) where each column is an attribute and each row is an observation	99
Unstructured data	Data that is not in a traditional row-and-column structure (e.g., emails, images, videos, social media posts)	100–101
Metadata		101

Term	Definition	Page
	Data about data; describes parts or properties of unstructured data (e.g., image size, image type, email subject)	
Data collection	Identifying already available data or collecting from appropriate sources for processing	101
Data storage	The process of storing data on storage devices so it can be retrieved later	102
Data processing	Converting raw data into meaningful information through the input-processing-output cycle	102
Outlier	An exceptionally large or small value in comparison to other values; can influence/affect mean and other statistical calculations	104
Mean (Average)	Sum of all numeric values divided by the count of values	104
Median	The middle value of sorted data; equally divides the data into two halves	104–105
Mode	The value that appears most frequently in the data; can be applied to numeric and non-numeric data	105
Range	Difference between the maximum and minimum values (M – S); a measure of dispersion for numeric data only	105–106
Standard Deviation (σ)	Positive square root of the average of squared differences of each value from the mean; considers all data values	106
DBMS	Database Management System; overcomes the limitations of file processing for data storage and retrieval	102
CSV	Comma Separated Values; a common digital format for storing structured data	101
Attribute	A column in structured data	99
Observation	A row in structured data	99
Tape Drive	Magnetic-tape secondary storage device	102
Memory Card	Removable flash storage device	102
Pen Drive	USB flash storage device	102
Information	Output of processing raw data	102
Data preparation	Input-stage activity that cleans and formats data for entry	103
Data entry	Input-stage activity that records data into the system	103
Frequency	Number of times a value appears in a dataset	105
Variance	Average of squared deviations from the mean (σ^2 before square root)	106
		104

Term	Definition	Page
Measure of central tendency	Single representative value of a dataset (mean, median, or mode)	
Measure of dispersion	Quantifies spread of data around the centre (range, standard deviation)	105-106

2.3 Diagrams / processes to remember

- **Figure 7.1 — Steps in data processing (p. 103):** Two-tier diagram showing (i) RAW DATA → Data Processing → INFORMATION and (ii) the Data Process Cycle: Input (Data Collection, Data Preparation, Data Entry) → Processing (Store, Retrieve, Classify, Update) → Output (Reports, Results, Processing System). Students must be able to place activities in the correct stage.
- **Figure 7.2 — Data-based problem statements (p. 103):** Three real-world cases — competitive exam admit card generation, bank ATM cash withdrawal, and train ticket issuance — each broken into Problem Statement / Inputs / Processing / Output columns. This figure is a prime source for case-based MCQs; remember the four-quadrant layout.
- **Table 7.1 — Structured data about kitchen items in a shop (p. 100):** Example of a structured dataset with columns ModelNo, ProductName, Unit Price, Discount(%), Items_in_Inventory. Used to illustrate how spreadsheet operations (summing a column, multiplying columns) work on structured data.
- **Table 7.3 — Standard deviation calculation table (p. 106):** Step-by-step working showing Height (x), $x - \bar{x}$, and $(x - \bar{x})^2$ columns, with $n = 9$, $\bar{x} = 101.33$, $\Sigma (x-\bar{x})^2 = 938.00$, $\sigma = 10.2$ cm. Understand the tabular method for computing σ.

2.4 Common confusions / NTA trap points

- **Mean vs Median for outlier-affected data:** Mean is sensitive to outliers; Median is not. NTA often presents a dataset with one extreme value and asks which measure of central tendency is most appropriate — the answer is Median (or Mode, if the question is about the most frequent value). Mean should be used only after removing outliers.
- **Range uses only two values; Standard Deviation uses all values:** A common distractor states that range considers all data values. It does not — range = max – min. Standard deviation considers every data point.
- **Mode can be non-numeric; Mean and Range cannot:** Mode is the only measure that works for non-numeric (categorical) data. NTA trap: asking which technique to use for "most popular car colour" — answer is Mode, not Mean.
- **Structured vs Unstructured mix-ups:** Emails and newspaper layouts are unstructured (no fixed format). A fee receipt or ATM record is structured. The defining criterion is whether the data fits a fixed row-column schema.

- **Data vs Metadata distinction (NCERT §7.1.2(B), p. 101).** Metadata describes data — image resolution is metadata of an image file, not the image data itself.
- **CSV is a structured-data format (NCERT §7.1.2(A), p. 101).** Each row is an observation; columns are attributes — qualifies as structured.
- **Median can be the average of two values (NCERT §7.5.1(B), p. 105).** For an even-length sorted list, median is the average of the two middle values.
- **Mode can be multimodal (NCERT §7.5.1(C), p. 105).** A dataset may have more than one mode.
- **Range is influenced by outliers (NCERT §7.5.2(A), p. 106).** Since it uses only max and min, a single extreme value distorts it.
- **σ vs Variance (NCERT §7.5.2(B), p. 106).** σ is the square root of variance; both measure spread.
- **DBMS over file processing (NCERT §7.3, p. 102).** Overcomes redundancy, inconsistency, isolation issues.

🎯 Practice MCQs

Q1. Which of the following is the correct definition of data as given in the NCERT chapter?

A. Processed facts that help in decision-making

B. A collection of characters, numbers, and other symbols that represent values of some situations or variables

C. Information extracted from a database using queries

D. Knowledge derived from observations and experiments

Q2. Which of the following is an example of STRUCTURED data?

A. A newspaper article with photographs

B. An email with attachments

C. An inventory table with columns ModelNo, ProductName, Unit Price, Discount, Items_in_Inventory

D. A social media post containing text and video

Q3. Consider the following statements about metadata: **Statement I:** Metadata is data about data. **Statement II:** For an image file, image size (in KB or MB), image type (JPEG, PNG), and image resolution are examples of metadata. Which of the following is correct?

A. Statement I is correct but Statement II is incorrect

B. Statement I is incorrect but Statement II is correct

C. Both Statement I and Statement II are correct

D. Both Statement I and Statement II are incorrect

📊 PYQ Alignment

Data handling aligns with the data handling and computer applications section of the CUET Computer Science syllabus; questions typically test classification of data types (structured vs unstructured), correct identification of the appropriate statistical measure for a given scenario, and recall of the data processing cycle stages. Calculation-based questions on mean, median, and standard deviation using small datasets (as in the NCERT examples) also appear in moderate frequency. See PYQ archive for Computer Science.