FREE EDITION · NOTES + 3 SAMPLE MCQS

CUET · COMPUTER SCIENCE · CLASS XI · CODE 308

Understanding Data

CUET unit: Understanding Data

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

WWW.UNIDRILL.IN

📌 Snapshot

- Data is the foundational concept here — what it is, why it matters, and the difference between raw data and processed information.
- Data is classified into structured and unstructured types, a distinction CUET directly tests through scenario-based identification questions.
- The data lifecycle — collection, storage, and processing — comes with real-world examples, making it a rich source of application-based MCQs.
- Five statistical techniques matter (Mean, Median, Mode, Range, Standard Deviation), each individually testable as a direct formula or scenario question.
- NTA favours "which statistical technique to use" scenario questions and definition-based discrimination between structured/unstructured data and measures of central tendency vs. variability.

📖 Detailed Notes

2.1 Core concepts

- **Data defined:** Data is a collection of characters, numbers, and other symbols that represent values of some situations or variables. The singular of "data" is "datum". Data need to be gathered, processed, and analysed for making decisions. (NCERT §5.1, p. 82)
- **Knowledge base:** A knowledge base is a store of information consisting of facts, assumptions, and rules which an AI system can use for decision making. (NCERT §5.1, p. 82 sidebar)
- **Importance of data:** Large amounts of data, when processed with a computer, reveal possibilities or hidden traits not otherwise visible. Examples include ATM transactions, meteorological satellite monitoring, dynamic pricing by airlines and cab apps, and market analysis by businesses. (NCERT §5.1.1, p. 82–83)
- **Examples of data:** Name/age/gender/contact details; banking and ticketing transaction data; images, graphics, audio, video; documents and web pages; online posts and messages; signals from sensors; satellite and meteorological data. (NCERT §5.1, p. 82)
- **Structured data:** Data organised in a well-defined format, usually stored in tabular (rows and columns) format where each column is an attribute/characteristic/variable

and each row is an observation. Example: inventory table with columns ModelNo, ProductName, UnitPrice, Discount(%), Items_in_Inventory. (NCERT §5.1.2(A), p. 84)

- **Unstructured data:** Data not in a fixed row-and-column structure. Examples include web pages, text documents, business reports, books, audio/video files, social media messages. Unstructured data are sometimes described with the help of metadata. (NCERT §5.1.2(B), p. 85)
- **Metadata:** Data about data. For an email: subject, recipient, main body, attachment. For an image file: image size (KB/MB), image type (JPEG, PNG), image resolution. (NCERT §5.1.2(B), p. 85)
- **Data collection:** Identifying already-available data or collecting from appropriate sources. Data may exist in a diary/register (needs digitising), already in a digital file such as CSV, or may need a new software system to record it. (NCERT §5.2, p. 85–86)
- **CSV:** Comma Separated Values — a digital format in which data can already be available and ready for use. (NCERT §5.2, p. 85)
- **Data storage:** The process of storing data on storage devices so that data can be retrieved later. Common storage devices include Hard Disk Drive (HDD), Solid State Drive (SSD), CD/DVD, Tape Drive, Pen Drive, Memory Card. File processing limitations can be overcome through DBMS. (NCERT §5.3, p. 86)
- **Data processing:** Raw data (numbers/text/image) is transformed through processing 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). (NCERT §5.4, p. 87, Figure 5.1)
- **Measures of central tendency:** A single value that gives an idea about the data. The three most common measures are Mean, Median, and Mode. (NCERT §5.5.1, p. 88)
- **Mean:** Average of numeric values of an attribute. Formula: sum of all n values divided by n. Mean is not suitable when there are outliers in the data. (NCERT §5.5.1(A), p. 88)
- **Outlier:** An exceptionally large or small value compared to other values; usually considered an error that can influence average calculations. (NCERT §5.5.1(A), p. 88 note)
- **Median:** When all values are sorted in ascending or descending order, the middle value is the Median. For odd number of values it is the middle position value; for even number of values it is the average of the two middle values. Median represents the central value at which the given data is equally divided into two parts. (NCERT §5.5.1(B), p. 89)
- **Mode:** The value that appears most number of times in the given data. Computed on the basis of frequency of occurrence. A dataset has no mode if each value occurs only once. There may be multiple modes if more than one value shares the highest

frequency. Mode can be found for numeric as well as non-numeric data. (NCERT §5.5.1(C), p. 89)

- **Measures of variability (dispersion):** Refer to the spread or variation of values around the mean. Two common measures are Range and Standard Deviation. Two data sets can have the same mean/median/mode but completely different levels of dispersion. (NCERT §5.5.2, p. 89)
- **Range:** Difference between the maximum and minimum values (M - S). Can be calculated only for numerical data. Badly influenced by outliers since it uses only two extreme values. (NCERT §5.5.2(A), p. 90)
- **Standard deviation (σ):** Positive square root of the average of the squared difference of each value from the mean. Considers all given data values (unlike Range). Smaller σ means less spread; larger σ means more spread. Formula: σ = sqrt(Σ(xi – x̄)² / n). (NCERT §5.5.2(B), p. 90)
- **Python for data analysis:** Python has libraries specially built for data processing and analysis. It is one programming tool used for efficient analysis of large volumes of data using statistical techniques. (NCERT §5.5, p. 91)
- **Why averages can mislead (NCERT §5.5.1(A), p. 88).** If one student in a class of 30 scores 100 and the rest score around 50, the mean rises significantly above 50 even though most students are around 50 — this is the classic outlier-distortion scenario. Switch to median in such cases. Mean is best when the data is symmetric and outlier-free.
- **Mode for categorical data (NCERT §5.5.1(C), p. 89).** The mode is the only one of the three measures of central tendency that meaningfully applies to categorical/non-numeric data, e.g., the most popular favourite colour in a class, the most viewed video category. Mean and median are inherently numeric.
- **Range vs IQR (NCERT §5.5.2(A), p. 90).** The NCERT only introduces range, not interquartile range; CUET sticks to NCERT scope. Range = max – min. Range can be zero when all values are identical (no spread).
- **Standard deviation interpretation (NCERT §5.5.2(B), p. 90).** Smaller σ means data points cluster around the mean; larger σ means high spread. σ uses every data point — that is why it is preferred over range as a measure of dispersion.
- **Data processing examples (NCERT §5.4, Figure 5.2, p. 87).** Three canonical cases: competitive-exam website (input = candidate details/test answers, output = score/rank report), bank ATM withdrawal (input = card + PIN + amount, output = receipt + cash), train ticket issue (input = passenger details/source/destination, output = ticket + seat assignment). These illustrate the IPO cycle.

2.2 Definitions to memorise

Term	Definition	Page
Data		82

Term	Definition	Page
	A collection of characters, numbers, and other symbols that represent values of some situations or variables	
Datum	Singular of data	82
Knowledge base	A store of information consisting of facts, assumptions, and rules which an AI system can use for decision making	82
Structured data	Data organised in a well-defined, tabular (rows and columns) format	84
Unstructured data	Data not in a fixed row-and-column structure (e.g., web pages, audio/video, social media messages)	85
Metadata	Data about data (e.g., image size, image type, email subject/ recipient)	85
CSV	Comma Separated Values; a common digital file format for storing data	85
Data storage	Process of storing data on storage devices so it can be retrieved later	86
DBMS	Database Management System; overcomes limitations of file processing	86
Data processing	Transformation of raw data into useful information through the Input-Processing-Output cycle	87
Mean	Average of numeric values; sum of all values divided by total number of values	88
Outlier	An exceptionally large or small value compared to other data values; influences average calculations	88
Median	Middle value when data is sorted in ascending or descending order	89
Mode	Value that appears most number of times in the data; applicable to numeric and non-numeric data	89
Range	Difference between maximum and minimum values (M - S); measure of dispersion for numerical data only	90
Standard deviation (σ)	Positive square root of the average of the squared differences of each value from the mean	90
Information	Output of processing data; usable form such as tables, charts, or reports	87
Attribute	A column in structured data representing a characteristic/ variable	84
Observation	A row in structured data representing a single record	84
Dispersion / Variability	How spread out data values are around the mean	89

Term	Definition	Page
Measure of central tendency	A single representative value for a dataset (mean, median, or mode)	88
Data Process Cycle	Input → Processing → Output sequence	87
HDD	Hard Disk Drive — magnetic secondary storage device	86
SSD	Solid State Drive — flash-based secondary storage	86
Tape Drive	Magnetic tape storage device used for large data backups	86
File processing	Storing data in flat files; limitations overcome by DBMS	86
Variance	The arithmetic mean of squared deviations from the mean (σ^2 before taking square root)	90
Frequency	Number of times a value appears in a dataset; basis for computing mode	89

2.3 Diagrams / processes to remember

- **Figure 5.1 — Steps in Data Processing (p. 87):** Shows two diagrams. First: RAW DATA (Numbers/Text/Image) → Data Processing → INFORMATION (In the form of table/chart/text). Second: Data Process Cycle with Input (Data Collection, Data Preparation, Data Entry) → Processing (Store, Retrieve, Classify, Update) → Output (Reports, Results, Processing System).
- **Figure 5.2 — Data Based Problem Statements (p. 87):** Three real-world scenarios (competitive exam website, bank ATM withdrawal, train ticket issue) each broken into Problem Statement, Inputs, Processing steps, and Output. Useful for identifying input-processing-output in application questions.
- **Table 5.1 — Structured data about kitchen items in a shop (p. 84):** Columns are ModelNo, ProductName, UnitPrice, Discount(%), Items_in_Inventory. Illustrates the attribute (column) and observation (row) structure of structured data.
- **Table 5.3 — Standard deviation of attendance of 9 students (p. 91):** Step-by-step calculation of σ for the height dataset [90, 102, 110, 115, 85, 90, 100, 110, 110] giving σ = 10.2 cm. Memorise the procedural steps: subtract mean from each value, square, sum, divide by n, take square root.

2.4 Common confusions / NTA trap points

- **Mean vs. Median for outliers:** Mean is sensitive to outliers (one extreme value distorts it); Median is not. NTA often presents a dataset with an extreme value and asks which measure is more appropriate — answer is Median.
- **Range vs. Standard deviation:** Both measure variability, but Range uses only two values (max and min) and is therefore badly affected by a single outlier. Standard deviation uses all values. A trap question asks which is a "better" or "more reliable" measure of dispersion — standard deviation is preferred.

- **Mode for non-numeric data:** Mean and Range can be calculated only for numeric data; Mode can be applied to both numeric and non-numeric data. A favourite NTA trap is to claim Mode is only for numbers.
- **Structured vs. Unstructured misclassification:** Students often label email body as structured data. The body is unstructured; the metadata (subject, recipient, attachment) gives it partial structure. Newspaper layout, tweets, audio files are all unstructured.
- **No mode vs. multiple modes (NCERT §5.5.1(C), p. 89).** A dataset where every value is unique has no mode. If two or more values share the highest frequency, the dataset has multiple modes. NTA may assert "a dataset always has exactly one mode" — this is false.
- **Mean has no necessary connection to any actual data point (NCERT §5.5.1(A), p. 88).** The mean might not appear anywhere in the dataset. NTA distractor: claims mean must be one of the observations.
- **Range only on numerical data (NCERT §5.5.2(A), p. 90).** Cannot compute range on categorical data. NTA tests this nuance.
- **CSV is just a file format, not a data type (NCERT §5.2, p. 85).** It is a structured representation but the values inside may be numbers or strings.
- **Information ≠ Data (NCERT §5.4, p. 87).** Information is the output of processing data — they are not interchangeable.
- **DBMS overcomes file processing limitations (NCERT §5.3, p. 86).** Flat files have redundancy/consistency issues; DBMS resolves them.
- **Median splits the data 50/50 (NCERT §5.5.1(B), p. 89).** Half of all values are ⩽ median and half ⩾ median.

🎯 Practice MCQs

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

A. Processed information presented in the form of tables or charts

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

C. A store of facts, assumptions, and rules used by an AI system for decision making

D. The middle value when all observations are arranged in ascending order

Q2. Consider the following statements about structured and unstructured data: **Statement I:** Structured data is stored in a tabular format where each column represents an attribute and each row represents an observation. **Statement II:** Unstructured data can sometimes be described with the help of metadata such as image size, image type, and image resolution. Which of the above statements is/are correct?

A. Only Statement I

B. Only Statement II

C. Both Statement I and Statement II

D. Neither Statement I nor Statement II

Q3. A school teacher records the following marks obtained by 7 students in a unit test: **45, 60, 55, 60, 70, 45, 60** What is the Mode of this dataset?

A. 45

B. 55

C. 60

D. 70

📊 PYQ Alignment

Data handling and statistics are a consistent contributor to CUET Computer Science papers, typically yielding 1–2 direct MCQs per year on data types (structured vs. unstructured), definitions of statistical measures, and scenario-based questions asking students to identify the appropriate statistical technique (mean/median/mode/range/standard deviation) for a described problem. The data processing cycle (Input-Processing-Output) and storage devices have also appeared as one-line identification or match questions. See PYQ archive for Computer Science.