IDS-Unit-2

18 June 2025 18:09

1. Compare Descriptive, Predictive, and Prescriptive Analytics. Provide real-world examples to illustrate the application of each type

Analytics can be broadly categorized into Descriptive, Predictive, and Prescriptive analytics. Each serves a different purpose in the decision-making process and is used at different stages of data analysis.

1. Descriptive Analytics

Purpose: Understand what has happened in the past.

Key Features:

- Summarizes historical data
- Uses dashboards, reports, and data visualizations
- Answers: *What happened?*

Real-World Example:

- Retail: A supermarket chain uses descriptive analytics to analyze last quarter's sales data. It identifies which products sold the most, peak shopping hours, and regional performance.
- Tool Example: Excel dashboards, Power BI, Tableau

2. Predictive Analytics

Purpose: Forecast what is likely to happen in the future.

Key Features:

- Uses statistical models and machine learning
- Identifies patterns and trends
- Answers: *What is likely to happen?*

Real-World Example:

- Banking: A bank uses predictive analytics to assess the likelihood of a customer defaulting on a loan based on their credit history, income, and spending behavior.
- Tool Example: Python (scikit-learn), R, SAS

3. Prescriptive Analytics

Purpose: Recommend actions to achieve desired outcomes.

Key Features:

- Uses optimization and simulation algorithms
- Suggests decision options
- Answers: *What should we do?*

Real-World Example:

- Logistics: A delivery company uses prescriptive analytics to determine the most efficient delivery routes, considering traffic, weather, and fuel costs.
- Tool Example: IBM CPLEX, Google OR-Tools, MATLAB

Comparison Table

Feature	Descriptive Analytics	Predictive Analytics	Prescriptive Analytics
Focus	Past	Future	Decision-making
Techniques	Reporting, Visualization	Statistical Modeling, ML	Optimization, Simulation
Question Answered	What happened?	What will happen?	What should we do?
Complexity	Low to Medium	Medium to High	High
Example Use Case	Sales reports	Customer churn prediction	Supply chain optimization

==========

2. Define kurtosis in the context of a dataset. What insights does it offer about the shape of a distribution?

Kurtosis is a statistical measure that describes the "tailedness" or peakedness of a probability distribution compared to a normal distribution.

Kurtosis is a statistical measure that describes the "tailedness" or peakedness of a probability distribution compared to a normal distribution.

Definition:

Kurtosis quantifies whether the data are heavy-tailed or light-tailed relative to a normal distribution. It helps identify the presence of outliers and the shape of the distribution's tails.

Types of Kurtosis:

1. Mesokurtic (Kurtosis ≈ 3):

- Normal distribution
- Moderate tails and peak
- Example: Standard bell curve

2. Leptokurtic (Kurtosis > 3):

- Heavy tails and sharp peak
- More outliers than a normal distribution
- Example: Financial returns with extreme gains/losses

3. Platykurtic (Kurtosis < 3):

- Light tails and flatter peak
- Fewer outliers
- Example: Uniform-like distributions

Note: Many statistical tools report "excess kurtosis", which is kurtosis minus 3. So:

- Excess kurtosis = 0 → Mesokurtic
- Excess kurtosis > 0 → Leptokurtic
- Excess kurtosis < 0 → Platykurtic

Insights from Kurtosis:

- High kurtosis: Indicates a high probability of extreme values (outliers).
- Low kurtosis: Suggests data are more evenly distributed with fewer outliers.
- Used in: Risk management, quality control, and anomaly detection.

==========

3. Explain the measures of central tendency (Mean, Median, and Mode) and measures of dispersion (Variance and Standard Deviation), including appropriate examples for each.

Here's a clear explanation of measures of central tendency and measures of dispersion, along with examples for each:

Measures of Central Tendency

These describe the center or typical value of a dataset.

1. Mean (Average)

- Definition: Sum of all values divided by the number of values.
- Formula:
  $\text{Mean} = \frac{\sum x}{n}$ $\text{Mean} = \frac{\sum xi}{n}$
- Example:
  For the dataset: [4, 8, 6, 5, 3],
  $\text{Mean} = \frac{4+8+6+5+3}{5} = \frac{26}{5} = 5.2$ $\text{Mean} = \frac{54+8+6+5+3}{5} = \frac{26}{5} = 5.2$

2. Median

- Definition: The middle value when data is sorted.
- Example:
  For [3, 4, 5, 6, 8], the median is 5.
  If even number of values, it's the average of the two middle ones.

3. Mode

- Definition: The value that appears most frequently.
- Example:
  For [2, 4, 4, 6, 7], the mode is 4.

◇ Measures of Dispersion

These describe the spread or variability of the data.

1. Variance

- Definition: Average of the squared differences from the mean.
- Formula:
  $\text{Variance} = \sum(xi - \bar{x})^2 n \text{Variance} = n\sum(xi - \bar{x})^2$
- Example:
  For [2, 4, 6], mean = 4
  $\text{Variance} = (2-4)^2 + (4-4)^2 + (6-4)^2 3 = 4+0+4 3 = 2.67 \text{Variance} = 3(2-4)^2 + (4-4)^2 + (6-4)^2 = 34+0+4 = 2.67$

2. Standard Deviation

- Definition: Square root of the variance. It's in the same unit as the data.
- Formula:
  $\text{Standard Deviation} = \text{Variance} \text{Standard Deviation} = \text{Variance}$
- Example:
  From above,
  $\text{Standard Deviation} = 2.67 \approx 1.63 \text{Standard Deviation} = 2.67 \approx 1.63$

☑ Summary Table

Measure	Purpose	Example Value (Dataset: [2, 4, 6])
Mean	Central value	4
Median	Middle value	4
Mode	Most frequent value	None (all unique)
Variance	Spread of data (squared units)	2.67
Standard Deviation	Spread of data (original units)	1.63

==
==========

4. If a fair coin is tossed four times, what is the probability of obtaining exactly two heads?Show your reasoning.

To find the probability of obtaining exactly two heads when a fair coin is tossed four times, we can use the binomial probability formula:

Binomial Probability Formula:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

Where:

- $n = 4$ (number of trials)
- $k = 2$ (number of desired successes, i.e., heads)
- $p = 0.5$ (probability of getting a head in one toss)

Screen clipping taken: 18-06-2025 18:25

Step-by-Step Calculation:

1. **Calculate the binomial coefficient:**

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4 \times 3}{2 \times 1} = 6$$

2. **Calculate the probability:**

$$P(X = 2) = 6 \cdot (0.5)^2 \cdot (0.5)^2 = 6 \cdot 0.25 \cdot 0.25 = 6 \cdot 0.0625 = 0.375$$

Screen clipping taken: 18-06-2025 18:26

☑ Final Answer:

0.3750.375

So, the probability of getting exactly two heads in four tosses of a fair coin is 0.375 or 37.5%.

==========

5. Describe the concept of covariance and explain how it differs from correlation. Include examples to support your explanation.

Covariance vs. Correlation

Both covariance and correlation measure the relationship between two variables, but they differ in scale, interpretation, and standardization.

Covariance

Definition:

Covariance measures how two variables change together. If both increase or decrease simultaneously, the covariance is positive. If one increases while the other decreases, it's negative.

Formula:

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Example:

Suppose we have:

- X = [1, 2, 3]
- Y = [2, 4, 6]

As X increases, Y also increases → **positive covariance**.

Limitations:

- Covariance is not standardized.
- Its value depends on the units of the variables, making it hard to compare across datasets.

Correlation

Definition:

Correlation measures both the strength and direction of a linear relationship between two variables. It is a standardized version of covariance.

Formula (Pearson correlation):

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Range:

- $r \in [-1, 1]$
- $r = 1$: Perfect positive linear relationship
- $r = -1$: Perfect negative linear relationship
- $r = 0$: No linear relationship

Screen clipping taken: 18-06-2025 18:30

Example:

Using the same X and Y:

- Correlation = 1 (perfect positive relationship)

✅ Key Differences

Feature	Covariance	Correlation
Scale	Unbounded	Between -1 and 1
Units	Depends on variable units	Unitless
Interpretation	Direction only	Direction **and** strength
Use Case	Intermediate step in correlation	Comparing relationships across data

Screen clipping taken: 18-06-2025 18:31

==========

6. What is hypothesis testing in data science? Discuss its purpose and general process with a relevant example.

What is Hypothesis Testing in Data Science?

Hypothesis testing is a statistical method used in data science to make inferences or decisions about a population based on sample data. It helps determine whether a certain assumption (hypothesis) about the data is likely to be true.

Purpose of Hypothesis Testing

- To validate assumptions or claims using data
- To compare groups or treatments
- To detect patterns or relationships
- To support decision-making with statistical evidence

From <https://copilot.cloud.microsoft/?fromcode=bingchat&redirectid=823852B9497E4A72B8F26930484FC937>

General Process of Hypothesis Testing

Formulate Hypotheses

- Null Hypothesis (H_0): Assumes no effect or difference
- Alternative Hypothesis (H_1): Assumes an effect or difference

Choose Significance Level (α)

- Common values: 0.05, 0.01
- Represents the probability of rejecting H_0 when it is actually true (Type I error)

Select the Appropriate Test

- Depends on data type and distribution (e.g., t-test, chi-square test, ANOVA)

Calculate Test Statistic

- Based on sample data

Determine p-value

- Probability of observing the data if H_0 is true

Make a Decision

- If p-value $\leq \alpha$, reject H_0 (evidence supports H_1)
- If p-value > α, fail to reject H_0 (insufficient evidence)

📊 Example: A/B Testing in Marketing

Scenario: A company wants to test if a new email subject line increases click-through rates.

- H_0: The new subject line has the same click-through rate as the old one.
- H_1: The new subject line has a higher click-through rate.

They send:

- Old subject line to 1,000 users
- New subject line to another 1,000 users

After collecting data, they perform a two-sample t-test to compare the means. If the p-value is less than 0.05, they conclude the new subject line is more effective.

==========

7. Identify and explain the various types of attributes used in data analysis. Provide suitable examples for each type.

In data analysis, attributes (also called features or variables) are characteristics or properties of data that help describe and analyze it. These attributes are classified based on the type of data they represent, and understanding them is crucial for selecting the right analytical methods.

Types of Attributes in Data Analysis

1. Nominal Attributes (Categorical)

- Definition: Represent categories with no inherent order.
- Examples:
- Gender: Male, Female, Other
- Colors: Red, Blue, Green
- Country: India, USA, Brazil

2. Ordinal Attributes

- Definition: Represent categories with a meaningful order, but the intervals between them are not uniform.
- Examples:
- Education Level: High School < Bachelor < Master < PhD
- Customer Satisfaction: Poor < Fair < Good < Excellent

3. Interval Attributes

- Definition: Numeric values with meaningful intervals, but no true zero point.
- Examples:
- Temperature in Celsius or Fahrenheit
- Dates (e.g., years like 1990, 2000)

4. Ratio Attributes

- Definition: Numeric values with meaningful intervals and a true zero point.
- Examples:
- Height, Weight, Age
- Income, Distance, Time

☑ Summary Table

Attribute Type	Nature	Order	Equal Intervals	True Zero	Examples
Nominal	Categorical	✗	✗	✗	Gender, Color, Country
Ordinal	Categorical	✓	✗	✗	Satisfaction, Education Level
Interval	Numerical	✓	✓	✗	Temperature (°C), Dates
Ratio	Numerical	✓	✓		

==========