ASPIRE
KNOWLEDGE & SKILLS

Unit 2

Basic Statistical Concepts

AI - Business Intelligence Analyst

Table of Content

Projected sales of main products in 2013
Distribution of market share among the major industry players
Share of market activity
Projected sales of main products in 2013
Passive market share

- ✔*Distinguish between different probability distributions such as Normal, Poisson, Exponential, Bernoulli, etc.*
- ✔*Identify correlation between variables using scatterplots and other graphical techniques*
- ✔*Apply basics of descriptive statistics including measures of central tendency such as mean, median and mode*
- ✔*Apply different correlation techniques such as Pearson's Correlation Coefficient, Methods of Least Squares etc.*
- ✔*Apply different techniques for regression analysis including linear, logistic, ridge, lasso, etc.*
- ✔*Use hypothesis testing to draw inferences and measure statistical significance*

Individuals at this job are responsible for performing different aspects of Business Analysis. S/he will be responsible for importing and preprocessing data and perform exploratory analysis to derive actionable insights. A BI analyst needs to have strong analytical skills and problem solving ability. S/he needs to have good communication skills to work with stakeholders across multiple teams such as marketing, sales, product development, etc.

Different Probability Distributions Such As Normal, Poisson, Exponential, Bernoulli, Etc.

Key Probability Distributions - "Foundations of Statistical Concepts"

Subtitle: Navigating the Landscape of Probability Distributions

Normal Distribution:

- Shape: Bell-shaped curve symmetric around the mean.
- Characteristics: Mean and standard deviation uniquely define the distribution.
- Applications: Widely observed in natural phenomena, such as height or IQ distributions.

Poisson Distribution:

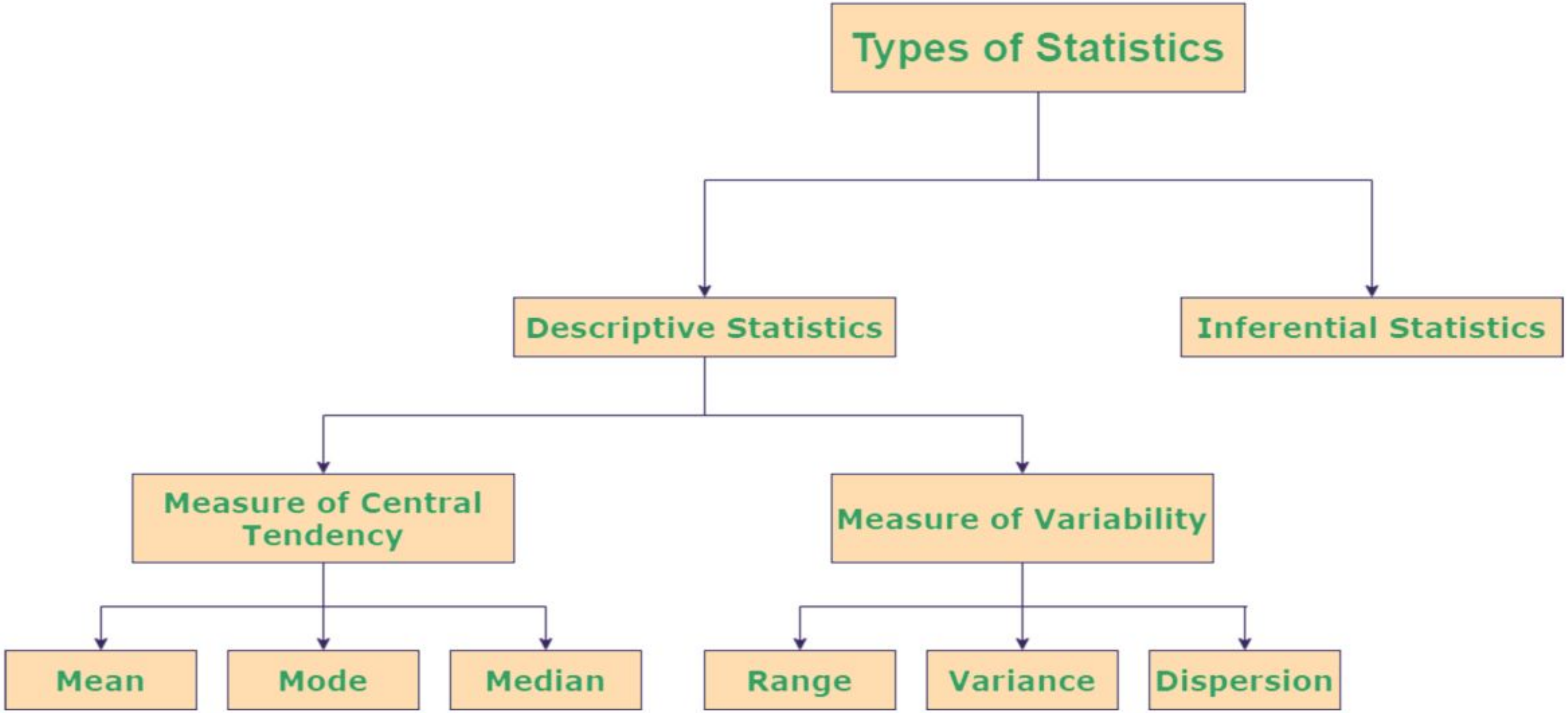
- Use Case: Models the number of events occurring in a fixed interval of time or space.
- Characteristics: Describes rare events with a known average rate of occurrence.
- Applications: Commonly used in fields like insurance, telecommunications, and healthcare for rare event prediction.

Exponential Distribution:

- Use Case: Models the time between events in a Poisson process.
- Characteristics: Memoryless property, where past events do not influence future events.
- Applications: Used in reliability engineering, queuing theory, and finance for modeling time between events.

Bernoulli Distribution:

- Binary Outcome: Represents a random variable with two possible outcomes (success or failure).
- Parameters: Probability of success (p) and failure (1-p).
- Applications: Commonly used for modeling binary events like coin flips or success/failure experiments.



Types of Statistics

Descriptive Statistics

Inferential Statistics

Measure of Central Tendency

Measure of Variability

Mean

Mode

Median

Range

Variance

Dispersion

Different Probability Distributions Such As Normal, Poisson, Exponential, Bernoulli, Etc.

Comparing Probability Distributions - "Selecting the Right Tool for the Analysis"

Subtitle: Tailoring Statistical Models to Analytical Objectives

Comparison of Distributions:

- Normal vs. Poisson: Normal for continuous data; Poisson for discrete event counts.
- Exponential vs. Normal: Exponential for modeling time between events; Normal for continuous data with symmetric distribution.
- Bernoulli vs. Poisson: Bernoulli for binary outcomes; Poisson for counting rare events over a fixed interval.

Central Limit Theorem:

- States that the sum or average of a large number of independent and identically distributed random variables approaches a normal distribution.
- Relevant when dealing with the distribution of sample means.

Applications in Business Intelligence:

- Decision-Making: Understanding the appropriate distribution aids in making informed decisions.
- Forecasting: Selecting the right distribution is crucial for accurate predictions.
- Risk Assessment: Probability distributions assist in assessing and mitigating risks in various business scenarios.

Model Selection Considerations:

- Data Type: Choose a distribution based on the nature of the data (continuous or discrete).
- Assumptions: Consider assumptions related to mean, variance, and distribution shape.
- Analytical Goals: Align the choice of distribution with the specific goals of the analysis.

Understanding Variable Relationships - "Visualizing Correlation with Scatterplots"

Subtitle: Illuminating Data Connections through Graphical Techniques

Scatterplots - A Visual Insight:

- Definition: A scatterplot is a graphical representation of individual data points on a two-dimensional plane, where each point represents the values of two variables.
- Axes: The horizontal (x-axis) and vertical (y-axis) axes represent the values of the two variables being compared.

Interpreting Scatterplots:

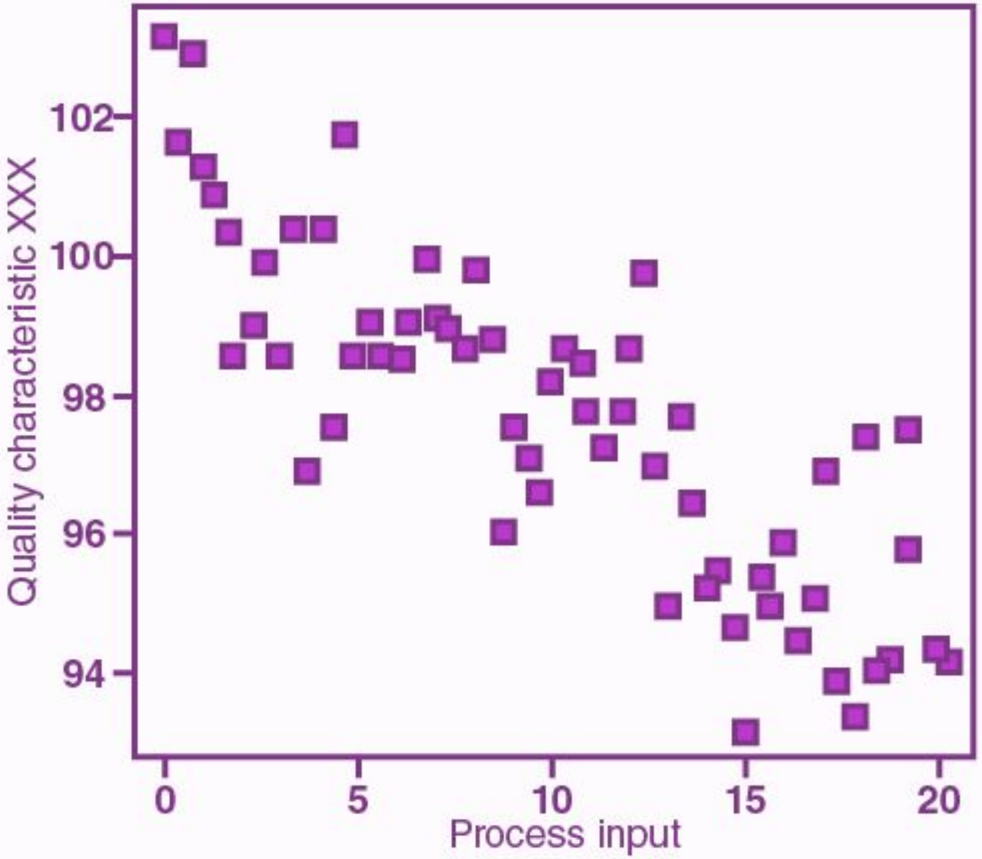
- Positive Correlation: When the points on the scatterplot tend to move upwards and to the right, there is a positive correlation. As one variable increases, the other also tends to increase.
- Negative Correlation: If the points on the scatterplot tend to move downwards and to the right, there is a negative correlation. As one variable increases, the other tends to decrease.
- No Correlation: A lack of a clear pattern in the scatterplot suggests no significant correlation between the variables.

Strength of Correlation:

- Closeness of Points: The closer the points are to forming a straight line, the stronger the correlation.
- Pattern Direction: The direction of the pattern (positive or negative) indicates the nature of the relationship.

Applications in Business Intelligence:

- Identifying Trends: Scatterplots assist in visually identifying trends or patterns in data relationships.
- Decision Support: Understanding correlations aids in making informed decisions based on the observed relationships.
- Data Exploration: Essential for exploring the connections between variables in the early stages of data analysis.



Scatterplot for quality characteristic XXX

Beyond Scatterplots - "Enhancing Graphical Techniques for Correlation"

Subtitle: Broadening Analytical Perspectives with Advanced Visualizations

Correlation Matrix:

- Definition: A correlation matrix provides a comprehensive view of correlations between multiple variables. Values range from -1 to 1, indicating the strength and direction of correlations.
- Applications: Useful for analyzing complex datasets with multiple variables and understanding relationships across the entire dataset.

Heatmaps:

- **Visualization Technique:** Heatmaps use color gradients to represent the strength of correlations in a matrix. Darker shades indicate stronger correlations.
- **Advantages:** Ideal for quickly identifying patterns and relationships, especially in datasets with a large number of variables.

Bubble Charts:

- **Representation:** In a bubble chart, each data point is represented by a circle (bubble), with the size of the bubble indicating a third variable.
- **Applications:** Useful when exploring correlations between three variables simultaneously, providing a dynamic and informative visualization.

Applications in BI Analysis:

- **Pattern Recognition:** Advanced graphical techniques enhance the analyst's ability to recognize complex patterns in data.
- **Multivariate Analysis:** Correlation matrices and heatmaps support the exploration of relationships in datasets with numerous variables.
- **Data-driven Insights:** Effective visualizations contribute to deriving actionable insights, guiding strategic decision-making.

Applying basics of descriptive statistics including measures of central tendency such as mean, median and mode

Introduction to Descriptive Statistics - "Unveiling Central Tendency"

Subtitle: Building a Foundation with Measures of Central Tendency

Descriptive Statistics Overview:

- Purpose: Descriptive statistics summarize and describe essential features of a dataset, providing insights into its central tendencies, variability, and distribution.
- Importance: Foundation for data analysis, aiding in understanding and interpreting data in a meaningful way.

Measures of Central Tendency:

- Mean (Average):
  - Definition: Calculated by summing up all values and dividing by the number of observations.
  - Characteristics: Sensitive to extreme values, providing a balance point for the dataset.
- Median (Middle Value):
  - Definition: The middle value when the data is sorted in ascending or descending order.
  - Characteristics: Less affected by outliers, offering a representation of the central position.
- Mode (Most Frequent Value):
  - Definition: The value that appears most frequently in the dataset.
  - Characteristics: Useful for identifying the most common observation.

Choosing the Right Measure:

- Mean: Preferred for symmetric distributions with no extreme values.
- Median: Suitable for skewed distributions or datasets with outliers.
- Mode: Effective for identifying the most prevalent category in categorical data.

Practical Applications - "Leveraging Central Tendency in BI Analysis"

Subtitle: Integrating Descriptive Statistics for Data-Driven Decisions

Data Exploration with Measures of Central Tendency:

- Analysis of Business Metrics: Mean, median, and mode assist in understanding the central values of key business metrics such as sales, revenue, or customer satisfaction.

Benchmarking and Comparison:

- Benchmarking Mean Values: Comparing mean values across different time periods or business units helps identify trends and performance variations.
- Median for Robust Comparisons: When dealing with data prone to outliers, median provides a more robust measure for comparison.

Strategic Decision-Making:

- Identifying Trends: Central tendency measures reveal patterns and trends within datasets, guiding strategic decision-making.
- Scenario Planning: Understanding the distribution of data helps in anticipating potential outcomes and risks.

Case Study - Revenue Analysis:

- Scenario: A company's quarterly revenue data.
- Application: Calculate mean, median, and mode to understand the central tendencies, providing insights into revenue patterns.

Challenges and Considerations:

- Outliers: Be mindful of extreme values that may skew the mean.
- Categorical Data: For categorical variables, mode becomes a crucial measure.

Applying different correlation techniques such as Pearson's Correlation Coefficient, Methods of Least Squares etc.

Correlation Techniques - "Unveiling Statistical Relationships"

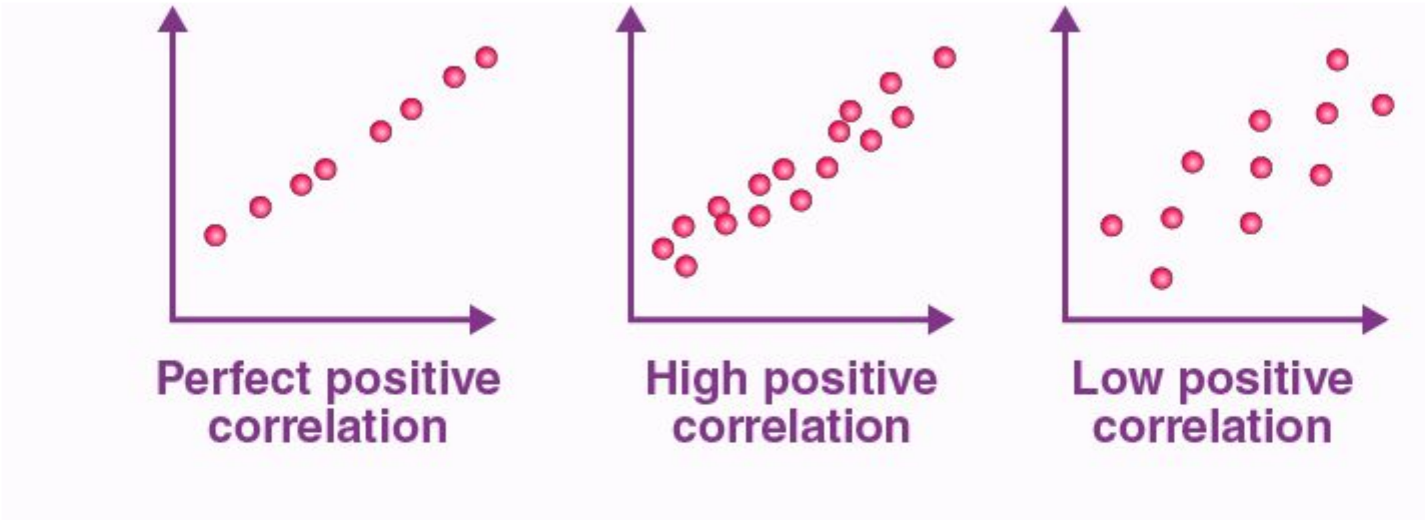
Subtitle: Employing Various Correlation Methods for In-Depth Analysis

1. **Pearson's Correlation Coefficient:**
   - Definition: A measure of linear association between two continuous variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation).
   - Interpretation:
     - ✔ +1: Perfect positive linear relationship
     - ✔ 0: No linear relationship
     - ✔ -1: Perfect negative linear relationship

2. **Spearman's Rank Correlation Coefficient:**
   - Non-parametric measure assessing strength & direction of monotonic relationships (not necessarily linear) between 2 variables.
   - Advantages: Robust to outliers, suitable for ordinal or ranked data.

3. **Kendall's Tau:**
   - Another non-parametric correlation measure, particularly useful for assessing associations in ordinal or ranked data.
   - Interpretation: Measures the number of concordant and discordant pairs of observations.

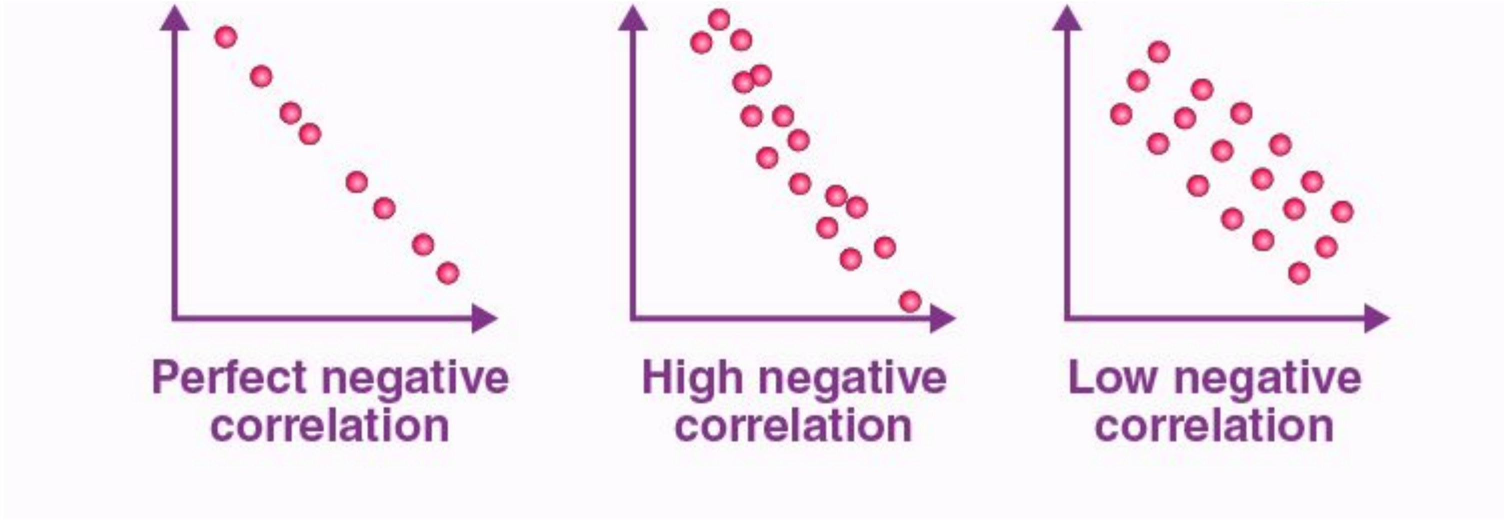
4. **Methods of Least Squares:**
   - A statistical approach to finding the line/curve that minimizes the sum of the squared differences between observed and predicted values.
   - Applications: Commonly used in linear regression to estimate the parameters of the regression line.



Perfect positive correlation

High positive correlation

Low positive correlation



Perfect negative correlation

High negative correlation

Low negative correlation

Applications in Business Intelligence - "Strategic Insights through Correlation Techniques"

Subtitle: Leveraging Correlation Techniques for Informed Decision-Making

Strategic Decision-Making:

- Understanding Market Trends: Correlation techniques assist in identifying relationships between market variables, aiding in strategic decision-making for marketing and sales strategies.

Financial Analysis:

- Risk Assessment: Employing correlation techniques helps in understanding the relationships between financial variables, facilitating risk assessment and portfolio management.
- Predictive Modeling: Utilizing correlation in financial data enables the creation of predictive models for market trends.

Customer Behavior Analysis:

- Product Associations: Identifying correlations in customer purchasing behavior aids in cross-selling and upselling strategies.
- Segmentation: Applying correlation techniques helps in segmenting customers based on their preferences and behaviors.

Applications in Business Intelligence - "Strategic Insights through Correlation Techniques"

Subtitle: Leveraging Correlation Techniques for Informed Decision-Making

Operational Efficiency:

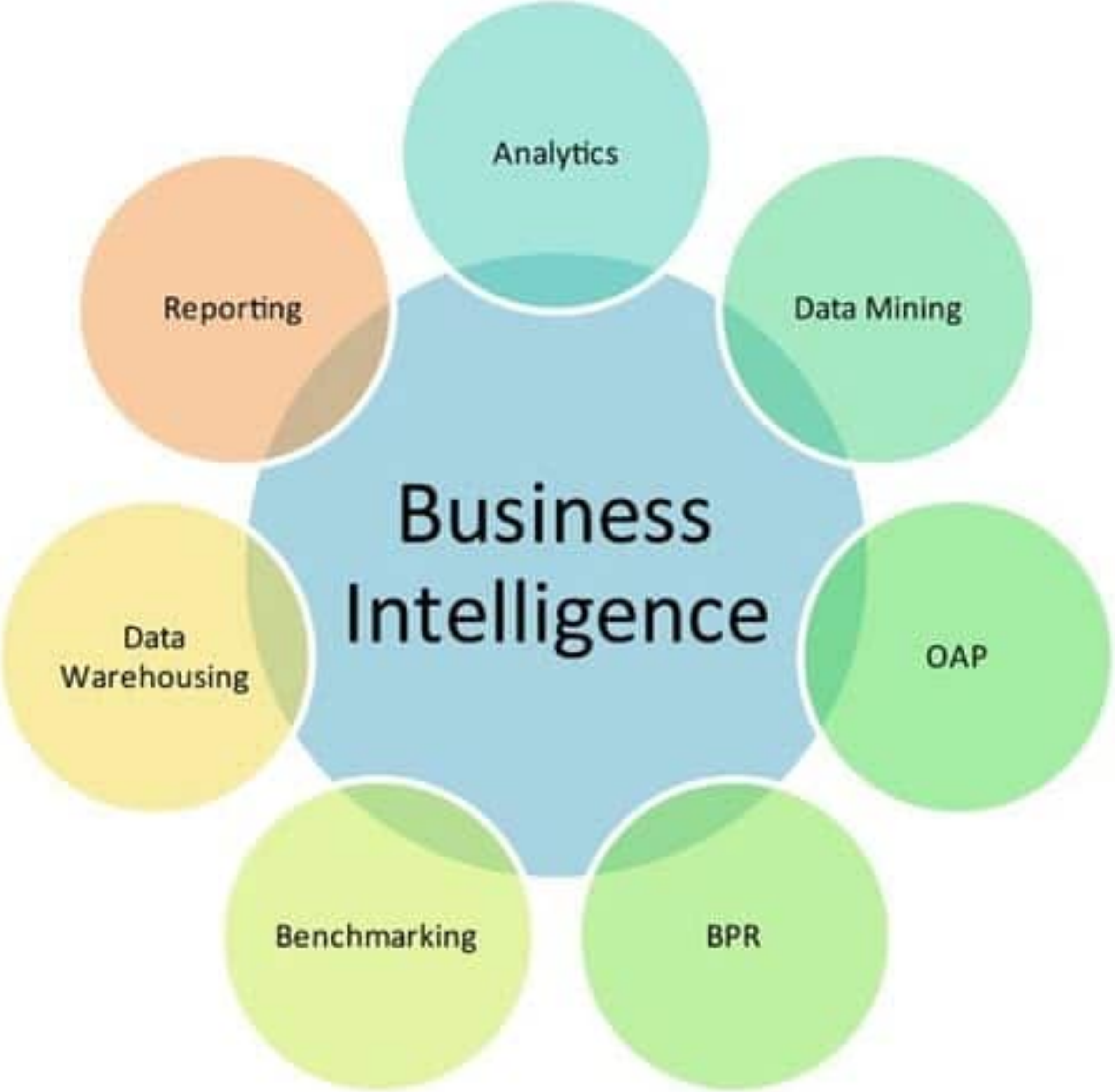
- Supply Chain Optimization: Understanding correlations between supply chain variables improves efficiency and reduces operational costs.
- Resource Allocation: Correlation techniques guide optimal resource allocation based on identified relationships.

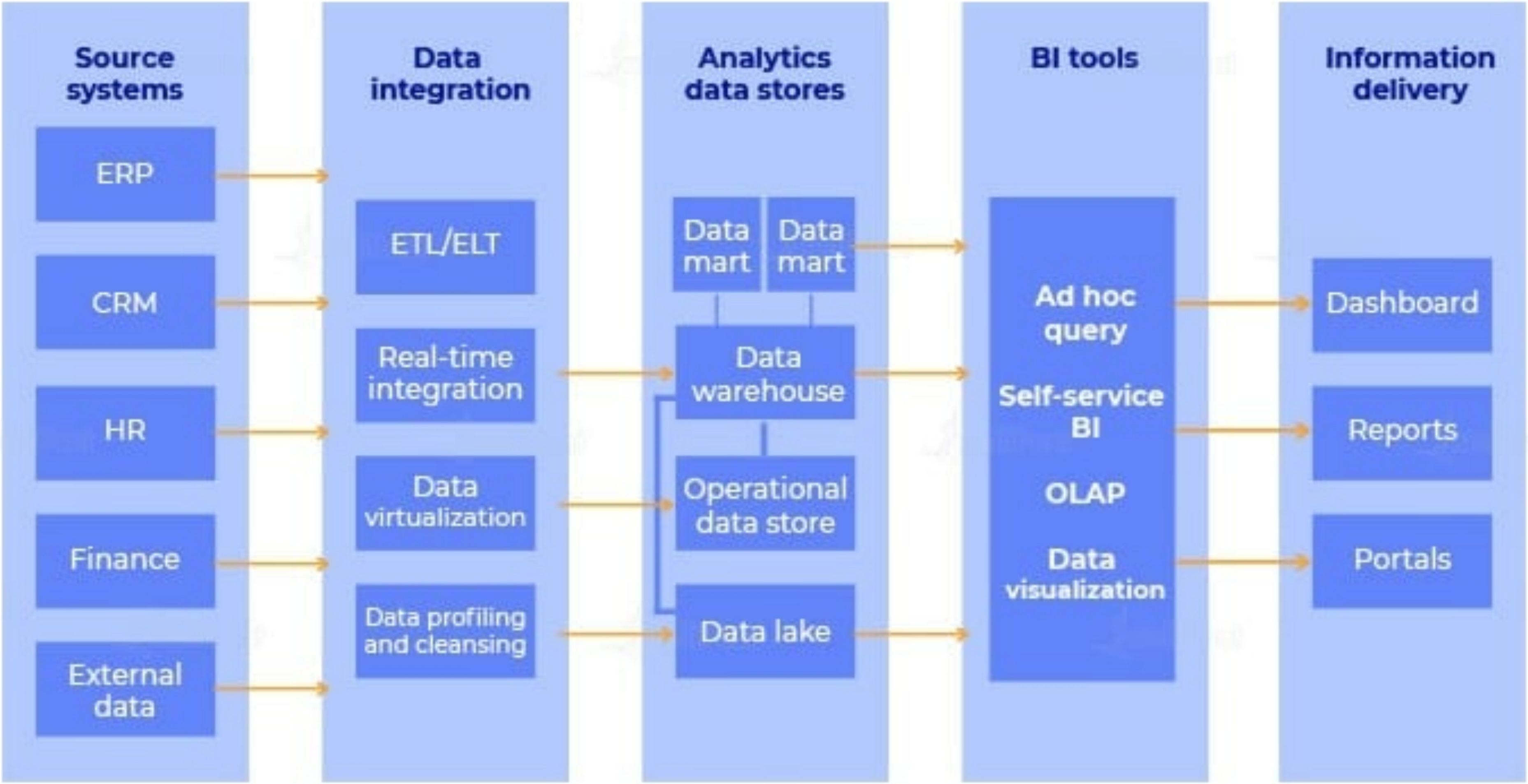
Case Study - Sales and Advertising Spend:

- Objective: Analyze the correlation between advertising spending and sales revenue.
- Method: Use Pearson's Correlation Coefficient to measure the strength and direction of the linear relationship.
- Outcome: Insights into the effectiveness of advertising campaigns on sales revenue.

Considerations and Best Practices:

- Causation vs. Correlation: Be cautious about inferring causation from correlation.
- Data Quality: Ensure data quality and completeness for accurate correlation analysis.


Analytics
Reporting
Data Mining
Business Intelligence
Data Warehousing
OAP
Benchmarking
BPR


Source systems
ERP
CRM
HR
Finance
External data
Data integration
ETL/ELT
Real-time integration
Data virtualization
Data profiling and cleansing
Analytics data stores
Data mart
Data mart
Data warehouse
Operational data store
Data lake
BI tools
Ad hoc query
Self-service BI
OLAP
Data visualization
Information delivery
Dashboard
Reports
Portals

Applying different techniques for regression analysis including linear, logistic, ridge, lasso, etc.

Regression Analysis Techniques - "Unraveling Insights through Regression Models"

Subtitle: Applying Various Regression Methods for Analyzing Relationships

1. **Linear Regression:**
   - Objective: Predicting a continuous dependent variable based on one or more independent variables.
   - Equation: $Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta kXk+\varepsilon$
   - Methodology: Least Squares minimization to find the line that best fits the data.

2. **Logistic Regression:**
   - Objective: Predicting the probability of an event occurring (binary outcome).
   - Equation: $logit(p)=\beta 0+\beta 1X1+\beta 2X2+...+\beta kXk+\varepsilon$
   - Application: Commonly used in classification tasks, such as predicting customer churn or fraud detection.

3. **Ridge Regression:**
   - Objective: A regularization technique to address multicollinearity by adding a penalty term to the linear regression equation.
   - Advantages: Prevents overfitting and stabilizes coefficient estimates.
   - Use Case: Useful when dealing with high-dimensional datasets.

Applying different techniques for regression analysis including linear, logistic, ridge, lasso, etc.

Practical Applications - "Strategic Insights through Regression Models"

Subtitle: Leveraging Regression Techniques for Business Decision-Making

1. **Forecasting and Trend Analysis:**
   - Linear Regression: Predicting future trends based on historical data.
   - Logistic Regression: Forecasting binary outcomes, e.g., predicting whether a customer will purchase a product.
2. **Customer Behavior Modeling:**
   - Logistic Regression: Analyzing factors influencing customer decisions, such as subscription renewal or product adoption.
   - Ridge Regression: Managing multicollinearity in datasets with numerous correlated variables.
3. **Risk Assessment and Fraud Detection:**
   - Logistic Regression: Identifying factors contributing to the likelihood of fraud occurrence.
   - Lasso Regression: Feature selection for effective risk assessment by shrinking less relevant variables.
4. **Marketing Effectiveness Analysis:**
   - Linear Regression: Evaluating the impact of marketing spend on sales revenue.
   - Lasso Regression: Identifying the most influential marketing channels for resource optimization.
5. **Case Study - Sales Prediction:**
   - Objective: Predicting monthly sales based on advertising spend, seasonality, and promotions.
   - Method: Apply Linear Regression to establish a relationship between variables.
   - Outcome: Insights into the impact of each factor on sales, aiding in strategic planning.
6. **Considerations and Best Practices:**
   - Data Preparation: Ensure data is cleaned and preprocessed before regression analysis.
   - Model Evaluation: Use appropriate metrics such as Mean Squared Error (MSE) or Area Under the Receiver Operating Characteristic (AUROC) for evaluation.

Introduction to Hypothesis Testing - "Unveiling Statistical Significance"

Subtitle: A Framework for Drawing Inferences in Data Analysis

Definition of Hypothesis Testing:

- Objective: Hypothesis testing is a statistical method used to make inferences about population parameters based on sample data.
- Key Components: Involves formulating a hypothesis, collecting and analyzing data, and drawing conclusions about the population.

The Hypothesis Structure:

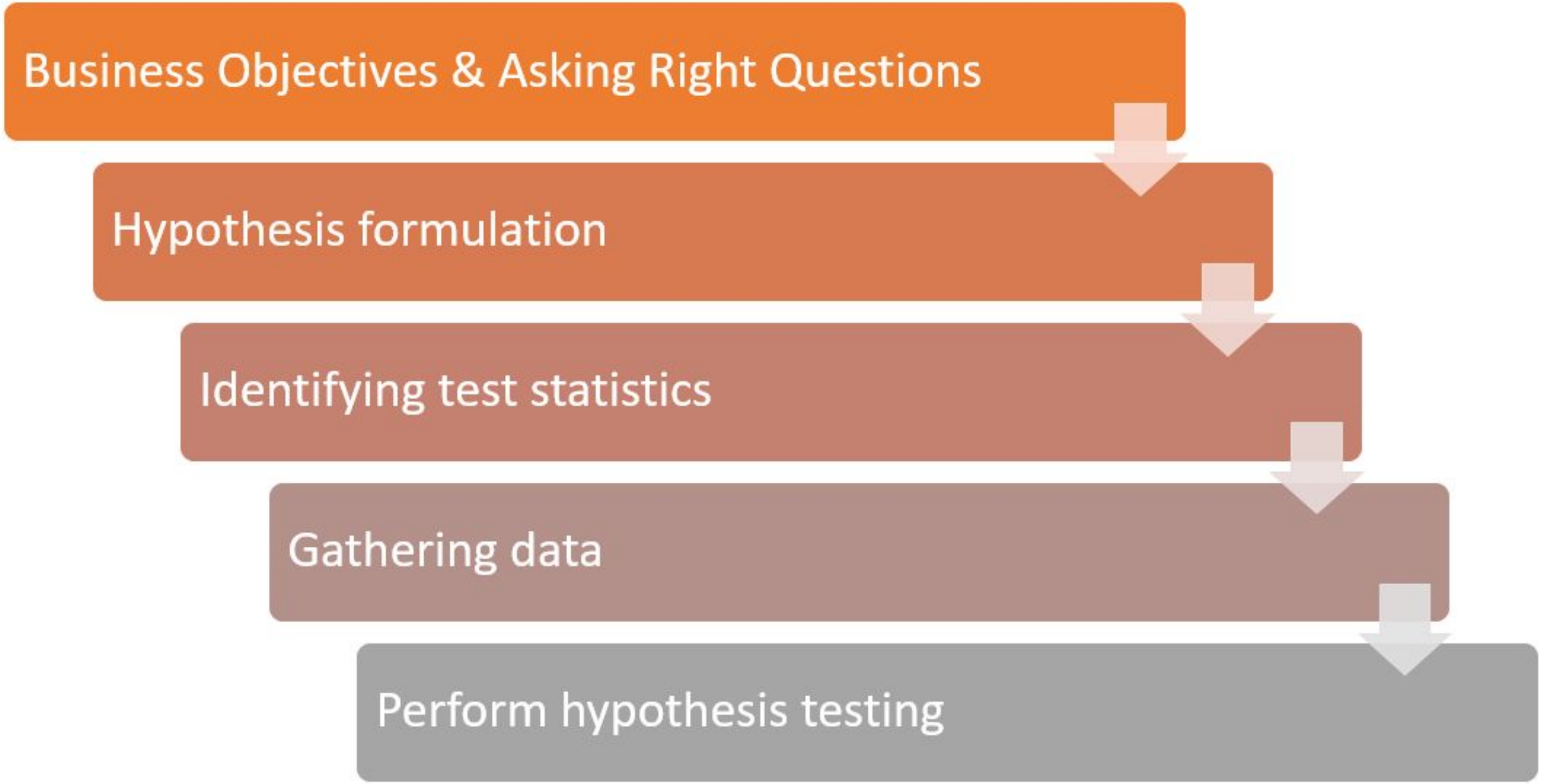
- Null Hypothesis (H_0): Represents a statement of no effect or no difference.
- Alternative Hypothesis (H_1): Opposes the null hypothesis, suggesting an effect or difference in the population.

Statistical Significance:

- P-Value: The probability of observing the data or more extreme results under the assumption that the null hypothesis is true.
- Significance Level (α): The predetermined threshold used to determine statistical significance (commonly set at 0.05).

Types of Errors:

- Type I Error (α): Incorrectly rejecting a true null hypothesis.
- Type II Error (β): Incorrectly failing to reject a false null hypothesis.



Business Objectives & Asking Right Questions

Hypothesis formulation

Identifying test statistics

Gathering data

Perform hypothesis testing

Practical Applications - "Leveraging Hypothesis Testing in Business Analysis"

Subtitle: Drawing Informed Conclusions for Strategic Decision-Making

1. A/B Testing in Marketing:
   - Scenario: Testing two versions (A and B) of a marketing campaign.
   - Hypotheses:
     - ✔ H_0: The versions have no significant difference.
     - ✔ H_1: One version outperforms the other.
   - Procedure: Collect data on user engagement, apply hypothesis testing to determine if there's a significant difference.

2. Employee Productivity Analysis:
   - Scenario: Evaluating the impact of a new productivity tool on employee performance.
   - Hypotheses:
     - ✔ H_0: The tool has no significant impact on productivity.
     - ✔ H_1: The tool improves employee productivity.
   - Procedure: Collect data on performance metrics, apply hypothesis testing to assess significance.

3. Customer Satisfaction Surveys:
   - Scenario: Analyzing the effectiveness of a customer service improvement initiative.
   - Hypotheses:
     - ✔ H_0: There is no significant improvement in customer satisfaction.
     - ✔ H_1: The initiative has a positive impact on customer satisfaction.
   - Procedure: Collect survey data, apply hypothesis testing to assess the significance of changes.

Practical Applications - "Leveraging Hypothesis Testing in Business Analysis"

Subtitle: Drawing Informed Conclusions for Strategic Decision-Making

4. Financial Decision-Making:
   - Scenario: Assessing the impact of a cost-cutting measure on profitability.
   - Hypotheses:
     - ✔ H_0: The cost-cutting measure has no significant effect on profitability.
     - ✔ H_1: The cost-cutting measure improves profitability.
   - Procedure: Collect financial data, apply hypothesis testing to determine if there's a significant change.

5. Best Practices:
   - Clearly Define Hypotheses: Ensure hypotheses are specific, testable, and relevant to the business context.
   - Understand the Data: Recognize the assumptions and limitations of the data used for hypothesis testing.
   - Consider Practical Significance: Statistical significance doesn't always imply practical significance; consider the magnitude of the effect.

Video

Types Of Distribution In Statistics

https://www.youtube.com/watch?v=Xg7ng3-Pm-8

Drawing Scatterplots & Finding Correlation of Data in Statistics

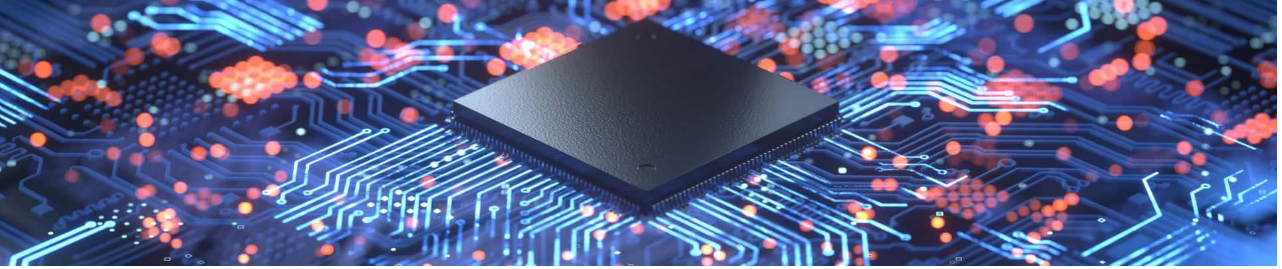
https://www.youtube.com/watch?v=8c6znSuxoRY

35 Types of Regression Models used in Data Science

https://www.youtube.com/watch?v=r2i7OBV4Y6A

Statistical Concepts and their Applications In Business Analytics

https://www.youtube.com/watch?v=6XX3pX0UXzU



ASPIRE
KNOWLEDGE & SKILLS

Thank You

Aspire Knowledge & Skills India Pvt Ltd.

1204, J.M Road, Kamala Arcade ,
Office No. 301-305, Opp. Bal Gandharva Rang
Mandir, Deccan, Pune – 411 004

Phone No: 020-25530291

Website: www.aspireks.com